Voice Conversion by modelling and transformation of extended voice characteristics

Stefan Huber



Doctoral Thesis

Voice Conversion by modelling and transformation of extended voice characteristics

Stefan Huber

Équipe Analyse / Synthése du Sons

Institut de Recherche et Coordination Acoustique/Musique (IRCAM)

Ecole Doctorale Informatique, Télécommunications et Electronique (EDITE)

Defended on September 11th, 2015, to fulfil the requirements for the degree of

Docteur de l'Université Pierre et Marie Curie (UPMC) Paris VI,

with specialization in Speech Signal Processing

Reviewers	Christophe d'Alessandro	HDR	LIMSI, CNRS (Paris, France)
	Yannis Stylianou	Professor	University of Crete (Heraklion, Greece)
Examiners	Antonio Bonafonte	Professor	TALP, Universitat Politècnica de Catalunya (Barcelona, Spain)
	Marius Cotescu	Ph.D.	R&D Acapela Group (Mons/Bergen, Belgium)
	Thomas Drugman	Ph.D.	Amazon Speech Research (Aachen, Germany)
	Olivier Adam	Professor	LAM, UPMC (Paris, France)
Supervisor	Axel Roebel	HDR	IRCAM, UPMC (Paris, France)

This work has been supervised by HDR Axel Roebel and Prof. Em. Xavier Rodet at IRCAM. The research was funded by a grant from ANR CIFRE contract 2011/2011 as a collaboration between the company Acapela Group and IRCAM. A follow-up funding was granted from the ANR ChaNTer project to integrate the proposed novel speech framework developed throughout this thesis work into the singing voice synthesizer of ChaNTer.

IRCAM - CNRS - UMR9912 - STMS
Sound Analysis / Synthesis Team
1, place Igor Stravinsky
75004 Paris
France

Acapela Group
Boulevard Dolez, 33
7000 Mons (Bergen)
Belgique (België)

Bibliography version
May 31, 2016

... to Life, the Universe, and Everything.

Abstract

Voice Conversion (VC) aims at transforming the characteristics of a source speaker's voice in such a way that it will be perceived as being uttered by a target speaker. The principle of VC is to define mapping functions for the conversion from one source speaker's voice to one target speaker's voice. The transformation functions of common STAte-of-the-ART (START) VC system adapt instantaneously to the characteristics of the source voice.

While recent VC systems have made considerable progress over the conversion quality of initial approaches, the quality is nevertheless not yet sufficient. Considerable improvements are required before VC techniques can be used in a professional industrial environment.

The objective of this thesis is to augment the quality of Voice Conversion to facilitate its industrial applicability to a reasonable extent. The basic properties of different START algorithms for Voice Conversion are discussed on their intrinsic advantages and shortcomings. Based on experimental evaluations of one GMM-based START VC approach the conclusion is that most VC systems which rely on statistical models are, due to averaging effect of the linear regression, less appropriate to achieve a high enough similarity score to the target speaker required for industrial usage.

The contributions established throughout the work for this thesis lie in the extended means to
a) model the glottal excitation source,
b) model a voice descriptor set using a novel speech system based on an extended source-filter model, and
c) further advance IRCAMs novel VC system by combining it with the contributions of a) and b).

a) Improvements to estimate the shape of the deterministic part of the glottal excitation source from speech signals are presented in this thesis. A STAte-of-the-ART method based on phase minimization to estimate the shape parameter R_d of the glottal source model LF has been considerably enhanced. First, the adaptation and extension of the utilized R_d parameter range avoids inconsistencies in the frame-based estimator. Second, the utilization of Viterbi smoothing suppresses unnatural jumps of the estimated glottal source parameter contour within short-time segments. Third, the exploitation of the correlation of other co-varying voice descriptors to additionally steer the Viterbi algorithm augments the estimators robustness, especially in segments with few stable harmonic sinusoids available where the phased minimization based paradigm is more error prone.

b) The estimation of the glottal excitation source is utilized to extract the contribution of the Vocal Tract Filter (VTF) from the spectral envelope by means of dividing the spectral envelope of the glottal pulse. This facilitates altering the voice quality of a given speech phrase by means of exciting the VTF with altered glottal pulse shapes. A novel speech system is presented which allows for the analysis, transformation and synthesis of different voice descriptors such as glottal excitation source, intensity, fundamental frequency and the voiced / unvoiced frequency boundary. The proposed speech framework *PSY* derives from *P*arametric Speech *SY*nthesis to indicate its fully parametric design to construct a speech phrase for synthesis. *PSY* is based on the separate processing of the voiced deterministic and the unvoiced stochastic part of a speech signal. Each voice descriptor and VTF or spectral envelope required for synthesis can be introduced from the same or different speakers. This flexibility allows for many voice modification possibilities or the generation of a human voice avatar.

c) To further advance the synthesis quality and the similarity score to the target speaker, a new VC system has been designed at IRCAM, being considerably different to all VC approaches known to date. The VC problem has been reformulated in such a way that no statistical model has to be trained on the corresponding source and target feature vectors. The new VC model does consequently not require to derive mapping functions expressing the correlation between voice features describing the voice identity of source and target speaker. With this, the well-known statistical over-smoothing effect of STAte-of-the-ART VC systems, introduced by the linear regression of the statistical models, is per se avoided by the definition of the VC system.

The VC system is based on techniques known from Unit Selection methods used in Text-To-Speech (TTS) synthesis systems. It is denominated by the abbreviation *coVoC* (concatenative Voice Conversion). Its quality can be close to typical TTS systems using Unit Selection, depending on the speaker pair chosen for conversion. The *coVoC* system has the advantage of requiring reasonably smaller speech databases than TTS systems. The current aim is to develop *coVoC* further such that it offers the same flexibility but a higher synthesis quality than HMM-based speech synthesis systems. Additionally beneficial is that it does not require a corpus for the source speaker since *coVoC* does not conduct the training of a statistical model. It requires only the phonetic annotation and the signal processing based estimation of voice feature descriptors of the target speakers' corpus, and the single source phrase desired to be converted into the perceptual characteristics of the target speaker. Moreover, *coVoC* allows for non-parallel speech corpora and provides thus the possibility to perform cross-lingual VC. The Concatenative Voice Conversion system *coVoC* is currently patent pending.

The novel *coVoC* system for Voice Conversion has been embedded in *PSY* to combine a Voice Conversion with a Voice Transformation approach. Only the spectral envelope representation to the target speaker is replaced by

the *coVoC* VC system. It maintains the excitation characteristic of the source speaker. The idea of combining *PSY* with *coVoC* being that the more voice descriptors of the target speaker are reflected in the construction of the converted from the source speech phrase, the more the resulting perception is comprised of the target speakers voice identity. The performance of *coVoC* alone or combined with *PSY* versus a START GMM-based VC system will be demonstrated. Both objective and subjective evaluation methodologies exhibit both in terms of synthesis quality of the converted-to-the-target signal and in terms of its similarity score to the target speaker substantial improvements to the baseline VC system.

Keywords: Voice Conversion, Voice Transformation, Voice Quality, Speech Analysis-Transformation-Synthesis, Glottal Excitation Source, Viterbi Smoothing, Statistical and Digital Signal Processing

Contents

Chapter 1

Voice Conversion (VC) - Introduction and overview

For He spoke, and it came into being. He commanded, and it came into existence.

THE HOLY BIBLE (PSALM 33:9)

1.1 An explanation of Voice Conversion

Voice Conversion (VC) aims at transforming the characteristics of the speech signal of a source speakers' voice in such a way that it will be perceived as being uttered by a target speaker. The VC technology can be described as the conversion process to transform a source speakers' voice indentity into the one of a target speaker. To date, common VC systems usually train a GMM-based statistical model to construct functions for the conversion from one source to one target speaker. These systems maintain the expressive intonation and prosody of the source speaker. The VC process can be separated into the following schematic topics:

1. **Analysis:**
   Estimation of the parameters describing the voice characteristics of a source and a target speaker from a speech signal. The descriptor extraction is usually based on quasi-stationary time-invariant signal segments like one spectral frame, and on their time-variant evolution over time.

2. **Training:**
   A statistical model is designed and trained to represent the relation between the source and target speakers features.

3. **Mapping and transformation:**
   A mapping function is deduced from the trained statistical model that allows the transformation of the parameters from the source to the target features. It expresses the relation between the source and target speaker according to the set of features that have been selected to define the voice characteristics.

4. **Synthesis:**
   The mapping function can be applied to the feature sequence of arbitrary speech signals uttered by the source speaker. This obtains a transformed feature sequence that corresponds to a speech signal containing the same text that is spoken with a similar style of expression, but being uttered by the target speaker. The transformation of the voice parameter vector that has been established in the parameter transformation stage needs to be maintained in the sound signal. A high quality transformation is required to preserve the naturalness of the original speech recording in the re-synthesized speech signal being converted from the source to the target speaker.

1.2 Voice Conversion applications

Voice Conversion has many interesting applications. The goals are promising for the different domains where speech and voices play an important role, such as Video Games, Video, Films, Animation, and Dubbing. Other areas of general audio processing and content creation like Music production, Multimedia in general or Speech-To-Speech translation are of interest to employ the VC technique. The relatively new technology has received increasing attention within the speech research community over the last years due to recent improvements in synthesis quality and the conversion score of a speaker identity. The reproduction of and / or transformation into specific voices may find use in:

- **Voice Re-Creation:**
  The re-creation of voices from deceased human persons based on old recordings.

- **Voice Dubbing:**
  As novel technology for movies or video games in order to dub a user's voice into the voice identity of another person, e.g. a famous celebrity. It saves movie producers expensive fees for a studio to rent and a star to speak all sentences. It enable to let a video player act vocally as a celebrity.

- **Cross-lingual VC:**
  - *Movie Dubbing:*
  The voice identity of a famous actor / actress in his or her native language as source language can be transformed into any other target language. The target language exhibits a different linguistic and contextual content, and may share a non-complete phonetic coverage as compared to the source language.
  - *Speech-to-Speech translation*:
  Preserve a speakers' voice identity in a recorded speech phrase being translated to another language

- **Text-To-Speech (TTS) corpora extension:**
  New voices for a TTS system can be created out of an existing corpus without the need to record, annotate and label a speaker's voice for a new TTS database. Currently the creation of a new voice is relatively costly and requires recordings of several hours. VC systems require a comparatively small amount of recordings. New voices could be achieved by means of converting an existing voice character database into the desired speaker. The creation of an expressive speech corpus out of a corpus constituting a normal speaking style is possible with VC, possibly combined with Voice Transformation.

- **Improvement of Voice Transformation Systems:**

  a) The desired transformation can first be executed roughly using Voice Conversion. Voice Transformation techniques can be applied afterwards. A male-to-female Voice Transformation could be executed by first applying VC using a pre-trained female target voice. The following fine-tuning of the converted target can be applied by means of Voice Transformation algorithms. The VC algorithm should be adjusted to produce a high signal quality to the expense of achieving a lower conversion score.

  b) The opposite way of first applying Voice Transformation and then using Voice Conversion is as well possible. The desired Voice Transformation algorithm is executed beforehand. Voice Conversion is applied afterwards to fine-tune the converted signal towards a pre-defined perceptual voice character.

- **Human Voice Avatar Generation:**
  The interactive generation of artificial voice characters is an extension of the Voice Conversion paradigm.

  a) *Digital Content Creation:*
  A new artificial voice identity can be created fitting to any digital human avatar personality. It could find use for personalized web 3.0 content creation, for marketing and advertising, or in virtual online communities.

  b) *Games:*
  A video player can construct interactively an artificial new voice character and talk with his own avatar voice identity.

- **Biometric Testing:**

  A biometric voice system can be tested against intrusion prevention by trying to confuse the front-end speaker identification or verification system.

- **Voice pathology**

a) *Speech Enhancement for Alaryngeal Voices:*
The voices of persons suffering from vocal disorders can be converted to normal speech to recover a natural sounding voice quality.

b) *Voice Training:*
A training interface for patients with vocal disorders can be established by means of a VC system.

1.3 Basic VC techniques

1.3.1 Parallel and non-parallel speech corpora

Voice Conversion frameworks differ generally in how speaker data is provided and processed. A parallel corpus requires to record the same sentences for both source and target speakers. The training on the data for the creation of the statistical model for a Voice Conversion framework utilizes these phonetically balanced sentences. The conversion model captures the correlation between source and target speaker and examines their acoustic correspondences or dissimilarities.

A non-parallel training corpus requires additional phonetic and linguistic information in order to map segmented frames into a feature space, cluster similar segments into groups and define phonetic categories [Machado and Queiroz, 2010]. The acoustical parameters of the source are then mapped within each category according to the similarity between source and target frames.

Parallel data can be generated from non-parallel data by clustering phonetically identical frames or segments. Similarly, unit selection can be utilized to match similar source and target phonemes, diphones, syllables or even words. The model adaptation to non-parallel data of a-priori known speaker voices creates as well utterance pairs for the training of a conversion model. The text-independent VC of [Duxans, 2006] employs either a modified EM algorithm with fixed co-variance matrices, or converts the non-parallel data set to transformed vectors of parallel data.

1.3.2 A blend of digital signal processing and statistical modelling techniques

Former research in VC proposed using codebook mapping to exchange centroid vectors defined by weighted sums between source and target feature sequences. Vector Quantization (VQ) reduced quantization errors from this hard-clustering approach. Nowadays, the most common approach in VC is to employ a statistical technique like Gaussian Mixture Models (GMM) [Stylianou, 1996, Kain, 2001] to establish the mapping function. Choosing a GMM as statistical model in order to learn how to map acoustic features from a source to a target voice allows a rather flexible configuration of the parameter space while achieving good results. Many other principles and techniques have as well been proposed [Stylianou, 2009, Machado and Queiroz, 2010], as for example other probabilistic models like Hidden Markov Models (HMM) [Wu et al., 2006, Zen et al., 2011], cognitive models like Artificial Neural Networks (ANN) [Desai et al., 2009, Desai et al., 2010], or a combination with signal processing means like Dynamic Frequency Warping [Erro and Moreno, 2007, Godoy et al., 2012]. The sections 4.6 and 4.8 will provide a more detailed overview of different STAte-of-the-ART statistical VC methods. The statistical modelling requires nevertheless signal processing means to analyze and synthesize the employed voice descriptor set describing a speakers voice identity. The chapters 2 and 3 will introduce several STAte-of-the-ART signal processing techniques like the Harmonic plus Noise Model and its different variants required to process speech by extended means.

1.3.3 Characteristic descriptors of voice identity

Most VC approaches consider only the transformation of the spectral envelope describing the vocal tract of a speaker. A huge part of the spectral envelopes represent the influence of the Vocal Tract Filter (VTF), the basic filtering element which forms the 'colour' of a voice identity. The VTF as part of the apparatus of human speech production will be introduced in section 2.1. Line Spectral Frequencies (LSF), introduced in section 2.6.2, are generally preferred for the transformation and interpolation of spectral envelopes. A precise envelope estimation using efficient cepstrum-based True-Envelope (TE) estimation is beneficial for the transformation quality [Villavicencio et al., 2006, Villavicencio et al., 2007].

It is interesting to note, however, that human voice impersonators use a different set of features to adapt their own voice to a given target voice. Notably, they adapt their prosody and to some extend the characteristics of the glottal source, because changing the vocal tract as a physical part of

the human speech apparatus is impossible. There exist few approaches that explicitly use these features in VC systems [Childers, 1995, Rentzos et al., 2003, Rentzos et al., 2004, Rao and Yegnanarayana, 2006, del Pozo and Young, 2008, Wu et al., 2010, Nose and Kobayashi, 2011]. The experimental comparison in [Rentzos et al., 2004] support the hypothesis that prosody related features have the potential to significantly improve the similarity of the converted and target speech characteristics.

1.4 Voice Conversion problems

Standard STAte-of-the-ART VC systems based on statistical models show in the literature on their evaluation that two main problems remain to be solved:
Conversion score - The insufficient similarity between the transformed source and the target voice identity.
Conversion quality - The artefacts that are present in the transformed signal.
The main sources of artefacts are related to inconsistencies in the transformed features. These may be inconsistencies between the vocal tract and the excitation source, as well as the inconsistencies introduced by an incoherent mapping of the features of consecutive signal frames.

1.4.1 Spectral over-smoothing

One current drawback of VC systems with statistical models like GMM is the over-smoothing effect of the parameter conversion. It is caused by the averaging effect of least-square error estimation techniques which degrade the quality by broadening formants in the converted spectra. This problem results from the ambiguities that are related to the inconsistent parameter relations between source and target speaker. The GMM generates for these ambiguous mappings a simple average mapping which results in the observed smoothing. This degrades the natural spectral envelope contour and eliminates specific spectral details [Machado and Queiroz, 2010]. The similarity score to the target speaker is reduced and artefacts in the synthesis are introduced. A certain muffled or nasalized effect can be noticed [Stylianou, 2009].

1.4.2 Trade-off between conversion quality and conversion score

The outputs of two VC systems are combined in [Toda et al., 2001], one using GMMs and the other Dynamic Frequency Warping (DFW). It increases the synthesis quality of the signal being converted to the target speaker, for the cost of decreasing the conversion score towards the target speakers voice identity. This observation is valid for most VC systems: Either a higher conversion score is achieved by loosing synthesis quality, or vice versa. The mutual influence between the quality and the conversion score of the converted target output stream leads to a trade-off. It results either in a low quality synthesis while gaining a high similarity between the synthesized conversion output and the desired target voice, or higher synthesis qualities come with the expense of a lower similarity to the desired target voice identity. The goal of each novel VC approach is therefore to establish a system being capable to increase simultaneously conversion score and synthesis quality.

1.4.3 Annotation

The annotation of a complete speaker corpus with at least ~5 to ~15 minutes of speech recordings of a spoken voice is usually regarded as too expensive to be done manually. Different algorithmic solutions exist in the literature to execute the required annotation of the recordings into single phonemes, diphones, syllables or words. Erroneous annotations leave the successive alignment algorithm and the complete VC system little possibilities to correct feature mismatches.

1.4.4 Alignment

Selected audio units, e.g. a phoneme or diphone pair from the source and the target speaker corpus, have to be aligned in time to map their corresponding voice descriptor content correctly for the training of a statistical model for VC. Dynamic Time Warping (DTW) or Hidden Markov Model (HMM) state alignment techniques are employed in the literature to account for time differences with which each speaker used to articulate each spoken phonetic content. Mel-Frequency Cepstral Coefficients (MFCC) can act as a distance metric for the time alignment. It cannot be assured that the alignment of voice features corresponds exactly to what a human listener would perceive if the features were synthesized. Local and global constraints may aid the sequence alignment

process of the underlying time series of the chosen voice descriptors to better approximate the true correspondence of both feature vectors to be mapped [Serrà, 2011]. Misalignments introduce feature mismatches which in turn will originate artefacts and dissimilarities.

1.5 Conceptual basis and objectives for this thesis on VC

There are clear theoretical arguments and some experimental evidence for the fact that the extended source filter modelling as well as an optimized statistical modelling have great potential to improve existing Voice Conversion strategies. Given that the work executed for this thesis targets an extended set of speaker dependent voice features leads to the promising outlook that at the end a significant performance improvement of existing Voice Conversion systems may be achieved. The advancements provided by the present thesis are intended to trigger further industrial use of Voice Conversion tools.

The objective of the presented thesis is to advance the VC performance in terms of conversion score and conversion quality by means of converting and transforming an extended set of voice descriptors from the source to target speaker such that the target speakers voice identity is better captured.

The extended voice descriptor set consists not just of the Vocal Tract Filter (VTF), but also of the deterministic part of the glottal excitation source, the unvoiced component, the Fundamental Frequency F_0, the Voiced / Unvoiced Frequency boundary F_{VU}, and the RMS-based energy E_{voi} for the voiced and E_{unv} for the unvoiced component of a speech signal.

The following three chapters 2, 3 and 4 discuss the relevant STAte-of-the-ART (STAR/START) found in the literature on which this thesis is build on. Each presented method constitutes to a huge extent a brilliant STAR to build upon an algorithmic universe for further signal processing advancements. The discussed algorithms establish additionally a good START to implement means contributing to the chosen thesis topic.

Chapter 5 presents enhancements of a START algorithm to estimate the deterministic part of the glottal excitation source. Chapter 6 establishes a novel speech analysis, transformation and synthesis system for spoken voices. It is intended to build the basis for the transformation and conversion of an extended set of voice features describing speaker identities. Section introduces a START Voice Conversion approach based on the means of concatenative synthesis. The rest of chapter 7 discusses means to advance this novel VC system.

Chapter 2

STAte-of-the-ART (START) in Speech Signal Processing

From the deep unfathomable vortex of that golden light in which the Victor bathes, All nature's wordless voice in thousand tones ariseth to proclaim: "Joy unto ye, O men of Myalba"

HELENA PETROVNA BLAVATSKY - THE VOICE OF THE SILENCE

This chapter presents some basic signal processing concepts required to understand the different algorithms discussed throughout this thesis. However, for the sake of brevity, not every signal processing technology is explained. The given informations are not too detailed for the same reason. A certain level of knowledge in Digital Signal Processing (DSP) for speech and audio is expected.

2.1 The human voice production system

Fig. 2.1 illustrates the physical configuration of the human speech production apparatus. The air pressures present at the areas of the larynx above and below the glottis are denominated supra-glottic and sub-glottic pressure. The sub-glottal pressure from the lungs causes an airflow passing through larynx and pharynx. The vocal folds led the glottis open and close [Baer et al., 1983]. This generates a glottal pulse once per glottal opening and closing [Baken, 1992]. Likewise, the Bernoulli effect [Bernoulli, 1738], here caused by constrictions in the larynx [Ladefoged, 1996], generates additional air turbulences. With this, the glottal pulses constitute the deterministic and the air turbulences the stochastic part of the glottal excitation source. Both pass through the vocal tract and radiate at the lips and nostril into the free air.

The vocal tract is comprised by the air passages above the larynx, namely the pharyngeal, the oral and the nasal cavity [Marasek, 1997]. The cavities of the vocal tract are the physical cause filtering the glottal excitation source signal. The Vocal Tract Filter (VTF) creates acoustic resonances and anti-resonances. The deterministic and stochastic signal parts of the glottal excitation source are convolved with the VTF impulse response. This describes the colouring effect of the VTF as part of the well-known Source-Filter model [Fant, 1960]. This model simplifies the acoustic theory of speech production into a linear system with a source signal exciting a filter [Fant, 1981].

Using the theory of Linear Time-Invariant (LTI) systems [Rabiner and Schafer, 1978], the production of voiced speech $v(n)$ can be interpreted by a pulse train $\delta_s(n)$ exciting a LTI system with impulse response $h(n)$ [Quatieri, 2002]:

$$v(n) = \delta_s(n) * h(n), \tag{2.1}$$

with $*$ denoting a convolution in the time domain.

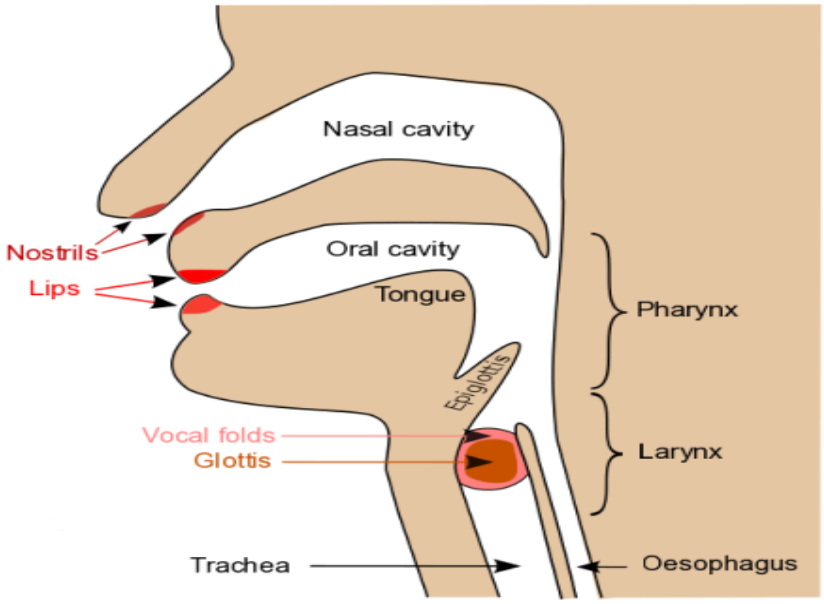


Figure 2.1: *Schematic diagram of the human speech production apparatus*

Similar to the interpretation of [Drugman, 2011] in the z-plane, a simplification of the voice production system can be described in the spectral domain. A speech spectrum $S(\omega)$ is generated by a convolution of the glottal excitation source $G(\omega)$ with the impulse response of the VTF $C(\omega)$, and the impulse response of the radiation filter $R(\omega)$ at lips and nostrils level:

$$S(\omega) = G(\omega) \cdot C(\omega) \cdot R(\omega). \tag{2.2}$$

Please note that here the periodic impulse train $\delta_s(n)$ to describe the sequence of glottal pulses is attributed to the glottal excitation source $G(\omega)$ for simplification purposes. The presented system is linear such that non-linear effects cannot be present. The mathematical description of the human voice production system in equ. 2.2 thus simplifies the involved physiological mechanisms and the resulting acoustic processes [Kane, 2013]. The signal representation unifies as in [Fant, 1960, Fant, 1981] the deterministic and the stochastic part of the glottal excitation source in the corresponding signal component $G(\omega)$. Several more detailed human voice production systems will be introduced in the chapters 3 and 6. which avoid this simplification. Some of these extended source-filter models treat the deterministic and stochastic component separately.

2.2 Processing of discrete-time audio signals

This section explains how a signal can be converted to minimum phase. It is required to understand the phase minimization paradigm of section 3.7.3 which is utilized to estimate the deterministic part of the glottal excitation source, introduced in chapter 5. First, the basic phase properties of discrete-time audio signals are presented in section 2.2.1. Second, section 2.2.2 shows how to compute the real and complex cepstrum which is required to further follow in section 2.2.3 the conversion to a minimum phase signal.

2.2.1 Phase properties

Different signal phase types of LTI systems are presented in this sections. The LTI system theory differentiates between linear, zero, minimum, maximum, and mixed phase systems. Table 2.1 summarizes the findings on the phase properties of LTI systems found in [Oppenheim and Schafer, 1975, Oppenheim et al., 1976, Oppenheim, 1978, Oppenheim et al., 1999, Smith, 2007]. A similar theory can be derived as in [Smith, ssed] for signals.

The time delay of each sinusoidal component can be expressed by the phase delay. The delay of the amplitude-envelope frequency-component is expressed by the group delay [Smith, 2007]. It is defined as the negative derivative of the phase of the Fourier Transform [Murthy and Yegnanarayana, 1991]. It can be used as a phase distortion measure [Oppenheim, 1978], with the deviation of the group delay from a constant indicating the degree of phase non-linearity. The phase delay and the group delay of a filter with linear phase response and slope α can be interpreted as a time delay [Smith, 2007]. A system with a linear phase response shifts an unit impulse, being expressed

by a Kronecker delta function δ_k for discrete time signals. The δ_k function is zero at each time but at time origin zero positive infinite with identity one [Dirac, 1930]. A non-linear phase response alters the relative phases of a sinusoidal signal. A zero phase signal is a special case of a linear phase signal with the phase slope $\alpha = 0$ being zero [Smith, 2007]. A minimum phase system is causal, stable and invertible, having all poles and zeros of $H(z)$ inside the unit circle. Every stable all-pole filter $H(z)=1/A(z)$ is minimum phase because the stability implies that $A(z)$ is minimum phase. A maximum phase system is the counterpart of a minimum phase system, having a phase response with maximum delay [Oppenheim et al., 1968]. Each minimum, maximum or mixed phase causal system requires the poles to be inside the unit circle for stability.

Table 2.1: *Phase properties of causal LTI systems*

Phase spectrum type	Group delay	Zeros	Impulse response	Stability
Linear phase	constant	inside, on and outside the unit circle	symmetric	causal, stable, non-invertible
Minimum phase	smallest, therefore minimum	inside and on the unit circle	energy maximally concentrated towards time zero	causal, stable, invertible
Maximum phase	maximum delay	outside the unit circle	maximum delay of energy	anti-causal, unstable if poles outside unit circle, non-invertible
Mixed phase	between minimum and maximum	inside and outside the unit circle	between minimum and maximum	non-invertible

2.2.2 Real and complex cepstrum

A discrete-time sequence $s(n)$ is represented in the spectral domain by $S(\omega)$. The real cepstrum $c(n)$ of $s(n)$ is given by the inverse Discrete-Time Fourier Transform (DTFT) of the logarithm of the magnitude spectrum $|S(\omega)|$ [Oppenheim and Schafer, 1975, Oppenheim, 1978]:

$$c(n) = \mathcal{F}^{-1}\left(\log(|S(\omega)|)\right). \tag{2.3}$$

The logarithm operator compresses the dynamic range and emphasizes the harmonic periodicity. The real cepstrum $c(n)$ can be transformed back to its spectral expression $S'(\omega)$ by inverse computing the real cepstrum as defined by equ. 2.4:

$$S'(\omega) = \exp\left(\mathcal{F}(c(n))\right). \tag{2.4}$$

The application of the absolute value $|S(\omega)|$ in the spectral representation of signal $s(n)$ removes its phase part $\angle S(\omega)$. The real cepstrum $c(n)$ as well as the signals $s(n)$ and $s'(n)$ related to the spectra $|S'(\omega)|$ and $|S(\omega)|$ are therefore zero phase signals [Degottex, 2010] if the signals $s(n)$ and $s'(n)$ are real. The discrete complex cepstrum $\hat{c}(n)$ defined by equ. 2.5 includes additionally to the logarithm of the magnitude spectrum $\log(|S(\omega)|)$ of equ. 2.3 the phase spectrum $\angle S(\omega)$. The following definition of $\hat{c}(n)$ can be found in [Quatieri, 2002, p.269]:

$$\begin{aligned} S(k) &= \sum_{n=0}^{N} s(n) \cdot e^{-j2\pi kn/N} \\ \hat{S}(k) &= \log(|S(k)|) + j\angle S(k) \\ \hat{c}(n) &= \frac{1}{N}\sum_{k=0}^{N-1} \hat{S}(k) \cdot e^{j2\pi kn/N} \end{aligned} \tag{2.5}$$

Similar definitions exist in [Oppenheim and Schafer, 1975, p.530] or [Bozkurt, 2005, Drugman et al., 2009a]. A mixed phase signal $s(n)$ is comprised of a minimum phase part $\hat{c}_{min}(n)$ and a maximum phase part $\hat{c}_{max}(n)$ [Drugman et al., 2011]. Consequential, the complex cepstrum $\hat{c}(n)$ of signal $s(n)$ can also be defined, e.g. as in [Oppenheim et al., 1976], by the corresponding addition of the minimum and maximum phase components in the cepstral domain:

$$\hat{c}(n) = \hat{c}_{min}(n) + \hat{c}_{max}(n) \tag{2.6}$$

The minimum phase part constitutes the causal part of the cepstrum. The maximum phase part is represented by the anti-causal part of the cepstrum [Doval et al., 2003, Smith, 2007]. The minimum phase part $\hat{c}_{min}(n)$ can be

extracted from the complex cepstrum $\hat{c}(n)$ by setting the negative quefrencies of the anti-causal cepstrum to zero, as defined by equ. 2.7. The terminology quefrency derives from the fact that the cepstral representation is nominally given in the time domain by the inverse Fourier transform. Inverting frequency results in quefrency.

$$\hat{c}_{min}(n) = \begin{cases} 0 & \forall \quad n < 0 \quad \text{(anti-causal part)} \\ \hat{c}(0)/2 & n = 0 \\ \hat{c}(n) & \forall \quad n > 0 \quad \text{(causal part)} \end{cases} \tag{2.7}$$

The maximum phase component $\hat{c}_{max}(n)$ can be extracted as defined by equ. 2.8 from the complex cepstrum $\hat{c}(n)$ by setting the positive quefrencies of the causal cepstrum to zero. The division by 2 at n=0 computes an equal distribution of the average spectral log amplitude.

$$\hat{c}_{max}(n) = \begin{cases} \hat{c}(n) & \forall \quad n < 0 \quad \text{(anti-causal part)} \\ \hat{c}(0)/2 & n = 0 \\ 0 & \forall \quad n > 0 \quad \text{(causal part)} \end{cases} \tag{2.8}$$

2.2.3 Minimum phase conversion

A mixed phase spectrum $S(\omega)$ can be converted to a minimum phase spectrum $S\_(\omega)$ by mirroring the anti-causal component onto the causal component in the complex cepstrum representation $\hat{c}(n)$ of $S(\omega)$ [Oppenheim et al., 1976, Degottex, 2010] [Oppenheim, 1978, p.794], as defined by equ. 2.9. The DC component is not divided by 2 because the minimum phase version $S\_(\omega)$ of $S(\omega)$ describes the same magnitude spectrum.

$$\hat{c}_{min}(n) = \begin{cases} 0 & \forall \quad n < 0 \quad \text{(anti-causal part)} \\ \hat{c}(0) & n = 0 \\ \hat{c}(n) + \hat{c}(-n) & \forall \quad n > 0 \quad \text{(causal part)} \end{cases} \tag{2.9}$$

Alternatively, the minimum phase part $\hat{c}_{min}(n)$ can be computed as defined by equ. 2.10 from the real cepstrum $c(n)$ by doubling the causal part and suppressing the anti-causal part, with N being the DFT size:

$$\hat{c}_{min}(n) = \begin{cases} c(n) & n = 0, N/2 \\ 2 \cdot c(n) & 1 \leq n < N/2 \quad \text{(the causal part)} \\ 0 & N/2 < n \leq N-1 \quad \text{(the anti-causal part).} \end{cases} \tag{2.10}$$

A signal $s(n)$ can be regarded as a minimum phase signal if all its poles and zeros lie within the unit circle [Oppenheim et al., 1999]. A minimum phase signal can be created by moving all zeros outside the unit circle $|z_i| > 1$ to lie inside unit circle $|z_i| < 1$. This operation simply requires to replace all z_i having $|z_i| > 1$ by its conjugate reciprocal $1/z_i$ [Oppenheim et al., 1999, Smith, 2007, p.281]. The operation is based on one property of the complex cepstrum:
- Each minimum phase zero in the spectrum rises a causal exponential in the cepstrum.
- Each maximum phase zero in the spectrum rises an anti-causal exponential in the cepstrum.

An example to compute the minimum phase representation of a signal without affecting its spectral magnitude is given by the following Matlab excerpt. It replaces after the transformation of the signal to the cepstrum the anti-causal components by the causal ones. With this, the original spectral phase is replaced by the minimum phase expression according to the spectral magnitude of the signal.

```
% Transform to the cepstral domain
compMagSpectrum = log(abs(mixedPhaseSpectrum));
cepstralSpectrum = real(ifft(compMagSpectrum));

% Filter quefrencies, ignore Nyquist bin
minPhaseCepstrum = fft([cepstralSpectrum(1); 2 * cepstralSpectrum(2:end/2)], fftlen);

% Transform back to the spectral domain
minPhaseSpecHalf = exp(minPhaseCepstrum(1:end/2+1));

% Complete to a full spectrum
minPhaseSpectrum = [minPhaseSpecHalf; conj(minPhaseSpecHalf(end-1:-1:2))];
```

Listing 2.1: Matlab excerpt explaining the transformation from mixed to minimum phase

2.3 Estimation of basic voice descriptors

2.3.1 Fundamental frequency F_0

Numerous methods exists in the literature to estimate the fundamental frequency F_0 from an audio signal. The YIN algorithm [de Cheveigne and Kawahara, 2002] employs as basic F_0 estimation algorithms of one time-invariant signal frame the Auto-Correlation Function (ACF) and the Average Magnitude Difference Function (AMDF). Several post-processing steps improve the peak picking on the time difference lag values τ which parameterize the time distance of the delayed versions of one signal frame. The F_0 estimate is simply the inverse of the estimated time lag τ. ACF and AMDF compare the frame with shifted versions of itself. Both are sensitive to octave and sub-harmonic errors. Applying both methods in the spectral domain improves their noise robustness.

Another F_0 estimation method which has gained popularity recently is the "Sawtooth Waveform Inspired Pitch Estimator" called SWIPE, described in [Camacho, 2007, Camacho and Harris, 2008]. The principle of SWIPE is to find the frequency maximizing the average peak-to-valley distance found at multiples of that frequency. Additionally it applies several normalizing processes to optimize the analysis of the signal content, such as harmonic weighting and the conversion to the ERB scale.

2.3.2 Voiced / Unvoiced Frequency boundary F_{VU}

The Voiced / Unvoiced Frequency boundary F_{VU} is the frequency at which the spectral representation of a signal is split into one deterministic frequency band below and one stochastic frequency band above the F_{VU}. Fig.

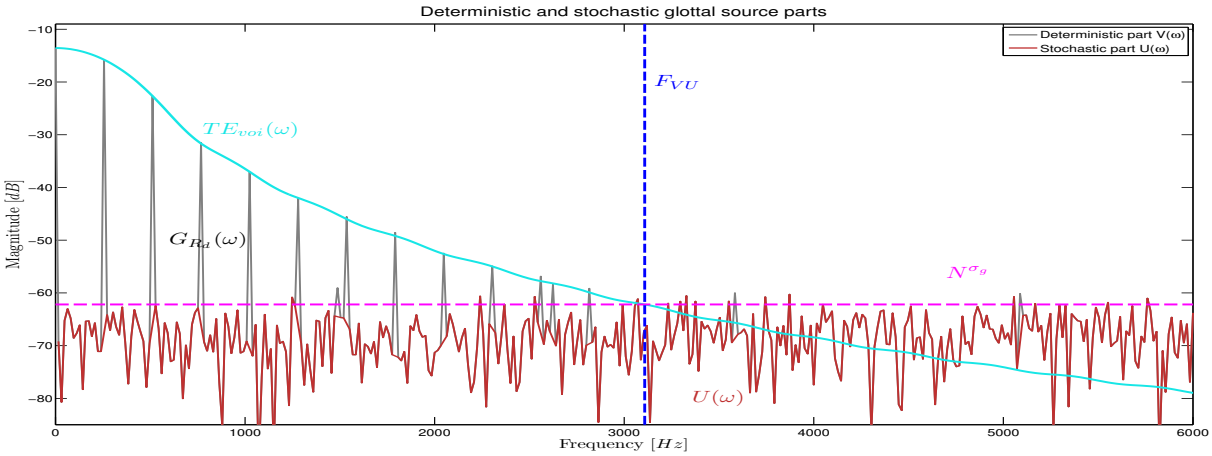


Figure 2.2: *Synthetic example of the glottal excitation source*

2.2 illustrates a signal segment of the glottal excitation source in the spectral domain. The example is a synthetically created signal, using techniques described in [Maeda, 1982, Huber and Röbel, 2013]. It is adapted from [Degottex et al., 2013] with kind permission of the author, and follows the SVLN approach explained in 3.8.4.2. The figure depicts F_{VU} as vertical dashed blue line. The noise level N^{σ_g} in horizontal dashed pink line is kept constant over the complete spectrum in this example. It describes mainly the highest peaks of the unvoiced part $U(\omega)$ of a speech signal, shown in red line in fig. 2.2. This noise component can also be denominated as the stochastic component $U(\omega)$ of a signal.

It can be comprehended from visual inspection of fig. 2.2 that the unvoiced part $U(\omega)$ perceptually masks above the F_{VU} the voiced part $V(\omega)$ which diminishes with higher frequency. In contrast, $V(\omega)$ perceptually masks below the F_{VU} the unvoiced part $U(\omega)$. The noise component $U(\omega)$ is constant with frequency for this example since it is a synthetically generated white noise signal which has not been convolved with a filter.

The estimation of the Voiced / Unvoiced Frequency boundary conducted throughout the work presented here follows the explanations detailed in [Röbel, 2010c]. Basically it divides the spectrum into narrow frequency bands. Each band contains spectral peaks which may be of stable sinusoidal or spurious noisy origin. Sinusoidal peaks are detected following the approach given in [Zivanovic et al., 2004, Zivanovic and Röbel, 2008]. The sinusoidal energy is measured on all extracted peaks being classified as sinusoidal. The Sinusoidal versus Noise Energy ratio (SNE) is measured in sub-bands having a fixed constant bandwidth. The F_{VU} is set to the highest frequency band for which the SNE measure lies above a given threshold θ_{SNE}. This approach to estimate F_{VU} works coherently up to an amplitude difference of ~6-12 dB between the sinusoidal versus the noise level. The measured F_{VU} contour

is smoothed using a median filter covering the time of half the length of the analysis window.

The F_{VU} frequency is furthermore commonly denominated as Maximum Voiced Frequency F_m. Other approaches to estimate F_{VU} or F_m can be found in [Stylianou, 2001, Ciobanu et al., 2012, Drugman and Stylianou, 2014].

2.3.3 Frequency dependent noise level estimation

A sinusoidal peak detection using a frequency dependent threshold is presented in [Every and Szymanski, 2004, Every and Szymanski, 2005, Every and Szymanski, 2006]. The result of a convolution of the utilized analysis window with the amplitude spectrum is divided from the latter. This emphasizes sinusoidal components sharing a similar shape with the spectral representation of the window. The emphasis suppresses noise influences of the stochastic signal part. The division subtracts all sinusoidal content and reveals the noise level contour. The correlation fails if sinusoidal peaks do not resemble the spectral shape of the windowing function. Therefore, an adaptive resonance bandwidth is employed in [Every, 2006] to span the main spectral lobe to different widths.

Another approach to estimate the noise floor is to establish a frequency-dependent noise level estimation. An adaptive threshold determination in [Zivanovic et al., 2004, Zivanovic and Röbel, 2008] is used to classify spectral peaks into stable sinusoids or spurious noisy peaks.

The probabilistic noise model represents in [Yeh, 2008] the noise floor as a frequency dependent spectral envelope. It employs the method of [Zivanovic and Röbel, 2008] as an initial classification of spectral peaks into sinusoids or noise. A further refinement assumes that the spectral magnitude of spurious noisy peaks follows one mode of a Rayleigh distribution. The Rayleigh distribution describes with different standard deviation modes the spectral distribution of filtered white noise. The magnitude distribution of each narrow band is fitted to the best matching Rayleigh distribution mode σ. A Probability Density Function of spectral peak magnitudes x is given by equ. 2.11. The probability of an observed peak to be a spurious noise component is high for magnitude peaks $x < \sigma$ and low for $x > \sigma$.

$$p(x) = \frac{x}{\sigma^2} \cdot e^{-\frac{x^2}{2\sigma^2}} \forall 0 \leq x < \infty, \sigma > 0 \tag{2.11}$$

2.4 Signal Models for Speech Processing

The processing of a speech spectrum $S(\omega)$ requires a model reflecting its intrinsic properties. The discussion given in the preceding sections 2.3.2 and 2.3.3 concerning the assembly of a speech spectrum $S(\omega)$ by the voiced deterministic $V(\omega)$ and the unvoiced stochastic $U(\omega)$ spectral parts suggests to model both components independently. Numerous speech models were proposed over time by the speech community. The models can be classified into families treating the sinusoidal content $V(\omega)$ and the noise content $U(\omega)$ by different means. Each model should adapt its feature dimensionality to the characteristics of the sound character. Voice transformations like pitch transposition require further modelling such as spectral envelope estimation for shape-invariant speech processing [Röbel, 2010b].

2.4.1 Sinusoidal modelling

The sinusoidal representation of speech in [McAulay and Quatieri, 1986] encodes a speech waveform with the following set of parameters, estimated on each sinusoid k :
- The amplitude A_k,
- the instantaneous frequency f_k,
- and the instantaneous phase ϕ_k.

The sinusoidal model of equ. 2.12 describes the voiced part of the speech signal $s(t)$ with the time-varying amplitudes $A(t)$ and the time-varying phases $\phi(t)$. Each time-varying phase $\phi_k(t)$ is described by $\omega_k(t) = 2\pi f_k t$, located at its corresponding instantaneous frequency $f_k(t)$, and shifted by the initial phase θ_k.

$$x(t) = \sum_{k=1}^{K} A_k(t) \cdot \cos(\phi_k(t)) = \sum_{k=1}^{K} A_k(t) \cdot e^{j(\omega_k(t)+\theta_k)} \tag{2.12}$$

The signal $s(n)$ is thus represented by a sum of K sinusoids. The localization of the sinusoidal amplitude peaks can be restricted to quasi-harmonic frequency intervals present in the Discrete Fourier Transform (DFT) spectrum [Almeida and Silva, 1984]. A simple peak-picking algorithm for the analysis of sinusoidal harmonic content

detects the presence of sinusoids [Serra, 1989, Serra, 1997]. Parabolic interpolation [Bonada, 2000] on the neighbouring DFT bins fits a parabola on each sinusoid k to retrieve its amplitude A_k and its instantaneous frequency f_k. Each instantaneous phase ϕ_k is obtained by linear interpolation of the phase spectrum [Serra and Smith, 1990].

2.4.2 Multi-Band Excitation (MBE)

The discussion of sections 6.2.3.3 and 2.3.3 suggest that one spectral frame of a speech signal may contain several F_{VU} boundaries. This signal interpretation is reflected by the Multi-Band Excitation (MBE) speech model proposed in [Griffin, 1987, Griffin et al., 1988]. MBE models speech as a product of an excitation spectrum $G(\omega)$ and a spectral envelope $\mathcal{T}_{sig}(\omega)$. Each excitation spectrum is specified by the instantaneous frequency f_k and its instantaneous phase ϕ_k per sinusoid k, and a frequency dependent voiced / unvoiced mixture function. MBE assumes that the spectrum consists of interleaved consecutive narrow-bands being either dominated by harmonic periodic and random aperiodic signal content. A decision constraint is therefore necessary to classify each band into harmonic or noise. Voiced bands contain comparably higher energies than unvoiced bands having lower energies. The division of a Short-Time Fourier Transform (STFT) [Griffin and Lim, 1983, Griffin and Lim, 1984] spectrum into multiple frequency bands centered around pitch harmonics reduces the number of parameters [Chan and Hui, 1996]. This allows the application of the Multiband Excitation Vocoder as an audio encoder for bit-rate reduction [Griffin et al., 1988]. The Multi-Band Excitation model is fully parametric. It allowed for high quality speech coding and provided a good synthetic voice quality at the time of publication. The well-known speech analysis, modification and synthesis system STRAIGHT, further discussed in section 2.5.1, constructs in [Kawahara et al., 2001] a mixed mode excitation signal for synthesis by controlling relative noise levels and the temporal envelope of the noise component.

[Dutoit and Leich, 1993] present an evaluation of the MBE model performances within the context of applying it to a High Quality Text-To-Speech synthesis. The analysis accuracy of MBE is influenced by the varying frequency and amplitude of single sinusoids over time since a constant frequency and amplitude is assumed during the complete analysis frame. Minor pitch changes and major amplitude variations affect the synthesis quality mostly in high frequency regions. It is perceived as high frequency noise, being typical for synthetic speech.

2.4.3 The Deterministic plus Stochastic Model (DSM)

The restriction to a quasi-harmonic modelling of a speech signal implied by the sinusoidal modelling of the preceding section 2.4.1 neglects the noise component $U(\omega)$. According to [Griffin et al., 1988], noise can be modelled by amplitude modulated Gaussian noise, convolved with the auto-regressive envelope of the sinusoidal part. The Spectral Modeling Synthesis (SMS) of [Serra, 1989, Serra and Smith, 1990, Serra, 1997] is based on a Deterministic plus Stochastic Decomposition. It is one of the first models introducing the modelling of the noise part [Röbel, 2010a].

The DSM speech model presented in [Stylianou, 1996] is a variant of the Harmonic plus Noise Model [Laroche et al., 1993b] of the following section 2.4.4. The deterministic part of DSM is defined by harmonically related sinusoids with piece-wise linearly varying complex amplitudes. The stochastic part is deduced as the residual signal received from subtracting the deterministic part from the spectrum $S(\omega)$. The residual contains consequently all errors received from a not precise estimation of the deterministic part. The spectral subtraction of the estimated deterministic part from $S(\omega)$ is applied over the whole spectrum. DSM operates thus full-band. Additionally it is robust against F_0 estimation failures since it allows for minor frequency deviations.

2.4.4 The Harmonic plus Noise Model (HNM)

The additional processing of the noise component $U(\omega)$ is as well reflected in the Harmonic and Noise Model (HNM). It was introduced in [Laroche et al., 1993b, Laroche et al., 1993a] and further exercised in [Stylianou et al., 1995, Stylianou, 1996, Stylianou, 2001]. An HNM assumes that $U(\omega)$ can be observed above the F_{VU} or F_m frequency, as shown in the sections 2.3.2 and 2.3.3. The spectrum is divided into a lower frequency band for sinusoidal and a higher frequency band for noise like content. The harmonic part of an HNM equals the deterministic part of DSM with zero spectral slope [Stylianou, 1996]. The noise part of an HNM is defined for all frequencies above the F_{VU} up to the Nyquist frequency.

[Saratxaga et al., 2010] bases a Harmonic plus Noise model on the Multiband Excitation Model discussed in section 2.4.2. It is extended with a novel phase information representation called Relative Phase Shift (RPS). The phase control techniques of RPS, proposed in [Saratxaga et al., 2009], provides structured phase patterns of the

harmonic components. It simplifies the manipulation of the perceptually important signal information contained in the phase part.

2.4.5 The Harmonic plus Stochastic Model (HSM)

Another variant of the HNM speech model of section 2.4.4 is the Harmonic plus Stochastic Model of [Stylianou, 1996]. HSM overlaps the sinusoidal harmonic part up to the Maximum Voiced Frequency F_m on the random noise part. The standard implementation of HSM is modified in [Erro and Moreno, 2007, Erro, 2008] to allow for phase manipulation procedures designed to work in pitch-asynchronous mode. This harmonic plus stochastic model variant is used to analyze, modify and synthesize the speech signal within the context of Voice Conversion for works presented in [Erro, 2008, Erro et al., 2010b].

2.4.6 The Quasi-Harmonic Model (QHM)

The Quasi-Harmonic Model (QHM), introduced in [Pantazis et al., 2008], addresses the sensitivity of sinusoidal models to frequency estimation errors. A frequency estimator within QHM corrects erroneous frequency estimations. The time-varying sinusoidal representation of QHM is thus robust against frequency mistakes [Pantazis et al., 2010b]. The sinusoidal content can in most cases be reduced to a set of quasi-harmonic frequencies. This applies if the analyzed sound segment contains a periodic waveform, as for example within voiced sound segments. The QHM describes the deterministic quasi-harmonic part of an analyzed audio segment as

$$s(t) = \left(\sum_{k=1}^{K} (a_k + t b_k) \cdot e^{j2\pi f_k t} \right) w(t), \tag{2.13}$$

with K components having complex amplitude a_k and complex slope b_k at frequencies f_k, windowed at time t with analysis window $w(t)$. The iterative estimation of the observed frequency interval in QHM is biased. The bias depends on the type and the length of the STFT analysis window w. QHM can correct frequency mismatches up to 135 Hz for a Hamming window of 16 ms. The bias is negligible small if the frequency error remains below one third of the bandwidth B of the squared analysis window [Pantazis et al., 2010b]. The adaptive Quasi-Harmonic Model (aQHM) of [Pantazis et al., 2010a, Pantazis et al., 2011] adapts its frequency basis to the time variations of the frequency components. This non-stationary frequency basis relieves the aQHM system from the restriction of the stationarity assumption to be harmonically related. The frequency adaptation of aQHM changes the QHM defined with equ. 2.13 to

$$s(t) = \left(\sum_{k=1}^{K} (a_k + t b_k) \cdot e^{j\hat{\phi}(t)} \right) w(t), \tag{2.14}$$

with $\hat{\phi}(t)$ being the instantaneous phase $\hat{\phi}(t) = 2\pi \int_0^t f_k(u)du$. Further extensions of this modelling approach can be found in [Kafentzis et al., 2012] introducing the extended adaptive Quasi-Harmonic Model (eaQHM), and in [Degottex and Stylianou, 2013] presenting the full-band adaptive Harmonic Model (aHM).

2.4.7 Extended noise and residual models

Recent research in the speech community has notably improved the speech synthesis quality by explicitly modelling as in [Drugman and Dutoit, 2012, Cabral and Carson-Berndsen, 2013] the deterministic and / or the stochastic component of the glottal excitation source, along with other voice descriptors.

[d'Alessandro et al., 1998] presents the decomposition of a speech signal into its periodic and aperiodic (PAP) components. The study investigates into the influences of additive random noise and modulation aperiodicities introduced by variations of F_0, Jitter and Shimmer. The PAP decomposition operates sufficiently robust for a wide range of F_0 variations and if larger amounts of Jitter and Shimmer are not present.

The time envelopes Triangular envelope, Hilbert envelope and Energy envelope model the temporal characteristics of random noise in an HNM model within an analysis / re-synthesis scheme on natural human speech in [Pantazis and Stylianou, 2008]. The energy envelope is modelled by a Fourier series with few harmonics. It approximates well to the energy distribution of the noise part and can be easily manipulated for pitch and time-scale modifications. It achieved the highest rating in a listening test.

The Deterministic plus Stochastic Model (DSM) for residual excitation of [Drugman and Dutoit, 2012] divides the residual spectrum after the deduction of the harmonic sinusoidal part into two distinct spectral bands, delimited

by the Maximum Voiced Frequency F_m. The deterministic component represents low frequency contents. The stochastic component consists of high pass filtered noise. An energy envelope modelled by the Hilbert transform modulates the time structure representing the interference of the fundamental periodicity originated at glottis level. The DSM of the residual approach aims at a compact representation of the excitation source to reduce the buzziness produced with parametric speech synthesizers. [Drugman and Dutoit, 2012] reports a similar synthesis quality being as good as the one achieved by the speech processing system STRAIGHT of section 2.5.1.

2.5 Other Models Signal for Speech Processing

Numerous other methodologies exist in the literature to alter speech by different means. The two in the following presented software framework are based on different means to conduct speech processing. The system STRAIGHT of section 2.5.1 relies on an advanced spectral modification scheme including a source-filter separation. The system SuperVP of section 2.5.2 is based on an extended phase vocoder version.

2.5.1 STRAIGHT

The STAte-of-the-ART speech analysis, modification and synthesis framework STRAIGHT [1] is freely available for research purposes and thus used in many scientific papers for comparison purposes. It is based on a source / resonator decomposition and allows for many advanced speech processing tasks with high quality [Kawahara, 1997, Kawahara et al., 1999, Kawahara et al., 2001]. The abbreviation STRAIGHT derives from "Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum".

A reformulation of STRAIGHT presented in [Kawahara et al., 2008] called TANDEM-STRAIGHT combines the estimation of the power spectrum without interfering temporal variations (TANDEM) with the spectral envelope estimation (STRAIGHT). STRAIGHT estimates the amplitude envelope non-iteratively in the frequency and the time domain using a smooth interpolation of spectral peaks. The representation of spectrum $S(\omega)$ in STRAIGHT is F_0-adaptive containing a surface reconstruction method in a combined time-frequency region. The included LF model and a phase manipulation method permits transforming the deterministic part of the glottal excitation source. Its corresponding stochastic part is generated by modulating and colouring white noise.

2.5.2 SuperVP

IRCAM's SuperVP software library [Liuni and Röbel, 2013] provides an extended version [Röbel, 2010b] of the standard phase vocoder of [Flanagan and Golden, 1966] and its further advancements like in [Puckette, 1995, Kawahara, 1997, Laroche and Dolson, 1999, Laroche, 2003]. Its executable implementation is accessible via command line from Matlab and Python environments [2]. SuperVP provides the basic means to execute many speech signal processing tasks required throughout the work here presented. In extension to speech processing applications, it facilitates the computation of other digital signal processing tasks for audio in general.

2.6 Spectral envelope estimation techniques

A constitutional overview of spectral envelope estimation techniques can be found in [Schwarz, 1998, Schwarz and Rodet, 1999]. Spectral envelope estimation determines the maximal amplitude per overlapping narrow-band frequency region found for one signal frame. The shape of the spectral envelope is described by the sequence of quasi-harmonic sinusoidal partials in magnitude and frequency below F_{VU}. Spurious harmonics and the noise floor describe the spectral shape above F_{VU}. The spectral envelope is approximated by a smoothing function which passes through prominent spectral peaks. The modelling of the spectral envelope suffers from aliasing effects caused by a non-accurate enough estimation. The spectral envelope is under-fitted if it doesn't pass through prominent spectral peaks caused by stable harmonic sinusoids. It is over-fitted if it passes through non-sinusoidal peaks originating from window sidelobes, spurious and noisy peaks etc.

[1] STRAIGHT: http://www.wakayama-u.ac.jp/kawahara/STRAIGHTadv/index_e.html
[2] SuperVP: http://anasynth.ircam.fr/home/english/software/SuperVP

2.6.1 Linear Predictive Coding (LPC)

Early approaches to estimate the spectral envelope were based on the Linear Predictive Coding (LPC) model [Makhoul, 1975, Vaidyanathan, 2008]. LPC provides a parametric estimation of the spectral envelope of a signal frame. It assumes the source-filter model of human voice production [Fant, 1981]. LPC models the vocal tract as an IIR filter by estimating the signal $y'(n)$ based on current samples $y(n)$ to minimize the prediction error $e(n)$ = $y'(n)$ - $y(n)$. Deriving the linear prediction model involves determining the filter coefficients a_0, a_1, ..., a_n in feed-forward and b_1, b_2, ... b_n in feed-backward direction. The difference equation 2.15 expresses the relation of the LPC filter coefficients a and b, the input samples x, the output samples y and the estimated output samples y':

$$y'(t) = \sum_{i=0}^{M} a_i \cdot x(t-i) - \sum_{j=1}^{N} b_j \cdot y(t-j) \tag{2.15}$$

The a-coefficients are commonly set to zero in linear prediction. Only previous output samples and the feed-backward filter coefficients b determine the estimation of the output signal. This difference equation of 2.15 is inserted into the auto-regressive all-pole model of equ. 2.16 to define the impulse response of the vocal tract:

$$H(z) = \frac{1}{1 - \sum_{j=1}^{N} b_j \cdot y(t-j)} \tag{2.16}$$

An auto-correlation of the signal frame with a delay version of itself is executed to predict the filter coefficients and to accurately approximate the real output. The LPC coefficients describe the impulse response of the VTF and with this the spectral envelope of an acoustic signal. The pole locations which correspond to the VTF formants can be derived by factoring the LPC coefficients. The pole angles define the formant frequencies. The LP residual signal represents the glottal source as the excitation source signal. It includes the fundamental frequency $F0$ and a stochastic noise component. The LPC order denotes the number of poles in the filter. Usually two poles are utilized to describe one formant. One problem of applying LPC is to choose the right order of the all-pole model such that the possibility of ill-conditioning is minimized [Makhoul, 1975]. Even the correct order may due to aliasing not achieve the desired perfect estimation of the spectral envelope. The Akaike Information Criterion (AIC) is used as order selection criterion in [de Waele and Broersen, 2003] to avoid a sub-optimal performance in using an AR model which could result in over- or under-fitting. The Mean Square Error (MSE) minimization of LPC is prone to produce a systematic error as a bias of the spectral peaks towards the harmonics [Villavicencio et al., 2007].

2.6.2 Line Spectral Frequencies (LSF)

The amplitudes, frequency positions and bandwidths of spectral formants of a speech signal are caused by the cavities of the physical vocal tract. The formant structure contributes to the perception of a speakers voice identity. However, the estimation of single formants [Snell and Milinazzo, 1993] can be erroneous and cumbersome [Helander et al., 2007]. The LPC analysis of the preceding section 2.6.1 works more reliable. The LSF coefficients can be directly derived from LPC coefficients. Additionally, the spectral envelope representation by means of Line Spectral Frequencies (LSF) provides a good performance for the transformation and interpolation of spectral envelopes [Paliwal, 1995, En-Najjary et al., 2003]. LSFs are therefore commonly used in VC systems [Percybrooks and Moore, 2007, Helander et al., 2008a, Hanzlíček and Matoušek, 2008]. LSF models narrow spectral formants with a higher precision and robustness than LPC [Bäckström, 2004]. The Line Spectral Pair (LSP) transformation has more powerful interpolation properties and is more robust against quantization errors, at the cost of higher computational complexity [Bäckström and Magi, 2006].

The auto-regressive model assumes that a signal is generated as an all-pole filter output $H(z)$ = 1 / $A(z)$, with $A(z)$ = $1 + a_1 \cdot z^{-1} + ... + a_p \cdot z^{-p}$ and p being the LPC analysis order. A LSP transformation decomposes the p^{th}-order linear predictor A(z) into a symmetrical and anti-symmetrical part, denoted by the polynomials $P(z)$ and $Q(z)$. The LSPs are the filter roots or zeroes on the z-plane. They determine uniquely the two polynomials $P(z)$ and $Q(z)$.

Line Spectrum Pairs:
$P(z) = A(z) + z^{-p+1} \cdot A(z^{-1})$
$Q(z) = A(z) - z^{-p+1} \cdot A(z^{-1})$

The LSPs lay on the unit circle in interlaced order. Zeroes on the unit circle are represented as vertical lines in the power spectrum. An all-pole model converts the zeroes into poles. The linear predictor $A(z)$ is expressed in terms of $P(z)$ and $Q(z)$ as $A(z) = \frac{1}{2} \cdot (P(z) + Q(z))$. The Line Spectrum Frequencies are their corresponding angular frequencies. The angular frequency properties are exploited for the calculation of characteristic spectral shapes. LSFs are according to [Erro, 2008] very well prone for the spectral representation and parameterization of speech data for Voice Conversion, due to the following properties:

- Good formant representation
- Good interpolation properties
- One wrongly converted coefficient affects merely a restricted data part
- Cepstral coefficients are parameterized in A(z)
- Inverse transformation of LSF coefficients into all-pole filters provides the filter response of the target speakers spectral envelope

2.6.3 Discrete All-Pole (DAP)

The Discrete All-Pole (DAP) model [El-Jaroudi and Makhoul, 1991] solves the Linear Prediction (LP) aliasing problem. It fits the all-pole model with a finite set of spectral locations, which are related to the harmonic frequency bins. DAP estimates the spectral envelope with lower distortion with respect to the harmonics, but suffers from filter instabilities. According to [Erro, 2008, Villavicencio et al., 2009] it is less robust in adapting the model order according to the harmonic peak information.

2.6.4 True Envelope ($\mathcal{T}$)

The cepstrum-based True Envelope $\mathcal{T}$ estimator of [Villavicencio et al., 2006, Röbel et al., 2007] optimally interpolates the observed spectral peaks. It applies iterative cepstral smoothing of the amplitude spectrum. With this it achieves the best band-limited interpolation of the major prominent spectral peaks. The $\mathcal{T}$ cepstral order can be adapted to the spectral characteristics for each frame. It defines at which point the real cepstrum representation of the spectrum is truncated. The iteration over the maximum of the original logarithmic spectrum and the previous cepstral representation lets the estimated envelope steadily grow until the true envelope contour is reached.

The LPC-based True-Envelope estimator $\mathcal{T}_{LPC}$ reduces the model mismatch of LPC producing over-smoothing and aliasing artefacts. $\mathcal{T}_{LPC}$ uses the $\mathcal{T}$ estimation as a target spectrum for the LPC auto-correlation matching criteria. A high order all-pole model enables the spectrum representation as LSF coefficients. The $\mathcal{T}$ cepstral coefficients can thus as well be transformed into an LSF representation [Villavicencio et al., 2006]. $\mathcal{T}_{LPC}$ uses local order selection to perform the $\mathcal{T}$ estimate, approaches faster to the minimum $\mathcal{T}$ error, and reduces the over-fitting effect by learning the order of the feature dimensionality.

In [Villavicencio et al., 2007] the optimal cepstral order selection is discussed in order to provide an envelope estimation with minimum error. The harmonic excitation spectrum samples the resonator filter with a sampling rate set to F_0. With this a nearly optimal cepstral order is defined depending on the maximum frequency difference between two relevant spectral peaks. Standard models provide an estimation of the envelope with a minimum error in dependence of the evaluated spectrum. Order selection adopts to the examined spectrum and has a resulting error close to the optimal order. The optimal estimation depends thus on the fundamental period and the envelope characteristics. Improvements in [Villavicencio et al., 2008] lead to the Mel-Frequency True-Envelope Linear Predictive Coding.

2.7 Conclusions

The explanations of section 2.1 how the human voice production system generates speech is inevitable to understand the means implemented for this thesis work to apply speech processing. Some basic properties of discrete-time audio signals and some basic voice descriptors are introduced in the sections 2.2 and 2.3 to comprehend further works discussed in the following chapters. The speech signal models of section 2.4 give further inside in the theoretical basis leading to the new speech model presented in chapter 6. The estimation of the spectral envelope of a speech signal discussed in section 2.6 is crucial for the VTF extraction being utilized in this thesis work.

Chapter 3

STAte-of-the-ART (START) in glottal excitation source modelling

And in the moment when the sound of Om touched Siddhartha's ear, his dormant spirit suddenly woke up and realized the foolishness of his actions.

Hermann Hesse - Siddhartha

3.1 Introduction

This chapter introduces the basic STAte-of-the-ART paradigms found in the literature to conduct the analysis, transformation and synthesis of the glottal excitation source. Further means to analyse the glottal excitation source achieved by the work for this thesis will be presented in the following chapter 5. The modelling of the glottal excitation source is required to alter the voice quality of a speech phrase. A voice quality transformation is beneficial in the Voice Conversion context to further transform the source into the target speakers voice identity. Means to transform the voice quality of a speaker will be presented in chapter 6.

3.2 The glottal excitation source

3.2.1 Time domain properties of glottal source shapes

The glottal excitation source consists of a deterministic and a stochastic component. Both are caused by the glottal flow, the air flow coming from the lungs and being modulated by the glottal area. Different modulation types lead to different vibration modes of the vocal folds. This in turn causes the different shapes of the deterministic part of the glottal excitation source. Fig. 3.1 depicts one characteristic shape example of the glottal flow and its derivative in the time domain. Table 3.1 explains the time instants given in fig. 3.1 which characterize the glottal flow and its derivative.

Open phase:

The open phase starts at time instant t_s=0 when the glottis starts to open. It is denominated the Glottal Opening Instant (GOI). The glottal opening occurs due the displacement of the vocal folds caused by the sub-glottal pressure from the lungs, depicted in fig. 2.1. The linear source-filter theory [Fant, 1960] states that airflow from the lungs passes due to the sub- and supra-glottal pressures the glottis and the vocal tract, until the open phase ends at time instant t_e [Fant, 1981]. However, the non-linear system behaviour of the underlying acoustic tube implies that the airflow may as well occur in opposite direction due to acoustic coupling [Titze et al., 2008]. The latter depends on the level of the glottal impedance and leads to oscillations of the airflow. The time instant t_e corresponds to Glottal

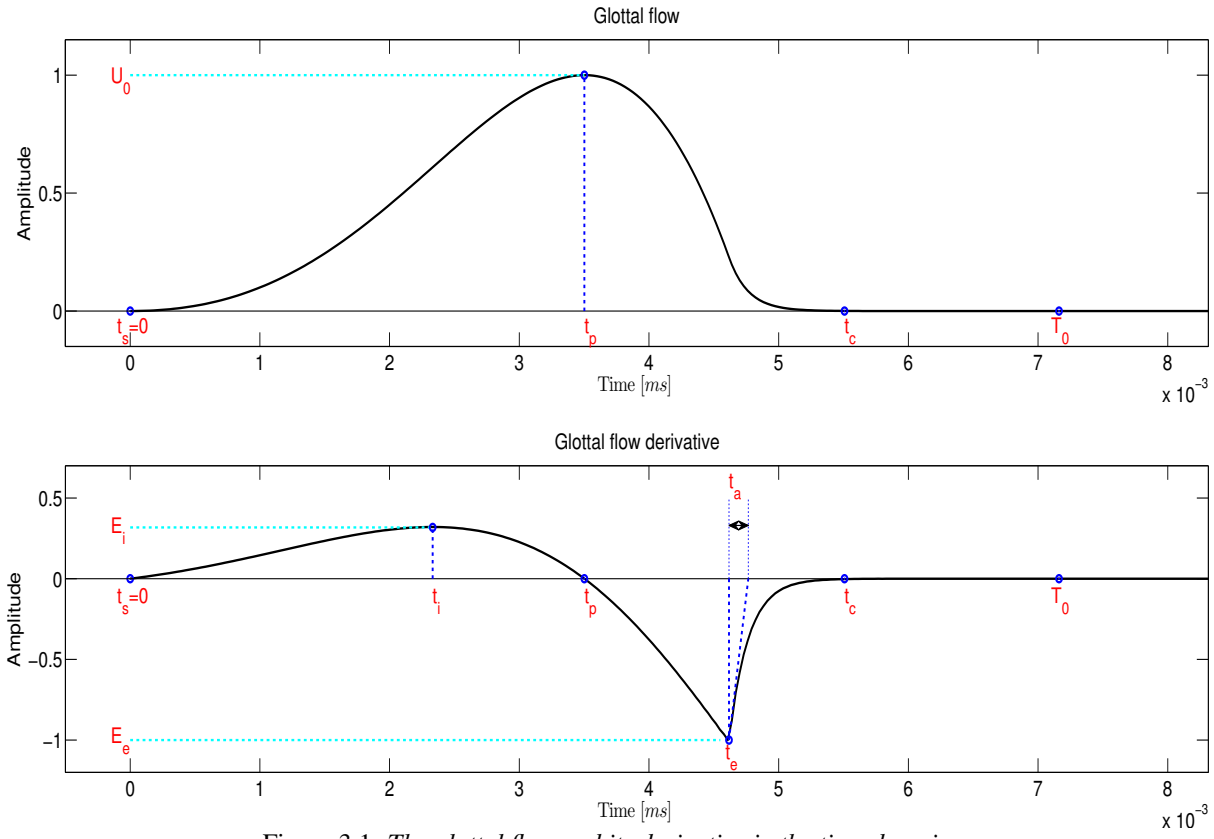


Figure 3.1: *The glottal flow and its derivative in the time domain*

Table 3.1: *Time instants of the glottal flow and its derivative*

Time instant	Description
t_s	Time of the start of the glottal flow and its derivative
t_i	Time of the maximum of the derivative
t_p	Time of the maximum of the glottal flow; the derivative crosses the x-axis
t_e	Time of the minimum of the derivative, corresponding to the GCI and the Open Quotient OQ
t_a	The effective return phase duration
t_c	Time of the end of the pulse; the glottis is completely closed
T_0	Time length of the whole pulse period

Closure Instant (GCI). It is the time instant when the glottal flow decreases at fastest speed and reaches maximum closure speed [d'Alessandro, 2006] such that the derivative of the glottal flow exhibits a negative maximum.

Return phase:
The vocal folds return to their initial relaxed displacement during the return phase t_a. The effective duration of t_a is expressed by the difference of t_e and the point in time where the projection of the derivative of the exponentially shaped return phase contour strikes the time axis, illustrated in fig. 3.1. The derivation of the return phase contour to determine its tangent can be deduced from the definition $E_2(t)$ given in equ. 3.1. The return phase duration t_a is a direct correlate to a first order low pass filter being active in higher frequency regions [Drugman et al., 2011]. The low pass filter models the spectral tilt of spectrum $S(\omega)$ [Doval et al., 2006]. Longer t_a durations express higher levels of attenuation in higher frequency regions and a less abrupt transition to the closed phase [Fant, 1995].

Closed phase:
The closed phase constitutes the time period when the glottis is completely closed. It reflects the time span between the end of the pulse t_c and the whole pulse period T_0.

Characteristic amplitudes:
The positive amplitude maximum U_0 of the glottal flow at time instant t_p is related to the amplitude of the voice fundamental [Fant, 1995]. The negative amplitude minimum E_e at time instant t_e is a basic determinant of formant amplitudes [Fant, 1995]. It relates to the harmonic sinusoidal amplitudes found at higher frequencies than the voice fundamental and correlates thus to the spectral tilt [Fant, 1997]. The time instant t_e corresponds to the GCI and is commonly known as the time instant of maximum excitation or maximal discontinuity [Murthy and Yegnanarayana, 1999]. The total sound pressure and the amplitudes of the vocal tract formants are

highly influenced by E_e [Fant et al., 2000]. Higher negative amplitudes E_e of the glottal flow derivative generate higher Sound Pressure Levels (SPL) of the resulting speech waveform [Alku et al., 1999]. The sub-glottal pressure reaches its maximum while the supra-glottal pressure attains its minimum such that the trans-glottal pressure is very high [Titze, 1984]. The glottal flow exhibits a maximum at time instant t_p at which its derivative is consequently zero and crosses therefore the time axis.

Please find a further discussion on the spectral correlates of these characteristic time instants and amplitude measures parameterizing the glottal excitation source in [Doval and d'Alessandro, 1999, Doval et al., 2006, d'Alessandro et al., 2007].

3.2.2 Glottal excitation source models

Many different models to describe the deterministic part of the glottal excitation source have been proposed by the speech community over the last decades. Such glottal source models describe the shape of the glottal flow and its derivative by parameterizing its characteristic time instants and amplitude values, introduced in the preceding section 3.2.1. A modification of such features causes variations in the re-synthesized speech signal like an altered length of the open phase of a pitch period or a modified spectral tilt in the spectral representation of a speech signal. This allows conducting voice quality transformations. Voice quality will be introduced in section 3.5. A transformation of voice quality in terms of vocal effort will be discussed in section 6.4.3.

Among others, the following glottal source models can be found in the literature:

- the R-Model by Rosenberg [Rosenberg, 1971],
- the F-Model by Fant [Fant, 1979],
- the LF-Model by Liljencrants and Fant [Fant et al., 1985],
- the FL-Model by Fujisaki and Ljungqvist [Fujisaki and Ljungqvist, 1986],
- the K-Model by Klatt and Klatt [Klatt and Klatt, 1990],
- the R++-Model by Veldhuis [Veldhuis, 1998],
- the CALM-Model by Doval [Doval et al., 2003],
- the H-Model by Hezard [Hezard et al., 2012],
- and a perceptually and physiologically motivated model proposed by Chen in [Chen et al., 2013a].

The listed models share a high similarity among each other since each of them describes the same signal type, the shape of a glottal pulse. All describe the time instant t_e and accordingly the Open Quotient OQ in relation to the fundamental period T_0. Differences appear for example in the modelling of the return phase. The chosen model for this work is the transformed LF model [Fant, 1995], introduced in the following section.

3.2.3 The LF glottal source model

The well-known glottal model of Liljencrants and Fant (LF) describes the shape of the deterministic part of the glottal excitation source [Fant et al., 1985]. The LF model parameterizes in particular the time derivative $g(t)$ of the glottal flow, depicted in fig. 3.1.

$$\text{g(t)}= \begin{cases} E_1(t) = E_0 \cdot e^{\alpha t} \cdot sin(\omega_g t) & \forall \;\; 0 < t <= t_e \;\; \text{(Open phase)} \\ E_2(t) = -E_0 \cdot \frac{1}{\epsilon t_a} \cdot \left[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)} \right] & \forall \;\; t_e < t < t_c \;\; \text{(Return phase)} \end{cases} \tag{3.1}$$

Equ. 3.1 defines the LF model which splits the modelling into two separate parts. The open phase is modelled by an exponential function being modulated by a sinusoid whose frequency is given by ω_g=$2\pi/t_p$. The return phase is modelled by another exponential function being parameterized by the time of the return phase t_a. The latter is the projection of the derivate of $E_2(t)$ at time instant t_e on the time axis. The return phase t_a constitutes a first order low pass filter with cutoff frequency F_a=1 / ($2\pi \cdot t_a$). The four LF parameters t_p, t_e, t_a and E_e uniquely determine the shape of a glottal pulse synthesized with the LF model. The requirement of area balance defines the condition that the net gain of flow equals zero within a fundamental period: $\int_0^{T_0} g(t)dt = 0$ [Fant et al., 1985]. The LF model parameters α and ϵ have to be derived from the three time instants t_e, t_p and t_a [Gobl, 2003, Degottex, 2010].

The transformed LF model of [Fant, 1995] normalizes the time instants t_p, t_e, t_a by the fundamental period T_0 to find the three R waveshape parameters R_g, R_k and R_a of the LF model. The R waveshape parameters R_a, R_k and R_g can as well be utilized to parameterize the LF model [Fant et al., 1994]. Table 3.2 lists the R waveshape parameters describing one synthetic LF glottal pulse. The R waveshape parameters are defined in terms of the characteristic time instants listed in table 3.1.

Table 3.2: *The R waveshape and characteristic glottal pulse parameters*

Denomination	Definition	Description
R_g	$T0/(2 \cdot t_p)$	The rising speed of the glottal pulse (T0-normalized)
R_k	$(t_e - t_p)/t_p$	The symmetry of the glottal pulse
R_a	$t_a/T0$	The duration of the return phase (T0-normalized)
OQ	$t_e/T0$	The duration of the open phase (T0-normalized)
α_m	t_p/t_e	The skewness of the glottal pulse

R_g is an inverse measure of the glottal pulse rise time [Fant et al., 2000].
R_a is a relative measure of the return phase duration which is correlated to the spectral tilt of natural human speech. R_k is defined by the ratio of the decay time to the rise time of the glottal flow and describes its asymmetry [Fant, 1997]. It is the inverse of the Speed Quotient SQ=$t_p/(t_e - t_p)$.

SQ expresses the ratio between opening and closing phase [Doval et al., 2006]. The Open Quotient OQ reflects the time duration of the open phase until time instant t_e. OQ expresses the time difference between GOI and GCI, normalized by the fundamental period T_0. The GOI equals time instant zero t_s=0 when the glottal flow or its derivative starts to rise above zero. The GCI equals time instant t_e when the glottal flow derivative reaches the negative minimum E_e. The asymmetry coefficient α_m expresses the skewness of the glottal pulse. The value of α_m is determined by the ratio of the time instants t_p and t_e. Lower (higher) values of α_m describe impulse-like (sinusoidal-like) glottal pulse shapes and reflect with this their degree of asymmetry [Doval et al., 2006]. The synthesis of an LF model requires additionally the fundamental frequency F_0, and the parameter E_e being the maximum negative amplitude of the glottal flow derivative measured at time instant t_e. The LF model is with this defined in total by five parameters.

The glottal formant F_g constitutes the representation of a glottal pulse in the spectrum. Bandwidth and frequency of a glottal formant F_g are controlled by α_m and OQ. The analysis of the perceptual relevance of the R waveshape parameters in [van Dinther, 2003, van Dinther et al., 2004] uses perceptual distance measures in listening tests to examine the required amount of variation to reach a JND.

3.3 Efficient glottal source parameterization

Voice qualities with a tense (pressed), modal (normal), or relaxed (breathy) phonation type are distinguishable by different shapes of the deterministic part of the glottal excitation source. Glottal source shapes can be efficiently described by one-dimensional parameterization techniques like the R_d parameter introduced in [Fant et al., 1994, Fant, 1995], or the Normalized Amplitude Quotient NAQ proposed in [Alku et al., 2002a].

3.3.1 LF regression parameter R_d

The parameter space of the Liljencrants and Fant (LF) model [Fant et al., 1985] is defined by its three R waveshape parameters R_g, R_k, and R_a. The LF shape parameter R_d reduces this parameter space to a one-dimensional parameterization [Fant, 1995]. The parameterization of R_d is determined by means of a statistical linear regression on values of the R waveshape parameter set [Fant and Liljencrants, 1994, Fant et al., 1994, Fant, 1995] that were observed for voiced natural speech in [Gobl, 1988, Karlsson, 1990]. It is based on exploiting systematic co-variations of the R waveshape parameters present in the studied speech corpora. The R_d regression curve of the LF shape parameters describes voice qualities on a continuum of tense, modal and relaxed voice qualities. Lower R_d values correspond to more tense and higher R_d values to more relaxed voice qualities.

Please note that the generic glottal flow derivative example shown in fig. 3.1 constitutes a synthetic signal generated with the LF model being parameterized with R_d=0.8. Please note additionally that the LF model sets in practice the time instant t_c to the fundamental period T_0: $t_c = T_0$. Other glottal source models may allow a different parameterization of t_c and T_0. An extensive analysis and discussion about the impact of the R waveshape parameters and its related parameters OQ, a_m, and t_a on the shape of the synthesized LF model in the spectral and in the time domain can be found in [Fant, 1995, Fant, 1997, Doval and d'Alessandro, 1999, Doval et al., 2006, d'Alessandro et al., 2007].

Fant introduced in [Fant and Liljencrants, 1994, Fant, 1995] equations to compute approximate R waveshape parameter values from an R_d value, denominated with the subscript p as 'predicted' R waveshape parameters R_{ap}, R_{kp}, and R_{gp}. In this work the term R_{*p} waveshape parameter set is employed to denote the predicted R waveshape parameter set. The equations deriving the parameters R_{ap} and R_{kp} from an estimated R_d value for the normal R_d range [0.3, 2.7] are given in [Fant, 1995]. A definition following the explanations given in [Fant, 1995] how to compute R_{gp} for the normal R_d range can be found in [Gobl, 2003, Degottex, 2010]. The R waveshape parameter

set for the upper R_d range]2.7, 5] was proposed by Fant in [Fant et al., 1994].

Fant proposed in [Fant, 1997] two possibilities to define the open quotient OQ from the R waveshape parameter set. The reduced form OQ_i defines the open quotient

$$OQ_i = t_e/T_0 = (1 + R_{kp})/2 \cdot R_{gp} \tag{3.2}$$

at the time instant of the maximum negative excitation t_e of the glottal flow derivative, normalized by the fundamental period T_0. The complete form OQ_e takes additionally into account the time of the return phrase t_a to define the open quotient

$$OQ_e = (t_e + t_a)/T_0 = (1 + R_{kp})/2 \cdot R_{gp} + R_{ap}. \tag{3.3}$$

The R_d parameter is highly correlated with the time instant t_e corresponding to OQ_i. Of high perceptual importance is the ratio of the peak U_0 of the glottal volume-velocity flow and the negative peak E_e of the glottal flow derivative [Fant et al., 1994]. It can be interpreted as effective pulse declination time $T_d = U_o/E_e$ by projecting the instants of both peaks in time to the time axis [Fant, 1997]. The now common naming convention R_d relates to T_d. The R_d parameter can be expressed as F_0-normalized glottal waveshape parameter [Fant, 1995]:

$$R_d = \frac{U_0}{E_e} \cdot \frac{F_0}{110} = \frac{1}{0.11} \cdot \frac{T_d}{T_0}. \tag{3.4}$$

The scaling factor 1/110 corresponds to F_0=110 Hz as a typical average in male speech [Gobl, 1988, Fant, 1997].

3.3.2 Normalized Amplitude Quotient NAQ

The Normalized Amplitude Quotient NAQ is a similar one-dimensional parameterization technique to describe the shapes of the deterministic part of the glottal excitation source over a range of tense, modal or relaxed voice qualities. The amplitude ratio of the positive peak U_0 and the negative peak E_e of the glottal flow derivative defines the T_d-related Amplitude Quotient (AQ) in [Alku and Vilkman, 1996] and the R_d-related NAQ in [Alku et al., 2002a]. The study of [Gobl, 2003] equals the NAQ parameter to AQ via a normalization by the fundamental period T_0. Equation 3.5 relates the NAQ to the R_d parameter by a scaling factor of 1000, normalized by the same fundamental frequency F_0 of 110 Hz found with equ. 3.4:

$$NAQ = AQ/T_0 = 0.11 \cdot R_d. \tag{3.5}$$

The NAQ parameter is used in [Campbell and Mokhtari, 2003] as measurement to examine voice qualities on a continuum from pressed to breathy phonation types. The NAQ measure exhibits significant correlations with interlocutor (who), speaking style (how), and speech act (what).

3.4 Glottal Closure Instant (GCI) estimation

Numerous algorithmic solutions exist in the literature to estimate the time instant of the glottal closure from a speech signal [Naylor et al., 2007, Degottex et al., 2009a, Sturmel et al., 2009, Thomas et al., 2009b, Degottex et al., 2010, d'Alessandro and Sturmel, 2011, Drugman et al., 2012b]. The GCI pulse position estimator of [Degottex et al., 2010, Degottex, 2010] is implemented in the SuperVP signal processing framework of section 2.5.2. It estimates for each analysis window position one single pulse position being closest to the window center. All estimated pulse positions are resolved over time using the fundamental period sequence and an R_d error measure to select a complete sequence of glottal pulses for one speech phrase. Setting the STFT window step size too large risks that individual pulses are omitted if the STFT window position is not placed close enough in time to the true GCI contained in the analyzed signal. The risk is minimized by defining the STFT analysis step size for the GCI estimator small enough such that it is set smaller than half of the shortest fundamental period T_0 expected in the analyzed speech phrase.

3.5 Voice quality

3.5.1 Definitions

Voice quality is a very broadly defined denomination to characterize different physiological, acoustical and perceptual phenomena of human voice production and perception. It is described very differently in the literature. The approaches to describe voice quality refer to features of the voice source. Voice quality can

be interpreted as the characteristic auditory colouring of a speakers voice [Laver, 1980]. A classification of voice quality into the three groups tense (pressed), modal (normal), and relaxed (breathy) can be found in [Drugman et al., 2011, Kane et al., 2013b]. Voice quality as a function of vocal effort can be expressed with a quiet, normal or loud voice [Liénard and Barras, 2013]. The loudness of a breathy voice quality is perceived lower compared to the loudness of normal speech. Loudness results in terms of voice quality from vocal effort [d'Alessandro, 2006]. The movement of the glottal folds and thus the shape of the glottal flow derivative is less abrupt around the time instant of glottal closure for relaxed speech [Thati et al., 2012].

Voice quality conveys paralinguistic information in speech. It should therefore be considered as prosodic characteristic [Pfitzinger, 2006]. According to [Laver, 1994], voice quality characteristics can be classified into the following four groups of settings: phonatory, articulatory, tension, and prosodic settings. The study of [Campbell and Mokhtari, 2003] proposes to add voice quality as a parameter of prosody, along with pitch, power and duration.

Different voice quality dimensions and phonation types can be organized using the following four prosodic dimensions: the voice register, noise, pressed-lax, and vocal effort [d'Alessandro, 2006]. Voice register dimensions can be used to stylize expressivity. The noise dimension can be related to breathiness and hoarseness. The pressed-lax dimension can serve as cue for speech emotions. The vocal effort dimension signals accentuation. Different descriptions of voice quality aim to describe these phonation types, e.g. by relating it to speaking styles and prosodic gestures [Childers and Lee, 1991, d'Alessandro, 2006].

The different phonation types or voice qualities of human speech production are generated by physiological mechanisms at glottis level [Gobl and Chasaide, 1992]. Descriptions of voice quality aim to explain these different vocal fold vibratory patterns [Childers and Lee, 1991]. The physiological correlates of voice qualities involve tension settings at laryngeal and supra-laryngeal levels, described in [Laver, 1980] with three parameters of muscular tension: Adductive tension, Medial compression, and Longitudinal tension. The latter is the tension of the vocal folds themselves. Medial compression is a result of adductive tension and reduces the length of the glottis. In general, the degree of tension rises in all three parameters from a breathy, lax, modal to a tense voice. The reference to all three parameters refers to the differentiation of the muscular tensions present at vocal tract and at larynx into three distinctive forces. A tense or relaxed voice quality results from adduction or abduction of the posterior part of the vocal folds [d'Alessandro, 2006].

Table 3.3: *Voice quality classification according to physiological mechanisms*

Voice Quality	Description
Tense voice	**High** degree of **tension** in all three muscular tension parameters; Tense voice perceived louder than other voice qualities; Typical aperiodicity of harshness not necessarily present; Extensive movements of tongue, lips and jaw.
Modal voice	**Moderate tension** in all three muscular tension parameters.
Lax voice	**Lower tension** in all three muscular tension parameters.
Breathy voice	**Minimal tension** in all three muscular tension parameters; The vocal folds do not come fully together.
Whispery voice	**Low** adductive **tension** and high medial compression; Laryngeal vibration very inefficient; High degree of audible frication noise.
Creaky voice	**High** medial compression, **high** adductive tension, **low** longitudinal **tension**; The vocal folds are relatively thick and compressed.

Table 3.3 lists a classification of physiological mechanisms into different voice qualities according to [Laver, 1968, Laver, 1980, Gobl and Chasaide, 1992]. Other voice quality types may for example be a soft, weak or falsetto voice [Klatt and Klatt, 1990]. The voice qualities jitter and shimmer describe the modulation in frequency and amplitude over a sequence of glottal pulses [Ghosh and Narayanan, 2011]. The phonation type jitter refers to the non-linear periodicity of the glottal pulses over time. Shimmer results from small variations of the glottal pulse amplitude of natural speech.

The studies of [Kreiman et al., 1992b, Kreiman et al., 1992a, Keating and Esposito, 2006] conduct perceptual tests to examine how listeners judge dissimilarity of acoustic feature descriptors related to voice quality. The features measured on natural human speech associated among others with voice quality are the fundamental frequency F_0, the frequencies of the first three formants F_1, F_2, F_3, the amplitude difference of the first two harmonic sinusoidal waves $H1^*$-$H2^*$, the amplitude difference of the second and fourth harmonic sinusoidal waves $H2^*$-$H4^*$, and the amplitude difference of the first harmonic sinusoid $H1^*$ versus the amplitude of the third formant

$A3^*$. However, it is not clear to the research community if F_0 constitutes a generally strong acoustic or perceptual correlate among different speakers and languages to be associated with the voice quality phenomena [Keating and Esposito, 2006]. The variation of these analyzed voice descriptors originates not just from changes of the spoken content such as different phonemes, but also by speech of different phonation types. These acoustic feature descriptors are associated with voice quality and are very well modelled by glottal source models like in [Fant et al., 1985, Klatt and Klatt, 1990].

3.5.2 Transformation

The voice quality of an original recording can be changed by means of transforming the deterministic and the stochastic part of the glottal excitation source [Gerratt and Kreiman, 2001]. The transformation is intended to alter the glottal excitation source in a way such that the re-synthesized speech is perceived to be uttered with a more tense or a more relaxed voice quality. However, a speech recording has to be discriminated into its stochastic and deterministic part such that a separate transformation of both components is facilitated. The deterministic glottal source part can be described by its estimated R_d contour. The stochastic part is the result of the subtraction of an estimated deterministic part from a speech phrase. Such classification into sinusoidal or noise peaks are discussed in sections 2.3.3 and 6.3.1.1.

The study of [Tooher and McKenna, 2003] analyzed the variation of estimated LF parameters across different vowels and changing phonetic context to derive dependencies and correlation. Advanced means are required if an expressive speech transformation is desired, e.g. to transform between emotional states of speech like sad, angry, happy and neutral. Further work presented in [Tooher et al., 2008] uses regression trees to model the different amount of variation of different glottal source parameters.

3.5.3 Just Noticable Difference (JND)

In [Henrich et al., 2003], different synthetic singing voice phrases have been presented as perceptual stimuli to listeners in different tests. The study examined which Just Noticable Difference (JND) thresholds are required to trigger the perceptual sensation of a change in voice quality within the perception of a human listener. 9 different sessions session were conducted. At each session, one stimuli was changed whereas the other stimuli were kept constant. The examined stimuli to synthesize different artificial singing voices consisted of 3 OQ values, 2 asymmetry coefficients α_m, 2 vowels and 2 fundamental frequencies F_0.

In general, OQ increases lead to higher JND thresholds whereas α_m increases lead to lower JND thresholds. This is in accordance with the R_d regression which implies a direct proportionality with OQ and an indirect proportionality with α_m: $Rd \sim OQ$, $Rd \sim 1/\alpha_m$. If a speaker changes the voice quality of his speech while speaking with a more tense voice, less changes in terms of ΔR_d are required by a speaker to arise the sensation of an altered voice quality in the perception of a listener. Contrariwise, changes in higher (lower) value regions of OQ (α_m)) range require longer distances of ΔOQ ($\Delta\alpha_m$) to arise a perceptual change at the listener. The speaker adapts subconsciously little voice quality changes to submit desired expressive information in a conversation.

A constant relative JND $\Delta OQ/OQ$ of 14 % for untrained and 10 % for trained listener is reported in [Henrich et al., 2003]. However, the variation of OQ was accompanied with an variation of the excitation amplitude E_e such that the measured JNDs of OQ mainly correspond to the sensation of vocal intensity variations. With only 2 sessions, a constant relative JND for α_m of 4 % change in $\Delta\alpha_m/\alpha_m$ has been determined such that α_m variations are more easily perceived than OQ variations. Both acoustical variations introduced with the change of OQ and α_m seem to depend on the sensations of perceiving different spectral slopes. No effect on JND could be determined for constant OQ values while changing the vowel or F_0.

A similar examination in [van Dinther et al., 2004] measures the perceptual distance of the R waveshape parameter space parameterizing the LF model. An excitation pattern distance of 4 dB is required on average to achieve a JND for the variation of the R waveshape parameter set. As well the JNDs are reported to be speaker dependent. On the one hand, the observation of a higher perceptual relevance in lower R_d value regions as with [Henrich et al., 2003] is not validated. On the other hand, the dependency of the required amount of parameter variation to achieve one JND on the parameter value is confirmed.

3.5.4 Creaky voice quality

Creaky voice is characterized by irregularly spaced glottal pulses corresponding to an aperiodic pitch period sequence [Gordon and Ladefoged, 2001]. The phonation type creaky voice is therefore also known as vocal fry voice

quality [Childers, 1995] since its auditory perception causes a certain perceptual sensation of a rough voice quality [Ishi et al., 2008]. Comparably lower OQ and very low F_0 values as well as adducted vocal folds are associated with creaky voice [Marasek, 1997]. The methods presented in [Ishi et al., 2008] measure the Intra-frame Periodicity (IFP) and the Inter-Pulse Similarity (IPS) of the glottal pulse sequence to detect creaky voice segments in a speech phrase. In [Kane et al., 2013a], two features extracted from the Linear Prediction (LP) residual signal are used as input to a decision tree classifier for the automatic detection of creaky voice segments: Secondary peaks found in time before the main excitation peak at the GCI and prominent impulse-like excitation peaks. Further work presented by the same authors in [Drugman et al., 2014] is based on an extended feature set including the previously mentioned ones to train an Artificial Neural Network (ANN) for the discrimination of creaky voice against other speech segments.

3.6 Source / filter separation

The production of voiced human speech can be approximately modelled by assuming the glottal source as excitation and the vocal tract as filtering element. The convolution of an impulse train defining the pulse positions with a time-varying impulse response corresponding to the glottal flow shapes define the deterministic part of the glottal excitation source. The glottal excitation source is completed by adding its stochastic part originating from noise turbulences at glottis level. As introduced in section 2.1, the convolution of the glottal excitation source waveform with the impulse response of the vocal-tract filter (VTF) and the filters defining the radiation at the lips and nostrils level results in the human speech signal. The resonances of the vocal tract can be modelled by an all-pole filter. The poles are expected to be stable due to the assumption that the vocal tract remains a passive and quasi-stationary medium within one evaluated signal frame. The coupling with the nasal cavities introduces pole-zero pairs. The zeros are assumed to lie inside the unit circle of the z-transform. The impulse response of the VTF is assumed to be a minimum phase signal. The vocal tract $C(\omega)$ is commonly modelled as an all-pole filter [Childers et al., 1977]. Zeros should be added if nasalized sounds are present [Drugman et al., 2011]. The glottal source $G(\omega)$ can in contrast be described by zeros only [Bozkurt et al., 2004].

The complete splitting of the human voice production model $S(\omega)$ into the characteristics of the acoustic excitation at the glottis level $G(\omega)$, the resonating filter of the vocal tract $C(\omega)$ and the nasal and lip radiation $L(\omega)$ can be found in [Degottex, 2010]. The splitting facilitates the independent manipulation of acoustical properties of the human voice. The parameterization of the excitation signal allows to include the related voice characteristics explicitly into the parameter transformation space. It facilitates additionally the parameterization and implementation of voice transformation algorithms such that the naturalness and coherence of the transformed voice signal can be improved.

The glottal formant constitutes a maximum in the amplitude spectrum of the glottal flow derivative. The modification of the formant position enables the simple control of the voice quality within an analysis-transformation-synthesis scheme. This permits the transformation between different voice qualities such as a pressed, normal or relaxed voice qualities [Vincent et al., 2007, Nordstrom et al., 2008, Tooher et al., 2008, Stylianou, 2009]. Moreover, the application of even extreme pitch transposition factors while maintaining high quality speech is possible with this technique [Degottex et al., 2011b].

The following section 3.7 presents means found in the literature to estimate the shape of the glottal excitation source. The subsequent section 3.8 discusses methods to transform and synthesize the glottal excitation source.

3.7 Estimation of the glottal excitation source

Much effort has been conducted by the speech research community to establish a reliable, robust and efficient method to extract the glottal excitation source from a recorded speech signal. Various algorithms have been proposed for this challenging task, as summarized in [Walker and Murphy, 2007]. How to estimate the glottal excitation source by robust means is still an open research question.

Section 3.7.1 discusses different inverse filtering based approaches which aim to cancel the contribution of the VTF from a speech signal such that the estimation of the glottal excitation source can be conducted more robustly on the remaining source signal. The different approaches of section 3.7.2 are based on analyzing the minimum and maximum phase parts of a speech signal to estimate the glottal excitation source. The phase minimization based method introduced in section 3.7.3 another methodology to estimate the glottal excitation source based on the phase properties of a speech signal. The method presented in section 3.7.4 aims to avoid the drawbacks of phase-based approaches by estimating the glottal excitation source directly in the amplitude spectrum.

Please find an extensive objective evaluation on natural human speech in section 5.6.4. It examines the methods phase minimization of section 3.7.3, the inverse filtering and dynamic programming (DyProg-LF) of section 3.7.1.4, and the amplitude spectrum measure (PowRd) of section 3.7.4.

3.7.1 Inverse filtering

Early approaches to inverse filter the contributions of the Vocal Tract Filter can be found in [Fant, 1961, Rothenberg, 1972].

3.7.1.1 Iterative Adaptive Inverse Filtering

A prominent technique to estimate and cancel single formants iteratively from a speech signal is the "Pitch Synchronous Iterative Adaptive Inverse Filtering" (PS-IAIF) method proposed by Alku in [Alku, 1992]. First, the spectral tilt effect of the glottal excitation source is eliminated from the speech signal. A pitch-synchronous linear predictive analysis of first order estimates LPC coefficients on the resulting signal to determine the VTF [Makhoul, 1975]. The LPC estimate is inverse filtered from the signal to obtain a second estimate of the glottal source. Second, the procedure is repeated with an LPC estimate of higher order to achieve more accurate estimates.

3.7.1.2 Inverse Filtering and Model Matching

The method "Simultaneous Inverse Filtering and Model Matching" (SIM) proposed in [Fröhlich et al., 2001] extends the means of IAIF. Its iterative procedure uses a multi-dimensional optimization technique to apply an error criterion to measure how well the inverse filtering performs. The LF model with the best matching parameterization determines the characteristic of the estimated glottal pulse shape. SIM is robust to phase disturbances and the window position of the analyzed frame with respect to the period in time.

3.7.1.3 Inverse Filtering and Convex Optimization

The glottal source KLGLOTT88 of [Klatt and Klatt, 1990] and the LF model of section 3.2.3 are used in [Pérez and Bonafonte, 2005] to simultaneously estimate the parameters describing vocal tract and glottal excitation source. It uses a quadratic programming algorithm by means of convex optimization to minimize a matching error. The latter results from matching synthesized glottal waves using the KLGLOTT88 model on the glottal waves obtained by inverse filtering. The convex optimization approach is inspired by a corresponding joint estimation of VTF and glottal excitation source described in [Lu and Smith, 1999]. The models are fitted on the inverse filtering results using the method of [Strik et al., 1993]. The LF model achieves better results on a test set of natural human speech in terms of the Root-Mean-Square Error (RMSE) since the KLGLOTT88 model does not consider a non-abrupt return phase. The LF model considers the return phase duration t_a with its corresponding parameter R_a. The continued work presented in [Pérez and Bonafonte, 2009] uses overlapping frames of several glottal cycles to increase the robustness and quality of the estimation. The KLGLOTT88 model is used as pre-initialization to estimate the LF model parameters. A good performance is reported on a synthetic data set including different SNR levels using amplitude-modulated white Gaussian noise.

3.7.1.4 Inverse Filtering and Dynamic Programming

The method introduced in [Kane et al., 2012, Kane and Gobl, 2013a] is denominated as "DyProg-LF". First, GCI locations are estimated using a modified version of the SE-DREAMS approach of [Drugman et al., 2012b] described in [Kane and Gobl, 2013a]. Second, the IAIF method of the preceding section 3.7.1.1 is used to compute the glottal source signal on which error values are calculated. The error criteria are based on measuring the correlation of glottal pulses synthesized with the LF model. The latter is parameterized over a grid of R_d values. The synthesized glottal pulses are matched to the source signals in the spectral and the time domain. The temporal and spectral errors form a target cost. A transition cost is computed from the continuity of the parameter trajectories over frames. Both costs are utilized in a dynamic programming algorithm to increase the robustness of the DyProg-LF algorithm to estimate the R_d contour of the glottal excitation source over time.

3.7.2 Minimum / maximum phase decomposition

The maximum phase component of a speech spectrum $S(\omega)$ is described by the anti-causal part of its complex cepstrum. The causal part of the complex cepstrum reflects the minimum phase component. Different approaches to estimate the deterministic part of the glottal excitation part are based on the hypothesis that the minimum phase part of $S(\omega)$ can be contributed to the VTF and the maximum phase part to the glottal source. Methods to conduct a minimum/maximum phase decomposition for the estimation of the deterministic part of the glottal source will be introduced in the following.

3.7.2.1 Causal-Anticausal Linear Model (CALM)

The first method exploiting the minimum / maximum phase hypothesis is the decomposition of a speech spectrum $S(\omega)$ into a periodic and aperiodic component proposed in [Doval et al., 1997]. The Causal-Anticausal Linear Model (CALM) model of [Doval et al., 2003] models the glottal flow as a pair of causal filter poles inside and anti-causal filter poles outside the unit circle. CALM is the only glottal source model operating in the spectral domain. The Open Quotient OQ reflects the anti-causal open phase. The causal return phase t_a is present from time instant t_e at the GCI until the time instant of the glottal closure t_c.

3.7.2.2 Zeros of the Z-transform (ZZT)

The Zeros of Z-Transform (ZZT) decomposition proposed in [Bozkurt et al., 2004] estimates a set of roots of the z-transform for each signal frame to conduct the causal-anticausal filter decomposition according to the CALM model. ZZT splits a mixed phase signal into the contributions of its maximum phase component corresponding to the glottal source, and its minimum phase component corresponding to the VTF. The all-zero signal representation of the ZZT requires an computationally expensive algorithm for the polynomial extraction of the roots [d'Alessandro et al., 2007].

3.7.2.3 Complex Cepstrum-Based Decomposition (CCD)

The Causal-Anticausal Decomposition (CCD) of [Drugman et al., 2009a] employs the Complex Cepstrum (CC) for the linear separation of the contributions of VTF and glottal source contained in a mixed phase spectrum $S(\omega)$. It is similar to the ZZT transform of the preceding section. CCD achieves the same decomposition quality than ZZT with a lower computational effort since the ZZT requires the factorization of high degree polynomials.
Return phase:
Please note that both ZZT and CCD can only estimate the open phase part as the maximum phase component of the glottal excitation source. However, the return phase part of the glottal excitation source is minimum phase [Drugman et al., 2011]. Both decomposition methods have to attribute therefore a possibly occurring return phase to the minimum phase vocal tract estimate [Kane, 2013]. In contrast, the glottal source can approximately be assumed to consist only of a maximum phase part if the time t_a of the return phase is very short which is more likely to occur for tense voice qualities.
Phase unwrapping:
The computation of the Complex Cepstrum requires contrary to the real cepstrum the additional modelling of the phase part [Oppenheim et al., 1968]. This implies the potentially error-prone task of phase unwrapping which can result in an erroneous phase representation [Bozkurt, 2005]. Problems arise due to possible phase ambiguities at the phase borders $[-\pi, \pi]$. Most often this depends on the current signal characteristic. Different algorithmic solutions exist in the literature to address the possible drawbacks of phase unwrapping for different signals [Tribolet, 1977, de Goetzen et al., 2000]. The study of [Childers et al., 1977] lists signals as potentially problematic having a linear phase component with a too huge slope, or signals having huge spectral notches. Moreover, aliasing due to a too low sampling rate and oversampling in the presence of noise are additional sources of phase unwrapping errors. Care has thus to be taken when designing a signal processing algorithm requiring phase unwrapping. The proposal of CCD in [Drugman et al., 2009a] addresses the mentioned phase unwrapping problems by applying a large FFT size such that a sufficiently high resolution facilitates robust phase unwrapping to minimize possible aliasing distortions.
GCI estimation and 2-period windowing:
The deconvolution quality of CCD is additionally depending on an exact estimation of the GCIs. This allows using a window covering only two fundamental periods. With this the anticausal maximum phase part of the open phase and the causal minimum phase part of the return phase can be precisely evaluated. The application of the Chirp Z-Transform (CZT) in [Drugman and Dutoit, 2010] adds robustness against GCI location errors. It better separates

the zeros because the z-transform is expressed on a spiral contour in the z-plane.
Glottal source transformation:
The presented CCD method synthesizes the estimated glottal excitation source in the time domain. A continuative processing of the split speech signal to transform the glottal source, for example in the context of a voice quality modification, requires consequently to match a glottal source model on the time domain signal. The Strik method of [Strik et al., 1993, Strik, 1998] facilitates the estimation of source parameters on a glottal source signal. This enables changing the estimated glottal source parameters before re-synthesis for different transformation purposes.

3.7.2.4 Causal-Anticausal extensions

Further approaches based on the Causal-Anticausal Decomposition to estimate the glottal source or to process speech can be found in the literature. A new operator to convert poles into zeros is used in [Hezard et al., 2013] within the context of a CALM-based source-filter separation. However, the method proved to be sensitive to noise and to T_0 estimation errors. In [Vondra and Vích, 2010b] a Complex Cepstrum-based approach obtains a mixed phase VTF. A higher estimation accuracy and a more natural speech re-synthesis compared to using a minimum phase VTF is reported. The method of [Vondra and Vích, 2010a] uses CC to estimate the maximum phase part of a speech signal. The frequency and the bandwidth of the glottal formant are estimated from the maximum phase information. A 2^{nd} order IIR filter is fitted on the estimated glottal formant. Modified versions of the IIR filter enable to alter the voice quality contained in the speech signal.

3.7.3 Phase Minimization

The estimated glottal excitation source $G(\omega)$ and the filter $L(\omega)$ representing the radiation at lips and nasal are removed in [Degottex et al., 2009a, Degottex et al., 2009b] from a speech recording to reveal the filter characteristic of the vocal tract $C(\omega)$. The results indicate that the proposed method is robust against the influences of glottal or additive noise.

The simultaneous estimation of the LF regression parameter R_d together with the temporal synchronization of the glottal source model and the estimation of Glottal Closure Instants (GCI) is proposed in [Degottex et al., 2010]. GCIs refer to the time instants of the maximum excitation energy E_e within a fundamental period. The time synchronized estimation improves the robustness of the R_d estimation. The latter reacts sensitive to failures of the temporal synchronization if it is not executed simultaneously.

The phase-based methods presented in the preceding sections deconvolve a mixed phase speech signal into the components of the maximum phase for the glottal source and the minimum phase for the VTF. However, the underlying hypothesis neglects that the glottal source signal does not exclusively constitute a maximum phase signal. A glottal pulse which contains a return phase t_a>0 introduces a minimum phase part in its signal representation.

A different phase-based method to estimate the deterministic part of the glottal excitation source contained in a speech recording is proposed in [Degottex, 2010, Degottex et al., 2011a]. It follows the hypothesis that the glottal pulse signal can be either a maximum phase signal if t_a=0 or a mixed phase signal if t_a>0. It matches synthesized LF models being parameterized by R_d against the speech signal. With this the maximum or mixed phase signal characteristic of the glottal source is taken into account by using the R_d parameter space. The glottal source signal can therefore be comprised by zeros lying outside for the maximum phase and inside the unit circle for the minimum phase part.

The R_d parameter estimation algorithms based on the objective functions for phase minimization were established in [Degottex et al., 2010, Degottex, 2010, Degottex et al., 2011a]. Synthesized LF models parameterized along the R_d value range are fitted against the speech spectrum $S(\omega)$. The objective function for phase minimization minimizes the mean squared phase error residual ϕ_ϵ resulting from each R_d-parameterized glottal pulse. The spectral representation of the latter is matched against a strictly harmonic representation of voiced speech of one spectral frame. The R_d value that is used to synthesize the glottal formant resulting in the lowest remaining phase error ϕ_ϵ is selected as estimated R_d value per frame.

The utilization of the LF model reflects the possible maximum or mixed phase behaviour of a glottal source signal. The phase minimization paradigm exploits the different properties of the phase spectra of the glottal source and VTF models. It employs contrary to the Complex Cepstrum-based method of section 3.7.2.3 the real cepstrum $c(n)$ defined in equ. 2.3 of section 2.2.2. It does therefore not have to perform the possibly error-prone task of phase unwrapping. Additionally, the real cepstrum is computationally faster. Moreover, the phase minimization paradigm is not depending on the window position relative to the GCI [Degottex et al., 2009b, Degottex et al., 2010]. However, the R_d parameterization restricts the synthesized LF model shapes to a subspace of its complete parameter space. Differences between the true glottal source shape contained in the signal and the evaluated LF shapes have

to be accepted. This may introduce a bias in the evaluation of the phase error residual ϕ_ϵ. Please find an exact description of the phase minimization approach to estimate the deterministic part of the glottal excitation source in section 5.3.

3.7.4 Amplitude spectrum measure (PowRd)

The PowRd method described in [Ó Cinnéide, 2012] operates upon the power spectrum since time domain and phase based methods are sensitive to phase distortion. To determine R_d it avoids unreliable phase information and high frequency information which can be corrupted by noise. A relative Itakura-Saito error criterion [Itakura and Saito, 1968] is used to determine the filter order and the coefficients of the VTF. The scale parameter E_e of the included LF glottal source model is determined in the power spectrum. The PowRd method is based on the SIM approach of the preceding section 3.7.1.2.

3.8 Extended source/filter-based speech models

The transformation of the glottal source part of one speech signal offers many advanced speech processing possibilities, e. g. voice quality alteration. A glottal source transformation from the source to the target speaker proved to contribute to the VC performance [Childers, 1995, del Pozo and Young, 2008, Pérez and Bonafonte, 2011]. This requires an extension of the speech model utilized within the VC software system such that parameters describing the glottal excitation source and their differences between speakers are reflected.

Section 3.8.1 describes early speech analysis-synthesis approaches. Three different speech analysis-synthesis algorithms being based on an extended source-filter model will be described in the following sections 3.8.2, 3.8.3, and 3.8.4. These extended speech models allow the transformation of the glottal excitation source for the synthesis of altered voice qualities or for general improvements of the synthesis quality. All three speech models utilize the LF glottal source model. This facilitates the transformation of relevant LF parameters before synthesis to conduct advanced speech processing tasks.

3.8.1 Linear Prediction analysis and synthesis

Common speech analysis and synthesis systems employ Linear Prediction (LP) analysis [Makhoul, 1975]. The estimated Linear Predictive Coding (LPC) coefficients represent the spectral envelope. Inverse filtering of the LPC filter applied to the speech signal reveals the LP residual. It includes the fundamental frequency F_0 and non-sinusoidal noise components generated at the larynx. The LP estimates can be employed to synthesize speech with the well-known source-filter model of human speech [Fant, 1981]. The LPC coefficients function as filter which is excited by the LP residual as its corresponding source [Nordstrom et al., 2008]. The latter is generated at the larynx and consists of the deterministic and stochastic signal parts of the glottal excitation source [Drugman and Dutoit, 2012]. The LP residual can thus be considered as the excitation source of human speech. Both the LP residual and the LPC coefficients proved to contribute to the recognition of speakers voice identities in speaker verification systems [Markov and S.Nakagawa, 1999, Prasanna et al., 2006, Chetouani et al., 2009].

The Glottal Excitation Linear Predictive (GELP) synthesizer in [Childers, 1995] models the glottal excitation source by a codebook-based method in the VC context. One glottal excitation codebook contains polynomials representing the glottal excitation waveform. One stochastic codebook stores unvoiced noise excitation entries. Another approach in [Childers, 1995] uses the LF glottal volume-velocity waveforms to model the different phonation types of voice quality.

3.8.2 Glottal Spectral Separation (GSS)

The GSS method [Cabral et al., 2008] consists of separating the glottal source effects from the spectral envelope. It facilitates the transformation between modal, breathy or tense voice qualities by modifying LF parameters. No poles and zeros of a speech model have to be calculated. This avoids possible drawbacks of the pole/zero estimation. The usage of the LF model clearly outperforms simple impulse excitation synthesis. The speech production model used for GSS is described as follows:

$$S(\omega) = D(\omega) \cdot G(\omega) \cdot V(\omega) \cdot R(\omega). \tag{3.6}$$

The spectrum $S(\omega)$ is represented by the Discrete-Time Fourier Transforms of an impulse train $D(\omega)$, a glottal pulse $G(\omega)$, a vocal tract transfer function $V(\omega)$, and the radiation characteristic $R(\omega)$.

3.8.2.1 Analysis

GSS produces speech automatically from LF model parameters trajectories and spectral features. The glottal flow derivative waveform v_g is estimated by inverse filtering and centered around GCIs. Inverse filter coefficients are calculated pitch-synchronously from the pre-emphasized speech signal. The LF model is fitted to a 4 kHz low pass and linear phase filtered version of v_g. The fitting estimates the timing parameters t_p, t_e, and t_a of the LF model. The fitted LF parameters are varied to minimize a Mean Square Error (MSE) criterion for optimization.

The spectral envelope $\hat{H}(\omega)$ of the speech signal is estimated via STRAIGHT [Kawahara et al., 2008], introduced in section 2.5.1. The spectral representation of the glottal pulse $E_{LF}(\omega)$ divides the spectral envelope $\hat{H}(\omega)$ to remove its influences like the spectral tilt from the signal. The result is used to estimate the Vocal Tract Filter (VTF) $V(\omega)$.

3.8.2.2 Synthesis

A variation of the LF timing parameters t_p, t_e, and t_a implies an alteration of the related parameters Open Quotient OQ, Speed Quotient SQ and Return Quotient RQ. Used for synthesis, this enables altering the voice quality. The GSS method applies a pitch-adaptive analysis window of two periods per STFT step windowing a synthesized sequence of glottal pulses. It is multiplied in the spectral domain with the VTF $V(\omega)$ and the radiation filter $R(\omega)$. The result is inverse Fourier transformed into the time domain to synthesize the resulting waveform, using Pitch-Synchronous Overlap-and-Add (PSOLA) [Valbret et al., 1992, Kortekaas and Kohlrausch, 1997].

3.8.3 ARX-LF Source-Filter Decomposition

The method "Auto-Regressive with eXogenous input" (ARX) described in [Vincent et al., 2005, Vincent et al., 2007] applies a joint estimation of the parameters describing the glottal excitation source and the VTF. The residual part is parameterized by a Harmonic plus Noise Modeling (HNM) analysis [Stylianou, 1996].. The deterministic part of the glottal excitation source is parameterized by the LF model. The ARX model assumes that the VTF originates from an auto-regressive (AR) system: $C(z) = 1/A(z)$. A least square error criterion selects the AR model order for this time-varying IIR system. Dynamic programming in form of the Viterbi algorithm is used to to optimize the parameter estimation. Modelling additionally zeros in the z-plane caused by nasalization via $C(z) = B(z)/A(z)$ constitutes an "AutoRegressive Moving Average with eXogenous input" model, abbreviated with ARMAX [Degottex, 2010]. The VTF estimation has no constraints on the positions of poles and zeros regarding the unit circle.

3.8.3.1 Analysis

The ARX model of [Ding et al., 1995] splits a speech signal into the contributions of the vocal tract α, the deterministic part of the glottal source Σ_{LF}, and a residual signal E. The latter captures the residual information which is not described by the estimated ARX-LF parameters. This residual may most likely be comprised of the remaining stochastic part of the glottal source. It may also contain mismatches between the real deterministic part of the glottal source and the LF waveform.

In [Vincent et al., 2005, Vincent et al., 2007], first a low band analysis estimates the following parameters mostly impacting low frequency regions: The Open Quotient OQ, the asymmetry coefficient α_m, the return phase quotient Q_a, the VTF order and GCI locations. GCIs are located by employing F_0 continuity constraints and an appropriateness measure on the ARX-LF model. The glottal source parameters are estimated by minimizing a least square error criterion over source and vocal tract parameters. A following full band analysis includes high frequency effects and refines the initial low band parameter estimation [Vincent et al., 2005].

Viterbi smoothing is used to regularize the sequence of LF source parameters Σ_{LF} to compute similar as in [Huber and Röbel, 2013] an optimized LF parameter sequence. The well-known auto-regressive model (AR) estimates the vocal tract parameters α. The residual is obtained after removing the glottal source and vocal tract effects. It is modelled using a HNM.

3.8.3.2 Synthesis

The synthesis algorithm of ARF-LF operates pitch synchronously. It passes the reconstructed glottal source signal through a time-varying filter. The glottal source is represented as a sum of the LF glottal waveform, a harmonic part

and a noise part. The noise part is synthesized by high pass filtering white Gaussian noise. The cutoff frequency set to the maximum voicing frequency F_m. Unvoiced frames contain only the noise part.

3.8.4 SVLN

The speech analysis, transformation and synthesis system denominated SVLN is derived from "**S**eparation of the **V**ocal tract with the **L**iljencrants-Fant model plus **N**oise" [Degottex, 2010]. The method estimates the R_d contour [Degottex et al., 2011a] using the phase minimization paradigm, introduced in section 3.7.3. Glottal pulses are synthesized in the spectral domain to divide it from the spectral representation of the speech signal. This extracts the contained Vocal Tract Filter [Degottex et al., 2011b]. SVLN facilitates advanced voice processing applications like pitch transposition or voice quality transformation while maintaining a high synthesis quality [Lanchantin et al., 2010, Degottex et al., 2011b, Degottex et al., 2013].

3.8.4.1 Voice production model

The SVLN method is based on a model representing the voice production of natural human speech with the following signal components:

$$S(\omega) = \left[H^{F_0}(\omega) \cdot G^{R_d}(\omega) + N^{\sigma_g}(\omega)\right] \cdot C^{\bar{c}}(\omega) \cdot L(\omega). \tag{3.7}$$

A quick overview explaining each involved signal component is given as follows:

Table 3.4: *Signal components of SVLN*

Denomination	Explanation
$H^{F_0}(\omega)$	Harmonic structure
$G^{R_d}(\omega)$	Synthesized LF model
$N^{\sigma_g}(\omega)$	Stochastic noise component of the glottal excitation source
$C^{\bar{c}}(\omega)$	Vocal Tract Filter (VTF)
$L(\omega)$	Radiation filter

The harmonic structure $\mathbf{H^{F_0}}(\omega)$ models a periodic impulse train of F_0: $H^{F_0}(\omega) = \sum_{k\in\mathbb{Z}} e^{j\omega k/F_0}$. The synthesized LF model $\mathbf{G^{R_d}}(\omega)$ is parameterized by the LF regression parameter R_d and the LF amplitude parameter E_e [Fant, 1995]. It represents the shape of the deterministic part of the glottal excitation source present in one single fundamental period. The stochastic part of the glottal excitation source, the random noise component $\mathbf{N}^{\sigma_g}(\omega)$, originates from air turbulences generated at the glottis. The noise is assumed to follow a Gaussian distribution with standard deviation σ_g. The Vocal Tract Filter $\mathbf{C}^{\bar{c}}(\omega)$ represents the resonances and anti-resonances of the vocal tract. The VTF is assumed to be minimum phase. It is parameterized by the cepstral coefficients vector $\bar{c}$. The radiation filter $\mathbf{L}(\omega)$ incorporates a filter representing the radiation at lips and nostrils level. It is assumed, according to [Markel and Gray, 1976], that it can be modelled as a time derivative by $L(\omega) = j\omega$. The shape of the spectral envelope $T(\omega)$ is influenced by the VTF $C(\omega)$, the radiation filter $L(\omega)$, and the glottal source $G(\omega)$.

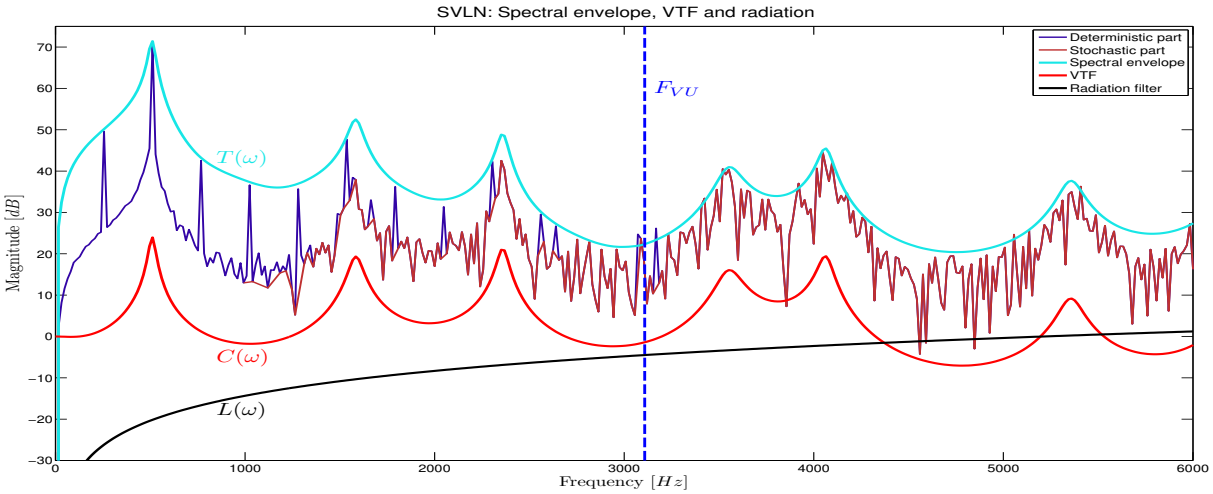


Figure 3.2: *Spectral envelope $T(\omega)$, VTF $C(\omega)$ and radiation $L(\omega)$*

3.8.4.2 Analysis

The voice production model of SVLN, defined in equ. 3.7, requires the analysis of the following voice descriptors: F_0, R_d, E_e, σ_g. Based on these estimations, the VTF can be extracted from the speech signal. The voice production model contains three interdependent gains. The gains E_e and σ_g control the energy of the deterministic and stochastic part of the glottal excitation source. The third gain is the mean log amplitude of the VTF. A constraint sets the mean log amplitude of the VTF to zero. This constraint reduces SVLNs energy modelling to the excitation amplitude E_e of the LF glottal model, and the noise level σ_g. The energy measured with the cepstral estimate of the VTF is attributed to E_e. The stochastic noise gain σ_g is deduced from the energy level present at the estimated F_{VU} frequency:

$$\sigma_g = |G^{R_d}(F_{VU})| \cdot \frac{\sqrt{2}}{\sqrt{\pi/2} \cdot \sqrt{\sum_t \text{win}[t]^2}}. \tag{3.8}$$

The measure $|G^{Rd}(F_{VU})|$ reflects the energy level of the signal measured at the F_{VU}. Since the spectral amplitudes of Gaussian noise obey a Rayleigh distribution, $|G^{Rd}(F_{VU})|$ is first converted to Rayleigh mode by $1/\sqrt{\pi/2}$. The result is further normalized by the standard deviation of the Gaussian distribution $\sqrt{2}$ [Yeh, 2008], and the energy of the analysis window $\sqrt{\sum_t \text{win}[t]^2}$ [Griffin and Lim, 1984].

Like illustrated in fig. 3.2, the VTF $C^{\bar{c}}(\omega)$ is constructed from the signal parts present in the deterministic frequency band $S(\omega < \omega_{F_{VU}})$ below the F_{VU}, and in the stochastic band $S(\omega > \omega_{F_{VU}})$ above the F_{VU}. The F_{VU} frequency is interpreted as angular frequency $\omega_{F_{VU}}$. The contributions of $L(\omega)$ and $G^{Rd}(\omega)$ are removed from $S(\omega < \omega_{F_{VU}})$ by spectral division. The True Envelope (TE) operator $\mathcal{T}(.)$ fits the harmonics of the division result to estimate the spectral envelope in the deterministic band. The spectral division of $L(\omega)$ and $|G^{Rd}(F_{VU})|$ in the stochastic band $S(\omega > \omega_{F_{VU}})$ ensures the continuity between the two frequency bands. The real cepstrum operator $\mathcal{P}(.)$ estimates the spectral envelope in the stochastic band. The application of factor $e^{0.058}$ on a linear scale to the mean log amplitude measured by $\mathcal{P}(.)$ converts the latter to Rayleigh mode [Yeh, 2008]. The normalization of the Rayleigh mean value ($\sqrt{\pi/2}$) computes the expected amplitude. Accordingly, the SVLN method constructs the VTF as follows:

$$C(\omega) = \begin{cases} \mathcal{T}\left(\frac{S(\omega)}{L(\omega)\cdot G^{R_d}(\omega)}\right) \cdot \gamma^{-1} & \text{if } \omega < \omega_{F_{VU}} \\ \mathcal{P}\left(\frac{S(\omega)}{L(\omega)\cdot G^{R_d}(F_{VU})}\right) \cdot \frac{\sqrt{\pi/2}}{\gamma \cdot e^{0.058}} & \text{if } \omega \geq F_{VU} \end{cases}, \tag{3.9}$$

with $\gamma = \sum_t \text{win}[t]/(F_s/F_0)$ normalizing for the number of periods of the analysis window. The VTF $C(\omega)$ is constructed by concatenating both envelopes $\mathcal{T}(\omega < \omega_{F_{VU}})$ and $\mathcal{P}(\omega \geq F_{VU}$ at the F_{VU}. A function $F_{sigmoid}$ is applied to transition between the harmonic $\mathcal{T}$ and the noise $\mathcal{P}$ envelope.

The division by the constant energy level $|G^{Rd}(F_{VU})|$ is applied throughout the complete stochastic band. This ignores the glottal pulse shape above the F_{VU}. The noise floor in the stochastic band is assumed to be white, with the noise level fixed to the energy measured with $|G^{Rd}(F_{VU})|$. A visual example of the synthesis of the random

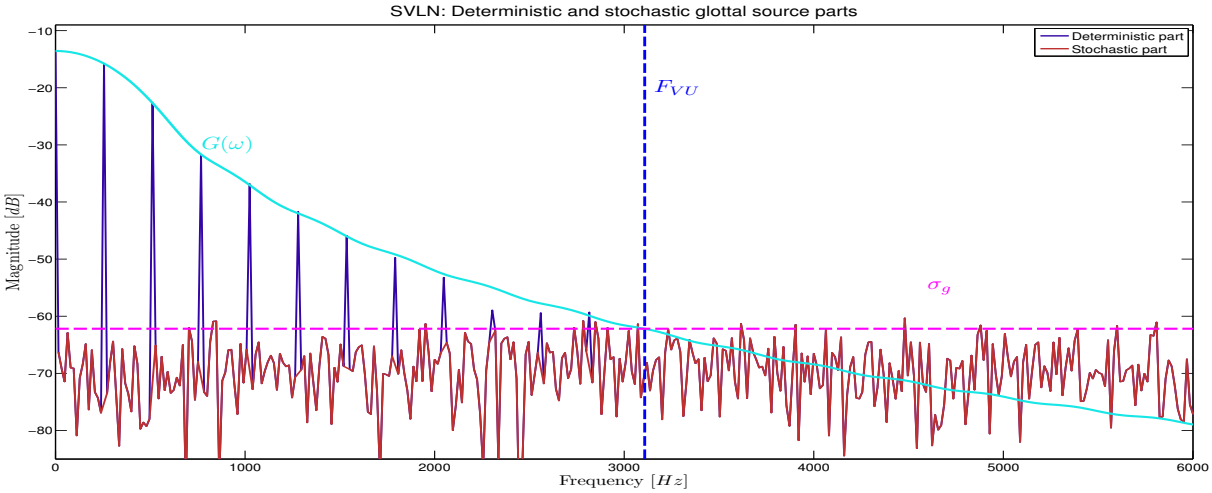


Figure 3.3: *Deterministic and stochastic parts of the glottal excitation source*

noise component $N^{\sigma_g}(\omega)$ and the deterministic component $G^{R_d}(\omega)$ of the glottal excitation source in SVLN is illustrated in fig. 3.3. A visual example of the VTF creation with SVLN is illustrated in fig. 3.4. Both figures were adapted from [Degottex et al., 2013] with kind permission of the author.

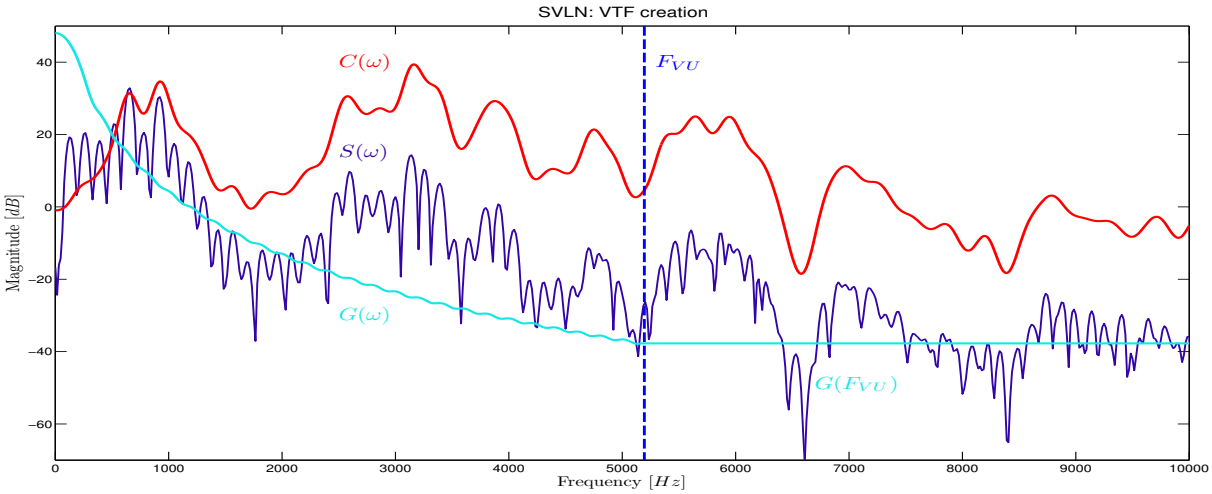


Figure 3.4: *SVLN VTF creation*

3.8.4.3 Synthesis

Amplitude modulation:
The stochastic noise component is filtered, modulated and windowed to achieve a natural sounding voice. Synthesized white noise without filtering and modulation would lead to a hoarse sounding voice. A high pass filter $F_{hp}^{F_{VU}}(\omega)$ with a cutoff frequency of F_{VU} Hz and a slope of 6 dB/kHz in the transition band is applied to a synthetically generated white noise signal.

An amplitude modulation function $v^{Rd}[t]$ according to [del Pozo and Young, 2008] is given as follows:

$$v^{R_d}[t] = \beta \cdot g^{R_d}[t] + (1 - \beta). \tag{3.10}$$

The constant factor β allows adjusting between a constant noise floor and the magnitude of the amplitude modulation. From informal listening tests, β is set empirically to 0.75. The term $g^{Rd}[t]$ refers to the LF model. It implies an amplitude modulation of the noise component synchronously to F_0.

Noise synthesis:
A cross fade between consecutive synthesized noise segments is applied. The stochastic noise component $N(\omega)$ is synthesized in the spectral domain for segment k by

$$N_k(\omega) = F_{hp}^{F_{VU_k}}(\omega) \cdot \mathcal{F}(v^{R_{d-k}}[t] \cdot \mathrm{win}_k[t] \cdot n^{\sigma_{gk}}[t]), \tag{3.11}$$

with the operator $\mathcal{F}(.)$ being the Discrete Fourier Transform. The zero-mean Gaussian random signal $n^{\sigma_{gk}}[t]$ has standard deviation σ_{gk}.

Mixing of voiced and unvoiced parts:
A transformation of the R_d contour used for the synthesis of $G^{R_d}(\omega)$ enables the modification of the glottal excitation source. The deterministic voiced part $G^{R_d}(\omega)$ of the glottal excitation source is added to the corresponding stochastic unvoiced part $N(\omega)$. Both are convolved with the Vocal Tract Filter $C(\omega)$ and the radiation filter $L(\omega)$. To synthesize the k^{th} speech segment $S_k(\omega)$ by

$$S_k(\omega) = (e^{-j\omega m_k} \cdot G^{Rd_k}(\omega) + N_k(\omega)) \cdot C^{\bar{c}_k}(\omega) \cdot j\omega, \tag{3.12}$$

the delay $e^{-j\omega m_k}$ places the time instant t_e of the GCI for the LF model at mark m_k.

3.9 Conclusions

The motivation to employ the modelling of the glottal excitation source within the context of Voice Conversion is indicated by the discussion of section 3.5 concerning the voice quality contained in any speech signal. Transforming the voice quality from a source to a target speaker shall contribute to render the converted speech phrase more towards the voice identity of the target speaker. This requires to estimate in particular the glottal pulse shape of the deterministic source part. Several START methods to estimate glottal pulse shapes are presented in section 3.7. The following sections 3.8 presents three START speech system to analyse, transform and synthesize the glottal excitation source. However, further means to transform and synthesize both the deterministic and the stochastic

part of the glottal excitation source are required to augment the analysis and synthesis quality. Several advancements to model and estimate the deterministic part with higher robustness will be presented in chapter 5. A novel speech framework to transform and synthesize both parts of the glottal excitation source along with an extended set of voice descriptors will be introduced in chapter 6.

Chapter 4

STAte-of-the-ART (START) in Voice Conversion

And the Lord spoke unto you from the heart of the fire: You heard the sound of His words, but did not see His similitude. There was only a voice.

THE HOLY TORAH (D'VARIM 4:12) / THE HOLY BIBLE (DEUTERONOMY 4:12)

4.1 Introduction

This chapter presents some of the most important STAte-of-the-ART methodologies in Voice Conversion found currently in the literature. More specific details about different VC approaches are explained here in continuation to the generic introductions given in chapter 1. The first section 4.2 gives insight into the problematic of having to account for the timing differences present between two speakers uttering the same phonetic content. The feature vectors used to describe the two speakers are commonly denominated as X for the source and Y for the target speaker. The means required to synthesize the converted-to-the target speaker feature vectors $\hat{Y}$ coherently in the converted speech phrase are mentioned in section 4.4.

The following section 4.5 introduces the initial VC approaches using Vector Quantization and Codebook mapping. A major advancement in VC was achieved by the utilization of statistical models, as presented in the subsequent section 4.6. However, the basic probabilistic approach proved to not achieve a VC performance being sufficient for industrial usage. Nowadays, many improvements and extensions to the conventional statistical models are proposed by the VC community. All proposed approaches aim to minimize the basic drawback of using statistical models for VC: The over-smoothing effect introduced by the linear regression of the probabilistic approach. The training of the model parameters has to average over the underlying speaker data. The approximation of the probabilistic modelling parameters to the training data of the utilized speaker pair loses consequently certain specific details. The extensions of the basic statistical model to minimize the over-smoothing effect are listed in section 4.6.3.

The conventional and advanced statistical models presented until now convert usually only the most important voice descriptor concerning a speakers voice identity: The Vocal Tract Filter (VTF). This single feature conversion demonstrates that even with the extended statistical modelling the means to convert a speech phrase to the target speaker are not satisfying enough. Certainly, there does not exist one single (golden) feature which uniquely identifies and unifies the complete voice identity of a speaker in the auditory perception of a human person. Different perceptual cues are required to arise the perceptual sensation of the target speakers voice identity in the human auditory cortex of a listener. Section 4.7 lists therefore the conversion of other voice descriptors like the residual, the fundamental frequency F_0 or the glottal excitation source.

However, it still appears that even with an extended set of converted voice descriptors and advanced statistical models the desired conversion score and synthesis quality is not high enough. The main reason being that the extended probabilistic means are not capable to overcome the basic deficiency of the over-smoothing effect. Therefore,

novel VC systems have been proposed over the last decade which aim to avoid or at least minimize the usage of a basic statistical model to map features between both speakers. With this the over-smoothing effect of the statistical model does not have to be addressed, as explained in section 4.8. Still these methods use to some extent probabilistic methods for different purposes.

Most of the presented VC approaches are based on parallel corpora such that source and target speaker have to utter the same sentences. The preparation of such corpora requires more effort and prohibits certain VC applications like the conversion of one voice into another language. Section 4.9 introduces non-parallel VC systems which enable the utilization of corpora having different phonetic content.

4.2 Annotation and alignment

The training of statistical models to derive conversion functions for the VC mapping is usually performed on time-aligned data. A speech recognizer is used to produce phoneme segments, as discussed in section 4.2.1. This forced alignment is then used to apply Dynamic Time Warping (DTW) within the phoneme borders of each source and target phoneme such that both are aligned, as presented in section 4.2.2. The phoneme border detection is more difficult to establish and has higher computational cost.

4.2.1 Automatic phonetic labelling and phoneme border detection

The Voice Conversion community generally reuses algorithms known from speech recognition to automatically detect the phoneme sequence for each speech phrase of a speaker corpus [Abe, 1991]. The "Sentence HMM" based alignment of [Arslan, 1999] uses phonetically balanced template sentence from source and target speaker. Several normalization techniques are applied to the corresponding feature vectors to provide a more robust spectral estimate for the training of Hidden Markov Models (HMM). The best state sequence is estimated for each utterance using Viterbi decoding.

A speaker-independent HMM-based speech recognizer is used in [Ye and Young, 2004b] to force align the target data. Each target speech frame is labelled with a state id. Each utterance is thus represented by a state sequence. The study reports phone detection errors which result in inappropriate transformations being learned causing reduced performance results.

ircamAlign:
In [Lanchantin, 2007, Lanchantin et al., 2008] the linguistic structure is extracted from the text and aligned to the speech signal at the phone level. The HMM-based alignment system uses parts of the HTK Speech Recognition Toolkit [Young et al., 2006] being optimised for maximum segmentation accuracy. A rule-based transcription of the whole text sentences is given by Liaphon [Bechet, 2001], a french rule based text-to-phone synthesizer. First, an HMM phoneme recogniser is optimised to a phoneme bi-gram language model without using text transcription. Second, a multi-pronunciation phonetic graph is synthesized by Liaphon from textual information. The alignment output is therefore available in different phonetic alphabets such as X-SAMPA (Extended SAMPA) [Wells, 1997] [1].

One general problem of the approach is that the method requires manual annotation of the training corpus before the system can learn the phoneme borders. Hand segmentation is far from being error-proof and needs careful verification. It is cumbersome to determine manually the "true" phonetic borders. One strategy is to set phoneme borders by visual inspection of the signal segment in the time and spectral domain where the signal changes the most its pattern. Supplementary listening of the phoneme transition may aid. But it remains cumbersome to identify a certain time instant by listening since the human perception and cognition requires longer time segments to evaluate audio signals properly in this case.

4.2.2 Dynamic Time Warping (DTW)

The Dynamic Programming algorithm called Dynamic Time Warping (DTW) aligns two varying time series. The non-linear sequence alignment method DTW is necessary to account for the natural timing difference present between two speech phrases and their contained phonetic content. DTW measures the similarity or the global distance between two time series to find their optimal matching path through a feature trellis. DTW warps the sequences non-linearly in time.

[1] www.phon.ucl.ac.uk/home/sampa/x-sampa.htm

Voice Conversion errors can partially be attributed to possibly erroneous or at least unfortunate time-alignment paths of the DTW algorithm [Stylianou, 1996]. Misalignments lead to erroneous mappings in the training which cause as well the over-smoothing effect [Godoy, 2011]. Care has therefore to be taken that the DTW algorithm provides the best possible time alignment. It is advisable for some cases to inspect the alignment visually to assure it works to a certain extent as expected. Depending on the underlying data structure, the theoretically best path is close to the diagonal. Definitely the alignment has to insert or omit certain frames and leave the diagonal. It is up to the user to assure that the chosen DTW parameterization works correctly.

Typically it is beneficial to apply some pre-processing steps to the data set, e.g. a zero-mean and unit variance normalization to compensate for offset and ambitus differences, as well as the removal of linear trends and noise. For VC, the removal of silent frames aids to the robustness of the DTW alignment [Helander et al., 2008b]. Manhattan or Euclidean Distance evaluate the distance $D(n,m)$ of the two time series X for source and Y for target having length n and m such that a n-by-m matrix is constructed [Keogh and Pazzani, 2001]. Local and global constraints like Monotonicity (no negative decrease, no time reversal), Continuity (no jumps) or Boundary Conditions (start at bottom, finish at top) steer the path search [Keogh and Ratanamahatana, 2005]. The basic operations insertion, deletion and substitution enable the walk through the trellis. The best path minimizes the total accumulative cost. The resulting warping path W maps X on Y.

DTW can also be applied without the forced alignment of section 4.2.1, especially if no speech recognizer is available. The sole application of DTW to the whole phrase implies higher risks that the path through the trellis saturates towards a global border. Adding forced alignment based on phonemes performs typically better than phrase level alignment [Helander et al., 2008b, Rajput et al., 2012].

4.3 The One-to-Many Mapping Problem

Speaking the same phonetic content within the same linguistic context results into dissimilar acoustic events [Godoy, 2011, Godoy et al., 2012], due to natural differences in prosody, speaking style, pronunciation and articulation. Even the same speaker speaking the same phrase repeatedly is not able to reproduce the exact same signal waveform. Also, different articulations are summarized under the same phonetic label. Moreover, the natural speaking differences between source and target speaker result into mappings where one source frames shares a high correlation with several target frames. This one-to-many mapping problem associates one source frame from one acoustic event with many target frames of different acoustic events.

The study of [Mouchtaris et al., 2007] proofs the existence of one-to-many relationships. It investigates into approaches how to exploit such relationships in the context of VC. If two or more target feature relations are given for one source feature, an optimal estimator will average over the features such that the converted speech sounds muffled. The proposed Constrained Vector Quantization (CVQ) approach assumes the presence of one-to-many relationships between the source and target feature space. The solution is to not average over many relations but to select one relation. CVQ links the source speaker codebook X having M entries to the target speaker codebook Y having $M \cdot K$ classes. One entry of X is linked to K entries of Y. Quantized vectors of the target space are conditioned using hard classification on quantized vectors of the source space. A diffusion metric proofs that the conditional Y-space is higher than the corresponding X-space. This validates the one-to-many relationship assumption. However, the paper does not propose a solution how to design a VC system on the findings.

4.4 Transformation and synthesis

STAte-of-the-ART Voice Conversion algorithms being based on a simple source-filter model and being limited in most cases to a transformation of the spectral envelope do not require advanced signal transformation algorithms. Care has to be taken before synthesis that the converted LSF feature vectors do not contain adjacent LSF values lying below the theoretical lower distance limit [Backstrom et al., 2007]. A straight-forward Overlap and Add approach or the usage of any variant of Pitch Synchronous Overlap and Add (PSOLA) is sufficient [Stylianou, 2009]. Advanced parametric models require extended analysis and synthesis means such as signal models like HNM, DSM et al. presented in section 2.4. The transformation of prosodic features and source excitation features requires on the other hand complex signal transformation algorithms, notably for dynamic transposition and time scale modification, as well as the conversion between periodic and aperiodic components of the excitation signal. Numerous algorithms exist for dynamic transposition and time scale modification of speech [Charpentier and Stella, 1986, Stylianou, 1996, Stylianou, 2001, Röbel, 2010b, Röbel, 2010c, Degottex et al., 2010]. The modification of the excitation quality, however, requires the development of additional signal processing operators and speech models like the START speech systems introduced in section 3.8. The novel speech framework *PSY* proposed in chapter

6 constitutes continued work based on the SVLN method explained in section 3.8.4 to transform voice descriptors like the glottal excitation source.

4.5 Vector Quantization (VQ) and Codebook mapping

One early approach to convert speech of one speaker to sound like that of another by mapping acoustical features was proposed in [Childers et al., 1985]. Initial research in VC employed codebook mapping to exchange centroid vectors defined by weighted sums between source and target feature sequences. Vector Quantization (VQ) reduced the quantization errors from this hard-clustering approach [Abe et al., 1988, Abe, 1991]. K-means is used as data clustering algorithm to partition the acoustic space. The technology suffers from discontinuities in the transformation function at transitions between classes.

The "Speaker Transformation Algorithm using Segmental Codebooks" (STASC) of [Arslan, 1999] applies weights to each codebook entry such that a distance minimization between source and target feature vectors can be applied for conversion. Many features compared to other approaches are contained in the codebook: vocal tract, excitation, intonation, energy, and duration characteristics. The STASC system applies three different transformation schemata separately: Excitation, Vocal Tract and Prosodic transformation. However, the codebook weighting results into formant broadening due to the interpolation of the LSF vectors. A bandwidth correction algorithm reduces this effect.

4.6 Statistical VC models

4.6.1 Gaussian Mixture Models (GMM)

Source density model:
A GMM-based statistical method divides in [Stylianou, 1996] the acoustic space of the source speaker into overlapping acoustic classes. The acoustic space is span by a p-dimensional vector of discrete MFCCs representing the spectral envelope. The GMM-based acoustic classes describe a continuous parameter space instead of the strict class separation applied by the clustering methods of the preceding section 4.5. This soft clustering technique improves the spectral conversion quality compared to hard clustering approaches. The probabilistic classification of the GMM represents acoustical features pertaining to one sound character with a certain probability of belonging to each of its modelled acoustic classes. Each Gaussian component constitutes a Gaussian normal distributions $N(x;\mu,\Sigma)$:

$$N(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}\sqrt{\Sigma}}\, exp(-\frac{1}{2}(x-\mu)^T\Sigma_i^{-1}(x-\mu)) \tag{4.1}$$

A GMM models the probability distribution of one voice character as a mixture of Q Gaussian components. A GMM-based statistical model P_{GMM} approximates the acoustic space of the source speakers data x as follows:

$$P_{GMM}(x;\alpha,\mu,\Sigma) = \sum_{q=1}^{Q}\alpha_q N(x;\mu_q,\Sigma_q)\,,\ \sum_{q=1}^{Q}\alpha_q = 1,\ \alpha_q \geq 0 \tag{4.2}$$

Each Gaussian component of the GMM represents an acoustic class by its weight α_q, the mean vector μ_q and the covariance matrix Σ_q. The weights α_q express the prior probability if a speaker's feature vector x has been generated by component q. This soft partitioning approach assigns to each data point a varying degree of membership to all local models, instead of a classification scheme with hard boundaries. The conditional probability $p(c_q|x)$ of a GMM component q given x is derived from Bayes' rule [Bishop, 2006]:

$$p(c_q|x) = \frac{\alpha_q \cdot N(x;\mu_q^x,\Sigma_q^{xx})}{\sum_{p=1}^{Q}\alpha_p \cdot N(x;\mu_p^x,\Sigma_p^{xx})} \tag{4.3}$$

The GMM parameters $\{\alpha,\mu,\Sigma\}$ maximize the global likelihood $P(X)$ and are estimated via the Expectation-Maximization algorithm (EM) [Bishop, 2006]. The iterative EM parameter estimation has to be chosen since no closed-form analytical solution exists [Erro et al., 2008] to solve the summation over Q components inside the logarithmic expression of the implied Maximum Likelihood (ML) [Dempster et al., 1977].

This GMM approach for VC is used to model the acoustic space of the source speaker [Stylianou and Cappe, 1998]. It is denominated as the source density model [Helander et al., 2010]. A least-squares optimization minimizes the

total squared conversion error on the target speakers data [Stylianou et al., 1998]:

$$\epsilon = \Sigma_{t=1}^{n} |y_t - \mathcal{F}(x_t)|^2, \tag{4.4}$$

with n being the number of aligned source and target vectors and $\mathcal{F}()$ expressing the GMM-based function to convert from source X to target Y. The error minimization computes the parameters of the mapping function to retrieve the corresponding target feature vectors.

Joint density model:
The source density model for VC was extended in [Kain, 2001] to a joint density model. It employs a joint GMM $\{\alpha_q, \mu_q^z, \Sigma_q^{zz}\}$ of aligned source and target vectors. This eliminates the need to compute the target features by means of the error minimization of equ. 4.4. The conversion function can be obtained directly from the GMM parameters [Helander et al., 2010]. The joint density model uses parallel data $Z = \{z_k\}, z_k = [x_k^T y_k^T]^T$ with the weights α_q, the mean vectors μ_q^x for the source and μ_q^y for the target speaker, the co-variance matrices Σ_q^{xx} for the source and Σ_q^{yy} for the target speaker, as well as by the cross-variance matrices Σ_q^{xy} and Σ_q^{yx} for both speakers.

A two-dimensional acoustic feature vector with x for the source speaker and y for the target speaker consists of D-dimensional static and / or dynamic features describing the spectral envelopes and other acoustics features of the human voice [Rentzos et al., 2004, Toda et al., 2005, Rao and Yegnanarayana, 2006, Erro, 2008, Machado and Queiroz, 2010]. Parallel data sets of source and target features are time-aligned using Dynamic Time Warping [Helander et al., 2008b, Rajput et al., 2012]. A joint Gaussian Mixture Model is trained on the joint probability density $p(X, Y|\lambda^{(Z)})$ of the feature vectors X and Y of the source and respectively the target speaker, conditioned on the model parameters λ with the weights α_q, mean vector μ_q^Z and the co-variance matrices Σ_q^Z over Q mixture sequences:

$$p(X, Y|\lambda^{(Z)}) = \sum_{q=1}^{Q} \alpha_q \cdot N(X, Y; \mu_q^{(Z)}, \Sigma_q^{(Z)}) \tag{4.5}$$

This joint model provides implicit information about the individual acoustic spaces X and Y as well as their cross-covariance matrices. The transformation function $F(x)$ for the conversion from one to another voice character can be directly derived from the trained GMM model:

$$F(x) = \sum_{q=1}^{Q} p_q(x) \cdot [\mu_q^y + \Sigma_q^{yx} \Sigma_q^{xx^{-1}} (x - \mu_q^x)] \tag{4.6}$$

Each Gaussian components q is evaluated on its posterior probability $p_q(x)$ that it has been produced by observation x. This results into the weighted sum of the utilized linear regression models of each Gaussian. The corresponding mean vector μ_q^x per component is evaluated on its deviation from the observation x. The deviation is normalized by the co-variance $\Sigma_q^{xx^{-1}}$ of the source speaker. The co-variance between source and target speaker is re-established by applying the cross-variance matrix Σ_q^{yx}. The resulting feature vector converted to the target is corrected by mean vector μ_q^y of the target speaker. The difference between source and target feature vectors weighs the conditional target mean vector for each mixture. The conversion vector is assembled by the weighted sum of the conditional mean vectors and the conditional probability of the source vector belonging to each mixture.

Covariance matrix types:
Three types of different conversion function can be employed according to the different handling of the co-variance matrices [Stylianou, 2001]. Not considering the co-variance matrices reduces the GMM approach to VQ which only reflects the mean of each GMM component. The restriction to diagonal co-variance matrices uses only the data contained on the diagonal of each matrix. This requires more GMM components [Helander et al., 2010] to model the correlation of source and target feature space properly since the mean target values are emphasized. Full co-variance matrices exploit all data learned while the GMM training. This requires less GMM components since it weights more importance on the target co-variance matrices.

GMM parameterization:
The aim is to optimize the log-likelihood such that the data points are generated by a mixture of Gaussians [Stylianou, 2001]. The EM algorithm may not necessarily converge to a log-likelihood value constituting the global optimum but to a local optimum or even only a stationary value [Wu, 1983, Ververidis and Kotropoulos, 2008]. Care has therefore be taken to apply the best performing number of EM iterations given the chosen order of GMM components and the order p of the employed feature vectors modelling the underlying speech data. An additional statistical stop criterion like the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) may aid the parameterization of the GMM training process. Data over-fitting may occur if a too high number of Gaussians is fitted on comparably too few data [Erro, 2008, Godoy, 2011]. Similarly, an under-fitting of the GMM

model parameters is as well possible. Choosing an optimal performing order of Gaussian components is difficult. The correct order may even change per frame. The approaches presented in section 4.6.3.2 address the dimensioning problem in GMM-based VC systems. Another straight-forward solution is to reserve some (ideal) target phrases for an a-posteriori comparison with several converted phrases being transformed by a different component dimensionality. The highest conversion and synthesis quality is achieved by the best performing order [Erro, 2008]. The initialization pattern before training can play a crucial role in the data approximation. It plays a minor role if diagonal co-variance matrices are used [Stylianou et al., 1998].

Posterior probabilities:
The posterior probability $p_q(x)$ is computed per frame and component q to examine the likelihood of the current source feature vector x to be generated by the evaluated GMM component component q. In practise, for the most (stable) part of one phoneme, only one component exhibits a very high probability close to 1.0 while all other components show very low probabilities close to 0.0 [Qiao et al., 2011]. The weighted sum of linear regressions reduces actually to a single one. This is in fact a positive behaviour of GMM-based Voice Conversion. This strongly biased probability weight changes on phoneme borders to another GMM component to properly reflect the next phoneme type. However, it is not assured that the GMM conversion function reflects properly the underlying correlation between source and target feature vectors while this transition. Artefacts are introduced in the converted speech phrase if the posterior probabilities change rapidly [Helander et al., 2010].

Data Sparseness:
The GMM modelling may suffer from sparse values in the co-variance matrices. The problem may even occur for STate-of-the-ART corpora with phonetically balanced and rich sentences of ~10 - 20 minutes of speech recordings. Especially the combination of the normalization by the source speakers co-variance $\Sigma_q^{xx^{-1}}$ and the weighting by the cross-variance matrix Σ_q^{yx} results into very small values for a huge amount of matrix entries [Chen et al., 2003]. The small values lead to a reduction of the modelled variance such that the conversion function is reduced to the mean values μ_q^y of the target. The variance reduction to mean values is a further reason for the over-smoothing effect.

4.6.2 Hidden Markov Models (HMM)

Their state transition property facilitates HMMs for the usage as time alignment method, and as model to reflect any feature evolution over time [Hsia et al., 2007]. Conventional GMM-based mapping for VC is combined with a Bigram HMM (Bi-HMM) in [Yue et al., 2008] to employ a second feature alignment stage in training and conversion. The estimated and DTW aligned source and target LSF vectors are employed as training data to construct the Bi-HMM. Viterbi decoding is used to derive from the Bi-HMM the best state sequence with maximum joint probability of the source and target vectors. The state chain of aligned features is used to train a conventional GMM-based VC system. The same procedure is used for conversion. Another Bi-HMM is constructed from the source speech phrase to be converted. Viterbi decoding obtains the best state chain from the Bi-HMM. The converted to the target LSF parameters are computed per state.

4.6.3 Statistical model optimization

The most uses parameterization in conventional GMM-based VC system as introduced in the preceding section is to convert the LSF parameterized spectral envelope between speakers. The reduced variance of the converted LSF vectors introduced by conventional GMM models results into the broadening of formants after the re-transformation to the LPC representation [Bäckström and Magi, 2006]. Specific statistical means try to minimize the over-smoothing by applying different constraints or extended statistical modelling algorithms to maintain a certain natural feature variance. Moreover, the time-independence assumption of GMM-based VC system facilitates the exclusive consideration of single frames which neglects the interrelated evolution of each feature over time [Hsia et al., 2007].

4.6.3.1 Phonetic GMMs

One approach being studied over the last decade by the VC community is to restrict the GMM modelling to single phoneme classes such that one dedicated GMM model reflects exclusively per phoneme class the correlation between source and target speaker [Godoy et al., 2009]. The idea being that one part of the over-smoothing effect can be attributed to the automatic partitioning of the acoustic space. The GMM components overlap to create artificial phonetic groups. A more precise modelling can be achieved by means of splitting the speech data into

single phoneme classes. Conversion functions are available for each phoneme class. The intention is to reduce the influence of the one-to-many mapping problem discussed in section 4.3.

Phonetic information like phoneme type, point of articulation, manner and voicing are used per frame in [Duxans et al., 2004] within a Classification and Regression Tree (CART) to organize the overlapping regions of the acoustic space in one acoustic class per leaf. A GMM with 1 to 5 components is trained on the data pertaining to one leaf. The objective and subjective evaluation reports improvements for this phonetic organization approach.

The analysis of linguistic structure in [Li et al., 2010] is employed to classify phonemes. The phoneme classes are base on phonetic similarity in energy distribution, the separability of individual phonemes and syllable duration. One GMM may cover an individual or a group of phonemes.

In [Godoy et al., 2009], contextual phonetic-linguistic knowledge is employed to structure acoustic classes and to guide frame classification. Frames are grouped into phonemes. One GMM component is used per phoneme in the training stage. The learning and classification on the phoneme level reduces the one-to-many mapping errors. However, the one-to-many problem still exists in Phonetic GMMs [Godoy, 2011].

4.6.3.2 Dynamic Model Selection (DMS)

Gaussian Mixture Models restrict VC to a constant dimensionality of the source and target vectors containing static and dynamic features of the converted audio content. On the other hand, the adequate representation of the audio streams is data dependent. The local F_0 value or the entropy of the formant structure interferes per frame the best performing order which most accurately approximates the underlying spectral envelope, or other voice descriptor selected for conversion. The GMM modelling enables different dimensionalities [Kain, 2001]. A frame-wise order selection of the number of GMM components using pre-trained models can be executed. The order adaptation should improves the data representation and the VC quality. According to [Erro, 2008], a higher resolution and quality is achieved with a higher-order filter, while a more stable and less erroneous conversion is secured with a lower filter order. The lowest filter order with highest quality should thus be chosen.

The employment of a local order selection in [Villavicencio et al., 2009] for cepstrum-based spectral envelope estimation and residual computation improved the synthesis quality. The parameter dimensionalities can as well be adapted to the underlying speaker characteristic. The $\mathcal{T}_{LPC}$ estimator, applied to the internal $\mathcal{T}$ estimator of its auto-regressive model, uses local order selection in order to model signal features adapted to its dynamic characteristics. A perfect joint correlation between the source and target feature vectors for GMM-based VC is achieved by using different dimensions for the source and target envelope parameters. The GMM-based linear regression adapts then better, reduces mismatches between both envelopes at the conversion and synthesis stage and retains the real spectral information. The linear regression model for conversion may partially be reduced to the mean target values when dimensions of finer detail correlate poor between different voice types and result in small values in the co-variance matrices. This depends on the number of mixture components and on the used co-variance matrix type.

The trade-off between the goodness of fit and the model complexity is addressed in [Lanchantin and Rodet, 2010, Lanchantin and Rodet, 2011] by selecting the best GMM model order over time. It assumes that the best model changes over time. Several models are used in parallel. The most appropriate model is selected according to the likelihood given the acoustic feature vector of the source speaker per frame. Problems occur while conversion if the feature values exhibit a lower likelihood of having been generated by the model. In such cases a general model of lower complexity is selected for conversion. If the source data proofs to be well reflected by the training data, than a more complex and precise model can be selected, leading to a more accurate conversion. Dynamic Model Selection improves the proximity to the target voice and the signal quality. DMS requires higher computational costs since several models of different order have to be trained. Subjective tests indicate an improvement in terms of conversion score and synthesis quality.

4.6.3.3 Maximum A-Posteriori (MAP) adaptation

The approach of [Chen et al., 2003] modifies the conventional GMM conversion function of equ. 4.6 by removing the co-variance $\Sigma_q^{xx^{-1}}$ and the cross-variance Σ_q^{yx} matrix due to the problem of data sparseness, discussed in section 4.6.1. Instead, a GMM modelling only adapts the source speaker. It is used to derive a target speaker GMM by means of Maximum A-Posteriori (MAP) adaptation. MAP is known from Speaker Verification systems. The same mixtures of source and target GMM are used to shift the source μ_l^x towards the target μ_l^y mean vectors by means of

the weighted distance metric:

$$F(x) = x + \sum_{l=1}^{L} p_l(x)(\mu_l^y - \mu_l^x) \tag{4.7}$$

The source and target variances are assumed to be identical with respect to the class means. A median and low-pass filter is employed to suppress discontinuities producing artefacts. The method exhibits improvements in both conversion score and synthesis quality compared to the joined-density GMM baseline.

4.6.3.4 Covariance correction

The method proposed in [Lanchantin et al., 2011b] allows correcting the co-variance matrix values such that the GMM-based over-smoothing effect can be alleviated. Please inspect the study online being available at [2] for further details. Only some basic explanations are given here to understand the co-variance correction algorithm.

The joint probability distribution of the source and target features vectors can be modelled by a Gaussian mixture having K components as follows:

$$p(z_n|\phi) = \sum_{k=1}^{K} \alpha_k \cdot \mathcal{N}(z_n; \phi_k), \tag{4.8}$$

with mixture weight α_k and $\mathcal{N}(z_n; \phi_k)$ containing the mean vector $\mu_k^z = [\mu_k^x, \mu_k^y]^t$ and the co-variance matrix $\Sigma = \begin{bmatrix} \Sigma_k^{XX} & \Sigma_k^{XY} \\ \Sigma_k^{YX} & \Sigma_k^{YY} \end{bmatrix}$. The observed loss of the global variance resulting in the spectral over-smoothing effect is explicitly given by the residual co-variance matrix $\Sigma_G^{res} = \Sigma_G^{yx} \left(\Sigma_G^{xx}\right)^{-1} \Sigma_G^{xy}$. This co-variance matrix represents the co-variance not explained by the Gaussian mixture regression. The idea is to correct the converted frame values according to Σ_G^{res}. A new value $\hat{y}_n^*$ of the converted vector is given by means of a Cholesky de-correlation and correlation method:

$$\hat{y}_n^* = \left[(\hat{y}_n - \mu_G^y)^t L_G^{yy} \left(L_G^{y|x}\right)^{-1}\right]^t + \mu_G^y. \tag{4.9}$$

L_G^{yy} and $L_G^{y|x}$ are upper triangular matrices, leaving:

$$\begin{cases} \left(L_G^{yy}\right)^t L_G^{yy} = \Sigma_G^{yy} \\ \left(L_G^{y|x}\right)^t L_G^{y|x} = \Sigma_G^{yy} - \lambda_G \cdot \Sigma_G^{res} \end{cases} \tag{4.10}$$

λ_G is the weight parameter governing the amount of co-variance correction. Please note that setting λ_G=0.0 sets $\hat{y}_n^*$=$\hat{y}_n$ such that the resulting co-variance matrix reduces to the conventional GMM-based one. The variance is again expressed as the target variance conditioned on the source variance. The feature vector converted to the target is likely to contain a degradation in variance introduced by the over-smoothing effect. But, setting λ_G=1.0 results into the co-variance matrix of the target speaker. Accordingly, the co-variance correction parameter λ_G allows to adjust the variance of the target speaker (λ_G=1) and the variance of the target speaker given the variance of the source speaker (λ_G=0). Informal tests in the laboratory and the evaluation conducted in [Lanchantin et al., 2011b] suggest that setting λ_G=0.9 achieves a good performance.

4.6.3.5 Dynamic feature consideration

A GMM-based VC approach is extended in [Duxans et al., 2004] to an HMM-based system such that dynamic characteristics can be modelled. All the states are interconnected to each other in the employed ergodic HMMs. Each state of their emission probability function constitutes a Gaussian function. However, no significant improvements versus the conventional GMM baseline could be reported.

4.6.3.6 Global Variance (GV)

Conventional GMM-based mapping approach are frame-based and neglect the natural evolution of the employed voice descriptors over time on the one hand. On the other hand, the natural feature evolution is intrinsically

[2] Lanchantin 2011 - Extended Conditional GMM and Covariance Matrix Correction for Real-Time Spectral Voice Conversion

contained by the source feature vectors used for mapping. Still the consideration of dynamic features as outlined in the preceding section 4.6.3.5 is able to improve the VC quality.

The over-smoothing problem is addressed in [Toda et al., 2005] by means of combining a joint GMM with the Global Variance (GV) of the converted spectra. It is intended to re-establish the natural variance of the feature evolution over time which is lost due to over-smoothing. GV is a statistical post-filtering method to increase the transformed data variance. The additional usage of delta features shall alleviate spectral discontinuities. The correlation between frames is evaluated to influence the parameter generation process. The Global Variance of static feature vectors in each utterance is denoted by $v(y)$, with v being the feature vector and y being the target static feature vector. The dynamic behaviour captured with Global Variance is modelled by the likelihood function L:

$$L = log\{p(Y|X, m, \lambda)^{\omega} \cdot p(v(y)|\lambda_v)\} \tag{4.11}$$

It expresses the conditional probability of the target feature vectors Y and the probability of the Global Variance $v(y)$ of the target static feature vectors. The constant ω weights the likelihood of the target feature vector sequence of the model parameters mean vector $\mu(v)$ and co-variance matrix $\sum^{(vv)}$ for the Global Variance vector $v(y)$. It defines the ratio of number of dimensions between $v(y)$ and Y. The mean vector $\mu(v)$ and co-variance matrix $\sum^{(vv)}$ are calculated using the target static feature vectors from the training data, and the parameters in each utterance from the given source and converted feature vector.

The GV probability serves as penalty term such that the likelihood becomes a function of the parameter Y [Godoy, 2011] to be estimated. Global Variance improves in ML-based VC systems the conversion quality and performs better than other a-posteriori-based spectral enhancement methods. GV is included in [Toda and Young, 2009] as criterion to improve the trajectory HMM method which will be introduced in section 4.6.3.8.

4.6.3.7 Spectral Parameter Trajectory (SPT)

The method called Spectral Parameter Trajectory (SPT) in [Toda et al., 2007] considers static feature vectors and their dynamic continuation over spectral sequences. It is used in a GMM-based training and VC scheme. A Maximum Likelihood (ML) based estimation for the SPT of the GMM mapping generates parameters with dynamic features through the correlation of feature vectors over frames. A sub-optimal mixture sequence m' given the input feature vector and the model parameter for both source and target speaker is defined by the following probabilistic function:

$$m' = arg\ max\ p(m|X, \lambda^{(Z)}) \tag{4.12}$$

The optimized mixture component is employed to satisfy the optimal likelihood of one target speaker feature sequence, given the source speaker feature sequences and the model parameters, defined as follows:

$$P(Y|X, \lambda^{(Z)}) \approx P(m|X, \lambda^{(Z)}) \cdot P(Y|X, m, \lambda^{(Z)}) \tag{4.13}$$

The converted static feature vector Y' can be retrieved to model correlations over frames of the target feature vectors, which in turn is influenced by the source feature vectors. SPT exhibits improvements of the synthesis quality.

4.6.3.8 Trajectory HMM/GMM

The Trajectory HMM/GMM technology of [Zen et al., 2004, Zen et al., 2007, Zen et al., 2008, Zen et al., 2011] advances the idea of SPT shown in the preceding section 4.6.3.7. A conjoint consideration of static and dynamic features is reflected in the model on utterance level. The approach exhibits an improved synthesis quality.

4.6.3.9 Gaussian process experts

A Gaussian process (GP) is a non-parametric Bayesian model which is more robust against over-fitting. Conventional GMM-based VC methods are prone to over-smoothing due to their poor modelling and may suffer from over-fitting due to a poor generalization. The utilization of Gaussian processes in [Pilkington et al., 2011] derives static and dynamic experts from a joint probability density function, given the source feature vector x. The converted target feature vector $\hat{y}$ can be predicted from the static and dynamic experts. The VC mapping function is now a sample from a Gaussian process. The objective evaluation using Mel-cepstral distortions in dB exhibits

the best results for GP without dynamic features. The usage of dynamic features in GP still outperforms a GMM baseline method with and without dynamic features, and the trajectory HMM/GMM approach of the preceding section 4.6.3.8. The study concludes that Gaussian process experts are robust to over-fitting and over-smoothing and predict thus the target spectra more accurately.

4.7 Conversion of remaining voice descriptors

Conventional VC systems treat only the conversion of the Vocal Tract Filter (VTF), constituting the perceptually most important voice descriptor to describe a speakers voice identity. The extended VC systems presented in the following include additionally other relevant voice descriptors contributing to speaker identity.

4.7.1 Transformation of the residual signal

Conventional VC systems only convert the VTF. Accordingly, the source signal contains the aperiodic and periodic signal components of the glottal excitation source.

3^{rd} speaker effect:
Converting the spectral envelope with standard VC techniques and applying the residual of the source speaker to the converted target spectrum as in [Kain, 2001] may arise the perception of a 3^{rd} speaker identity. The converted phrase is perceptually neither close to the source nor to the target speakers voice identity. Other listening tests for VC report as well the perception of a 3^{rd} speaker if the residual was not considered at all for VC [Sündermann et al., 2005a, Sündermann et al., 2005c, Sündermann et al., 2005b, Erro, 2008, Lanchantin et al., 2011b]. Prosodic as well as acoustical and segmental cues are important for the identification of a speakers voice identity and the perceptual discrimination of a 3^{rd} speaker [Felps et al., 2009]. This indicates that the residual component contains speaker-dependent information. The reason being that the spectral features and the corresponding residuals are correlated. Achieving a VC of highest quality requires consequently the consideration of the residual.

The target residual can be constructed from the converted target speakers LSF vector with a prediction mechanism using a residual codebook [Kain and Macon, 2001]. A residual codebook of the target speaker provides perceptually close residuals for the converted target speaker frames. Each codebook entry has usually different weights assigned. A residual can be obtained from a weighted linear combination of codewords [Arslan, 1999, Uriz et al., 2011], or from an unit selection algorithm known from speech synthesis [Sündermann et al., 2005b]. The statistically closest residual can be retrieved by minimizing the distance or optimizing the global cost. Optimal residual computation uses prediction techniques [Sündermann et al., 2005a] to improve the residual processing. Still, the artificial construction of sinusoidal content from converted spectral envelopes combined with predicted residuals may lead to phase mismatches causing artefacts [Kain, 2001, Ye and Young, 2004a]. A final smoothing of the concatenated residual units is required to minimize possible artefacts. However, a certain level of naturalness may be lost.

In [Sündermann et al., 2005a] the modelling of the residual part using different variants like codebooks, to predict the residual from the converted feature vectors or to select the best matching residual from the target database is investigated. The best performing method is the residual selection technique with additional smoothing. The following works of [Sündermann et al., 2005b, Sündermann et al., 2005c] propose unit selection known from concatenative speech synthesis which outperforms residual selection and smoothing. A residual codebook is created by determining the residual for each mixture component. The selection of a perceptually closest target residual codebook entry predicted from the converted spectral envelope outperforms the compared baseline approach in [Duxans and Bonafonte, 2006, Ye and Young, 2004a]. This indicates a stronger intra-speaker correlation of the source and filter representation compared with the correlation of the residual between different speakers [Erro, 2008]. The residual prediction technique of [Percybrooks and Moore, 2007] employs transition probabilities between clusters to improve the probabilistic selection of codebook entries using GMMs. Further work presented in [Percybrooks and Moore, 2012] models the temporal dependencies present between frames in consecutive speech with an HMM.

4.7.2 Glottal excitation source modelling for VC

The mapping function of STAte-of-the-ART VC systems is usually conditioned on spectral envelope features only. This representation mixes the characteristics of the glottal excitation source with the vocal tract transfer function. This mixture of features leads to an increasing complexity in the parameter space that is used for the training of

the feature mapping. Taking into account a separated feature representation of glottal source and VTF parameters is considered as an important and challenging factor for VC systems [Rao and Yegnanarayana, 2006].

Recently new algorithms for the estimation of glottal pulse parameters [Vincent et al., 2007, Drugman et al., 2009a, Degottex et al., 2011a, Huber and Röbel, 2013] to separate the contribution of the vocal tract transfer function from an spectral envelope estimation [Röbel et al., 2007, Villavicencio et al., 2009] have been proposed. The speech processing systems discussed in section 3.8 provide means to transform the glottal excitation source.

Recently quite a few VC systems have used glottal pulse parameters [Childers, 1995, Rentzos et al., 2004, del Pozo and Young, 2008, Pérez and Bonafonte, 2011] for the separation of the glottal pulse in the spectrum from the spectral envelope features describing the vocal tract transfer function. The idea being that the mapping between source and target features should change with the phonemes, and that the decoding into phonemes can be performed by means of clustering spectral envelope features. Due to the fact that the glottal pulse and especially the glottal formant is part of the spectral envelope, the phonetic decoding in current VC systems is sub-optimal. An improvement can be expected if the effect of the glottal pulse is extracted from the spectral envelope before the statistical model is trained. Glottal pulse parameters can then be added separately as parameter and could be transformed explicitly.

The approach presented in [del Pozo and Young, 2008] employs GMM-based modelling to convert glottal source parameters between speakers in the context of Voice Conversion. The glottal excitation strength E_e, the normalized R waveshape parameters R_g, R_k, and R_a of the LF model and the energy of the aspiration noise N_e are utilized as feature set for the glottal source. The study reports an increase of the synthesis quality while the conversion score towards the target speaker does not improve compared to the baseline method.

The same glottal source feature set is utilized in [Pérez and Bonafonte, 2011]. Extended means are applied to analyze the aspiration noise. A CART-based decision tree classifies the VTF data into phonetic categories. A GMM is trained explicitly on the data attributed to each category. General improvements in synthesis quality are reported. The conversion score augmented only for one out of four source and target speaker pairs.

4.7.3 Modelling of prosodic features and F_0

Conventional VC systems are limited to deal with the short-term timbre modelling of the VTF. The frame-based means of conventional GMM-based VC systems neglect the variations in prosody present between source and target speaker. Speaking characteristics contributing to the perception of voice identities such as speaking rate, duration and timing are usually not or only to a minor extent reflected in the conversion. However, the long-term speech prosody variation is a relevant part contributing significantly to the perceived voice identity. Recently, studies intend at integrating global and short-term speech prosody variations into VC systems [Wu et al., 2010, Nose and Kobayashi, 2011]. A rich description of prosodic characteristics are: Intensity, Intonation, Timing, Voice Quality, and Degree of Reduction [Pfitzinger, 2006].

Intensity can be reliably measured by means of the Root-Mean-Square (RMS). RMS will be used in the proposed speech system of chapter 6 as basis of its energy modelling, introduced in section 6.4.1. Intensity can also be measured by smoothing the instantaneous amplitude of the Hilbert transformation. Short-term amplitude variations maximize speech loudness. Intensity is important to account for a natural speech synthesis.

Intonation can be characterized by different voice descriptors being related to speaking gestures and the expressive behaviour of human speech. The fundamental frequency F_0 plays an important role in this context.

Timing being interpreted as perceptual local speech rate (PLSR) constitutes another prosodic contour. Syllable durations and speech pauses characterize the timing structure of speech. Speech rate relates to the duration ratios between voiced and unvoiced speech segments [Lanchantin et al., 2011a].

Voice quality is described in detail in section 3.5. The classification of voice quality into the common major groups tense, modal, and relaxed relates voice quality as a function of vocal effort [Liénard and Barras, 2013]. Other voice quality groups exist as well.

Degree of reduction refers to the phonological mechanisms present when uttering stressed or unstressed vowels, syllables or words.

This rich description of prosody can be used to model by statistical means the speaking style of different speakers. The modelling of speaking style generally limits to the short-term instantaneous or the global mean and variance variations of the voice descriptors describing the voice identity of a speaker. The studies of [Rentzos et al., 2003, Rentzos et al., 2004] report that the transformed voice similarity did benefit from the inclusion of prosodic features into the VC system. In order to integrate properly the prosodic information for VC, the most appropriate approach

is to model the local and global speaking style of a speaker, with respect to the F_0, timing, and energy contours. The modelling of prosody should reflect as much as possible the prosodic features discussed in this section. The approach of [Yutani et al., 2008] employs a DP-GMM to model two different length sequences facilitating the simultaneous conversion of spectrum and duration. The conversion from neutral to expressive speech within the context of Voice Conversion in [Wu et al., 2006] uses an HMM including duration-embedded characteristics to achieve prosodic conversion using an expressive style-dependent decision tree. A neural network model is trained in [Rao and Yegnanarayana, 2007] on phonological, positional and contextual information to predict the durations of syllables.

The most simple method to transform the F_0 contour from source to target is a normalization in mean and variance. The approach of [Wu et al., 2010] uses histogram equalization to convert F_0 between speakers. It outperforms the conventional mean and variance normalization in objective and subjective tests. A syllable-based prosodic codebook uses linguistic information and segmental durations in [Helander and Nurminen, 2007]. It outperforms the baseline method using GMM-based F_0 conversion in a VC listening test.

4.8 Other VC approaches

4.8.1 Direct modelling of Spectral Peak Parameters

A Peak-HMM is proposed in [Godoy et al., 2010b, Godoy et al., 2010a] to narrow the feature mappings to the spectral peak level within frames. It transforms individual peaks instead of the vocal tract formants described by the spectral envelope. The method exhibits an increase in the transformed data variance to alleviate the over-smoothing problem. However, the focus on narrow mappings and transformations of individual peaks proved to be cumbersome and ineffective since the converted speech quality suffers from significant artefacts [Godoy, 2011].

4.8.2 Dynamic Frequency Warping (DFW)

VC approaches based on Frequency Warping maintain the natural signal contour describing formants. A higher synthesis quality is achieved. However, the conversion score improvements are limited and the voice identity of the target speaker is not fully converted.

The early approach of [Valbret et al., 1992] finds a frequency warping function $w'(\omega)$ of the spectra $X(\omega)$ and $Y(\omega)$ which minimizes the spectral distance between $X(w'(\omega))$ and $Y(\omega)$. The spectral tilt is deleted by eliminating the effect of the glottal excitation source. The acoustic space is divided by means of VQ since different warping functions are estimated for different phonemes.

4.8.2.1 Vocal Tract Length Normalization (VTLN)

The fast speaker adaptation technique denoted Vocal Tract Length Normalization (VTLN) accounts for length differences in the physical vocal tract by means of linearly warping the frequency axis [Panchapagesan and Alwan, 2009]. It is a standard technique for speaker normalization in speech recognition to compensate for the effect of speaker-dependent vocal tract lengths [Elenius and Blomberg, 2010]. The studies of [Sündermann and Ney, 2003, Sündermann et al., 2003] divide the acoustic feature space into classes to apply different types of VTLN-based frequency warping functions. Smoothing avoids discontinuities between different classes. It achieves a high synthesis quality while the speaker identity is not fully converted. VTLN is successfully applied to the LPC residual in [Sündermann et al., 2006b].

4.8.2.2 Weighted Frequency Warping (WFW)

The VC approach called Weighted Frequency Warping (WFW) of [Erro and Moreno, 2007, Erro et al., 2008, Erro, 2008] combines frequency warping with GMM-based modelling in a hybrid GMM-DFW approach [Godoy et al., 2012]. Frequency warping calculates optimal warping functions for each Gaussian component. The method enhances the VC performance both in terms of synthesis quality and conversion score. WFW estimates its frequency warping functions $W(f)$ for source vector x as a weighted combination of m basis functions $W_i(f)$ with GMM-based probability weights $p_i(x)$:

$$W(f) = \sum_{i=1}^{m} p_i(x) \cdot W_i(f) \tag{4.14}$$

Mean LSF vectors represent the central formant frequencies and define the piecewise-linear frequency-warping functions $W_i(f)$ per Gaussian component i. The soft classification from the GMM probabilities produces a smooth evolution of the transformation function. LSF vectors are only used to model and classify the WFW weights. The WFW conversion function $W(f)$ is applied to the original complex spectral envelope $S(f)$:

$$S_\omega(f) = S\left(W^{-1}(f)\right) \tag{4.15}$$

The application of equ. 4.15 reallocates the formants on the frequency axis but leaves intensity, bandwidth and spectral tilt remain unmodified such that a different energy distribution remains over frequency bands. The final converted spectrum

$$S'(f) = G(f) \cdot S_\omega(f) = \left(\left|\frac{S_g^{(k)}(f)}{S_\omega^{(k)}(f)}\right| * B(f)\right) \cdot S_\omega(f) \tag{4.16}$$

is computed as in [Erro et al., 2010b] by applying the energy correction filter $G(f)$. It modifies the magnitude spectrum, the spectral tilt and energy distribution but does not affect small spectral details. $B(f)$ represents a triangular shaped smoothing-in-frequency window whose shape controls the similarity-quality trade-off.

Automatic mapping of formants (AMF):
No one-to-one correspondence of formants can be assured despite the possibly high correlations of source and target formant structures. The AMF method minimizes a spectral distortion measure D to search for pole-pair combinations between $X(\omega)$ and $Y(\omega)$. AMF operates on an all-pole representation in the continuous spectrum and uses formant positions in radian frequency ω as pivot points.

4.8.2.3 Dynamic Frequency Warping with Amplitude scaling (DFWA)

The DFWA approach proposed in [Godoy et al., 2011, Godoy et al., 2012] employs the method to pick peaks from the magnitude spectrum presented in section 4.8.1. It does not require contrary to WFW a separate GMM transformation to adjust the spectral power after DFW [Erro, 2008]. DFWA mappings are established on a global acoustic class level using histograms to model the statistical distribution of formant frequencies and amplitudes. DFWA offers a flexible and versatile VC framework without having to rely on aligned speaker frames. Extensive objective and subjective evaluations suggest that DFWA outperforms existing GMM and DFW-based methods. It maintains spectral details in the transformed spectral envelopes but only slight improvements in the conversion scores are reported.

4.8.2.4 Correlation-based Frequency Warping (CFW)

The Frequency Warping approaches WFW and DFWA of the preceding sections are based on different piece-wise linear transformation in the spectrum. The DFW transformation function is based on determining pairs of spectral magnitude peaks, AMF uses formant pairs. The approaches minimize the spectral distance of the warping path through a trellis span by the peak or formant sequence. However, the spectral pairing of peaks or formants may be error-prone due to misleading associations. Additionally, the accurate estimation of formants is cumbersome [Fulop, 2003]. The Correlation-based Frequency Warping (CFW) proposed in [Tian et al., 2014] learns frequency warping functions by maximizing the correlation between the converted and the target spectra. CFW is thus more robust to inaccurate formant estimation. The method outperforms a DFW and an AMF based baseline approach in synthesis quality and conversion score on a VC listening test.

4.8.3 Frame Selection

Several VC approaches based on Frame Selection have recently been proposed by the VC community. The algorithms combine several techniques known from unit selection for TTS synthesis.

4.8.3.1 Unit selection for TTS-based speech synthesis

Unit selection as known from speech synthesis implies the knowledge about each database entry. Corpus-based concatenative synthesis employs a unit selection algorithm which selects units or a unit sequence best matching the target specification by utilizing sound character descriptors [Hunt and Black, 1996]. Selected sound units can also be transformed to better match the target description.

Larger DBs accomplish higher qualities in the perceived sound output due to a higher availability of sample units for each selection step to match the target more precisely. The unit with the lowest descriptor distance or the highest spectral match is selected. Path search algorithm search through the DB space to approximate the optimal path.

Statistical models destroy the natural contour and evolution of the utilized feature vectors. The idea of many Frame Selection approaches for VC is to extract features directly from the training database. This should avoid any feature smoothing such that the target speakers voice identity and a high synthesis quality is preserved.

4.8.3.2 Proof-of-concept analysis

The selection of larger units such as diphones requires a huge amount of training data. The paper of [Helander et al., 2007] analyzes if a reasonably VC performance can be achieved by means of Frame Selection using smaller databases. As experimental setup, LSF vectors are taken as the perfect achievable LSF conversion from the actual target phrase. A frame-based selection on the target database is executed on the perfectly converted LSF vectors. This demonstrates which best performance can be achieved with the frame-based selection method concerning the upper bound of the given the data set. The Frame Selection is based on a weighted squared error criterion and on perceptual weights approximating the mechanisms of the human ear [Paliwal and Atal, 1993]. However, the study concludes that the selection of single LSF frames is not suitable for small DBs. Even the best results of the experimental design to test the upper bound did not exhibit results below a spectral distortion measure of 2 dB. The following sections present further research using the idea of Frame Selection for VC, despite the proof-of-concept analysis of [Helander et al., 2007] denies its feasibility.

4.8.3.3 Frame Selection with trade-off parameterization

The unit selection approach of [Sündermann et al., 2006a] employs single speech frames and is independent of additional linguistic information. It is based on minimizing the following cost function between the source feature sequence x_1^M and a non-parallel target feature sequence y_1^N:

$$\hat{y}_1^M = \underset{y_1^M}{arg\,min} \sum_{m=1}^{M} \left\{ \alpha \cdot S(y_m - x_m) + (1-\alpha) \cdot S(y_{m-1} - y_m) \right\} \tag{4.17}$$

The target cost $\alpha \cdot S(y_m - x_m)$ evaluates the Euclidean distance between source and target frames. The concatenation cost $(1-\alpha) \cdot S(y_{m-1} - y_m)$ reflects the Euclidean distance between neighbouring target frames. The parameter α adjusts for the trade-off between the fitting accuracy of evaluated source and target frame versus the a spectral continuity criterion.

4.8.3.4 Frame Selection using GMM pre-conversion

Three different Frame Selection setups for VC are examined in [Dutoit et al., 2007]:
1. Frame Selection without GMM-based feature mapping,
2. Frame Selection with GMM-based feature mapping, utilizing the source speakers excitation, and
3. Frame Selection with GMM-based feature mapping synthesizing an excitation converted to the target.

The proposal consists of two independent system blocks:
X -> Y': Conversion from source X to target Y' using as GMM-based mapping function the source density model of section 4.6.1.
Y' -> Y'': Speech-to-speech synthesis from target Y' to target Y'' to achieve a more accurate approximate of target Y.

The proposed system is similar to unit selection using Viterbi smoothing in concatenative speech synthesis. The presented algorithm employs 32 ms frames as feature for Frame Selection, instead of diphones or phones as in TTS-based unit selection systems. A pre-clustering divides the target database into 256 group of frames to reduce complexity while retaining quality.

The Frame Selection algorithm uses Viterbi decoding to select frames Y from the target database. This minimizes the overall distance between frames and the output sequence Y'. The overall distance is a combination of target and concatenation cost. Method 3 achieves the highest conversion score (2.77) but the lowest MOS sound quality rating (2.56) among the evaluated approaches.

The systems proposed in [Gu and Tsai, 2014, Gu and Tsai, 2015] employ similar means as in [Dutoit et al., 2007]. First, an initial GMM-based feature mapping from source X to target Y' is executed. Second, Frame Selection

converts from target Y' to target Y''. The difference being that a speech recognizer selects one out of 39 segmental GMMs to optimize the first mapping stage. Additionally, the pitch is converted between source and target speaker.

4.8.3.5 Frame Selection using K-Histograms

The first proposal in [Uriz et al., 2008] is followed by a series of further studies using Frame Selection for VC. The systems discussed in the following share similarity to the early VC approach of [Abe et al., 1988], with continuity constraints to avoid concatenation artefacts. The Frame Selection approach of [Uriz et al., 2008] is build on minimizing the distance between source and target feature frames:

$$min_{\hat{x}} = \left[\sum_i d(x_i, \hat{x}_i) + d(\hat{y}_{i-1}, \hat{y}_i) \right] \tag{4.18}$$

A feature vector sequence x of the source speaker is used to find an optimal feature vector sequence $\hat{x}$ from the training data of the source speaker. The feature vector $\hat{y}$ constitutes the corresponding feature vector sequence retrieved from the target speakers training data. The cost function is basically the same as equ. 4.17 without the trade-off parameter α. This reduces the VC mapping function to a lookup table of source-target feature pairs. The distance term to the left can be interpreted as a source matching cost: Find for the source feature vector X the optimal source feature vector X' in the training data. The distance term to the right can be interpreted as a target concatenation cost: Minimise the discontinuities between the corresponding target feature vectors in the training data to reflect their dynamic behaviour. This minimizes discontinuities introduced by the one-to-many problem [Godoy et al., 2009] at which one source frame may have many corresponding target frames. The standard Frame Selection method FS uses Viterbi decoding to find according to the source sequence, the target and the concatenation cost the optimal target sequence. The excitation of the source speaker and the converted features are used for synthesis. The privileged method $FS\,opt$ employs instead of the source the target sequence to find the closest target feature vector in the training data. $FS\,opt$ measures the highest sound quality and conversion score achievable by FS.

$FS\,opt$ receives in a listening test a similarly high conversion score than FS. A corresponding objective evaluation by means of the P distance measure proposed in [Kain, 2001] exhibits a high difference between both methods: $FS\,opt$ ~0.7, FS ~0.25. The objective measure reflects the expectation of the outperformance of the privileged method $FS\,opt$ while the listeners could not confirm the expectation. However, both methods receive comparably low synthesis ratings when compared to a GMM baseline method and the WFW method of section 4.8.2.2.

A mean and standard deviation normalization of the source speakers LSF vectors according to the target speakers distribution is added in [Uriz et al., 2009c] to the Frame Selection algorithm. Improvements of the introduced normalization are not clearly comprehensible due to the usage of a different P distance measure compared to [Uriz et al., 2008].

The algorithms K-Histograms and K-Means to cluster categorical and respectively numerical data are examined for VC in [Uriz et al., 2009b]. The LSF coefficients are clustered by means of K-Histograms in [Uriz et al., 2009a]. The transformation between histograms is performed by means of a Cumulative Distribution Function (CDF). The VC system based on Frame Selection is evaluated with and without K-Histograms. The usage of K-Histograms achieves a higher conversion score and sound quality rating in a listening test compared to the conventional Frame Selection method, to the Frame Selection approach of 4.8.3.4 and to the GMM baseline. However, since this approach used the excitation of the source speaker, the final system of this series proposed in [Uriz et al., 2011] uses additionally residual conversion and averaging over concatenated segments to reduce concatenation differences and to augment the synthesis quality.

4.8.4 Unit selection

Several VC approaches based on unit selection known from TTS systems have recently been proposed by the VC community. The Frame Selection based variants presented in the preceding section 4.8.3 reduce the selection of whole units to its nomenclature, the selection of one single frame. Conventional GMM-based VC systems and the Frame Selection approaches assume the independence of consecutive frames. Reflecting only one frame to calculate the concatenation cost ignores temporal information and does not consider a smooth frame-to-frame transition in the target space. It results into discontinuities at the frame concatenation points which affects the perceptional quality of the synthesized speech phrase.

The most promising perception of how to enforce Voice Conversion by robust and effective means is found in the literature to the authors knowledge in [Wu et al., 2013] for the time being. It selects units being called exemplars which span over multiple frames. The algorithm works roughly as follows:

1. *Data alignment:* Find source-target exemplar pairs from parallel training data, align source X to target Y by means of DTW explained in section 4.2.2 without using transcription information to detect phonemes as in section 4.2.1.
2. *Trellis generation:* Select several target candidate exemplars found in the target corpus for each source exemplar found in the source phrase.
3. *Cost function:* Calculate target and concatenation cost.
4. *Sequence estimation:* Apply Viterbi decoding to find the optimal target exemplar sequence which minimizes the overall target and concatenation cost.
5. *Parameter generation:* Generate the converted speech parameters from overlapping exemplars by considering a smoothing window as temporal information constraint.

A mel-cepstral distortion metric is utilized to tune the frame amount per exemplar and the smoothing window length. The proposed system outperforms a Joint density GMM and a conventional Frame Selection method. The approach alleviates the selection of frames from the target database having different prosodic and phonetic content.

4.9 Non-parallel VC

4.9.1 Text-independent VC

A non-parallel corpus differs in the phonetic content spoken by source and target speaker. The sentences do not have to be the same as in the parallel VC case. This special VC use case is denominated text-independent VC. It adds more freedom and flexibility to apply VC by relieving the constraint which requires both speakers to utter the same phonetic content. Text-independent VC usually transforms a non-parallel into a quasi-parallel training corpus such that conventional VC means can be applied afterwards. However, many proposed text-independent VC system found in the literature suffer from a degradation of the conversion score and the synthesis quality [Duxans et al., 2006, Erro, 2008, Erro et al., 2010a, Silén et al., 2013]. The degradation results from the means required to establish a mapping of corresponding features. The feature space of each corpus is segmented into frames or units. They can then be clustered into similar groups to define artificial phonetic categories which may or may not coincide with single phonemes [Machado and Queiroz, 2010]. The artificial grouping allows the mapping of the selected voice descriptors from the source into the target speakers feature space. An advantage of the artificial phonetic categories is that it does not require a-priori phonetic or linguistic information.

4.9.2 Cross-lingual VC

The non-parallel design of text-independent VC systems allow that source and target speaker talk in different languages. It requires that the same or at least similar phonemes exist in both languages. However, most language pairs suffer from the fact that a certain subset of both phonetic alphabets is not covered. Strategies have to be implement known from the text-independent mapping of artificial phoneme groups to provide data for the subset of missing phonemes [Machado and Queiroz, 2010].

Units covering complete phonemes instead of single frames can straight-forwardly be selected from the target corpus. Unit selection is a well-known technique and the quasi standard in TTS speech synthesis [Hunt and Black, 1996, Taylor, 2009]. Unit selection techniques are employed in [Duxans et al., 2006] for intra-lingual and cross-lingual VC to avoid the need for a parallel corpus. A TTS speech synthesis system is employed to select acoustic units from the target speakers corpus such that they can be aligned with the corresponding units of the source speaker. Then conventional VC means are applied to convert between a speaker pair using a trained GMM along with different algorithmic variants like residual selection or prosody transformation. Another version utilized for comparison is the direct concatenation of selected target speaker units given by the TTS system. It does not apply a GMM for conversion and includes the source speakers prosody for synthesis. It is surprisingly rated with a lower speaker identity score than the GMM-based versions. One explanation could be the automatic unit segmentation which may be erroneous such that artefacts are introduced [Duxans, 2006].

4.10 Source speaker selection

Informal work on speech encoding suggests that some voices seem to be better prone for their parameterization by LPC coefficients. Similarly, the choice of the selected speaker pair for a VC setup is important and has significant impact on the VC performance [Turk and Arslan, 2005]. Since each VC application desires to convert into the voice identity of one specific target speaker like a celebrity person, the question remains which source speaker is best suited to achieve the best VC performance. The perceptual evaluation of choosing one voice imitator from a huge database is time consuming and expensive. Moreover, it does not assure that the VC algorithm performs best with the perceptually chosen speaker. An objective criteria aids the selection process by comparing acoustical features of source and target speaker.

The method proposed in [Turk and Arslan, 2005, Turk and Arslan, 2009] and filed under the patent of [Turk et al., 2007] determines the objective distances of different acoustical features per source-target speaker pair. A rich set of acoustical features is employed: VTF, Pitch, Duration, Energy, Spectral Tilt, Open Quotient, Jitter and Shimmer, harmonic energy ratio between a low to high frequency region, $H1$-$H2$, and the shape of an EGG recording. The objective distances are used to estimate the subjective quality of the Voice Conversion output. An Artificial Neural Networks (ANN) is trained to learn the regression between the acoustical distance measures and subjective scores derived from corresponding listening tests. The trained ANN algorithm selects and ranks each source speaker in terms of the expected output quality for the conversion to a specific target voice. The rankings predicted by the algorithm are compared in [Turk and Arslan, 2009] with listeners preferences. A ten-fold cross-validation between the ANN-based rankings and listening test results exhibits an average correlation of 0.84 for a male-to-male and 0.58 for a female-to-female conversion.

A large-scale automatic voice casting system based on the measurement of voice similarities is proposed in [Obin et al., 2014]. It classifies speech into classes, instead of expressing voice similarity directly in the acoustic space. The concatenated output probabilities per class form a vector representing the vocal signature of a speech recording. A similarity search is performed on the vocal signatures. A set of target actors being the most similar to the source actor is determined. The multi-label system clearly outperforms standard speaker recognition systems.

4.11 Objective and subjective evaluation methodology

Several methods for the objective measurement of the signal quality and the similarity between the pre-defined and the converted target voices exist in the literature to evaluate the performance related to the individual components and the different Voice Conversion methods. The results of the objective evaluation measures should validate the corresponding subjective listening tests in order to exclude discrepancies between signal processing based measures and the human auditory perception. Both evaluation methods evaluate if the converted sound character is perceived as being realistic with intelligibility, naturalness and without artefacts.

The objective evaluation needs to identify and define relevant features which properly reflect the Voice Conversion performance. However, it remains cumbersome to define objective distance measures being perceptually meaningful [Machado and Queiroz, 2010]. A parallel data approach using utterance pairs of two voices speaking the same sentence can be evaluated by comparing the converted output target vector with the original target vector. Distance measures determine the acoustic alteration by calculating the differences introduced by the conversion from the source to the target sound character. This measure can only be as accurate as the natural differences in timing and prosody decrease the similarity between speakers. Such differences exist even if the same sentence is spoken by two speakers having a similar prosodic behaviour since even the same speaker would naturally differ in timing when repeating the same sentence. Objective evaluation measures are especially helpful for the evaluation and the continuous monitoring of algorithmic changes.

Subjective perceptual tests are the ultimate test for a VC system, since the objective measure can only approximately simulate the human auditory perception and cognition.

4.12 Conclusions

Spectral envelope parameterization:
The widespread usage of parameterizing the LPC encoded spectral envelope by means of LSF vectors implies one drawback. The same indices of LSF coefficients do not necessarily capture the same formants or respective frequency regions [Godoy, 2011]. The modelling of the acoustic spaces of both speakers and the employed conversion

function may be systematically biased such that erroneous mappings are applied. The direct usage of parameters describing the spectral envelope such as cepstral coefficients or True Envelope is prohibitive due to the too huge dimensionality which complicates the statistical modelling and feature mapping.

Conventional VC systems:
Conventional statistical models used in VC systems are prone to over-smoothing and over-fitting due to poor modelling and poor generalization. This destroys the coherent formant interrelation in the converted spectral envelope sequences.

VC with dynamic features:
The utilization of dynamic features extracted from successive frames tries to maintain or re-establish the coherent natural formant contour. It reduces but does not suppress the degradation introduced by the over-smoothing effect and the frame-by-frame conversion. The synthesis quality is still not satisfying and the converted phrase is still not perceived as the target speakers voice identity.

VC using an extended feature set:
Residual: The residual constitutes an important part of a speakers voice identity. Its consideration in a VC should alleviate the perception of a 3^{rd} speakers voice. Residual modelling augments the synthesis quality and conversion score.
Glottal excitation source: The additional separate conversion of glottal parameters has been successfully addressed and represents important means in VC.
Prosody: Basic means have been addressed by the VC community to transform the speaking characteristics. However, further work is required with a higher level of abstraction to provide a more detailed prosodic feature conversion.

VC using Frequency Warping:
Frequency-warping approaches maintain the natural formant structure of human speech such that a high synthesis quality is achieved. However, the voice identity of the target speaker is still not sufficiently captured.

VC using Frame and Unit Selection:
The novel approaches to extract data directly from the target database are promising. The coherent natural formant contour per spectral frame is maintained. However, the feature succession over frames is not sufficiently preserved. Many concatenation points may appear if the selection algorithm matches frames from segments having different prosodic and phonetic content.

The following chapter 5 proposes means to robustly estimate the deterministic part of the glottal excitation source. The reason being that the conversion of glottal pulse shapes contributes as shown in section 4.7.2 to the VC conversion score. The subsequent chapter 6 presents a novel speech framework designed to analyse, transform and synthesize an extended set of voice descriptors. Both novel contributions, the glottal pulse shape estimator and the speech framework, are utilized in chapter 7 which presents a novel VC system developed at IRCAM. This new VC system has been further advanced by combining it with the contributions of the chapters 5 and 6.

Chapter 5

Contribution - Glottal excitation source modelling

Every wave is related to every other wave.

Sogyal Rinpoche - The Tibetan Book of Living and Dying

5.1 Introduction

This chapter presents an extended version of the work conducted for the two publications [Huber et al., 2012, Huber and Röbel, 2013] published while this work and listed in section 9.1. It summarizes the results of investigating into the estimation of the shape of the deterministic component of the glottal excitation source from a speech recording. Means are presented to extend the STAte-of-the-ART method of section 3.7.3 based on the phase minimization criterion.

The Liljencrants-Fant (LF) model of [Fant et al., 1985, Fant, 1995] and 3.2.3 describes the deterministic glottal source component by modelling the glottal volume-velocity flow and its derivative. The glottal source shape parameter R_d, introduced in [Fant and Liljencrants, 1994, Fant et al., 1994, Fant, 1995, Fant, 1997] and section 3.3.1, parameterizes the LF model by an efficient regression reducing the LF parameter space from a three-dimensional to a one-dimensional one. The R_d regression parameter describes along its range the transition in voice quality from a tense to a modal to a breathy voice character.

The adaptation and range extension of the R_d parameter regression was introduced in [Huber et al., 2012], further discussed in [Huber and Röbel, 2013] and is here presented in section 5.2. It enables to coherently assess the normal [Fant, 1995] and the upper R_d range [Fant et al., 1994]. Two variants to adapt and extent the R_d range of are discussed.

The basic model for the human speech production is introduced in section 5.3.1. The phase minimization based methods to estimate the R_d parameter are outlined in section 5.3. An optimized R_d estimation using Viterbi smoothing is explained in section 5.4. A novel attempt called Viterbi steering to optimize Viterbi smoothing is presented in section 5.5. Two extensive objective evaluation tests analyze the performance of the presented algorithms to estimate the deterministic part of the glottal excitation source. The evaluation results on a synthetic test set and on natural human speech are presented in the sections 5.6.3 and respectively 5.6.4. A summary and conclusions about the advancements in estimating the deterministic part of the glottal excitation source are given in section 5.7.

5.2 The adapted and extended R_d range

Informal experimental evaluations with natural human speech signals show the importance to cover more extreme adducted and abducted phonations. This requires to extend the normal R_d range [0.3, 2.7] defined in [Fant, 1995] to lower R_d values up to R_d=0.1 (extremely tense adducted phonation) and to higher R_d values up to R_d=6.0

(extremely relaxed abducted phonation). The upper R_d range defined in [Fant et al., 1994] for R_d>2.7 is required to describe abducted phonations occurring mainly at phoneme transitions as well as at word and speaking pause boundaries. The R_d range extension covers more glottal source shapes contained in the analyzed speech signal and augments thus the robustness of the R_d estimation. The experimental findings show that the R_d range extension is especially beneficial to estimate and synthesize R_d for abducted phonations observed for more breathy voice qualities and at word or speaking pause boundaries. It will be shown on two voice quality transformation tests presented in section 6.6 that the synthesis of extremely tense adducted phonation types in the lower R_d range [0.1, 0.3] may lead to an unnatural sounding re-synthesis or artefacts.

The R waveshape parameters R_a, R_k and R_g are introduced in section 3.3.1. Their definition and accordingly their parameterization of the LF model requires to properly reflect the proposed adaptation and R_d range extension. The prediction R_{*p} waveshape parameters R_{ap}, R_{kp} and R_{gp} are defined by the corresponding prediction function to derive their value from an R_d value. The prediction of the waveshape R_{*p} parameter set for the normal R_d range is given in [Fant, 1995]. The only definition found in the literature for the prediction of the waveshape R_{*p} parameter set for the upper R_d range $R_d > 2.7$ is given by the equations 8 to 11 in [Fant et al., 1994].

Unfortunately, the equations defining the R_d regression [Fant et al., 1994, Fant, 1995] do not produce smooth contours of the R waveshape parameters when changing R_d continuously between the normal and the upper R_d range. The equations proposed in [Huber et al., 2012, Huber and Röbel, 2013] coherently cover the extended R_d range [0.1, 6.0]. Fig. 2 of [Fant, 1995] depicts the contour of the R waveshape parameters and for OQ for the R_d range [0.3, 5], using only 10 sampling points. It is the sole figure found in the literature illustrating these contours for both the normal and the upper R_d range established by Fant. However, joining the curves for the normal and the upper R_d range of the parameters R_{kp}, R_{gp} and OQ reveals a discontinuity at the interconnection point R_d=2.7, shown in fig. 5.1.

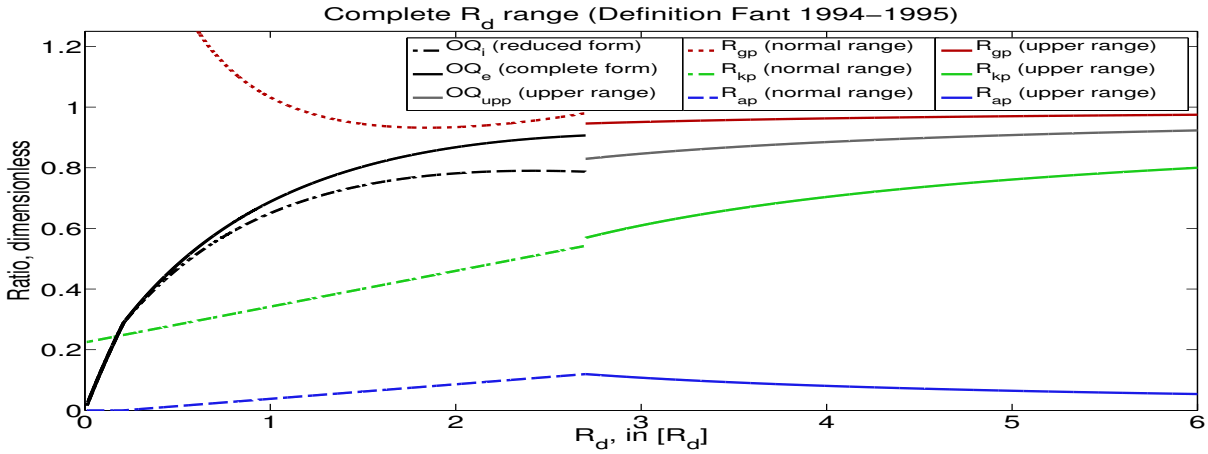


Figure 5.1: *Original waveshape R_{*p} parameter contours*

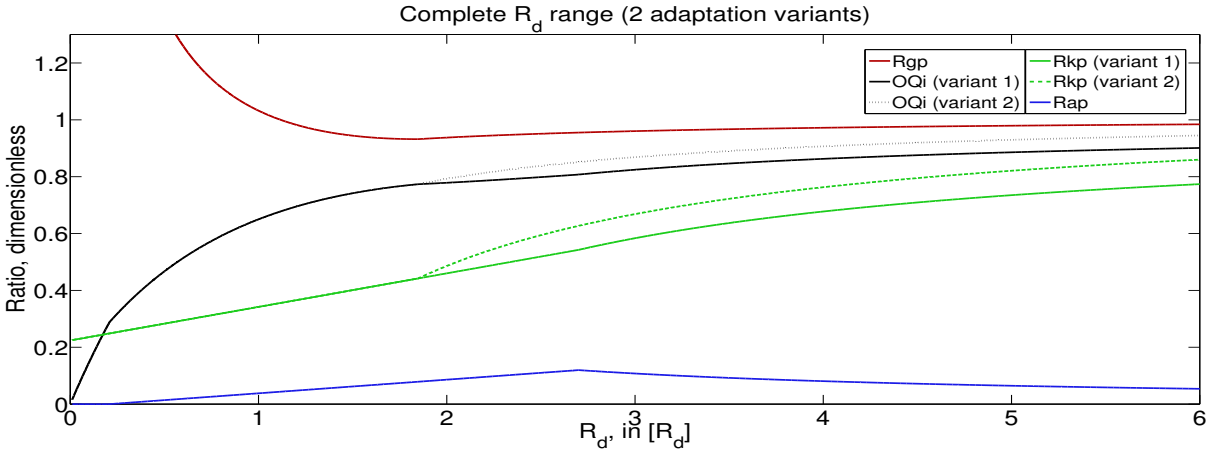


Figure 5.2: *Adapted waveshape R_{*p} parameter contours*

The parameter contour of OQ for the upper R_d range [Fant et al., 1994] does not fit to OQ for the normal R_d range [Fant, 1995], neither to OQ_e in complete form nor to OQ_i in reduced form [Fant, 1997]. An adaptation of the equations defining the computation of the R_{*p} waveshape parameter set to establish continuous parameter curves when changing R_d between both ranges were proposed in [Huber et al., 2012, Huber and Röbel, 2013].

Additionally, the R_d range was extended to [0.01, 6] to cover more extreme tense or breathy voice qualities. This requires to set R_{ap} for R_d<0.21 to zero to avoid a negative return phase t_a. The R_{gp} curves of the normal and the upper R_d range are adapted at the minimum of the convex function of R_{gp} for the normal R_d range at R_d=1.8476. An offset of $9.3552 \cdot 10^{-3}$ has to be added to the R_{gp} contour of the upper R_d range to compensate for a remaining difference. The equations 5.1, 5.2, 5.3, and 5.4 follow the adaptation proposed in [Huber et al., 2012], denoted here as variant 1. It requires to add an offset of $-4.2753 \cdot 10^{-2}$ to $R_{kp_{2.70}}$ of the upper R_d range.

Another adaptation variant 2 was introduced in [Huber and Röbel, 2013] taking into account that R_{kp} for the upper R_d range depends on R_{gp}. This dependency can be comprehended in [Fant et al., 1994]. R_{kp} is adapted to depend on the upper range equation not at R_d=2.7 as defined in [Fant et al., 1994] but already at R_d=1.8476. This renders a more exact conformance with the R_d regression adaptation of R_{gp} in [Huber et al., 2012]. R_{kp} for adaptation variant 2 is consequently denominated as $R_{kp_{1.85}}$. An offset compensation of $+4.2753 \cdot 10^{-2}$ adapts $R_{kp_{1.85}}$ accordingly at the upper R_d range. The contours of the R_{*p} parameter adaptation for both variants are shown in fig. 5.2.

Please note that the OQ contour of the reduced form OQ_i, if derived from the original R_d regression of [Fant, 1995, Fant, 1997], exhibits for the normal R_d range a maximum of OQ_i=0.790 already at R_d=2.42 and a lower value of OQ_i=0.787 at R_d=2.70. But, OQ should increase over the R_d range from lower values for tense adducted to higher values for breathy abducted phonations [Henrich et al., 1999, Doval et al., 2006]. The decrease of OQ_i for the R_d range]2.42 2.70] introduces ambiguities into pulse parameter estimation algorithms. The R_d regression adaptation variants suppress these ambiguities by establishing a strictly increasing OQ_i contour over the whole R_d range. The reduced OQ_i form defined in equ. 3.2 exhibits maximum values of OQ_i=0.90 at R_d=6.0 for variant 1 and OQ_i=0.95 for variant 2. The following set of equations defines the adaptation of the predicted R_{*p} waveshape parameters for the regression of the extended R_d range:

$$R_{ap} = \begin{cases} 0 & \forall\ 0.01 \le R_d < 0.21 \\ (-1 + 4.8 \cdot R_d)/100 & \forall\ 0.21 \le R_d \le 2.70 \\ (32.3/R_d)/100 & \forall\ 2.70 < R_d \le 6.00 \end{cases} \tag{5.1}$$

$$OQ_{upp} = \quad 1 - 1/(2.17 \cdot R_d) \quad \forall\ 1.8476 \le R_d \le 6.00 \tag{5.2}$$

Variant 1:

$$R_{kp_{2.70}} = \begin{cases} (22.4 + 11.8 \cdot R_d)/100 & \forall\ 0.01 \le R_d \le 2.70 \\ (2 \cdot R_{gp} \cdot OQ_{upp}) - 1.0428 & \forall\ 2.70 < R_d \le 6.00 \end{cases} \tag{5.3}$$

$$R_{gp} = \begin{cases} 0.25 \cdot R_{kp_{2.70}}/(\frac{0.11 \cdot R_d}{0.5 + 1.2 \cdot R_{kp_{2.70}}} - R_{ap}) & \forall\ 0.01 \le R_d \le 1.8476 \\ 9.3552 \cdot 10^{-3} + (596 \cdot 10^{-2}/(7.96 - 2 \cdot OQ_{upp})) & \forall\ 1.8476 < R_d \le 6.00 \end{cases} \tag{5.4}$$

Variant 2:

$$R_{kp_{1.85}} = \begin{cases} (22.4 + 11.8 \cdot R_d)/100 & \forall\ 0.01 \le R_d \le 1.8476 \\ (2 \cdot R_{gp} \cdot OQ_{upp}) - 0.9572 & \forall\ 1.8476 \le R_d \le 6.00 \end{cases} \tag{5.5}$$

$$R_{gp} = \begin{cases} 0.25 \cdot R_{kp_{1.85}}/(\frac{0.11 \cdot R_d}{0.5 + 1.2 \cdot R_{kp_{1.85}}} - R_{ap}) & \forall\ 0.01 \le R_d \le 1.8476 \\ 9.3552 \cdot 10^{-3} + (596 \cdot 10^{-2}/(7.96 - 2 \cdot OQ_{upp})) & \forall\ 1.8476 < R_d \le 6.00 \end{cases} \tag{5.6}$$

Both adaptation variants establish a continuous contour of the predicted R_{*p} waveshape parameters over the extended R_d range. The latter can be defined with a lower limit which has to be higher than R_d=0.0. The upper limit can be chosen in practice with an R_d value in the R_d range [3, 6]. R_d upper limit values and R_d values being higher than R_d>6.0 may not reflect different glottal pulse shapes in-between them being distinctive enough to discriminate to a perceptually significant extent different voice qualities. Please note that the equations defining R_{ap} and OQ_{upp} are the original equations as defined by Fant for the normal R_d range in [Fant, 1995] and the upper R_d range in [Fant et al., 1994]. The equations defining R_{ap} and OQ_{upp} for lower R_d values R_d<0.3 result from the R_d range adaptation, proposed in [Huber et al., 2012, Huber and Röbel, 2013].

Fig. 5.3 illustrates for the R_d adaptation variants 1 and 2 the relation of the Open Quotient OQ versus the return phase t_a and the asymmetry coefficient α_m. Both examples are given for the extended R_d range [0.1, 6]. Fig. 5.4 exemplifies the relation of the return phase t_a versus the asymmetry coefficient α_m if both are derived from a parameterized using the adapted and extended R_d range. The three-dimensional interrelation of the return phase t_a, the asymmetry coefficient α_m and the Open Quotient OQ using the parameterization of the adapted and extended R_d range is depicted for both adaptation variants in fig. 5.4.

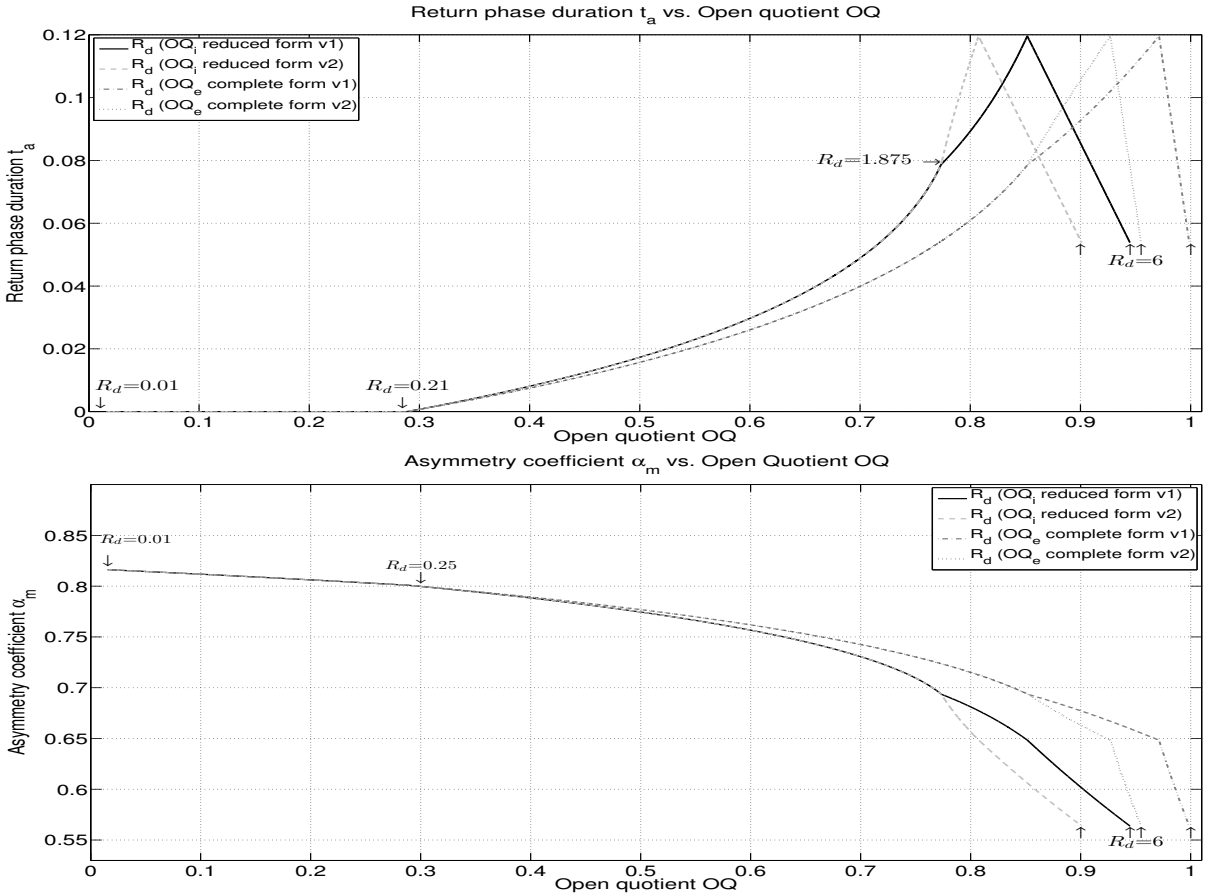


Figure 5.3: *Relation of return phase t_a and asymmetry coefficient α_m versus Open Quotient OQ*

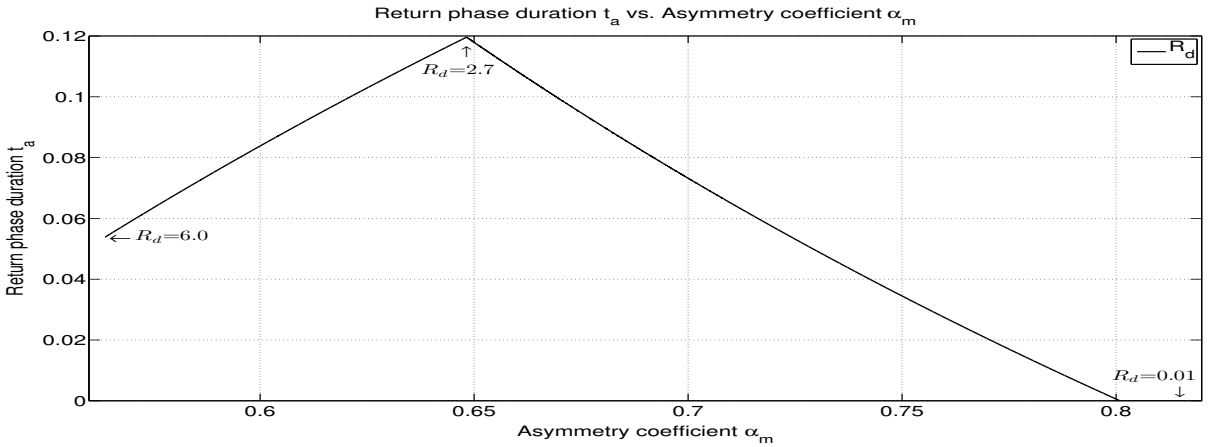


Figure 5.4: *Relation of return phase t_a vs. asymmetry coefficient α_m*

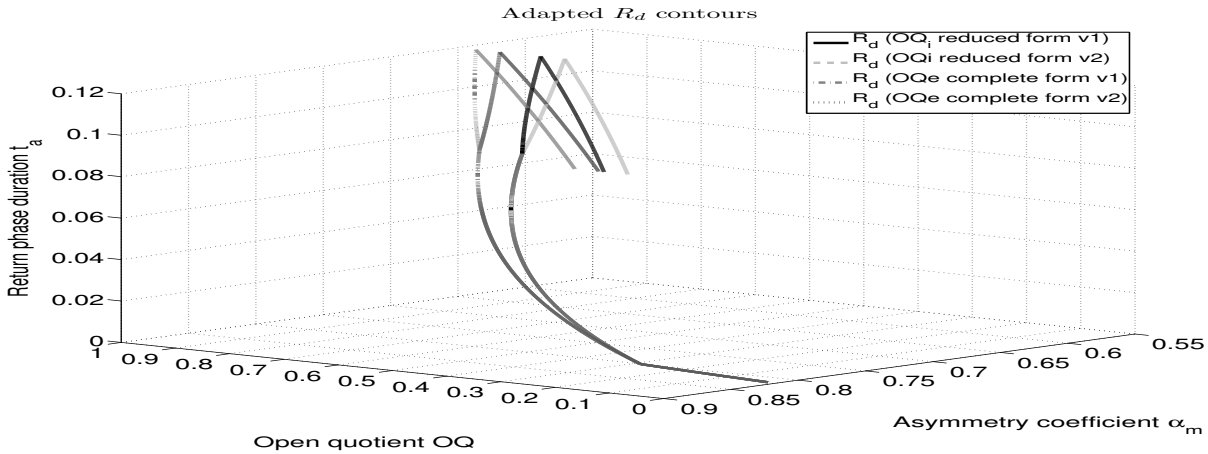


Figure 5.5: *Interrelation of return phase t_a, asymmetry coefficient α_m and Open Quotient OQ*

5.3 Glottal source estimation using phase minimization

Three phase minimization variants extend the baseline phase minimization method proposed in [Degottex, 2010, Degottex et al., 2011a] and presented here in section 3.7.3. The phase minimization variants presented in this section are build on the baseline method called MSPD2I1 in [Huber et al., 2012]. The variants called MSPD2I0, MSPD2I2 and MSPD2IX were as well proposed in [Huber et al., 2012]. The first two proposed methods MSPD2I0 and MSPD2I2 extent the phase minimization paradigm by applying a different number of consecutive integration steps. The third proposed method MSPD2IX achieves a more robust estimation of the glottal shape parameter R_d by means of superimposing the evaluation errors calculated by the different phase error methods. Different tests presented in the evaluation section 5.6 suggests that MSPD2IX achieves the most reliable R_d estimation.

5.3.1 A deterministic-only voice production model

The deterministic part of the voice production model used for the analysis consists of an extended source-filter model for stationary speech in the spectral domain

$$S(\omega) = G_{R_d}(\omega) \cdot C(\omega) \cdot L(\omega) \cdot H(\omega, F_0, D), \tag{5.7}$$

with ω being the angular frequency. Equation 5.7 defines the deterministic part of the voice production model which is composed of a representation of the following components. The LF shape parameter R_d of [Fant, 1995] parameterizes the Liljencrants-Fant (LF) glottal pulse model of [Fant et al., 1985] of the glottal excitation source $G_{R_d}(\omega)$. The vocal tract transfer function is denoted as $C(\omega)$. It is assumed to have a minimum phase filter response. The radiation at lips and nostrils level is given by an approximate representation being $L(\omega) = j\omega$. The harmonic structure $H(\omega, F_0, D)$ is parameterized by the fundamental frequency F_0 and the delay D. The fundamental frequency F_0 is estimated using the approach presented in [Yeh and Röbel, 2004]. The delay between the glottal pulse sequence and the frame center of the applied (Hanning) window is expressed in terms of the phase delay D of the fundamental.

5.3.2 The phase minimization paradigm

The phase minimization algorithm estimates first a sinusoidal model for each speech signal frame. It is transformed into a harmonic model describing a single pitch period with the discrete spectrum S_k which follows equ. 5.7. Each bin k represents a single quasi-harmonic sinusoidal partial k. The partials are estimated from a Fourier transform of a windowed speech signal. The harmonic model is assumed to be noise free for each harmonic k up to K. The highest harmonic sinusoidal partial K is determined by rounding the ratio 8 kHz to F_0 to the nearest integer value. The procedure described in [Stylianou, 2001] is utilized to construct S_k from a signal frame by means of finding the parameter set having minimum error. According to equ. 5.7 the voice production model of the deterministic component of the speech signal can be simplified into:

$$S_k = G_k^{R_d} \cdot C_{k-} \cdot L_k \cdot e^{jk\phi} \quad k \in [0, 1, \ldots, K] \tag{5.8}$$

The linear phase term $e^{jk\phi}$ defines the time position of the glottal pulse in the period. $G_k^{R_d}$ represents the LF glottal model being parameterized by the R_d parameter. The vocal-tract filter C_{k-} is assumed to be minimum phase. The term L_k represents the radiation at the lips and nostrils level. According to [Markel and Gray, 1976] the filter L_k can be approximated by a time derivative and is thus set to $L_k = jk$.

The algorithm proceeds by means of testing the minimum phase property of the VTF spectrum by division in the frequency domain of the glottal model that is obtained for a sufficiently compact grid of R_d values:

$$C_k^{Rd} = \mathcal{E}_-\left(\frac{S_k}{G_k^{R_d} \cdot jk}\right) \tag{5.9}$$

The operator $\mathcal{E}_-(.)$ is the minimum phase realization of its argument that is calculated by means of using the real cepstrum $c(n)$ introduced in section 2.2.2. C_k^{Rd} will represent for the correct R_d parameter the minimum phase transfer function of the VTF [Degottex et al., 2011a].

The convolutive residual $R_k^{(\theta,\phi)}$:
The VTF expression C_k^{Rd} of equ. 5.9 is inserted into the voice production model of equ. 5.8. It forms the mathematical basis for the convolutive residual $R_k^{(\theta,\phi)}$ defined in equ. 5.10:

$$R_k^{(\theta,\phi)} = \frac{S_k}{e^{jk\phi} \cdot G_k^\theta \cdot jk \cdot \mathcal{E}_-(S_k / G_k^\theta \cdot jk)} = e^{-jk\phi} \cdot \frac{S_k}{\mathcal{E}_-(S_k)} \cdot \frac{\mathcal{E}_-(G_k^\theta \cdot jk)}{G_k^\theta \cdot jk} \tag{5.10}$$

The shape of the glottal pulse is denoted by θ. ϕ refers to the position of the glottal pulse with respect to the fundamental period in the time domain [Degottex, 2010]. The division of S_k, G_k^θ and jk by their respective minimum phase versions flattens their amplitude spectrum. The remaining convolutive residual $R_k^{(\theta,\phi)}$ is thus all-pass for any chosen glottal model. Its modulus is of unit amplitude: $|R_k^{(\theta,\phi)}| = 1 \; \forall \; k, \theta, \phi$. Therefore, a mismatch of the model parameters to describe the observed speech signal affects only the phase spectrum of $R_k^{(\theta,\phi)}$. The result is that the better the estimate of the fitted voice model S_k, the closer is the convolutive residual $R_k^{(\theta,\phi)}$ to a Dirac delta function with a flat amplitude and zero phase spectrum [Degottex et al., 2011a]. Hence, the smaller the phase spectrum of $R_k^{(\theta,\phi)}$ the closer is the R_d value utilized to synthesize the glottal model G_k^θ to the true glottal shape contained in the observed signal [Huber et al., 2012]. This solution is unique as long as the glottal pulse that is present in the speech signal is covered by the R_d parameter space.

The matching of the different glottal pulse shapes θ over a tense R_d grid deletes to a different extent all components of the signal such that a phase error ϕ_ϵ due to matching differences (desired) and a linear phase component $e^{-jk\phi}$ (not desired) remains. The objective of the following section is to determine which phase minimization variant best deletes the non-desired linear phase component $e^{-jk\phi}$ to optimally evaluate the desired matching differences such that the glottal pulse shape parameterized by R_d is most robustly estimated.

5.3.3 The phase minimization variants

The objective function of each phase minimization algorithm estimates R_d by means of minimizing the deviation of the convolutive residual from a minimum phase transfer function. An additional constant factor is introduced as error by the simplification of $L(k)$ into jk which does not affect the results.

The phase minimization on the convolutive residual $R^{(\theta,\phi)}$ is dependent on the remaining linear phase component ϕ reflecting the time position of the current window position. The solution is to apply a 1^{st} differentiation which sets the linear phase slope to a constant, and to apply a 2^{nd} differentiation which sets the constant to zero. However, the differential operator introduces a high pass effect which can be minimized by consecutive integral operations. Section 5.6 presents different evaluations examining which phase minimization variant achieves the highest performance in estimating R_d.

Differentiations:
The underlying deterministic-only voice production model of equ. 5.7 is restricted to the observed number K of harmonic sinusoidal partials X. The difference operator $\Delta\angle X_k$ approximates the frequency derivative of the phase for each harmonic k:

$$\Delta\angle X_k = \angle X_{k+1} - X_k, \tag{5.11}$$

with $\angle$ denoting the phase angle as principal value of a complex number X.

The main problem with the convolutive residual $R_k^{(\theta,\phi)}$ is its dependency on the pulse position ϕ because an arbitrary delay D is introduced into the voice production model of the equations 5.7 and 5.8. It depends on the delay between the pulse position and the frame center in terms of the phase delay D of the fundamental [Huber et al., 2012]. This dependency can be removed as in [Degottex et al., 2011a] by means of applying a 2^{nd} order difference operator:

$$\Delta^2\angle X_k = \angle\frac{X_{k+1} \cdot X_{k-1}}{X_k^2}. \tag{5.12}$$

$\Delta^2\angle X_k$ is centered on each of the harmonics k of the convolutive residual $R_k^{(\theta,\phi)}$ in the complex plane. This removes the linear phase component of the observed phase spectrum and removes therefore the dependency to ϕ. Only the deviation from a linear phase trend remains. The phase of the convolutive residual $R_k^{(\theta,\phi)}$ can be compared to the optimal target value 0 to find the optimal R_d parameter. The phase minimization based glottal source estimation is with this operation independent to the window position relative to the pulse position in time.

Integrations:
Please note that the 2^{nd} order difference operator of equ. 5.12 not only removes the linear phase. It also applies a high pass filter to the phase difference that will be used to determine the optimal R_d parameter. Thus, subsequent integrations are required to suppress the influence of the high pass filter. To compensate this high pass filter a phase integration according to equ. 5.13 can be applied:

$$\Delta^{-1}\angle X_k = \angle\prod_{n=1}^{k} X_k \tag{5.13}$$

This inverts the high pass filter without re-establishing the linear phase trend. Each integration step leads to a different weighting of the phase errors ϕ_ϵ of the convolutive residual $R_k^{(\theta,\phi)}$.

A different number of integration steps L can be applied on the remaining phase errors ϕ_ϵ of the convolutive residual $R_k^{(\theta,\phi)}$, with L being in the set [0,1,2]. An objective function minimizes each differentiation and integration result. The algorithm selects the lowest phase error residual ϕ_ϵ being considered as the most likely R_d estimate. The denomination Mean Squared Phase Differentiation Integration specifies the acronym MSPDI. The number of differentiation steps is indicated by a subsequent number after D. Similarly, the number of integration steps is indicated by a subsequent number after I. The application of first 2 differentiations and then 0, 1, and respectively 2 integration steps on the phase errors ϕ_ϵ of the convolutive residual $R_k^{(\theta,\phi)}$ is contained in the acronyms of the three phase minimization methods MSPD2I0, MSPD2I1, and MSPD2I2 [Huber et al., 2012]. The different error measures of this three methods are combined by the method MSPD2IX in form of a weighted sum. The summing of the phase errors ϕ_ϵ augments the robustness of the method MSPD2IX. If one phase error measure is erroneous it is likely to be suppressed by the other two phase errors ϕ_ϵ. The different objective functions described as [MSPD2I0, MSPD2I1, MSPD2I2] present a different and not necessarily correlated error surface. The error surfaces drawn from confusion matrices for each objective function will be given in section 5.6.1.

The objective function MSPD2I0:
The objective function to minimize the results of equ. 5.12 is the method MSPD2I0:

$$\text{MSPD2I0}(\theta, N) = \frac{1}{N}\sum_{k=1}^{N}\left(\Delta^2 \angle R_k^\theta\right)^2 \tag{5.14}$$

Please note that only the 2^{nd} order difference operator of equ. 5.12 is applied to the convolutive residual $R_k^{(\theta,\phi)}$ of equ. 5.10.

The objective function MSPD2I1:
An anti-difference operation (Δ^{-1})

$$\Delta^{-1}\Delta^2 \angle X_k = \angle \prod_{n=1}^{k} \frac{X_{n+1} \cdot X_{n-1}}{X_n^2} \tag{5.15}$$

applied to the second order phase difference of equ. 5.12 performs an integration according to equ. 5.13. This retrieves again the first order frequency derivative representation.

The results of equ. 5.15 are evaluated by the corresponding objective function named MSPD^2 in [Degottex et al., 2011a]. The consistent naming convention given in this context is MSPD2I1. Its objective function is defined by equ. 5.16:

$$\text{MSPD2I1}(\theta, N) = \frac{1}{N}\sum_{k=1}^{N}\left(\Delta^{-1}\Delta^2 \angle R_k^\theta\right)^2 \tag{5.16}$$

The method MSPD2I1 constitutes the baseline method of the presented phase minimization variants. It is used by the original implementation of SVLN [Degottex, 2010] presented in section 3.8.4 to estimate R_d.

The objective function MSPD2I2:
Applying two anti-difference operators Δ^{-2} to the second order phase difference of equ. 5.12 computes the twice differentiated and twice integrated phase term:

$$\Delta^{-2}\Delta^2 \angle X_k = \angle \prod_{n=2}^{k}\prod_{n=2}^{k} \frac{X_{n+1} \cdot X_{n-1}}{X_n^2} \tag{5.17}$$

The corresponding objective function to minimize the results of equ. 5.17 is the method MSPD2I2:

$$\text{MSPD2I2}(\theta, N) = \frac{1}{N}\sum_{k=1}^{N}\left(\Delta^{-2}\Delta^2 \angle R_k^\theta\right)^2 \tag{5.18}$$

MSPD2I2 is the most selective and most distinctive among the different phase minimization methods. It weights slight differences of the matched glottal model to the observed glottal source the most.

The objective function MSPD2IX:
A linear superposition of the error surfaces of each preceding objective functions MSPD2I0, MSPD2I1 and MSPD2I2 results into the combined objective function called MSPD2IX:

$$\text{MSPD2IX(w0,w1,w2)} = \text{w0} \cdot \text{MSPD2I0} + \text{w1} \cdot \text{MSPD2I1} + \text{w2} \cdot \text{MSPD2I2}$$

The weights $w0$=$w1$=$w2$ defined in equ. 5.19 allow to adjust the influence of each objective function MSPD2I0, MSPD2I1 and MSPD2I2 on their combination MSPD2IX. The motivation for the linear superposition of MSPD2IX and suggestions on the weighting will be discussed in detail in section 5.6.1. Each objective function is analyzed visually on their theoretical performance of the error surfaces retrieved from confusion matrices.

5.3.4 A summary of drawbacks estimating the glottal excitation source

Please note that if the duration of the impulse response of the VTF is close to or above the period the evaluation of the minimum phase property of the VTF becomes problematic. Ambiguous solutions may arise in these cases which may lead to erroneous R_d (contour) estimates. Therefore, higher fundamental frequencies F_0 decrease as reported in [Huber et al., 2012] the robustness and the accuracy of the R_d estimation.

The major conditions that are likely to result in an erroneous R_d estimation are discussed in this section. The findings are based on the analysis of a large number of speech signals, conducted for the evaluation shown in chapter 5.6. The R_d parameter estimation operates frame-based by selecting at each analysis step the glottal pulse shape corresponding to the lowest remaining phase error residual ϕ_ϵ. Errors may arise from:

- environmental or aspiration noise,
- general ambiguities from non-linear phase distortions present in the phase residual [Walker and Murphy, 2007, Ó Cinnéide et al., 2011],
- situations where the R_d parameterization of the LF model restricts the synthesized and estimated glottal source shape to an subspace of the LF model parameter space which does not cover the true glottal source contained in the signal,
- the fact that the precise minimum phase impulse response of the vocal tract cannot be observed with the real cepstrum used by the phase minimization methods [Degottex, 2010] from signal parts where only few stable harmonic partials are available before being masked by noise. This situation occurs predominantly for higher fundamental frequencies, at phoneme transitions or at word and speaking pause boundaries. Moreover, the stationarity of the vocal-tract filter over the length of the analysis window may not be anymore valid at these situations. The phase minimization paradigm [Degottex et al., 2011a] may systematically be mislead in such segments.

The analysis of these findings result into means increasing the robustness of the R_d parameter estimation, presented in the sections 5.4 and 5.5.

5.4 Viterbi smoothing

Many different approaches have been proposed by the speech research community over the last decades to solve the problem of estimating all components that are used in a model of voice production, e.g. as defined in equ. 2.2, from a speech recording. However, for the moment none of these algorithms is sufficiently robust to allow for a reliable analysis of natural human speech. This section presents an approach to add robustness to the estimation of the glottal shape parameter R_d. First it is based on the phase minimization variants introduced in section 5.3.3. But additionally it utilizes the Viterbi algorithm of [Forney, 1973] to address the problems in estimating the glottal excitation source as discussed in section 5.3.4. The attempt shall provide means to obtain a physiologically consistent estimate of the glottal pulse shape parameter R_d from synthetic and natural human speech signals. The smoothing with the Viterbi algorithm suppresses unnatural jumps and avoids local instabilities of the R_d estimator within short-time segments.

The random influences listed in section 5.3.4 can partially be reduced by smoothing over time with the Viterbi algorithm, as long as these problems are present over a relatively short-time segment. The probabilistic model of standard Viterbi smoothing is defined as follows:

Observation probability $P(O|X)$:
The speech production model is approximate on a reasonably small grid of R_d values. Each of the N_{R_d} R_d values represents a hidden state X_i of a finite-state Markov process that defines the random process to establish the Viterbi algorithm. The phase error ϕ_ϵ of the convolutive residual determines the log-likelihood of the observation. The probabilistic distribution of the observation is configured so that the minimum error of the residual phase E_{R_d}=0 has maximum probability. The emitted observations over time span up the lattice over which the Viterbi algorithm determines the optimal path representing the lowest overall error.

Transition probability $P(X_n|X_{n-1})$:
The transition probability is described as a function of the R_d parameter slope $\Delta R_d/\Delta_n$, with Δ_n representing the

time difference between two analysis frames such that the transition probability can consistently handle different STFT analysis step sizes. The probabilistic distribution of the transition is modelled as Gaussian with zero mean and variance σ_T^2.

Optimal Viterbi smoothing path:
The sequence of observations is segmented into regions of voiced speech. The voicing decision is based on the presence of frames containing valid glottal closure instants (GCI). The SIGMA algorithm of [Thomas and Naylor, 2009] is utilized to detect the GCIs. The R_d estimation evaluation on natural human speech signal of section 5.6.4 is restricted to only consider voiced segments having at least five consecutive voiced frames. The R_d sequences having maximum probability are determined by applying the Viterbi algorithm independently to each voiced segment. The log-likelihood of each sequence is

$$L(p) = \sum_n \log(P(O|X_p(n)) \cdot P(X_p(n)|X_p(n-1))), \tag{5.19}$$

where n is the discrete time and p is a path through the state space of the process. The log probability function E_{R_d} of the observation probability $P(O|X)$ is inserted into equ. 5.19. Its distribution is scaled with parameter α_a. The probabilistic distribution $\Delta R_d/\Delta_n$ of the transition probability $P(X_n|X_{n-1})$ is included with its scale parameter γ_g to find

$$\bar{L}(p) = -\sum_n \alpha_a \cdot E_{R_{d_n}} + \gamma_g \cdot \frac{\Delta R_d}{\Delta_n} + c_d. \tag{5.20}$$

The term c_d is a constant gathering all the contributions of the constant scaling factors of the distributions. This constant term can be ignored by the Viterbi algorithm. The scaling factor γ_g of the log-likelihood of the transition is factored out. This leaves the parameter $\alpha = \alpha_a/\gamma_g$ as control parameter. It defines the probability function that is used to perform Viterbi smoothing on all sequences p:

$$\bar{\bar{L}}(p) = -\sum_n \alpha \cdot E_{R_{d_n}} + \frac{\Delta R_d}{\Delta_n}. \tag{5.21}$$

The experimental setup of section 5.6.4 examines which value for α creates the best Viterbi smoothing results without the application of the novel Viterbi steering.

5.5 Viterbi steering

Each of the phase minimization based R_d estimators can under the conditions discussed in section 5.3.4 be systematically skewed. In addition, the Viterbi smoothing of the preceding section 5.4 cannot correct possibly skewed or biased R_d contours in longer time segments. A possible systematic bias for each R_d estimator occurs predominantly in regions where only few stable harmonic sinusoids are available, e.g. at phoneme transitions, word and speaking pause boundaries or for higher fundamental frequencies. However, the phase minimization paradigm of [Degottex et al., 2011a] requires a precise estimation of the minimum phase response of the VTF from the observed partials. This condition may not be given if only few stable harmonic sinusoids are observable.

The Viterbi steering approach presented in this section shall provide means to correct a possible systematic bias of the R_d estimation in the mentioned problematic regions. The steering of the Viterbi algorithm improves Viterbi smoothing by exploiting the co-variation of other voice descriptors. The Viterbi steering algorithm is based on GMMs representing the joint density of the voice descriptors and the Open Quotient (OQ_{EGG}) estimated from corresponding electroglottographic (EGG) signals. A conversion function derived from the trained mixture model predicts the second OQ_{GMM} estimate from the voice descriptors. Converted to R_d it constitutes a second predicted R_d estimate. It is employed to define an additional prior probability to adapt the partial probabilities of the Viterbi algorithm accordingly.

Additional voice descriptors are examined on their co-variation in terms of a strong positive or negative correlation with a robust OQ estimate. The latter defines partially the shape of the deterministic part of the glottal excitation source. The recordings of natural human speech of the speakers BDL, JMK and SLT of the CMU Arctic speech database are used, introduced in [Kominek and Black, 2004]. This speech database simultaneously provides the recorded speech waveforms and their corresponding EGG signals. The DECOM method of [Henrich et al., 2004] estimates OQ_{EGG} from the corresponding EGG signals. The OQ_{EGG} contours are derived from all available phrases of all three speakers for all voiced segments using the voicing decision described in section 5.4.

The novel steering of the Viterbi algorithm constitutes an extension of standard Viterbi smoothing presented in the preceding Section 5.4. It serves as a proof-of-concept for the possibility to exploit specific speech signal features with a machine learning approach to aid Viterbi smoothing for the estimation of the LF shape parameter R_d. The usage of a statistical model exploits the information measured from additional voice descriptors that are correlated with OQ_{EGG}. The utilized features and the OQ_{EGG} originate both from the same underlying glottal gestures which reflect the physiological mechanisms of human speech production at the larynx [Laver, 1980, Gobl and Chasaide, 1992].

5.5.1 Exploiting the co-variation of voice descriptors

Several voice descriptors are determined from an extensive analysis of different voice descriptors and combinations in-between them on the CMU databases. Each selected feature demonstrates to be highly positively correlated with the OQ_{EGG} reference. Only the Voiced/Unvoiced Frequency boundary F_{VU} introduced in section 2.3.2 shares a negative correlation [Huber and Röbel, 2013]. All proposed voice descriptors are not influenced by a lower number of stable harmonic sinusoids and are therefore well suited to exploit their co-variation with the shape of the glottal excitation source.

$H1$-$H2$:
The amplitude difference in dB of the first two harmonic sinusoidal partials $H1$ and $H2$ ($H1$-$H2$). According to [Henrich et al., 2001] it is a reliable spectral correlate of OQ. $H1$-$H2$ generally depends on OQ [Liénard and Barras, 2013] and on the asymmetry α_m of the glottal flow or it's derivative [Doval et al., 2006]. $H1$-$H2$ proved to contribute to the discrimination between breathy and tense voice qualities in [Scherer et al., 2012].

The $H1$-$H2$ amplitude difference used in this work is measured as the direct relation in the magnitude spectrum. No inverse filtering is applied to measure $H1^*$-$H2^*$ from the corresponding glottal source signal as in [Fant, 1995, Hanson, 1995, Henrich et al., 2001]. This avoids possible problems with inverse filtering [Rothenberg, 1972, Alku, 1992, Drugman et al., 2008]. However, the direct measure of $H1$ and $H2$ in the magnitude spectrum is according to [Keating and Esposito, 2006] influenced by the first formant F_1. To smooth the direct $H1$-$H2$ measure a median filter of order 5 is applied.

3 MFCC bins:
The sum of the 3^{rd}, 4^{th} and 6^{th} MFCC bin [Ellis, 2005] models the spectral slope [Scherer et al., 2012] or the spectral tilt [Murphy, 2001] to reflect the amplitude continuation of the spectral envelope being correlated to a tense, modal or breathy phonation. However, a regression on the slope of the spectral peaks as in [Scherer et al., 2012] did not exhibit a high correlation to the OQ_{EGG} reference. In addition, as well the spectral tilt measures R_{14} or R_{24} as in [Murphy, 2001] and other related measurements of the spectral tilt or slope did not correlate feasibly high enough with OQ_{EGG}. In contrast, the summed combination of the 3^{rd}, 4^{th} and 6^{th} MFCC bin [Ellis, 2005] achieves a suitable high correlation with OQ_{EGG}. Apparently, the proposed summation of the three MFCC bins is less influenced by the variation of the vocal tract formants.

F_0:
The fundamental frequency F_0 shares according to [Fant et al., 1994] systematic dependencies with U_0 and E_e, and hence with R_d. U_0 has a close relation to the amplitude of the voice fundamental. E_e is the basic determinant of formant amplitudes. In [Laver, 1968] the laryngeal settings are categorized into phonation types, pitch ranges and loudness ranges. Therefore, the larynx as physiological origin of the voice does not only give rise to different voice qualities but also affects the fundamental frequency F_0 and the sound pressure level (SPL) of speech signals. Speakers favour a particular pitch range for each phonation type [Laver, 1968, Childers and Lee, 1991]. However, different studies as in [Laver, 1980, Maddieson and Hess, 1987, Hanson et al., 1990] have shown that the relation between pitch and different phonation types is speaker-dependent. The F_0 estimation uses the monophonic F_0 implementation based on the principles described in [Yeh and Röbel, 2004].

F_{VU}:
The F_{VU} boundary correlates with the voiced / unvoiced energy ratio, and the bandwidth of the glottal formants. The F_{VU} is thus related to E_e which determines the amount of generated sinusoidal energy [Fant, 1995]. The noise energy level is as well related to E_e and F_{VU}. It originates from turbulences created at the glottis, for example due to an imperfect glottal closure and a high airflow rate [Childers and Lee, 1991]. A tense voice, parameterized by lower R_d values, originates a broad excitation spectrum with sinusoidal content present in higher frequency regions. A relaxed voice, parameterized by higher R_d values, originates only few harmonic sinusoidal partials in lower frequency regions [Fant, 1995]. The F_{VU} estimation follows the principles outlined in section 2.3.2.

5.5.2 GMM-based prediction model

The establishment of a formula that uses the proposed voice descriptor feature set to predict R_d is very difficult. For example, the empiric formulation of [Fant, 1997] expresses the relation between the $H1^*$-$H2^*$ measure with OQ by an exponential function. It implies a lower limit of -5.73 dB for $H1^*$-$H2^*$ at OQ=0.3. However, an informal examination on the CMU Arctic databases proved that $H1$-$H2$ and OQ_{EGG} values below both limits exist. In [Henrich et al., 2001, Doval et al., 2006] it is shown that the relation between $H1^*$-$H2^*$ and OQ is additionally influenced by the asymmetry coefficient α_m representing the skewness of the glottal pulse. Moreover, recent studies like [Kreiman et al., 2012a, Kreiman et al., 2012b, Chen et al., 2013b] suggest that the relationship is speaker-dependent. This leads to positive and negative correlations, or situations where one parameter remains relatively constant while the other varies considerably.

No analytic formulation expressing the relation of OQ with any of the other voice descriptors can be found in the literature, and surely not for the complete set of the proposed voice descriptors. Thus, a GMM is employed to model the relation of the co-variation feature set with the OQ_{EGG} reference. A similar approach using Gaussian mixture modelling to predict glottal source signals or speech features has already been proposed in [Darch et al., 2007, Thomas et al., 2009a, Gudnason et al., 2012]. Per speaker one GMM is trained on the co-variation feature combination estimated on each voiced segment of the other two speaker databases and their corresponding OQ_{EGG} estimates.

The Viterbi steering approach is examined on two distinct voice descriptor feature sets D:
Feature set D_1 refers to the utilization of the voice descriptors $H1$-$H2$, F_{VU} and the sum of the 3^{rd}, 4^{th} and 6^{th} MFCC bin.
Feature set D_2 additionally includes the fundamental frequency F_0.

The GMM modelling is based on a modified version of the Voice Conversion system described in [Lanchantin and Rodet, 2010, Lanchantin and Rodet, 2011]. The joint probability density $p(D, R|\lambda^{(Z)})$ is expressed as

$$p(D, R|\lambda^{(Z)}) = \sum_{q=1}^{Q} \alpha_q \cdot N(D, R; \mu_q^{(Z)}, \Sigma_q^{(Z)}), \tag{5.22}$$

with D being either the feature set D_1 or D_2, and R being the OQ_{EGG} reference, conditioned on the model parameters λ with the weights α_q, mean vector μ_q^Z and the co-variance matrices Σ_q^Z over Q mixture sequences. Q is set to 6 mixture components for model D_1, and to respectively 8 mixture components for model D_2. The function

$$F(d) = \sum_{q=1}^{Q} p_q^d d \cdot [\mu_q^r + \Sigma_q^{rd} \Sigma_q^{dd^{-1}} (d - \mu_q^d)] \tag{5.23}$$

for the prediction of OQ_{GMM} from a feature set D is derived from the trained GMM model $\mathcal{M}$, with the conjoint data set of D and R being expressed by $Z = \{z_k\}, z_k = [d_k^T r_k^T]^T$.

5.5.3 Viterbi steering model

The prediction of OQ_{GMM} values from each GMM model per speaker defines the additional prior probability P_{prior} for the Viterbi algorithm. It is used to steer the standard Viterbi smoothing approach according to the prediction of OQ_{GMM} using either the feature sets D_1 or D_2.

Prior R_d probability $P(X|D)$:
The OQ_{GMM} prediction is transformed into the R_d value range $R_{d_{GMM}}$. It is modelled as Gaussian with variance σ_P^2. It defines an additional prior R_d probability from the $R_{d_{GMM}}$ prediction given the voice descriptor feature set D that is used to steer the Viterbi algorithm. A possibly occurring mean offset between the predicted $R_{d_{GMM}}$ and the estimated R_d of each phase minimization method has to be compensated per voiced segment. The probabilistic modelling P_{prior} is configured to have maximum probability if the value of $R_{d_{GMM}}$ and the R_d value estimated by our phase minimization methods are congruent to each other. Differences between predicted and estimated R_d lead to lower P_{prior} probabilities.

Optimal Viterbi steering path:
The prior R_d probability is inserted into equ. 5.19 to find

$$L(p) = \sum_n \log(P(O_{R_d}|X_p(n)) \cdot P(X_p(n)|D(n)) \cdot P(X_p(n)|X_p(n-1))). \tag{5.24}$$

The scale parameters of the log-likelihood function L of the distribution is defined as follows. Parameter α_a represents the scale parameter of the error function E_{R_d} of the observation probability O_{R_d}. The log-likelihood P_{prior} of the prior R_d probability is scaled by parameter β_b. Parameter γ_g scales the distribution of the transition probability $P(X_n|X_{n-1})$.

$$\bar{L}(p) = -\sum_n \alpha_a \cdot E_{R_{d_n}} + \beta_b \cdot M_{prior_n} + \gamma_g \cdot \frac{\Delta R_d}{\Delta_n} + c_d \tag{5.25}$$

Again, the constant term c_d gathers all the contributions of the constant scaling factors of the distributions and can be ignored by the Viterbi algorithm. The scaling factor γ_g is factored out to define the scaling factors $\alpha = \alpha_a/\gamma_g$ and $\beta = \beta_b/\gamma_g$ as control parameters of the Viterbi steering approach on all sequences p.

$$\bar{\bar{L}}(p) = -\sum_n \alpha \cdot E_{R_{d_n}} + \beta \cdot M_{prior_n} + \frac{\Delta R_d}{\Delta_n} \tag{5.26}$$

The evaluation of the novel Viterbi steering of equ. 5.26 is presented and discussed in the following section 5.6. The evaluation tests examine which values of α and β result in the highest performance to estimate R_d contours on natural human speech. The speech corpora utilized as evaluation data set are taken from three CMU Arctic databases [Kominek and Black, 2004].

5.6 Evaluation

This chapter presents an extensive evaluation on the performances of the phase minimization based methods explained in section 5.3.4, the different variants to adapt and extent the R_d regression shown in section 5.2, and the Viterbi smoothing as well as the Viterbi steering approach of section 5.4 and respectively section 5.5. The evaluation is conducted on synthetic confusion matrices shown in section 5.6.1, one extended synthetic test presented in section 5.6.3, and on one test of natural human speech discussed in section 5.6.4. The latter evaluation is similar to [Fröhlich et al., 2001, Ó Cinnéide, 2012, Kane and Gobl, 2013a]. For the test on natural human speech, the OQ_{R_d} estimates are derived from the estimated R_d curves and compared with the OQ_{EGG} estimates derived from an analysis of synchronously to the audio waveforms recorded EGG signals. The objective of the evaluation is to determine the best performing and most robust parameterization for Viterbi smoothing, Viterbi steering, and the phase minimization variants.

5.6.1 Evaluation of error surfaces from confusion matrices

This section presents an theoretical proof-of-concept analysis of the different phase minimization variants presented in section 5.3.3. The different corresponding objective functions produce similarly as in [Degottex, 2010] confusion matrices along the R_d range. The confusion matrices serve to detect ambiguities of the functions for phase minimization in estimating R_d. The error surfaces of the confusion matrices illustrate the sensitivity of the objective functions for phase minimization according to the following influences:
-the variation of R_d over its complete range,
-the fundamental frequency F_0,
-the first formant of the vocal tract F_1, and
-the glottal formant F_g.

Test setup:
The experimental evaluation examines the accuracy of each phase minimization method to distinguish between the shapes of a fitting or mismatching glottal formant F_g of the synthetic model $G^{R_d}(\omega)$, under the influence of the first formant F_1.

The objective functions examine the remaining error residuals obtained by synthesizing and matching a set of voice descriptors according to the presented experimental setup. The test uses the following parameterization:
- The fundamental frequency F_0 is set to 80 Hz.
- The vocal tract formant F_1 is modelled by a 2-pole filter, having a pole position at 800 Hz and a radius of 0.98 close to the unit circle.
- The synthesis of the LF model is parameterized by R_d and F_0 to generate the glottal formants F_g. This results in

different shapes of the glottal formant F_g in the spectrum.
- The same formant F_1 is convolved with each synthesized glottal pulse $G^{R_{d2}}(\omega)$.

The spectra resulting from the convolution of formant F_1 and different glottal pulses F_g contain therefore different shapes of F_g. Each F_g is parameterized by an R_{d2} value. A vector of R_{d2} values is derived from a sufficiently dense grid over the R_d range.

Each synthesized glottal pulse $G^{R_{d2}}(\omega)$ to be convolved with F_1 thus parameterized by the glottal source shape parameter R_{d2}. Each glottal pulse $G^{R_{d1}}(\omega)$ under evaluation is parameterized by the glottal source shape parameter R_{d1} to be evaluated. Each confusion matrix shown in fig. 5.6 is the result of the corresponding objective function to evaluate each R_{d1} value against all other R_{d2} values along the R_d range. An ideal error surface of any confusion matrix would exhibit a tiny error valley in blue colour on the diagonal, and large error mountains in red colour elsewhere away from the diagonal.

Test evaluation:
Fig. 5.6 depicts the error surfaces of the confusion matrices retrieved from each R_d parameter estimation method of each phase minimization variant. The figures depict from top to down the variants MSPD2IX, MSPD2I0, MSPD2I1 and MSPD2I2. An ideal error surface would have a tiny blue error valley at the matching diagonal axis, with the rest of the error surface exhibiting large error mountains in red colour elsewhere away from the diagonal. A dark blue colour indicates a complete match, and a dark red colour a complete mismatch of the synthesized $G^{R_{d2}}(\omega)$ versus the evaluated glottal pulse $G^{R_{d2}}(\omega)$. It is not predictable how many stable sinusoidal partials are observable from the speech signal for each frame. Therefore only 7 sinusoidal partials N are employed as a realistic expectation for the R_d evaluation before the harmonic content is masked by noise. Please note that the results are qualitatively the same for other numbers of harmonic partials N.

By visual inspection of fig. 5.6 it can be observed that each integration step leads to a more tiny error valley in blue colour being delimited by broader error hills in red or respectively yellow colour. The 2^{nd} figure shows the error surface for no integration step applied with MSPD2I0, followed by one integration step of MSPD2I1 and the lowest figure with two integration steps for MSPD2I2. Broader error valleys appear more at the upper R_d range $R_d > 2.7$. This may lead to possibly unnatural broad steps when estimating R_d especially at word or pause boundaries of a continuous speech signal.

Analysis per variant:
MSPD2I0 exhibits the broadest error valley at the lower normal R_d range $R_d <= 2.7$. It is delimited by high error values (red) for higher values of R_d. Additionally it exhibits the broadest error valley at the upper R_d range $R_d > 2.7$. **MSPD2I1** may suffer from ambiguities from the additional error valleys for low R_d values $R_d < 0.5$ versus higher R_d values $R_d > 3$ at the upper left and lower right. **MSPD2I2** may be mislead by several appearing side minima in blue colour being especially present in the normal R_d range for $R_d <= 2.7$. Its error surface shows the smallest error valley of all variants in the middle. No significant side minima can be observed on the R_d range borders for **MSPD2IX**. Continuous side minima are only observable close to the ideal error valley in the middle. This may lead to little unnatural jumps of the R_d estimation around the ideal error valley.

Analysis between variants:
MSPD2I1 exhibits a smaller error valley than MSPD2I0 but may suffer from ambiguities from the additional error valleys for low R_d values $R_d < 0.5$ versus higher R_d values $R_d > 3$ at the upper left and lower right. MSPD2I2 has a more distinguishing error valley than MSPD2I0 and MSPD2I1, with only very minor secondary minima present. The combinatorial error surface of MSPD2IX exhibits the least ambiguities compared to the other phase minimization variants. It shows a close to be quasi-ideal small error valley in the middle. However, the size of its error valley increases with increasing R_d values.

MSPD2IX error combination and weighting:
MSPD2IX slightly improves the robustness of each single phase minimization variant because the linear superposition introduces a combinatorial effect. It may cancel out in may cases one misleading side minima present in one variant by non-misleading error hills present in the same R_d regions of the other two variants. A combination of all error functions is thus able to emphasize the true minima since the appearing side minima result from non-linear fluctuations present for only one phase error function and not from an analytic functional behaviour. A non-informal test conducted on the synthetic test set presented in the following section 5.6.3 indicates that a more refined variation of the weighting sequence $w0{=}w1{=}w2$ defined in equ. 5.19 for MSPD2IX does not lead to statistically significant improvements [Huber and Röbel, 2013]. An equal weighting is therefore set for MSPD2IX.

However, a different behaviour concerning the MSPD2IX weighting sequence can be observed for the analysis of natural human speech. The summary given in section 5.3.4 lists that for natural human speech each phase error function might exhibit stronger side minima than the true minima due to influences originating from noise, higher frequencies F_0 resulting in an impulse response of the VTF being longer than the fundamental period, or the non-stationarity of the VTF at transient regions. Further informal tests conducted throughout the work for the

Confusion matrices illustrating the error surface of each phase minimization variant

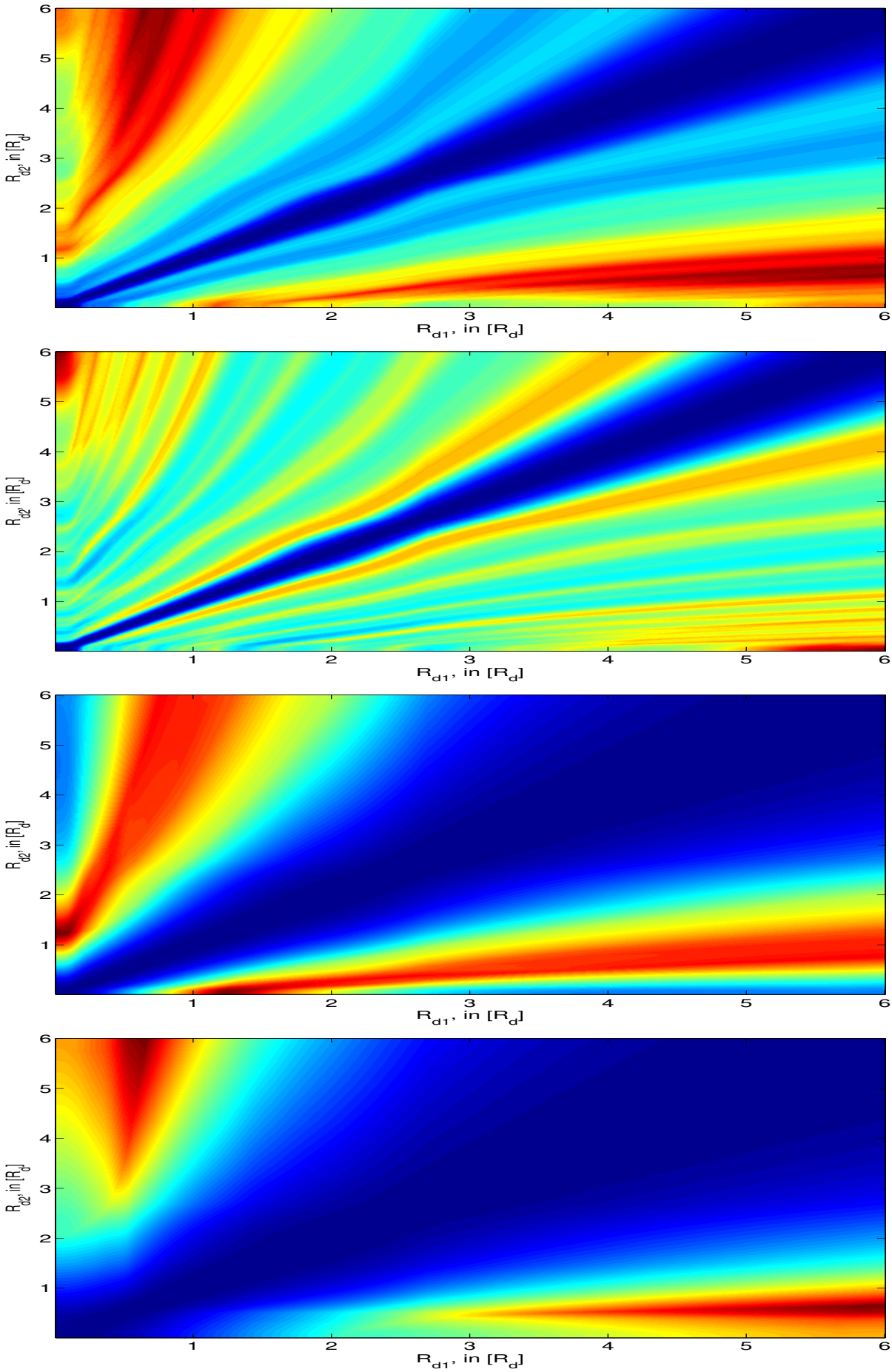


Figure 5.6: *Error surfaces from top to down: MSPD2IX, MSPD2I2, MSPD2I1, MSPD2I0*

novel speech system of chapter 6 demonstrate a different behaviour on natural human speech. The tests indicate that a weighting in favour of the phase minimization variants MSPD2I1 applying a 1^{st} and MSPD2I2 applying a 2^{nd} integration is beneficial for the robustness of the R_d estimation. The reasoning being that each integration step removes further the effect of the high pass filtering introduced by the preceding differentiation steps. This evaluates more lower frequency regions which minimizes the noise influences present in higher frequency regions. These suggestions are validated by the error surface evaluation on natural human speech of section 5.6.4.2 corresponding to the theoretical error surface analysis presented in this section. In addition, the results of the objective evaluation on natural human speech given in section 5.6.4 show that MSPD2IX performs best, followed by MSPD2I2, afterwards MSPD2I1 and finally MSPD2I0 performing the worst.

5.6.2 Spectral distortion effect

An explanation of the R_d estimation errors is given by the fact that the complete VTF cannot always be observed because some sinusoidal partials may be covered by noise. The synthetic test presented in this section examines how many stable sinusoidal partials N_{harms} from the harmonic model are required to reliably construct the minimum phase spectrum of the first N bins of a discrete spectrum S_k of a single period. With this the problem of the calculation of the minimum phase variant of a filter given a handled version of the filter transfer function is examined. The calculation depends as well on the position of the formants and the fundamental frequency F_0. Both are however not considered in this experiment such that its result is of limited validity.

The number of sinusoids to evaluate the phase errors ϕ_ϵ is set to N=7. The amount of N_{harms} is varied to measure the mean error of the R_d estimation.

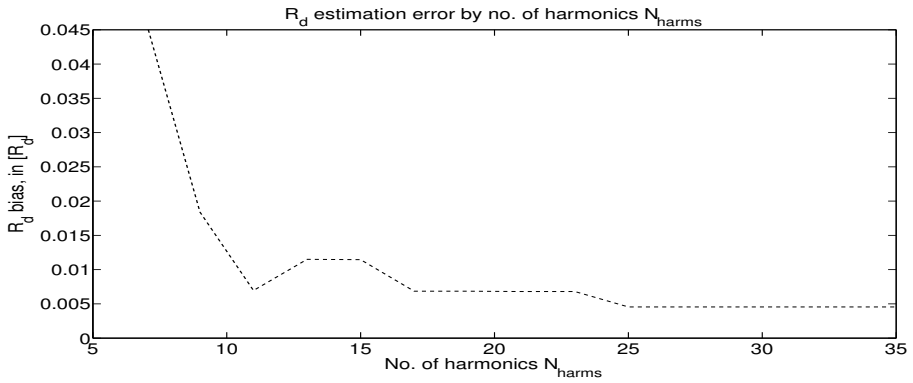


Figure 5.7: *R_d estimation error by number of sinusoidal harmonics N_{harms}*

The y-axis reflects the R_d estimation error bias in terms of the R_d value unit. The results depicted in fig. 5.7 show that for N_{harms}=11 the error function is already reasonably attenuated because the boundary effects that are introduced at the spectral border have sufficiently diminished.

5.6.3 Objective evaluation on a synthetic test set

This section presents the examination on the performance of the phase minimization algorithms to estimate R_d on a synthetic test set similar to [Degottex et al., 2011a, Huber et al., 2012]. The four R_d estimation methods MSPD2I0, MSPD2I1, MSPD2I2 and MSPD2IX [Huber et al., 2012] are evaluated with respect to their dependency on some characteristics present in speech signals. The influences of the fundamental frequency F_0, the number of observed stable harmonic sinusoids N, different configurations of the VTF, the glottal source noise n^{σ_g} and the environmental noise n^{σ_e} are investigated.

The estimation of glottal source characteristics depends as reported in [Drugman et al., 2008] on the position of the glottal formant F_g to the VTF formants, notably to the first formant F_1. The VTF influence is simulated using 16 synthetic vowels C_{VTF}. Maeda's digital simulator [Maeda, 1982] synthesizes each C_{VTF} using one of 10 different F_0 values within the range [80, 293] Hz. Each C_{VTF} is convolved with each glottal formant parameterized by an R_d value within the R_d range [0.1, 6] on a grid of step size R_d=0.1. 6 Gaussian noise levels n between [-50, -25] dB are added to the voiced signal. Each noise level is applied to both noise influences $n^{\sigma_g}[n]$ and $n^{\sigma_e}[n]$.

The ratio of F_{VU} boundary to F_0 determines how many stable harmonic sinusoids N are observable before being masked by noise. The influence of N for the range [3, 8] is evaluated by restricting the R_d estimation algorithm

to observe N partials. The number of harmonics N_{harm} used to calculate the minimum phase version of the VTF is limited to the ratio of 8 kHz by F_0. Additionally the position of the window with respect to the period in time is simulated on a grid of 4 different delays ϕ^* covering the range $[-0.5 \cdot T\,,\, 0.5 \cdot T]$. One synthetic test per R_d regression variant and phase minimization method consists in total of 1,382,400 single tests (60 R_d · 10 F_0 · 6 $n^{\sigma_g}[n]$ / $n^{\sigma_e}[n]$ · 16 C_{VTF} · 6 N · 4 $\phi^{\cdot}$ values).

Four fundamental periods found in the middle of each synthesized test waveform are windowed. Each windowed segment is examined by each phase minimization variant on the contained glottal pulse shape. Each remaining phase error ϕ_ϵ is used to estimate the glottal pulse shape parameter R_d per test set. The test results are presented in the following in a compact manner by adding up the bias and standard deviation of the R_d estimation errors as a function of the examined parameters. The different parameters under evaluation are the fundamental frequency F_0, the number of stable harmonics N and the glottal pulse shape parameter R_d itself.

5.6.3.1 Examination on dependency in F_0

Fig. 5.8 illustrates how higher fundamental frequencies F_0 lead to a constant increase in the error amount of wrongly estimated R_d glottal source pulse shapes. The y-axis reflects the amount of R_d estimation error bias and standard deviation, summarized over all test results for each utilized F_0 value depicted on the x-axis.

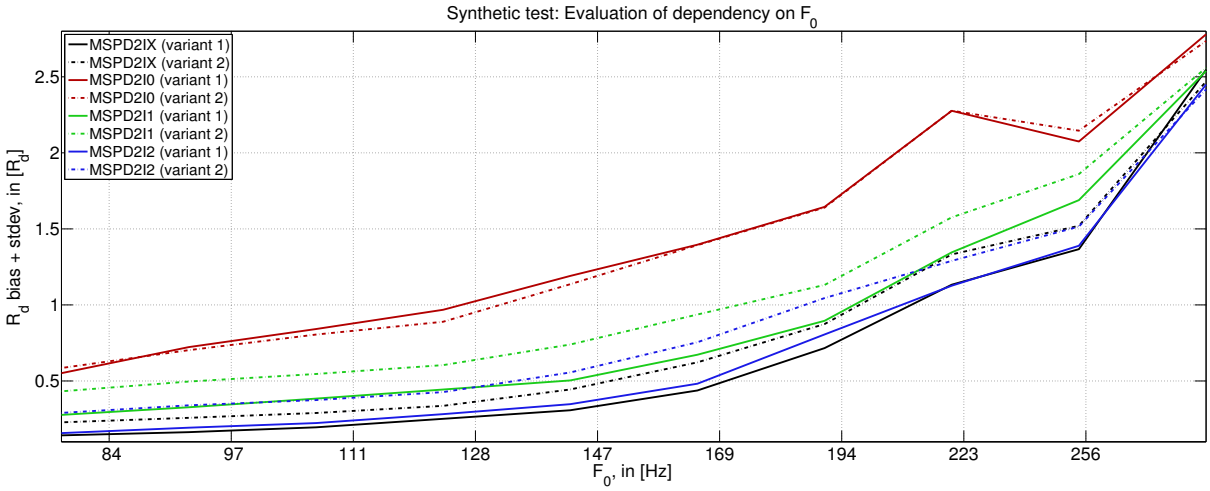


Figure 5.8: *Evaluation of R_d estimation, F_0-dependency*

The R_d regression adaptation variant 2 (depicted for each method in dash-dotted lines) performs in general worse than variant 1 (solid lines). The overall best performing methods over the complete frequency range of 80 to 293 Hz are MSPD2IX and MSPD2I2 under the utilization of adaptation variant 1. The method MSPD2IX using the less performant adaptation variant 2 even outperforms the baseline method MSPD2I1 using adaptation variant 1 over the complete F_0 range. The latter yields overall similar results as MSPD2I2 using the less good performing adaptation variant 2. The method MSPD2I0 performs in general worst. The R_d estimation results are relatively robust up to frequencies of ~150-200 Hz, above which more severe perturbations of the estimation accuracy are apparent [Drugman et al., 2008, Huber et al., 2012]. If the duration of the impulse response of the VTF is close to or above the period the estimation of the VTF minimum phase property is less accurate. This leads to less robust estimations with increasing F_0.

5.6.3.2 Examination on dependency in harmonic partials

The same effect of a comparatively lower number of stable harmonic sinusoids N and a less accurate estimation of the minimum phase property of the VTF is introduced with higher noise levels. These influences are simulated by varying N, shown in fig. 5.9. The y-axis reflects as in fig. 5.8 the amount of R_d estimation error bias and standard deviation, here being summarized over all test results for each utilized number N of stable harmonic sinusoids depicted on the x-axis.

Both the glottal source noise n^{σ_g} and the environmental noise n^{σ_e} are set to n=1 for the variation of N depicted in fig. 5.9. This corresponds to the lowest noise level of -50 dB used in the synthetic test. The characteristics of natural human speech are well reflected with this low noise level. The misleading interference of noise is suppressed. This is required to properly examine a different number of stable harmonic sinusoids N. The method MSPD2I0 is

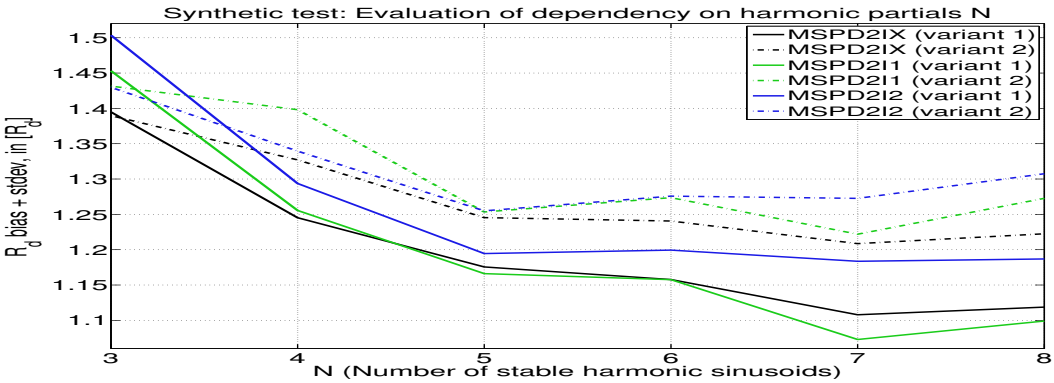


Figure 5.9: *Evaluation of R_d estimation, N-dependency*

omitted in fig. 5.9 due to its weak performance. The best performing methods MSPD2IX and MSPD2I1 indicate most obviously that a lower number of stable harmonic sinusoids N leads to a higher amount of R_d estimation errors. However, each method exhibits the lowest overall accumulated amount of R_d estimation errors at N=7. This is justified by the fact that the harmonic sinusoids of $N >= 8$ may already be covered by noise at level -50 dB. Again, the R_d adaptation variant 2 leads to less good R_d estimation results than adaptation variant 1.

5.6.3.3 Examination on dependency in voice quality

This section explains how the objective function of each phase minimization based method is dependent on the phase differences of the LF model. The y-axis reflects again the amount of R_d estimation error bias and standard deviation. The latter is summarized over all test results for each evaluated R_d value on the x-axis. The LF model latter is thus parameterized by different R_d values over the complete R_d range. A too high self-similarity of LF models parameterized by an R waveshape parameter set being close in value leaves the estimation method with little differentiation possibilities.

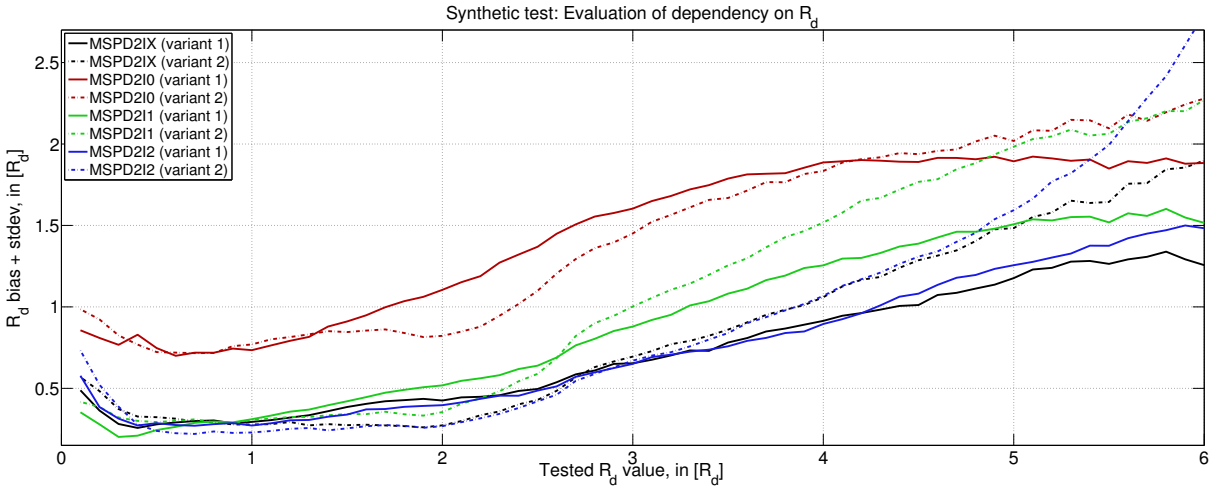


Figure 5.10: *Evaluation of R_d estimation, dependency on voice quality measured in R_d*

Fig. 5.10 indicates that glottal source shapes in the R_d range of [0.3, 2] are more dissimilar to each other than the glottal source shapes of the upper R_d range of]2.7, 6] or tense phonations parameterized by the R_d range extension of 5.2 below the lower limit $R_d < 0.3$ of the normal R_d range. This reflects to a certain extent the observations concluded with the R_d confusion matrices of [Huber et al., 2012]: Broader error valleys of each objective function for phase minimization lead to a less robust R_d estimation. The conceptual equivalent to the R_d confusion matrices is shown by Fig. 5.12. The R_d estimation error surfaces are spanned up frame-wise over time and reflect the behaviour of each objective function. The overall less good R_d estimation performance of R_d adaptation variant 2 results from the fact that it renders higher OQ values which proves for this synthetic test to suffer from a higher error rate. Examining fig. 5.10 by visual inspection reveals that the phase minimization methods using adaptation variant 2 perform better for lower R_d values in the normal R_d range [0.3, 2.7] and gradually under-perform more for higher R_d values above $R_d > 2.7$.

5.6.4 Objective evaluation on natural human speech

In this section the four phase minimization methods of [Huber et al., 2012] are evaluated on recordings of natural human speech. Each method has estimated R_d on each voiced segment of all available phrases of the CMU Arctic speech databases [Kominek and Black, 2004] BDL, JMK, and SLT.

5.6.4.1 Test basis on EGG measurements

No reliable ground truth is known to date to evaluate the estimation of glottal source parameters from natural human speech. A measurement of the movement of the glottal folds is given by EGG signals recorded simultaneously to speech signals. It is considered to form the basis of a more robust glottal source parameter estimation compared to estimations based on recorded audio signals of human speech. EGG waveforms are regarded as valid indicator of the vocal fold contact area to measure glottal activity [Baer et al., 1983]. The differentiated EGG (DEGG) can be considered as reliable indicator of the time instant of glottal closing (GCI) [Henrich et al., 2004].

However, it is not yet validated that the glottal opening and closing events extracted from an EGG signal reflect exactly the time instants of the physiological contact of the vocal folds muscles [Baer et al., 1983, Childers et al., 1986, Orlikoff, 1991, Marasek, 1997]. Furthermore, the EGG-based time instants may not exactly match the start and end of the glottal air flow [Fröhlich et al., 2001]. Moreover, despite the general acceptance to provide more reliable estimates, the EGG-based measurements can still be inaccurate [Colton and Conture, 1990, Marasek, 1997, Sapienza et al., 1998]. A reliable and exact determination of the time instant of glottal opening (GOI) can be more difficult and erroneous than the estimation of GCIs [Baken, 1992, Baer et al., 1983]. The GOI estimation on EGG waveforms can especially be error-prone if strands of mucus bridge the glottis while the opening of the vocal folds [Titze and Talkin, 1979, Childers et al., 1986, Dromey et al., 1992]. Other vocal fold vibratory motions than the modal register may lead as well to less robust estimations [Childers et al., 1986]. Difficult to estimate are phonation types with a continuously open glottis. For such cases the variation of the impedance measured at the larynx does not correspond to the glottal area [Marasek, 1997]. Moreover, higher F_0 values may result in EGG waveforms with lower Signal-To-Noise ratios (SNR) [Hanson et al., 1990] which poses more difficulties to reliably extract the time instants when the vocal folds open and close. The study of [Herbst, 2004] illustrates that each analyzed algorithm to estimate OQ from an EGG signal introduces a bias, either by having to choose a certain threshold to measure the short-term peak-to-peak amplitudes of the EGG signal or by having to pick one of possibly several peaks from the DEGG signal appearing while the glottal opening phase [Childers and Lee, 1991].

Despite the mentioned problems the test scenario for natural human speech to compare the OQ estimations from EGG and audio recordings is chosen because of its reasonable reliability in contrast to other methods and its relatively easy setup. The example of fig. 5.11 shows the curves of the frame-based R_d estimator in dotted lines

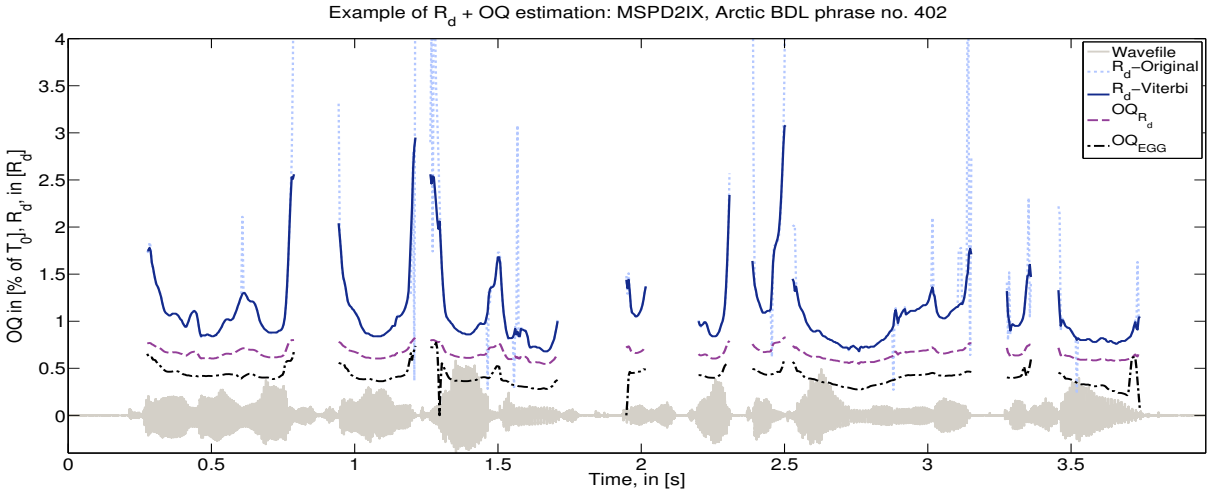


Figure 5.11: *Speaker BDL phrase 402 - R_d estimation with MSPD2IX and Viterbi smoothing, α=0.47*

with light blue colour, the R_d estimation after Viterbi smoothing in solid lines with dark blue colour, the from it derived OQ_{R_d} contour in dashed lines with purple colour, and the OQ_{EGG} reference in dash-dotted lines with black colour. Each OQ contour OQ_{R_d} derived from each R_d estimate is compared with each OQ_{EGG} contour estimated on the corresponding EGG signal using the DECOM method of [Henrich et al., 2004]. A general non-constant offset between the OQ_{R_d} and OQ_{EGG} contours can be observed as in [Childers and Lee, 1991, Herbst, 2004] due to

the mentioned systematic bias of the OQ estimation by the EGG-based technique. The phrase shown in fig. 5.11 exemplifies that the EGG measure can be error-prone as the physiologically impossible jumps of OQ_{EGG} around ~1.3 and ~1.95 seconds illustrate.

5.6.4.2 Error surface evaluation on natural human speech

This section discusses a similar evaluation metric as it has been conducted in section 5.6.1 with the evaluation of the error surfaces from confusion matrices. The difference being that here the error surface is derived from natural human speech signals instead of a synthetic test setup. Fig. 5.12 depicts examples of how Viterbi smoothing suppresses unnatural jumps of each frame-based R_d estimator. The error residuals of the phase error functions of each phase minimization based objective function generate an error curve per frame. The error curve is span over the evaluated R_d range [0.1, 6.0]. The y-axis reflects for each frame the computed error curve as R_d estimation error in brightness. Dark colours constitute error minima. Bright colours constitute error maxima. Frames over time span up the illustrated error surfaces. Completely black segments are set as unvoiced.

The error lattices of the voiced segments define the observation probability of the noise-robust Viterbi algorithm. The Viterbi algorithm computes the highest probability which best explains the observation sequence O and which determines the optimal state sequence X per voiced segment. The optimal state sequences X of glottal source shapes R_d of the standard Viterbi smoothing approach are illustrated as dashed grey lines. The initial R_d estimates are illustrated in white colour and reflect the R_d value where the frame-based phase error function exhibits the lowest error. Tiny error valleys in black around these initial R_d estimates are very well developed for the methods MSPD2IX and MSPD2I2. MSPD2IX has distinctive error hills which are plotted with a brighter contrast and leaves little confusing side minima to its objective function to minimize the error of the phase error function. MSPD2I2 shares are similarly robust error surface for this natural human speech example of speaker BDL. Side minima appear e.g. at ~0.6 seconds at R_d≈2.0 and R_d≈5.0. No unnatural jumps occur since these side minima are higher than the overall lowest error value. The latter are present for MSPD2I2 at ~2.9 seconds where the initial R_d estimate in white jumps three times from the apparently true R_d contour and its obvious error valley at R_d ≈1.0 to misleading side minima at R_d ≈2.5 and R_d ≈4.0. The method MSPD2I1 exhibits as well clear valleys which are broader and less distinctive than the ones of MSPD2IX and MSPD2I2. Its occurring side minima are more developed. This results in a higher probability to produce physiological impossible jumps of the R_d estimate. MSPD2I1 suffers for example at ~0.6 seconds from a misleading side minima which got suppressed for MSPD2IX and MSPD2I2. The reason why the method MSPD2I0 performs worst is apparent when examining its error surface shown in Fig. 5.12. No clear error valleys in black for the underlying glottal excitation source contained in the analyzed speech phrase are established. Its original R_d estimates in white and the smoothed R_d contours in grey do not follow the true shape of the glottal source.

5.6.4.3 OQ test across speakers

Without Viterbi smoothing:
The Pearson product moment correlation coefficient r [Pearson, 1900] normalizes the co-variance of OQ_{R_d} and OQ_{EGG} by the product of its standard deviations. The Pearson coefficient r is used as correlation metric to examine how well the OQ_{R_d} derived from each R_d estimate correlates with OQ_{EGG}. It is defined in the range [-1, 1] with -1 expressing a perfectly negative correlation, +1 a perfectly positive correlation, and 0 no correlation. The Root-Mean-Square Error (*rmse*, RMSE) serves as second evaluation metric. The mean between the OQ estimates for each voiced segment is removed to avoid any impact of the bias that is present in the EGG-based OQ_{EGG} estimate for the *rmse* measure. Please note that the calculation of r implies the removal of a possible bias between both evaluated sample distributions. In the following, the R_d adaptation variants 1 and 2 of section 5.2 are evaluated together with each phase minimization variant of section 5.3.3.

Table 5.1: *OQ test results, without Viterbi smoothing, Rd adaptation variant 1*

Metric	MSPD2IX	MSPD2I0	MSPD2I1	MSPD2I2
r	0.2958	0.1599	0.2910	0.3305
rmse	0.0835	0.1928	0.0804	0.0772

The results for each objective function estimating R_d without applying Viterbi smoothing are listed in the tables 5.1+5.2. The r-correlation maxima and the *rmse*-error minima are shown per method. The method MSPD2I2 achieves the highest correlation and a smaller error between its estimated and the EGG-based OQ contours. The results of MSPD2IX and MSPD2I1 are slightly worse. MSPD2I0 performs worst by a substantial margin. Please

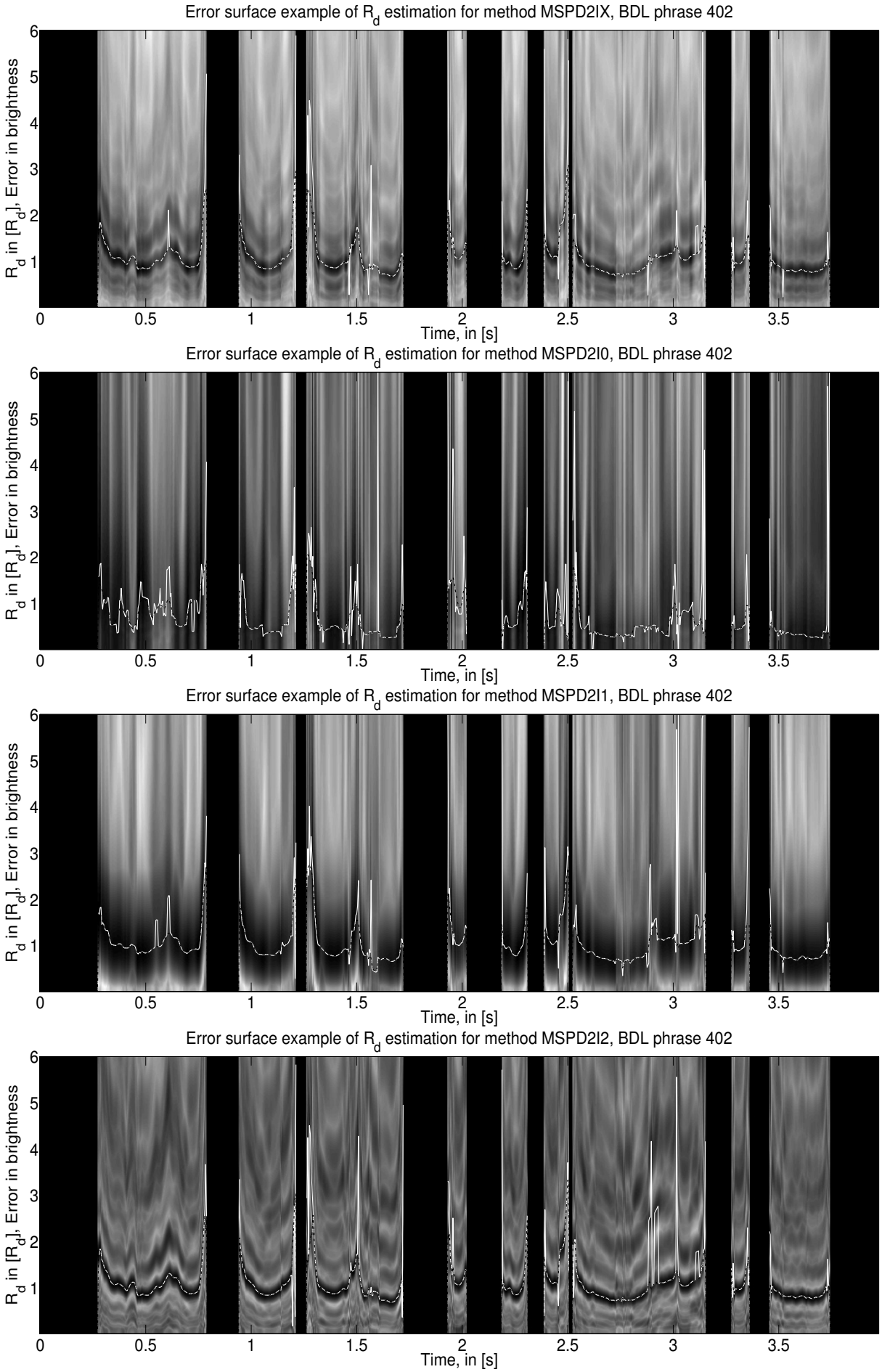


Figure 5.12: *Speaker BDL phrase 402 - R_d error surface examples, 4 phase minimization methods, α=0.47*

Table 5.2: *OQ test results, without Viterbi smoothing, Rd adaptation variant 2*

Metric	MSPD2IX	MSPD2I0	MSPD2I1	MSPD2I2
r	0.3189	0.1788	0.3138	0.3457
rmse	0.0765	0.1846	0.0747	0.0724

note that the baseline method MSPD2I1 of [Degottex et al., 2011a] constraint to the normal R_d range [0.3, 2.7] and without Viterbi smoothing achieves r=0.23. By visual inspection its estimated R_d and OQ curves appear to fluctuate more. However, due to the constraint range the failures are less weighted. The R_d adaptation variant 2 obtains better results for each method than variant 1. Only the R_d adaptation variant 2 will be shown in the following since it outperforms variant 1 for each test set on natural human speech.

Smoothing with a moving average filter:
Different moving average filter types were evaluated on their ability to suppress the local instabilities of each frame-based R_d estimator that are present within short-time segments. The best results were achieved by a median

Table 5.3: *OQ test results, median smoothing, order 5*

Metric	MSPD2IX	MSPD2I0	MSPD2I1	MSPD2I2
r	0.3438	0.1776	0.3343	0.3711
rmse	0.0710	0.1336	0.0774	0.0635

filter with order 5, shown in table 5.3. It improves the estimated R_d contours of each phase minimization method only to a marginal extent. In the following an investigation into means of establishing a more robust correction of the estimated R_d contours by utilizing different configurations of the Viterbi algorithm will be presented.

Standard Viterbi smoothing
The results of Viterbi smoothing without the utilization of the novel OQ_{GMM} prediction-based Viterbi steering are summarized in table 5.4. The improvements of applying a dynamic programming algorithm to smooth the

Table 5.4: *Viterbi smoothing (optimal α-values in parentheses)*

Metric	MSPD2IX	MSPD2I0	MSPD2I1	MSPD2I2
r	0.5327 (0.07)	0.2404 (0.13)	0.4894 (0.07)	0.5241 (0.09)
rmse	0.0507 (0.03)	0.0564 (0.01)	0.0515 (0.03)	0.0513 (0.01)

estimated glottal source shape curves are apparent when comparing its results with the ones of the tables 5.1 and 5.2 (without the application of Viterbi smoothing) and table 5.3 (smoothing with a moving average filter). Especially the best performing methods MSPD2IX, MSPD2I2 and MSPD2I1 benefit enormously from Viterbi smoothing while the improvements for the worst method MSPD2IO are limited. The corresponding α-values to scale the observation probability of Viterbi smoothing are given in parentheses. One global maximum for the correlation r and one global minimum for the error *rmse* exists for each method concerning the Viterbi parameter α. The r-maxima occur for α in the range [0.07, 0.13] and lie with a maximal offset of α=0.10 to the *rmse*-minima.

Viterbi smoothing summary

This section summarizes the results of the OQ comparison test without Viterbi steering in terms of the Pearson r correlation coefficient. The baseline MSPD2I1 achieves with the restriction to the normal R_d range [0.3, 2.7] and without Viterbi smoothing a Pearson r correlation of r=0.23.

Table 5.5: *OQ comparison results, Pearson r, (optimal α)*

Index	**MSPD2IX**	**MSPD2I0**	**MSPD2I1**	**MSPD2I2**
1	0.2958	0.1599	0.2910	*0.3305*
2	0.3189	0.1788	0.3138	**0.3457**
3	0.3153	0.1598	0.3063	*0.3542*
4	0.3438	0.1776	0.3343	**0.3711**
5	*0.5266* (0.11)	0.2224 (0.13)	0.4804 (0.07)	0.5165 (0.13)
6	**0.5327** (0.07)	0.2404 (0.13)	0.4894 (0.07)	0.5241 (0.09)

Table 5.5 summarizes the results with the following indices:

Index 1: R_d adaptation 1, without Viterbi smoothing
Index 2: R_d adaptation 2, without Viterbi smoothing
Index 3: R_d adaptation 1, moving average smoothing filter
Index 4: R_d adaptation 2, moving average smoothing filter
Index 5: R_d adaptation 1, Viterbi smoothing
Index 6: R_d adaptation 2, Viterbi smoothing

The phase minimization variant MSPD2I2 achieves for the tests without Viterbi smoothing and with a moving average smoothing filter the highest *r*-correlations. The additive combination MSPD2IX of the phase minimization variants achieves the overall best performance of r=0.5327 with Viterbi smoothing and R_d adaptation 2. The R_d adaptation variant 2 in bold outperforms the R_d adaptation variant 1 in italic letters for the best performing methods. The R_d adaptation variant 2 performs as well for each other test better than R_d adaptation variant 1.

In the following the results of the novel steering of Viterbi smoothing will be presented in the same manner. Each algorithm variant exhibits one global *r*-maxima and one global *rmse*-minima concerning the scaling parameters α or respectively β of the Viterbi algorithm, introduced in section 5.4. The objective of the following tests is to determine the best overall values a) for the scaling parameters α and β of Viterbi smoothing and steering, and b) for each phase minimization method.

Viterbi steering using GMM prediction
The GMM-based OQ prediction by means of a 3-fold leave-one-out cross-validation on the training and test sets corresponding to each speaker database will be evaluated. The *rmse*-error and the *r*-correlation of each OQ_{GMM}

Table 5.6: *Validation on training and test sets per speaker, GMM model 1*

Speaker	*rmse* (train)	*rmse* (test)	*r* (train)	*r* (test)
BDL	0.0837	0.1140	0.5508	0.4705
JMK	0.0837	0.0787	0.7280	0.3500
SLT	0.0574	0.1114	0.9054	0.3956
ALL	0.0749	0.1014	0.7281	0.4054

prediction model using the feature set D_1 of GMM model 1 are shown in table 5.6. Table 5.7 lists respectively the results using the feature set D_2 of GMM model 2. Please note that by the straightforward training of a GMM using

Table 5.7: *Validation on training and test sets per speaker, GMM model 2*

Speaker	*rmse* (train)	*rmse* (test)	*r* (train)	*r* (test)
BDL	0.0781	0.0592	0.6273	0.7285
JMK	0.0742	0.0917	0.7926	0.3478
SLT	0.0529	0.1072	0.9209	0.4030
ALL	0.0684	0.0860	0.7803	0.4931

the OQ_{EGG} contours and the corresponding voice descriptor feature combination of two speakers, the predicted OQ_{GMM} contours for the test set on the third speaker achieves *r*-correlations shown in the tables 5.6+5.7 being close to the performance of the signal processing based R_d estimation methods. The potential of the proposed OQ prediction using voice descriptors is indicated by the corresponding results of *r* (train) on the training test sets. It outperforms the results discussed in the following tests. However, further examination conducted after the publication of these results in [Huber and Röbel, 2013] indicate that the comparably high *r*-correlation and low *rmse* error values are interfered by the estimation being examined on two speakers. The different means between two speakers introduce a bias into both evaluation metrics.

The remaining principal problem of Viterbi steering is to overcome the speaker-dependency of the modelling to generalize better over speaker specific characteristics. This could be solved by employing more speaker corpora and a more sophisticated handling of the feature combination. The higher prediction accuracies of the training versus the test sets for the speakers JMK and SLT indicate that the utilized feature sets do not generalize optimally on their data sets. However, the OQ_{GMM} prediction model 2 for speaker BDL is able to predict more precise OQ_{GMM} contours on the BDL test set than on its own training set of the speakers JMK and SLT.

The intrinsic characteristics of each speaker data set of the utilized CMU Arctic databases will be analyzed more in detail in section 5.6.4.4. The following sections present the utilization of the models trained to evaluate the test set errors. This reflects the later application of the proposed novel Viterbi steering to estimate R_d where no training data will be available.

a) Viterbi steering, OQ_{GMM} prediction model 1:

The results of the four phase minimization methods using Viterbi smoothing and its auxiliary GMM prediction based Viterbi steering for model 1 of section 5.4 are shown in table 5.8. The Viterbi scaling parameter α is varied

Table 5.8: *Viterbi steering, model 1 (optimal α-values in parentheses)*

Metric	MSPD2IX	MSPD2I0	MSPD2I1	MSPD2I2
r	0.5348 (0.07)	0.4202 (0.05)	0.5047 (0.09)	0.5227 (0.03)
$rmse$	0.0502 (0.05)	0.0536 (0.07)	0.0510 (0.07)	0.0507 (0.03)

for this test while the Viterbi scale β remains fixed to a constant value of 1.0. The values in parentheses illustrate the α values of the r-maxima and $rmse$-minima. A one-way ANOVA comparison [Hill and Lewicki, 2007] with the correlation results of standard Viterbi smoothing presented in table 5.4 validates that the improvements of Viterbi steering using model 1 are statistically significant for method MSPD2I1 at significance level 1 % (p-value < 0.01) and for method MSPD2I0 at significance level 0.1 % (p-value < 0.001). No statistically significant improvements could be validated for the methods MSPD2IX and MSPD2I2. The best β-values for each method are

Table 5.9: *Viterbi steering, model 1 (optimal β-values in parentheses)*

Metric	MSPD2IX	MSPD2I0	MSPD2I1	MSPD2I2
r	0.5350 (0.82)	0.4203 (1.06)	0.5071 (0.67)	0.5233 (0.76)
$rmse$	0.0502 (1.06)	0.0527 (0.40)	0.0510 (0.91)	0.0506 (1.50)

determined by fixing α to the maximum values of the r-correlation while varying β to scale the prior R_d probability M_{prior} of section 5.4, shown in table 5.8. The one-way ANOVA analysis between the r-correlation and $rmse$-error distributions of the maximum α-scale parameters of table 5.8 with the maximum β-scale parameters of table 5.9 shows no statistical significant improvement for any of the four phase minimization based methods. This validates the widespread distribution of the scale parameter β concerning the r-correlation maxima and $rmse$-error minima over a bigger value range than α. The ANOVA analysis demonstrates that β has no statistically significant influence on each of the evaluated phase minimization variants.

b) Viterbi steering, OQ_{GMM} prediction model 2

The OQ_{GMM} prediction model 2 uses as additional voice descriptor the fundamental frequency F_0 to further augment the robustness of the R_d estimation on natural human speech. A comparison of each $rmse$-error and each r-correlation value for each speaker between table 5.6 (model 1, without F_0) and table 5.7 (model 2, with F_0) shows that the consideration of F_0 contributes to the robustness of the GMM estimation model to predict OQ_{GMM}. On the one hand, only the correlation on the data test set for speaker JMK deteriorates to a marginal extent from r=0.3500 for model 1 to r=0.3478 for model 2. On the other hand, the correlation on the data test set for speaker BDL augments by employing F_0 from r=0.4705 for model 1 to r=0.7285 for model 2 to a significant extent. Again,

Table 5.10: *Viterbi steering, model 2 (optimal α-values in parentheses)*

Metric	MSPD2IX	MSPD2I0	MSPD2I1	MSPD2I2
r	0.5437 (0.05)	0.4623 (0.01)	0.5234 (0.05)	0.5337 (0.01)
$rmse$	0.0498 (0.03)	0.0515 (0.03)	0.0501 (0.05)	0.0499 (0.01)

the scale parameter β remains fixed to 1.0 while the scale parameter α is varied. Table 5.10 illustrates the results of the α-optimization for the Viterbi steering of model 2. A one-way ANOVA analysis evaluates between the results of the corresponding test to optimize the scale parameter α of Viterbi steering using model 1 (illustrated in table 5.8) with Viterbi steering using model 2 (listed in table 5.10). It demonstrates improvements to a statistically significant extent for method MSPD2I2 at significance level 5 % (p-value < 0.05), for method MSPD2I1 at significance level 1 % (p-value < 0.01), and for method MSPD2I0 again at significance level 0.1 % (p-value < 0.001). No statistically significant improvement could be measured for the overall best performing method MSPD2IX when using model 2 compared to using model 1. Please note that the r-correlation improved for MSPD2IX slightly from r=0.5348 (listed in table 5.8) to r=0.5437 (listed in table 5.10). Fixing the determined α-maxima in terms of the measured r-correlations, depicted in table 5.10, to optimize the scale parameter β, whose results are illustrated in table 5.11, does not exhibit statistically significant improvements. The β-optimization of table 5.11 shows the overall best performance of the OQ comparison test to motivate the following one-way ANOVA analysis. It validates that applying Viterbi steering is statistically significant for all evaluated phase minimization variants. The

Table 5.11: *Viterbi steering, model 2 (optimal β-values in parentheses)*

Metric	MSPD2IX	MSPD2I0	MSPD2I1	MSPD2I2
r	0.5438 (1.06)	0.4625 (1.03)	0.5246 (0.76)	0.5343 (0.76)
rmse	0.0498 (1.50)	0.0502 (0.20)	0.0501 (1.00)	0.0499 (1.09)

evaluation distributions of standard Viterbi smoothing illustrated in table 5.4 are compared with the overall best Viterbi steering results, the β-optimization of model 2 shown in table 5.11. The improvements are statistically significant for method MSPD2IX at significance level 5 % (p-value < 0.05), for method MSPD2I2 at significance level 10 % (p-value < 0.1), as well as for the methods MSPD2I1 and MSPD2I0 at significance level 0.1 % (p-value < 0.001).

c) Viterbi steering summary

The utilization of the novel Viterbi steering approach to augment the robustness of Viterbi smoothing when applied to smooth the estimated R_d contours demonstrates improvements to a statistically significant extent when comparing the results with standard Viterbi smoothing
a) for both employed GMM models (with and without F_0),
b) for each of the four phase minimization methods of section 5.3.3 and,
c) for both R_d regression adaptation variants of section 5.2.

Please note that only the results for R_d regression adaptation variant 2 was presented. The adaptation variant 1 proved to be as well statistically significant. The usage of the OQ_{GMM} prediction model 2 outperforms model 1 for each R_d estimator and each R_d regression variant on each evaluation metric (*r*-correlation and *rmse*-error). This suggests that the fundamental frequency F_0 as the first dimension of prosody [Fujisaki et al., 1981] correlates to a reasonably extent with the R_d contours. F_0 can thus additionally be exploited as co-variation feature with other voice descriptors to train a model for the R_d prediction. However, this has to be further examined on a bigger test set employing more speakers to evaluate its speaker-dependency.

The R_d regression adaptation variant 2 achieves the overall better results compared to variant 1 on the OQ comparison test. The overall best performing R_d estimation method is MSPD2IX, followed by MSPD2I2, MSPD2I1 and respectively MSPD2I0. It was shown that the worst performing method MSPD2I0 profits the most from the auxiliary steering of the Viterbi algorithm while the best performing method MSPD2IX profits the least from Viterbi steering. On the one hand, this suggests that the OQ_{GMM} prediction exploiting the co-variation of other voice descriptors is a relatively robust manner to estimate R_d. On the other hand, this conclusion indicates that MSPD2IX and to some extent the good performing methods MSPD2I2 and MSPD2I1 may already estimate comparatively robust R_d curves by the utilization of standard Viterbi smoothing without the additional steering.

The EGG-based technique is not error-free, as shown in fig. 5.11. The employed evaluation metric comparing the OQ_{EGG} with the OQ_{Rd} values presented in this evaluation chapter is therefore limited. It prevents the potential to achieve higher correlation and lower error measurements. The relatively lower performance improvements of Viterbi steering compared to Viterbi smoothing suggest furthermore that the estimation results may already be close to a certain upper boundary given the utilized data set and employed estimation algorithms.

***OQ* test of other methods across speakers:**
The results of the following other glottal source estimation algorithms are examined on the same test set of the CMU Arctic databases. It provides an objective comparison to the presented algorithm variants of the preceding sections. The same evaluation metrics *r* and *rmse* are applied.

DyProg-LF:
The first method chosen for comparison is the DyProg-LF method described in section 3.7.1.4. It uses Inverse Filtering and Dynamic Programming to estimate R_d.

Strik-LF:
The second method chosen for comparison is called Strik-LF, proposed in [Strik et al., 1993, Strik, 1998]. For this work the same glottal source signals estimated by the inverse filtering method of the DyProg-LF algorithm were utilized. The method estimates LF model parameters and its amplitude measures directly on glottal source signals in the time domain. A two part optimization procedure improves the LF model parameter estimation. First, the Nelder and Mead simplex optimization algorithm is applied being insensitive to large errors in the initialization. Second, a steepest descent optimization algorithm further refines the LF model fit.

PowRd:
The third evaluated method called PowRd is the power spectrum based method of [Ó Cinnéide, 2012], introduced in section 3.7.4.

Other methods results:
As well a non-constant offset between the OQ curves estimated by each of the three comparison methods versus

Table 5.12: *Comparison results of other methods*

Metric	DyProg-LF	Strik-LF	PowRd
r	0.3721	0.1215	0.1776
$rmse$	0.0716	0.1217	0.1760

the OQ curves derived from the corresponding EGG recordings was inspected. It confirms the observation of a possible systematic OQ bias by the EGG-based method discussed in section 5.6.4.1.

However, since the author of [Kane and Gobl, 2013a] provided the OQ estimation results for the methods Strik-LF and DyProg-LF using the method described in [Drugman et al., 2012b] to estimate the required GCIs, the GCI time instants and voicing decisions were not completely congruent to the DECOM basis utilized here. It lead to a slightly different evaluation metric. The Strik-LF and the DyProg-LF method are based on the same estimation of the glottal source signal from inverse filtering. The better performance of the DyProg-LF compared to the Strik-LF method shown in table 5.12 confirms to a certain extent the conclusions drawn from our the presented Viterbi smoothing. Dynamic programming immensely improves the results of a glottal source shape parameter estimation. It suppresses unnatural jumps in short-time segments. On the other hand, parts of the better performance of the DyProg-LF method can be assigned to the conjoint optimization in the time and the spectral domain.

The evaluated variant of the PowRd method is frame-based without the utilization of a dynamic programming approach. Its estimation robustness could therefore be augmented by employing as well a probabilistic smoothing algorithm.

5.6.4.4 OQ test per speaker

The preceding sections of the OQ comparison test illustrated the performance of the employed algorithms generalized over different speakers to estimate R_d. However, it is of vital interest to examine the estimation robustness of each method in dependency to the intrinsic peculiarities of each speaker.

Speaker analysis:

Table 5.13: *Mean μ and standard deviation σ of speaker characteristics*

Measure	BDL	JMK	SLT
F_0^{μ}	121.83Hz	112.42 Hz	174.19 Hz
F_0^{σ}	17.16 Hz	13.91 Hz	17.38 Hz
OQ_{EGG}^{μ}	0.41	0.62	0.54
OQ_{EGG}^{σ}	0.09	0.07	0.10

Table 5.13 shows the mean μ and standard deviation σ for F_0, and OQ derived from the EGG signals. Both are measured on all voiced segments and all phrases for each speaker of the employed three CMU Arctic speakers. The two male speakers BDL and JMK exhibit a comparatively lower mean pitch F_0^{μ} than the female speaker SLT. Speaker JMK has the least variance σ^2 in his F_0 contour while SLT and BDL exhibit larger F_0 variations. The observation for BDL corroborates the findings of [Drugman et al., 2012a]. JMK demonstrates the highest mean open quotient OQ_{EGG}^{μ} with the lowest variance σ^2.

Informal listening tests suggest that BDL has a very clear articulation and speaks with a high vocal effort. BDL has an overall modal phonation but uses quite often creaky voice offsets [Drugman et al., 2012a] which may degrade the R_d estimation accuracy due to the non-modal phonation of the creaky voice quality. JMK has a clear articulation but speaks with a weak vocal effort which partially results in a whispered [Obin, 2012] and creaky voice quality [Titze, 1994] with a bit of nasality. SLT under-articulates and speaks with a low vocal effort. She has a rather modal phonation with a bit of nasality. All three speakers talk with a low pulmonic pressure [Catford, 1977]. BDL and to a less extent STL exhibit a pressed voice quality. JMK in contrary has a more relaxed voice quality which can lead according to the evaluation of the synthetic test set in section 5.6.3 to a less robust glottal source shape parameter estimation performance.

Per speaker test setup:
The OQ comparison test results per speaker are presented in the following. Only the results for each speaker using R_d adaptation variant 2 are shown. It demonstrates the better performance throughout the whole OQ comparison test. The discussion examines only Pearsons correlation metric r.

Please note that the r-correlations as well as the α- and β-values listed for each speaker in the following sections correspond to the maxima measured for each test set. Each preceding test was executed for all speakers to determine the best scale parameter α for Viterbi smoothing and the best scale parameter β for Viterbi steering per

algorithmic variant and globally over speakers. Therefore, not the optimal α-maxima per speaker but the optimal α values over speakers were fixed for the subsequent β variation tests. This global setting does partially render not optimal speaker specific results for the tests presented in the following. The values which will be given in the following sections to each correlation r in parenthesis correspond to the following indices of each algorithmic variant:

1. Without Viterbi smoothing
2. Standard Viterbi smoothing
3. Viterbi steering, model 1, α variation, β fixation
4. Viterbi steering, model 2, α variation, β fixation

OQ test results for speaker BDL:
The results of the synthetic test discussed in section 5.6.3 associate higher F_0 values as well as higher R_d and OQ values with a lower performance in estimating glottal source shape parameters. Speaker BDL presents among

Table 5.14: *BDL r-correlation results (α- or β-values in parentheses)*

Smoothing	MSPD2IX	MSPD2I0	MSPD2I1	MSPD2I2
r (1)	0.5165	0.4761	0.5005	0.5233
r (2)	0.7263 (0.05)	0.6029 (0.21)	0.6820 (0.03)	0.7043 (0.11)
r (3)	0.7312 (0.07)	0.6323 (0.13)	0.6989 (0.09)	0.7021 (0.03)
r (4)	0.7771 (0.01)	0.7479 (0.03)	0.7708 (0.03)	0.7568 (0.01)

the three evaluated speakers the highest r-correlations, listed in table 5.14. He has the lowest mean open quotient OQ^{μ}_{EGG} and a comparatively low F_0. His high vocal effort and clear articulation contribute to the ease of estimating his glottal excitation source shape. The utilization of F_0 within the voice descriptor set for prediction model 2 contributes to the R_d estimation robustness for speaker BDL. The other three methods DyProg-LF, Strik-LF, and

Table 5.15: *BDL comparison results of other methods*

Metric	DyProg-LF	Strik-LF	PowRd
r	0.6606	0.3268	0.3315

PowRd achieve as well the best R_d estimation results for speaker BDL, illustrated in table 5.15.

OQ test results for speaker JMK:
Speaker JMK poses not before expected problems to estimate the shape of his glottal excitation source. Despite the lowest mean F_0 of all speakers his high mean open quotient OQ^{μ}_{EGG} measured and his perceived weak vocal effort lead to less robust R_d estimation results. The EGG-based OQ_{EGG} reference of speaker JMK exhibits more physiological impossible movements compared to the other speakers OQ_{EGG} reference. The Viterbi steering of

Table 5.16: *JMK r-correlation results (α- or β-values in parentheses)*

Smoothing	MSPD2IX	MSPD2I0	MSPD2I1	MSPD2I2
r (1)	0.1902	-0.0183	0.1755	0.2478
r (2)	0.4267 (0.05)	0.0322 (0.05)	0.3922 (0.07)	0.4235 (0.03)
r (3)	0.4299 (0.03)	0.3307 (0.01)	0.3900 (0.07)	0.4242 (0.01)
r (4)	0.4344 (0.05)	0.2646 (0.01)	0.3937 (0.09)	0.4274 (0.03)

model 2 using F_0 demonstrates only slight improvements for all but the worst performing method MSPD2I0, depicted in table 5.16. The latter does not benefit from the exploitation of the F_0-covariation but from the utilization of Viterbi steering in general. Moreover, the three methods employed for comparison have even greater problems

Table 5.17: *JMK comparison results of other methods*

Metric	DyProg-LF	Strik-LF	PowRd
r	0.0879	-0.1106	0.0797

to establish performant R_d estimation results for speaker JMK, shown in table 5.17. The PowRd method without the utilization of dynamic programming achieves a similar performance to the DyProg-LF approach which uses

dynamic programming.

***OQ* test results for speaker SLT:**
Speaker SLT with the highest F_0^{μ} achieves similar R_d estimation results as speaker JMK with the lowest F_0^{μ}. No

Table 5.18: *SLT r-correlation results (α- or β-values in parentheses)*

Smoothing	MSPD2IX	MSPD2I0	MSPD2I1	MSPD2I2
r (1)	0.2485	0.0764	0.2636	0.2651
r (2)	0.4486 (0.15)	0.0943 (0.53)	0.4005 (0.13)	0.4541 (0.11)
r (3)	0.4497 (0.15)	0.3929 (0.03)	0.4280 (0.15)	0.4500 (0.11)
r (4)	0.4444 (0.15)	0.3772 (0.03)	0.4197 (0.13)	0.4454 (0.11)

method could benefit from F_0 as voice descriptor for the OQ_{GMM} prediction model 2, shown in table 5.18. It indicates that the rather large difference in F_0^{μ} between the employed training data set of the male speakers BDL and JMK versus the test data set for the female speaker SLT requires more speaker data with higher pitch to train the prediction model 2 using F_0. The DyProg-LF method listed in table 5.19 benefits from the application of its used

Table 5.19: *SLT comparison results of other methods*

Metric	DyProg-LF	Strik-LF	PowRd
r	0.3638	0.1206	0.1204

dynamic programming approach and achieves nearly a similar performance compared to the phase minimization variants using Viterbi smoothing and steering.

***OQ* test results summary per speaker:**
The results of the best performing phase minimization variant MSPD2IX are summarized per CMU Arctic speaker for the following algorithmic variants being indexed by the same number as shown in fig. 5.13:
1. No Viterbi smoothing
2. Standard Viterbi smoothing
3. Viterbi steering, model 1, α-optimization
4. Viterbi steering, model 1, β-optimization
5. Viterbi steering, model 2, α-optimization
6. Viterbi steering, model 2, β-optimization

The *OQ* test across speakers of section 5.6.4.3 reports no statistically significant improvements for MSPD2IX when conducting an ANOVA analysis between the estimation results of algorithmic variant number 2 (Standard Viterbi smoothing) with the following Viterbi steering variants of number 3 to number 5. Only the β-optimized Viterbi steering with model 2, listed here as algorithmic variant number 6, achieves statistically significant improvements for method MSPD2IX at significance level 5 % (p-value < 0.05) when compared to standard Viterbi smoothing. Clear improvements can be visually inspected in fig. 5.13 for speaker BDL in terms of higher mean and higher values of the horizontal whiskers which reflect the variability of the results outside the upper and lower quartiles. Also the outliers shown in red dots exhibit in general higher r-correlation values. However, such improvements aren't or only to a very little extent visible for the speakers JMK and SLT. The improvement of Viterbi steering (numbers 3 to 6) compared to Viterbi smoothing (numbers 2) is therefore limited. In contrast, the sole application of Viterbi smoothing shows significant improvements compared to the frame-based MSPD2IX estimator without smoothing (number 1).

5.6.4.5 Viterbi steering extension

Additional voice descriptors:
This paragraph presents further investigations to improve Viterbi steering by means of utilizing voice descriptors being higher correlated to the OQ_{EGG} reference than the voice descriptor presented in the preceding sections. The idea being that higher co-varying voice descriptors should improve the GMM data modelling and the corresponding OQ_{GMM} prediction to increase the Viterbi steering performance accordingly.

Table 5.20: *Pearsons r-correlation of selected voice descriptors versus the OQ_{EGG} reference*

Metric	$H1$-$H2$ $[dB]$	3 MFCC bins	F_{VU} $[Hz]$	F_0 $[Hz]$	Loudness $[dB]$	Skewness	Kurtosis
r	0.3496	0.3401	0.2627	0.2153	0.4750	0.4907	0.4563

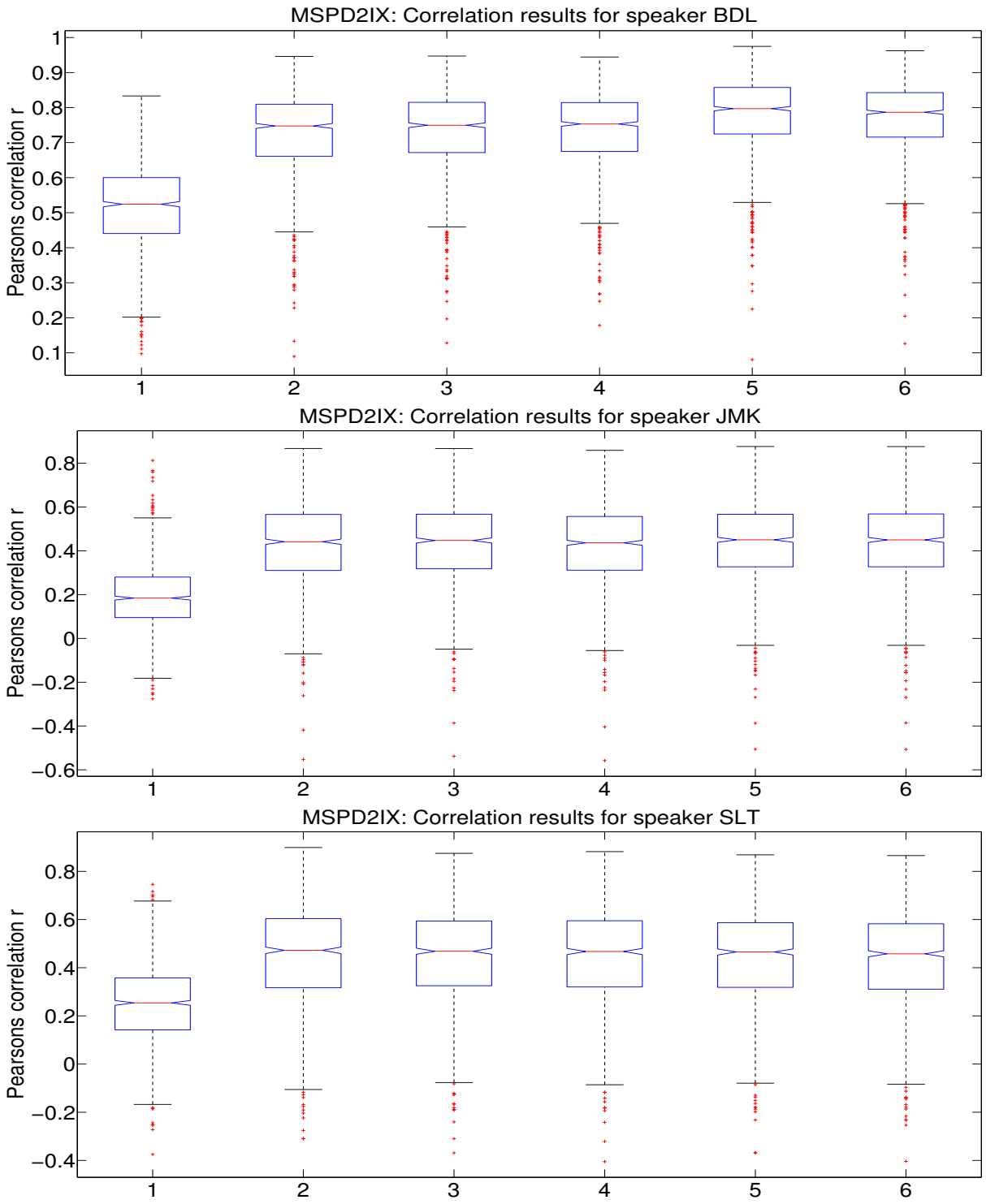


Figure 5.13: *OQ test results summary per speaker*

The first four columns of table 5.20 lists the voice descriptors which have been employed for the study of [Huber and Röbel, 2013]. $H1$-$H2$, the 3 MFCC bins, F_{VU} and F_0 exhibit a comparably lower r-correlation than the voice descriptors of the last three columns. The Skewness and the Kurtosis of the spectral representation of signal $S(\omega)$ and its Loudness in dB co-vary to a higher extent with OQ_{EGG}. A large set of 1285 voice descriptors available from the IRCAM descriptor library [Peeters, 2004] was examined on their correlation with OQ_{EGG}. The two new selected voice descriptors Skewness and the Kurtosis exhibit a higher r-correlation since they reflect better the spectral changes implied with the continuously changing glottal gestures. The Loudness measure reflects the vocal intensity originating from the sub-glottal pressure [Laver, 1968, Vilkman et al., 1999, Fant and Kruckenberg, 2005]. All voice descriptor particularities are reflected to a different extent in the signal $S(\omega)$ and the OQ_{EGG} reference.

Two new OQ_{GMM} prediction models 1 and 2 are trained on their corresponding voice descriptor set:
GMM model 3 - Feature set D_3:
D_3 = $H1$-$H2$, 3 MFCCs, $log_{10}(F_{VU}^{0\mu})$, $log_{10}(F_0^{0\mu})$.
GMM model 4 - Feature set D_4:
D_4 = $H1$-$H2$, 3 MFCCs, $log_{10}(F_{VU}^{0\mu})$, $log_{10}(F_0^{0\mu})$, Loudness dB, Spectral Skewness, Spectral Kurtosis.

The feature set D_3 of GMM model 3 equals D_2 of section 5.5.2 with the difference that the two frequency measures F_{VU} and F_0 have zero mean and are interpreted on the logarithmic scale. GMM model 4 employs in its feature set D_4 additionally the new voice descriptors Loudness dB, the Spectral Skewness and Kurtosis being listed in table 5.20.

Standard 3-fold cross validation:

Table 5.21: *GMM prediction results, training set, 3-fold*

Speaker	r (train, Model 3)	r (train, Model 4)
BDL	0.5604	0.6410
JMK	0.7852	0.7946
SLT	0.9115	0.9230
ALL	0.7524	0.7862

Table 5.21 lists the Pearson r-correlation measured on the data set to train each GMM by means of the 3-fold leave-one-out cross-validation as performed with the preceding models. The removal of the mean for F_0 and F_{VU} and the application of the logarithmic scale does not contribute to higher r-correlations of the GMM prediction. The column for Model 3 (train) lists lower r-correlations per speaker than the corresponding 3^{rd} column r (train) of table 5.7. The additional consideration of the three highly co-varying voice descriptors Loudness, Spectral Skewness and Spectral Kurtosis in the feature set D_4 achieves slightly higher r-correlations for the training test set of GMM Model 4, shown in the 3^{rd} column of table 5.21. However, the higher correlations of r=0.7862 for Model 4 is only marginal better compared to r=0.7803 for Model 2, shown in table 5.7. Table 5.22 lists the r-correlation

Table 5.22: *GMM prediction and Viterbi steering results, test set, 3-fold*

Method	r (test)
Model 3 (predict)	0.4244
Model 4 (predict)	0.3845
Model 3 + MSPD2IX	0.5200 (0.07)
Model 4 + MSPD2IX	0.5105 (0.05)

results not per speaker but being averaged over the three used speakers. The desired improvements of utilizing additional highly correlating voice descriptors do not manifest in the corresponding evaluation on the test set. The performance of the sole GMM prediction using the models 3 and 4 on the test set is with r=0.4244 and r=0.3845 lower than r=0.4931 for the corresponding prediction of model 2 listed in table 5.7. The probabilistic surface of the GMM prediction is combined with the error surface of the phase minimization algorithm MSPD2IX to form a probabilistic surface for the Viterbi algorithm. Global r-maxima exist for the optimization of the Viterbi weighting parameter α, shown in parentheses. The results for the presented extension of Viterbi steering listed in the 4^{th} and 5^{th} row of table 5.22 lie below r=0.5438 of the preceding Viterbi steering model 2 shown in table 5.11.

5-fold cross validation without speaker dependency:
One reason for the failing improvement of the Viterbi steering models 3 and 4 may be in general the utilized 3-fold leave-one-out cross validation. The GMM prediction model is trained on two speakers to predict OQ_{GMM} on the third speaker. Speaker value differences may introduce a bias into the data modelling and prevent the model to generalize well. The following investigation presents therefore a 5-fold cross validation. The GMM training

data is comprised of 80 % of all three speakers corpora, which the remaining 20 % building the test data from all three speakers. The higher r values reported in table 5.23 of the 5-fold approach with sole GMM prediction

Table 5.23: *GMM prediction results, training set, 5-fold*

Speaker	r (train, Model 3)	r (train, Model 4)
ALL	0.8255	0.8445

on the training set confirm that value differences between speakers are present in the 3-fold cross validation. The

Table 5.24: *GMM prediction and Viterbi steering results, test set, 5-fold*

Method	r (test)
Model 3 (predict)	0.5081
Model 4 (predict)	0.5041
Model 3 + MSPD2IX	0.5641 (0.07)
Model 4 + MSPD2IX	0.5701 (0.05)

evaluation results of the 5-fold cross validation shown in table 5.24 constitute the highest r-correlations on the test data for both the sole GMM prediction listed in the 2^{nd} and 3^{rd} row and the actual Viterbi steering in combination with phase minimization listed in the 4^{th} and 5^{th} row. The better results of table 5.23 on the training data set and of table 5.24 on the test set versus the 3-fold leave-one-out approach confirm that a speaker dependency is present in the utilized GMM modelling. This indicates that more speaker data is required such that the GMM-based data modelling generalizes well over different speakers and their intrinsic peculiarities.

5.7 Conclusions

R_d range adaptation and extension:
A continuous coverage of voice qualities from a tense to a normal to a relaxed phonation type is provided by the proposed parameterization of the extended R_d range presented in section 5.2. It requires a parametric adaptation of the equations defining the R_d regression of the parameter space of the LF model. The R_d range extension covers more adducted and abducted phonations found outside the normal R_d range [0.3, 2.7]. The R_d adaptation variant 2 achieves a more precise distinction between glottal source shapes for the normal R_d range. This can be the reason for its better performance on the test set of natural human speech of section 5.6.4 where R_d values occur predominantly in the normal R_d range.

The objective evaluation on a synthetic test set of section 5.6.3 by interpreting the test results as in section 5.6.3.3 in terms of voice quality shows that higher values than e.g. R_d=6.0 to limit the R_d upper range may not be beneficial to the R_d estimation. One reason is that the estimation of higher R_d values is more error-prone because the (perceptual) similarity of different glottal pulse shapes in higher R_d value regions is higher than in lower R_d value regions. The perceptual impact in the upper R_d range for abducted termination is lower than for adduction in the lower and normal R_d range. This is additional corroborated by the test results of test on natural human speech of section 5.6.4. The adaptation variant 2 achieves a better R_d estimation performance since it samples the LF parameter space of the predicted R_{*p} waveshape set with a higher resolution in the R_d sub-range R_d=[1.8476, 2.7]. of the perceptually more important normal R_d range [0.3, 2.7]. Moreover, higher R_d values reflect more relaxed voice qualities with an increased breathiness and thus a higher noise level with less stable harmonic sinusoids available.

Phase minimization variants:
The results of section 5.6 demonstrate that the two novel phase minimization variants MSPD2IX and MSPD2I2 explained in section 5.3.3 and proposed in [Huber et al., 2012] outperform the baseline method MSPD2I1 of [Degottex, 2010, Degottex et al., 2011a] in estimating R_d. The results of the synthetic test of section 5.6.3 confirm their promising proof-of-concept explained with the analysis of the error surfaces shown in fig. 5.12. The error surfaces estimated on natural human speech shown in fig. 5.6 verify the findings. The better performance of MSPD2IX and MSPD2I2 is completely validated by the two objective evaluation tests presented in sections 5.6.4 and 5.6.3. The estimation of voice qualities with a comparatively higher relaxed phonation and less vocal effort poses more difficulties.

Viterbi steering:
The Viterbi steering attempt explained in section 5.5 was implemented as a proof-of-concept to add more ro-

bustness to the glottal source estimation algorithms which use already Viterbi smoothing. It demonstrates certain advancements to a statistically significant extent compared to the Viterbi smoothing baseline approach of section 5.4. The exploitation of co-varying voice descriptors is able to increase the R_d estimation performance. The utilization of a machine learning approach to predict R_d from a set of voice descriptors is able to aid and to possibly outperform the signal processing paradigms known to date to estimate glottal source signals in the future.

The continuative extension of Viterbi steering discussed in section 5.6.4.5 tries to provide means leading to a higher performant R_d estimation compared to its baseline models 1 and 2 proposed in [Huber and Röbel, 2013]. The utilization of three additional highly correlated voice descriptors provided by the IRCAM descriptor library [Peeters, 2004] could not improve the Viterbi steering method. Despite model 5 improves the GMM prediction on the training data it results into a comparably lower performance on the test data set. Any executed test trial (e.g. the selection of different new voice descriptors) to improve the GMM prediction and Viterbi steering resulted in improvements for two and in a deterioration for another speaker. This indicates that the utilized CMU Arctic data set in combination with the evaluated algorithms and evaluation metrics is already saturating on a certain upper border.

Viterbi steering requires more speech corpora from more speakers of different age and gender covering more speaker characteristics in the trained models to achieve a robust GMM-based prediction which is able to generalize properly its data modelling on any analyzed speaker. The underlying data model of the utilized GMM models requires to cover any possible feature combination present in any analyzed speech phrase. The 5-fold cross validation presented in section 5.6.4.5 confirms that the utilized 3-fold leave-one-out cross validation shown in the preceding sections 5.6.4.3 and 5.6.4.4 suffers from a speaker dependency.

Another drawback of the presented Viterbi steering is that the probability surfaces of the GMM prediction and phase minimization may compete against each other. The simple superposition of both probability surfaces does not assure that their combined probability constitutes only one maximum per evaluated frame. Contrariwise, the sole application of either GMM prediction or phase minimization proved to be less performant.

The achieved improvements of Viterbi steering are of minor impact compared to the comparably bigger amount of required data modelling and higher computational costs. The R_d estimator using standard Viterbi smoothing is therefore the default glottal source estimation algorithm used throughout this thesis. The novel speech framework introduced in chapter 6 and the novel VC system discussed in chapter 7 employ MSPD2IX along with standard Viterbi smoothing to estimate R_d.

Please note that the partially very promising results of the GMM-based predictor reported in the tables 5.6 and 5.7 for the training test set may be biased. The employed Pearson r-correlation metric was evaluated over two speakers used in the training. The evaluation metric analyzes therefore additionally the mean difference present between two speakers. This is not the case for the evaluation on the test set where only the prediction results for one speaker are examined. The correlation values for the test set express therefore only specific details between the predicted R_d contour and the R_d contour derived from the EGG comparison basis. In contrast, the correlation values for the training set may be biased by the mean difference present between two speakers. This behaviour was realized while the non-published continuation work on Viterbi steering presented in section 5.6.4.5, after the publication of [Huber and Röbel, 2013].

Viterbi smoothing:
The importance to smooth estimated glottal source parameters over time was validated as in [Vincent et al., 2007, Kane et al., 2012, Kane and Gobl, 2013b] by the experimental findings. The results shown in table 5.4 and in fig. 5.13 confirm that contour smoothing by means of dynamic programming immensely increases the glottal source estimation performance compared to the frame-based R_d estimator, shown in table 5.2.

Estimator parameterization:
The optimal parameter adjustment poses further difficulties especially for the analysis of natural human speech. Different parameterizations of the R_d estimator lead to different estimated R_d contours. A missing ground truth in real world application leaves the user to choose visually which R_d contour could resemble the most the true underlying R_d contour contained in the input signal $S(\omega)$. A user adjusts the parameterization of the R_d estimator by means of an iterative trial-and-error parameter optimization accordingly. Crucial parameters are the α (and β) weights of Viterbi smoothing (and steering) determining the influence between transition and (the two) error surface(s). Also the chosen minimum and maximum amount of available stable harmonic sinusoids to execute phase minimization is influential. One trick to optimize the parameterization is to examine the estimated GCI locations visually versus the repetitive minimal amplitude locations present in each short-time segment of the time domain waveform according to the local fundamental period. If both time locations differ to a huge extent over longer time segments like a diphone or a syllable A too huge time difference of both time locations present over longer time segments like a single phoneme or a complete syllable indicates an erroneous R_d estimation which in turn leads to the erroneous GCI estimation.

Chapter 6

Contribution - PSY: A flexible Parametric Speech SYnthesis system

My sheep listen to my voice. I know them, and they follow me.

The Holy Bible (John 10:27)

6.1 Overview

6.1.1 Introduction

Building set framework:
A new flexible speech framework for the advanced analysis, transformation and synthesis of spoken voices is presented in this chapter. It is based on an extended source-filter model introduced in the following section 6.1.2. The novel speech system is openly designed for the utilization as a basic building block system allowing for advanced Voice Transformation and Voice Conversion purposes to alter voices. The analysis sub-system of *PSY* allows for the automated estimation of an extended voice descriptor set for a source and target speaker pair. Its synthesis sub-system depends on a set of designated voice descriptors to construct the synthesis of a spoken phrase. It can process the required voice descriptors from the same or different voice sources. The extended voice descriptor set consists of a voiced part $V(\omega)$ and an unvoiced part $U(\omega)$. $V(\omega)$ comprises the Vocal Tract Filter (VTF) $C(\omega)$, the glottal pulse $G(\omega)$, the fundamental frequency F_0, the Voiced / Unvoiced Frequency boundary F_{VU}, and its RMS-based energy E_{voi}. $U(\omega)$ consists of the unvoiced component signal itself and its RMS-based energy E_{unv}.

The novel speech system is denoted "**PSY**" to refer to **P**arametric Speech Analysis, Transformation and **SY**nthesis. For the time being, *PSY* is primarily suited for advanced transformations of the glottal excitation source of speech signals, and the transformation of the employed voice descriptor set from a source to a target speaker within the context of VC.

It facilitates additionally the conversion and synthesis of spectral envelope and VTF sequences in combination with the novel VC system introduced in the subsequent chapter 7. A future and not yet tested application is the transformation of the fundamental frequency F_0 and other prosodic characteristics.

SVLN continuation:
Recent research in the speech community [Drugman and Dutoit, 2012, Cabral and Carson-Berndsen, 2013] has notably improved the speech synthesis quality by explicitly modelling the deterministic and stochastic component of the glottal excitation source. Advanced source-filter decomposition strategies, presented in section 3.8 and proposed in [Vincent et al., 2007, Cabral et al., 2008, Degottex et al., 2013], address finer details defined by extended voice production models for human speech. These approaches analyze an extended voice descriptor set to model

their transformation and synthesis. The extended voice descriptor sets consist of: the Vocal Tract Filter (VTF), the shapes and the GCI positions of glottal pulses, the random noise component, and energies. The parametric analysis, transformation and synthesis framework *PSY* is build on the means of the SVLN system presented in chapter 3.8.4 since SVLN was developed in the same laboratory. In *PSY*, the underlying model of SVLN is notably extended by
a) advanced means to estimate the unvoiced stochastic component $U(\omega)$ introduced in chapter 6.3,
b) an additional model being more robust to control energy changes required to facilitate voice quality transformations and VC,
c) more robust means to handle the VTF by relieving the model from the dependency on the F_{VU}, and
d) the modelling of finer details of speech signals such that a certain naturalness is maintained in the synthesized speech waveform.
The *PSY* framework is explained in the subsequent sections and chapters. At certain points in the following it will be partially compared to its SVLN baseline approach. Example illustrations of particular technical steps are given in the following discussion for the CMU Arctic speaker BDL [Kominek and Black, 2004, Huber and Röbel, 2013] to better visualize the algorithmic behaviour.

Chapter 6.5 presents different schemata to synthesize a speech signal. Different synthesis possibilities cover different requirements imposed by different speech processing tasks like VC or the transformation of voice quality. For this reason, *PSY* offers to process speech with a wide-band or a full-band model without frequency limits [Degottex and Stylianou, 2013]. Future applications like F_0 transformation or the generation of a human voice avatar have not yet been implemented.

PSY operates with only little constraints imposed on the utilized voice descriptor set. The SVLN baseline approaches of [Degottex, 2010, Drugman and Dutoit, 2012, Degottex et al., 2013] assume a constant voice quality over longer time segments like one phoneme, one syllable or one word. In contrast, *PSY* attempts to capture and reproduce finer voice quality details such that synthesized phrases do not contain buzzy and muffled sounding effects. However, the robustness and synthesis quality of *PSY* is for the time being depending on the accuracy of the estimated GCI times, along with the corresponding R_d^{gci} estimation at each closure time of the glottal excitation pulse.

Voiced and Unvoiced separation:
In *PSY*, one main conceptual design separates the processing into a voiced deterministic $V(\omega)$ and into an unvoiced stochastic $U(\omega)$ component. The separation is based on a deterministic plus stochastic model (DSM) approach [Serra, 1989]. The modelling of $V(\omega)$ by glottal pulses convolved with VTFs facilitates the superimposition of auxiliary or the suppression of present sinusoidal content in the voiced component. Such handling of the sinusoidal content is for instance required to transform the deterministic part of the glottal excitation source, according to a transformed $R_d^{'}$ or $R_d^{'gci}$ contour. Within the context of the separation into a voiced and an unvoiced part, a novel method to estimate the stochastic component is presented. An informal subjective listening test conducted within the laboratory suggests that the re-synthesized stochastic component carries, in comparison to the re-synthesized sinusoidal part, more perceptual information of a speakers voice identity and information required to render a synthesized speech signal intelligible. The unvoiced component $U(\omega)$ is as the voiced component $V(\omega)$ coloured by the resonances of the vocal tract formants. A similar observation concerning the LPC residual representing the source excitation is discussed in [Sündermann et al., 2006b]. The residual obtained by inverse filtering is discussed in [Erro, 2008]. It contains among noise and the glottal source perceptually important formant and phase information contributing to a speakers voice identity.

Pulse-based speech modelling:
One major difference of the proposed system lies in its liberation of not having to model amplitude A_k, instantaneous frequency f_k and instantaneous phase ϕ_k of each quasi-harmonic sinusoid k in voiced segments [Serra, 1989]. This freedom is achieved by simply exciting a sequence of VTF envelopes with a glottal pulse sequence, synthesized at the corresponding glottal closure time instants. The localization of the pulses at the GCIs liberates the *PSY* framework of having to model the phase and frequency continuation of single sinusoids for synthesis [Bonada, 2008]. The correct amplitude values A are determined by the convolution of glottal pulses with the VTF, along with the energy management of *PSY* which will be presented in chapter 6.4.1. The complete sinusoidal analysis and synthesis to model speech is with this transferred to the spectral envelope estimation of the signal and the VTF, as well as the analysis and synthesis of a pulse sequence of the glottal excitation source. However, the sinusoidal parameter estimation may perform more robust than the GCI [Degottex, 2010] and the glottal pulse shape parameter estimation [Huber and Röbel, 2013].

The processing of the unvoiced stochastic component via the white noise excitation of spectral envelopes is described section 6.3. Its corresponding energy model is introduced in section 6.4.1. Both relieve the speech processing system of having to model explicitly the noise modulations in time, frequency and amplitude. The evaluation presented at the end of this chapter indicate that this constitutes an improvement compared to the synthesis of an ar-

tificially constructed noise component as in [Klatt and Klatt, 1990, d'Alessandro et al., 1998, Lu and Smith, 2001].

Excitation:
A more simple synthesis possibility is to excite a spectral envelope sequence with the impulse train $\delta_s(n)$ of a Dirac delta function $\delta(n)$. Different means to use an impulse train for F_0 modification in a LPC vocoder are discussed in [Cotescu and Gavat, 2010]. The utilization of a glottal source model within a speech synthesis system not just permits to alter F_0 but also the voice quality in terms of a tense, modal or breathy voice. Results given in [Cabral et al., 2008] indicate that the glottal excitation source outperforms impulse excitation.

Publication:
Please note that the works presented in this chapter were published as a basic version in [Huber and Röbel, 2015a] and as an extended version in [Huber and Röbel, 2015b].

6.1.2 Voice production model

PSY operates upon the following generic interpretation of the human voice production in the time domain:

$$s(n) = u(n) + v(n) \tag{6.1}$$

$$= u(n) + \sum_i g(n, P_i) * \delta(n - P_i) * c(n, P_i) \tag{6.2}$$

A speech signal $s(n)$ can be approximated as a superposition of an unvoiced stochastic component $u(n)$, and a voiced deterministic component $v(n)$. The deterministic component contains a sequence of glottal pulses that are located at the time positions P. Glottal Closure Instants (GCI) are estimated by the means described in [Degottex et al., 2010]. Each estimated GCI defines at index i the corresponding time position P_i. Each glottal pulse is represented by the glottal flow derivative $g(n, P_i)$. The LF glottal source model introduced in section 3.2.3 is used to synthesize $g(n, P_i)$. The one-dimensional LF shape parameter R_d discussed in section 3.3.1 parameterizes the LF model to describe the pulse shape of $g(n, P_i)$. The glottal flow derivative $g(n, P_i)$ describes as well the effect of the radiation filter $r(n)$ at lips and nostrils level [Fant, 1981]. $g(n, P_i)$ is convolved with a Dirac impulse at the GCI P_i and the Vocal Tract Filter (VTF) that is active for the related position $c(n, P_i)$. The VTF $c(n)$ is supposed to be minimum phase [Maia and Stylianou, 2014].

The speech signal model given in equ. 6.2 is processed in the spectral domain using the Short-Time Fourier Transform (STFT) for being able to make spectral domain manipulations. Please note that for brevity the sequence of a few consecutive glottal pulses $g(n, P_j)$ will be denoted as $g_s(n) = \sum_j g(n, P_j) * \delta(n - P_j)$ in the following. The summation over the GCI index j is set to comprise a signal segment of a few glottal pulses being covered by the Hanning window $w_h(n)$ of the STFT. Each glottal pulse is related to a slightly different VTF. The glottal pulse shape and the VTF are assumed to not change within the Hanning window w_h. Both are expected to be approximately given by the corresponding parameters in the window center.

Assuming that the filtering processes implied by each convolutional operation between the signal components of equ. 6.2 is involving impulse responses that are shorter than the length of window w_h allows interpreting the voice production model in the spectral domain. The STFT of the speech signal with each signal component denoted in upper case is given by:

$$S(\omega, k) = U(\omega, k) + V(\omega, k) \tag{6.3}$$

$$= U(\omega, k) + G(\omega, k) \cdot \Delta(\omega, k) \cdot C(\omega, k) \tag{6.4}$$

Each STFT frame k reflects the position of the window center. The frequency variable of the Discrete-Time Fourier Transform (DTFT) is denoted by ω. The dependency of all signal spectra with respect to k will be dropped in the following for easier illustration purposes. The STFT frame index k will still be utilized where needed. The unvoiced $U(\omega)$ and the voiced $V(\omega)$ components are the DTFT of the windowed unvoiced and voiced signals from equ. 6.2. It is assumed that g and c and the corresponding DTFT spectra $G(\omega)$ and $C(\omega)$ are quasi-stationary within the window. The spectral representation $\Delta(\omega)$ of the Dirac impulse sequence $\delta(n - P_i)$ reflects the GCI sequence of the glottal pulses time instants.

A quick overview explaining each signal component involved in the presented voice production model of *PSY* is given in table 6.1. Denotations in upper case indicate the spectral domain, denotations in lower case the time domain. More detailed explanations of each involved signal component are listed as follows:

$U(\omega)$: The unvoiced stochastic component $U(\omega)$ is extracted from the spectral representation $S(\omega)$ of signal $s(n)$ by deleting the sinusoidal content. The precise procedure to estimate $U(\omega)$ using different algorithmic variations are explained in chapter 6.3.

Table 6.1: *Signal descriptors used in PSY*

Spectral domain	Time domain	Signal type
$U(\omega)$	$u(n)$	Stochastic component, unvoiced or noise part
$V(\omega)$	$v(n)$	Deterministic component, voiced part
$R(\omega)$	$r(n)$	Radiation filter at lips and nostrils level
$C(\omega)$	$c(n)$	Vocal Tract Filter (VTF)
$G(\omega)$	$g(n)$	Glottal flow derivative signal, deterministic part of the glottal excitation source
$G_s(\omega)$	$g_s(n)$	Glottal flow derivative sequence, covering a few consecutive glottal pulses $g(n, P_j)$
$G^{gci}_{R_d}(\omega)$	$g^{gci}_{R_d}$	One glottal pulse at an associated GCI position, parameterized by the LF shape parameter R_d
$\Delta(\omega)$	$\delta_s(n)$	Impulse train representing a sequence of Dirac delta functions
$W_h(\omega)$	$w_h(n)$	Hanning window

$V(\omega)$: The voiced deterministic component $V(\omega)$ is estimated to synthesize purely harmonic sinusoidal content by means of exciting a VTF $C(\omega)$ by a glottal excitation pulse sequence $G(\omega)$. Another preliminary approach implemented in *PSY* for comparison purposes is given by the excitation of a spectral envelope sequence $\mathcal{T}_{sig}$ estimated on the signal $s(n)$ with the impulse train of a Dirac delta function $\delta(n)$. The precise procedure to construct $V(\omega)$ and its different algorithmic variations are explained in chapter 6.5.3.

$R(\omega)$: The radiation filter $R(\omega)$ in the spectral and $r(n)$ in the time domain reflects as explained in chapter 3.8.4.1 the radiation at lips and nostrils level. It is described as the time derivative $r(n)$, being in the spectral domain described by $R(\omega) = j\omega$ [Markel and Gray, 1976, Degottex et al., 2011a]. Please note that the radiation filter at lips and nostrils level $R(\omega)$ is not explicitly addressed in the voice production model of *PSY* described by equ. 6.2 and equ. 6.4. $R(\omega)$ is implicitly processed in the presented speech system *PSY*. For the voiced component $V(\omega)$, the glottal flow derivative $G(\omega)$ is a result of the convolution of $r(n)$ with the glottal volume-velocity flow. For the unvoiced component $U(\omega)$, the convolution of $R(\omega)$ with the stochastic part is already contained in the analyzed signal $s(n)$. The DSM-based extraction of $U(\omega)$ contains therefore implicitly the radiation $R(\omega)$.

$C(\omega)$: A Vocal Tract Filter $C(\omega)$ describes the filter of human voice production, as discussed in the sections 2.1 and 3.8.4.1. An Auto-Regressive (AR) model represents the VTF resonances exclusively by poles $A(z)$: $C(z) = 1/A(z)$. The AR Model considers the VTF as an all-pole filter [Mathews et al., 1961]. Adding the coupling with the nasal cavity introduces pole-zero pairs. The additional modelling with zeros $B(z)$ results in an Auto-Regressive and Moving Average (ARMA) model: $C(z) = B(z)/A(z)$ [Makhoul, 1975]. Two different means to extract the VTF from a speech signal are presented in chapter 6.2.3. Both are based on the AR model without considering the possibility to attribute zeros to the VTF.

$G(\omega)$: $G(\omega)$ and $g(n)$ represent the glottal flow derivative signal as part of the human voice production model of *PSY*. The glottal pulses constitutes the deterministic part of the glottal excitation source.

$G_s(\omega)$: $G_s(\omega)$ and $g_s(n)$ represent the windowed combination of several consecutive glottal pulses. Its length is determined by the window size of the STFT. The exact deployment of $G_s(\omega)$ and $g_S(n)$ is detailed in chapter 6.2.2.2.

$G^{gci}_{R_d}(\omega)$: $G^{gci}_{R_d}(\omega)$ and $g^{gci}_{R_d}$ represent a single glottal flow derivative, synthesized on a time instant corresponding to an estimated GCI being associated to the pulse. The spectral $G^{gci}_{R_d}(\omega)$ and time domain $g^{gci}_{R_d}$ waveforms represent the shape of the glottal pulse. The latter is described with the LF model [Fant et al., 1985] which is parameterized by the LF regression parameter R_d, introduced in chapter 3.3.1.

$\Delta(\omega)$: The Dirac delta function $\delta(n)$ represents in the time domain in a similar manner as explained in section 3.8.4.1 the pulse interval generated by the opening and closing of the glottis. The latter is the source of each pulse in the human voice production, causing the deterministic part in speech signals [Degottex, 2010]. $\Delta(\omega)$ constitutes the spectral representation of the Dirac impulse sequence $\delta_s(n)$. Please note that contrary to the usage of the harmonic structure $H^{F_0}(\omega)$ in section 3.8.4.1, $\Delta(\omega)$ expresses in the voice production model of *PSY* described in equ. 6.4 the Dirac impulse sequence. The latter causes the voicing of the sinusoidal content. $\Delta(\omega)$ does not reflect a harmonic structure for irregular spaced pulse sequences such as speech segments containing creak [Drugman et al., 2013, Kane et al., 2013a]. However, the Dirac impulse train $\delta_s(n)$ reflects for many voiced speech segments a quasi-periodic pulse interval sequence. The quasi-harmonic periodicity is for these cases expressed by $\Delta(\omega)$. The pulse continuation over time given by the time intervals of the GCI

sequence defines for these cases the fundamental periodicity which can be expressed by the fundamental frequency F_0.

6.1.3 *PSY* analysis system layout

Fig. 6.1 illustrates the different analysis blocks to estimate the voice descriptor set utilized in the speech framework *PSY*. The estimation of the voice descriptors F_0, F_{VU} and R_d^{gci} allows to synthesize a sequence of glottal pulses as glottal pulse waveform $g(n)$ on which the spectral envelope sequence $T_g(\omega)$ is estimated. The Vocal Tract Filter $C(\omega)$ results from a division of the spectral envelope sequence $T_{sig}(\omega)$ by $T_g(\omega)$. The latter is estimated on the original speech signal $s(n)$. The spectral envelope sequence $T_{unv}(\omega)$ is estimated on the extracted unvoiced component $u(n)$. The RMS-based energy contour E_{unv} of $u(n)$ is subtracted from the speech signals energy contour E_{sig} to express the energy contour E_{voi} of the voiced component $v(n)$ as their simple linear difference. The voiced E_{voi} and unvoiced E_{unv} energy contours are estimated over a whole speech corpus to provide training data for the corresponding voiced M_{voi} and unvoiced M_{unv} GMM-based energy models. The latter is used to model the energy variations implied with the voice quality transformation, presented in section 6.4.3. The VTF sequences $C(\omega)$ and the spectral envelope sequences $T_{unv}(\omega)$ are employed within the VC context , introduced in section 7.3. The different steps required to analyse all illustrated voice descriptors, spectral envelopes and the VTF sequence of one speech phrase will be introduced in the following.

Block diagram illustrating the different analysis parts in *PSY*

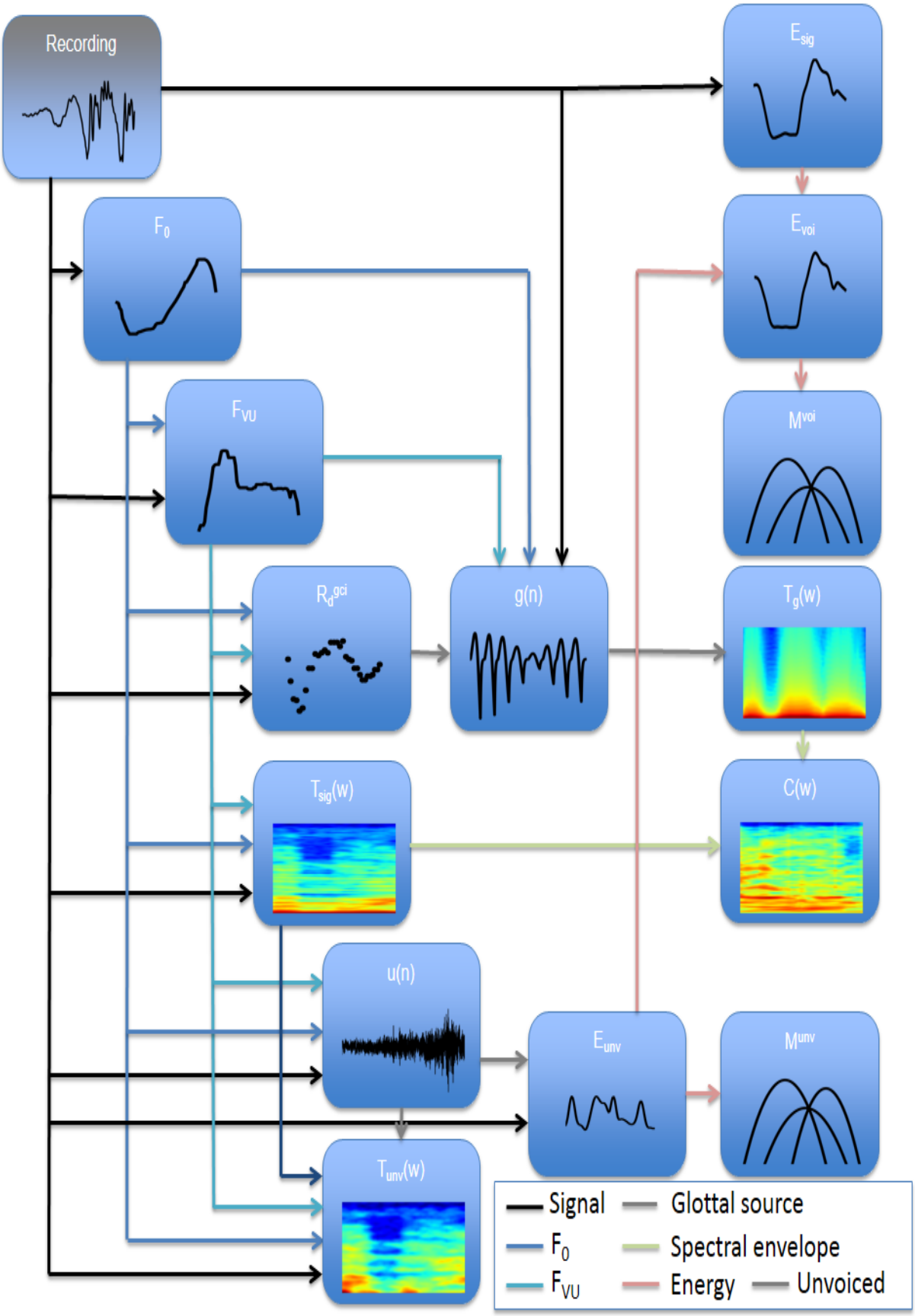


Figure 6.1: *System overview of the analysis stage in PSY*

6.2 Analysis I - The voiced deterministic component

6.2.1 Voice descriptors

6.2.1.1 Fundamental Frequency F_0

A monophonic version of the algorithm to estimate multiple fundamental frequencies proposed in [Yeh et al., 2010] constitutes the basic means to estimate F_0 throughout the work being here presented. It is implemented in IRCAMs basic signal processing system SuperVP, introduced in chapter 2.5.2. SuperVP enables further parameterization and post-processing steps to configure the F_0 estimation process.

First, an estimated F_0 value is only considered valid if the noise level estimation N^{σ_g} of the corresponding analysis frame lies below the noise threshold θ_{sn}. In *PSY*, θ_{sn} is set to a lower value of only 20 dB. This allows a less conservative F_0 estimation with a higher probability of available F_0 estimates within voiced segments. It leads to a lower rate of false negative F_0 estimations in voiced segments. However, more false positive F_0 estimations may occur in unvoiced segments.

Second, stable harmonic sinusoids are inevitably required for a robust analysis of the harmonicity, an attribute representing the degree of acoustic periodicity [Boersma, 1993] which is implicitly contained in the deterministic component $V(\omega)$ of most audio signals. The evaluated frequency range is thus delimited by its upper border L_{high}, set to 1.7 kHz in *PSY*. This diminishes the interference of unstable noisy peaks which increasingly mask the sinusoidal content in higher frequency regions.

Third, a harmonicity score P_{harm} of the F_0 estimate serves as voiced / unvoiced confidence measure. The F_0 estimate is set to zero if the P_{harm} measure lies below the significance level α_{harm}. As with the noise threshold θ_{sn}, the significance level α_{harm} is set as well to a relatively low value of 20 % to lower the rate of false negative F_0 estimations in *PSY*.

Fourth, a median smoothing filter of order $M_s = 8$ is applied to the estimated F_0 contour.

6.2.1.2 Voiced / Unvoiced Frequency F_{VU}

As indicated in chapter 2.3.2, the basic F_{VU} estimation follows in *PSY* the explanations given in [Röbel, 2010c]. Same as with the F_0 estimation, the F_{VU} estimator is available in SuperVP to allow for further configurations. The error tolerance ϵ_{sn} for noise vs. sinusoidal peak classification is set to a relatively high value of 30 % to lower the rate of false negative F_{VU} estimations in *PSY*. A median smoothing filter covering a signal segment of 2.5 analysis windows is applied to the initial F_{VU} estimate.

6.2.1.3 LF shape parameter R_d

The LF regression parameter R_d efficiently parameterizes the LF glottal source model [Fant, 1997, Huber and Röbel, 2013]. It is estimated using the previously established F_0 and F_{VU} contours as additional signal information. The R_d estimation is connected with an additional GCI estimation [Degottex et al., 2010] to restrict the R_d estimation to R_d^{gci}. It describes the R_d contour as R_d^{gci} exclusively on the estimated sequence of glottal closure time instants. The estimation of the R_d^{gci} and R_d contours in *PSY* is explained by the following steps.

Step 1 - MSPD2IX:
The phase minimization variant MSPD2IX, introduced in chapter 5.3.3 and proposed in [Huber et al., 2012], builds the basic means to estimate R_d in *PSY*. According to some informal heuristic tests on the speech recordings presented in the evaluation chapter 6.6, the weight for each phase minimization of MSPD2IX in *PSY* is parameterized as follows: MSPD2I0 = 25 %, MSPD2I1 = 25 %, MSPD2I2 = 50 %. This setting is slightly contrary to the proposal given in [Huber and Röbel, 2013], being presented in this study in section 5.6.3. This considerably huge synthetic test set indicated that a variation of the single weights for MSPD2IX does not alter the estimation results to a statistically significant extent. Therefore, a default equal weighting of MSPD2I0 = 33 %, MSPD2I1 = 33 %, MSPD2I2 = 33 % has been proposed. However, further informal tests on natural human speech suggest, without the possibility of having a ground truth for comparison available, that a weighting scheme in favour of the second integrator MSPD2I2 leads to a more robust R_d estimation.

Step 2 - Viterbi smoothing:
Standard Viterbi smoothing, introduced in chapter 5.4 and proposed in [Huber and Röbel, 2013], is applied on the initial MSPD2IX estimate to suppress non-natural jumps within short-time segments [Vincent et al., 2007, Kane and Gobl, 2013b].

Step 3 - GCI bound:
The initial R_d estimate in *PSY* using SuperVP is bound to GCIs and results first in a R_d^{gci} contour.

Step 4 - Fade R_d^{gci} in/out at voiced segment borders:
One major problem in estimating the deterministic part of the glottal excitation source appears when only few stable harmonic partials can be observed [Drugman et al., 2008, Degottex et al., 2011a]. As explained in chapter 5, such situations may occur at phoneme transitions or at word and speaking pause boundaries, as well as for higher F_0 values [Drugman et al., 2008, Huber et al., 2012].

R_d border hypothesis:
One general expectation on an estimated R_d contour is the hypothesis that the R_d values should comparably increase on word and speaking pause boundaries, and as well as on phoneme transitions. The same hypothesis expects that the R_d contour should generally be comprised of lower R_d values at the more stable parts of a speech phrase, for example in the middle of a vowel. The reason for that hypothesis being that the muscles of the vocal chord should be in a more relaxed state at speaking pause boundaries. Before starting and after ending an utterance, the vocal chord muscles rest inactive. To enter into a strained vibration mode, the muscles have to transition from completely resting via relaxed and modal to finally tense. Please note that this tense or pressed voice quality is relatively independent from vocal effort [d'Alessandro, 2006]. A tense voice quality is described by lower R_d values. A relaxed voice quality is related to higher R_d values.

The tests on natural human speech of the CMU Arctic data set [Kominek and Black, 2004] conducted for the study of [Huber and Röbel, 2013] are explained in chapter 5.6. Informal visual inspections on this test results confirm in general this hypothesis for the comparison basis OQ_{EGG}. However, the OQ_{R_d} contour, derived from the phase minimization based R_d estimates [Huber et al., 2012], appears to sometimes contradict the hypothesis by approaching comparably lower instead of higher R_d values on some speaking pause borders. As outlined in chapter 5.6, the R_d estimation based on phase minimization is more error prone if only a lower number of stable harmonic sinusoids N is observable before being masked by noise. This situation occurs predominantly at speaking pause borders and at transitions between some phoneme pairs.

Therefore, a simple algorithm to fade in/out the R_d^{gci} contour is implemented in *PSY*. The algorithm works according to the following explanations given by the pseudo-code **Algorithm 1**. The algorithm assures that the first three R_d^{gci} estimations found at voiced segment borders do not decrease at segment ends and do not increase at segment starts. The algorithm fades the R_d^{gci} contour in at the start time t_{start} of a voiced segment, and fades R_d^{gci} out at the end time t_{end} of a voiced segment.

Fig. 6.2 illustrates the impact of fading in and out the R_d^{gci} contour at voiced segment borders. The original R_d^{gci} curve is shown by a solid line in cyan colour. The corrected R_d^{gci} curve is shown by point markers in red colour. Examples where the extrapolation resulted in the desired increase of the R_d^{gci} contour can be inspected around the time instants ~0.15 and ~3.30 seconds. Please note that the denomination $R_{d_{gci}}$ given in the fig. 6.2 denotes R_d^{gci}.

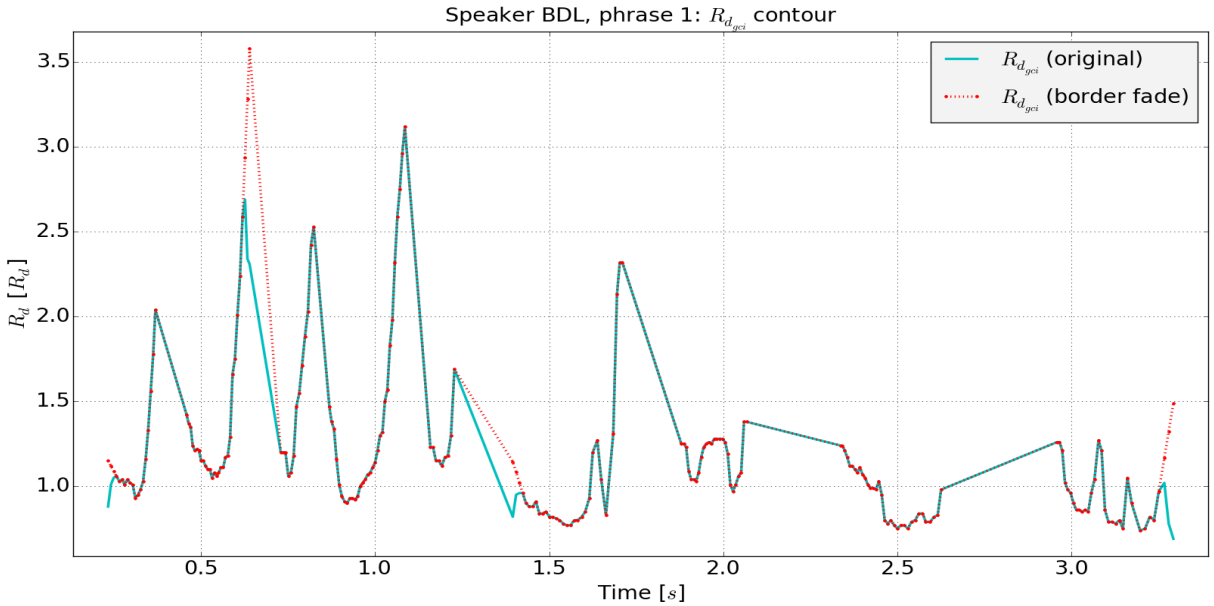


Figure 6.2: *Example of R_d^{gci} fade in/out at voiced segment borders*

Algorithm 1 - Fade R_d^{gci} in/out on voiced segment borders

Compute voicing decision with frequency threshold $F_{min} = 50\ Hz$:

$$\text{voiced} = \begin{cases} \text{True} & \forall \quad (F_0 \wedge F_{VU}) >= F_{min} \\ \text{False} & \forall \quad (F_0 \mid F_{VU}) < F_{min} \end{cases}$$

Delete all R_d^{gci} estimations found in unvoiced segments:

$R_d^{gci}[\sim voiced] = []$

Compute all indices K'_{start} and K'_{end} corresponding to start times T_{start} and end times T_{end} of voiced segments over all STFT time steps K and time instants T:

$T_{start}, K'_{start} = t_k, k \quad \forall$ voiced[k-1] == False $\wedge$ voiced[k] == True

$T_{end}, K'_{end} = t_k, k \quad \forall$ voiced[k-1] == True $\wedge$ voiced[k] == False

for (t_{start}, k_{start}) in (T_{start}, K'_{start}) **do**
  if any($R_d^{gci}[k_{start}$:k_{start}+2]) does not decrease compared to its predecessor from k_{start} to k_{start}+2: **then**
    Extrapolate: $R_d^{gci}[k_{start}$:k_{start}+2] from $f\left(R_d^{gci}[k_{start}\text{+3:end}]\right)$
  end if
  if (still after extrapolation) any($R_d^{gci}[k_{start}$:k_{start}+3]) does not decrease (fallback): **then**
    Saturate: $R_d^{gci}[k_{start}$:k_{start}+2] to $R_d^{gci}[k_{start}$+3]
  end if
end for

for (t_{end}, k_{end}) in (T_{end}, K'_{end}) **do**
  if any($R_d^{gci}[k_{end}$-2:k_{end}]) does not increase compared to its predecessor from k_{end}-2 to k_{end}: **then**
    Extrapolate: $R_d^{gci}[k_{end}$-2:k_{end}] from $f\left(R_d^{gci}[\text{begin:}k_{end}\text{-3}]\right)$
  end if
  if (still after extrapolation) any value in $R_d^{gci}[k_{end}$-2:k_{end}] does not increase (fallback): **then**
    Saturate: $R_d^{gci}[k_{end}$-2:k_{end}] to $R_d^{gci}[k_{end}$-3]
  end if
end for

Step 6 - GCI time correction:
The informal visual inspections mentioned for the previous step 5 indicate as well that not just the R_d^{gci} values estimated at voiced segment borders are more likely to be erroneous. Also the estimated GCI times may be wrong if only few stable harmonic sinusoids N are observable.

A simple algorithm in *PSY* permits the correction of the estimated GCI time instants t_{gci} at voiced segment borders. Each t_{gci} is compared to the time instant t_{minsig} found at the minimum of the original signal $s(n)$ within a time distance of half a fundamental period $t_{search} = t_{gci} \pm 1/(F_0 \cdot 2)$. If the time difference t_δ^{gci} between t_{minsig} and t_{gci} exceeds the GCI time threshold θ_{gci}, t_{gci} is considered as wrongly estimated. A wrong t_{gci} estimate is only replaced by t_{minsig} if the latter is located within the time distance constraint θ_{dist}^{gci}. The constraint imposed by θ_{dist}^{gci} assures that t_{minsig} lies within a relatively similar time distance to the preceding GCI t_{gci-1} and / or the subsequent GCI t_{gci+1}. This guarantees that the quasi-harmonic continuation of the fundamental periods, set by the estimated GCI contour, remains approximative to the F_0 contour.

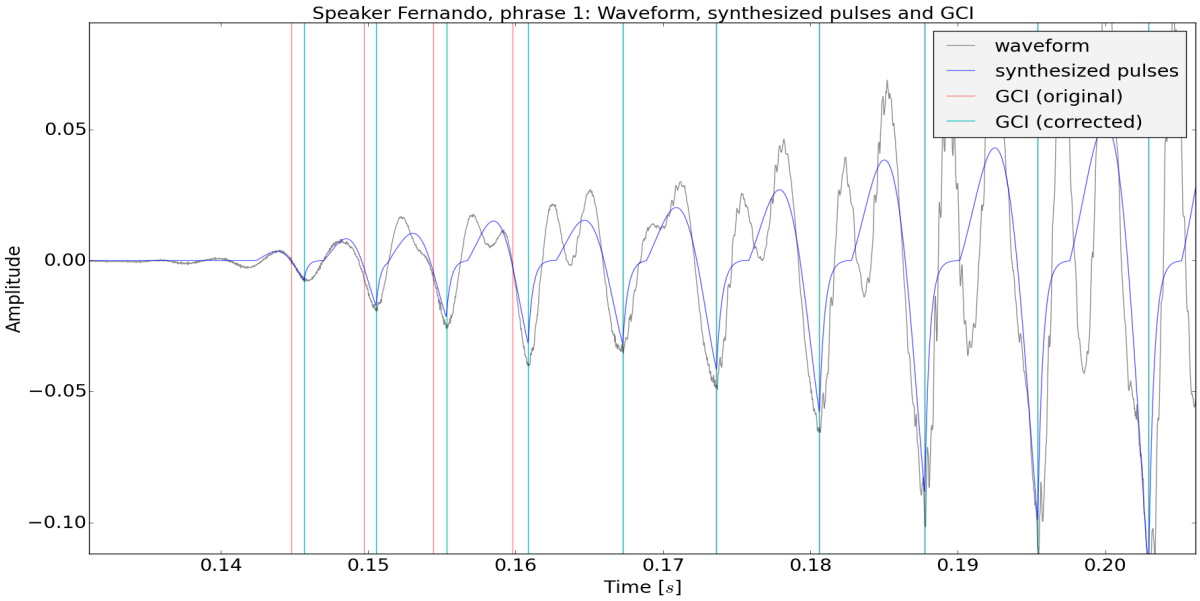


Figure 6.3: *Speaker Fernando - Example of GCI timing correction*

Fig. 6.3 illustrates the correction in time of wrongly estimated GCIs. The original estimated GCI time instants are depicted in red colour. Correctly estimated GCIs are superimposed by the finally utilized GCI time instants shown in cyan colour. Wrongly estimated GCI time instants are consequently visible in red colour. The estimated GCIs in red colour are clearly set in a wrong manner to time instants where the speech waveform depicted in grey colour does not contain a local minimum. The GCI time correction algorithm identifies such GCI misplacements. It searches for a local minimum within a quasi-harmonic distance to the preceding GCI and relocates the wrong GCI in red to the final GCI in cyan colour.

Please note that this algorithmic step is by default de-activated in *PSY*. It is activated for experimental usage only if erroneous GCI estimation lead to obvious sound artefacts on voiced borders. The main reason being that the simple minimum t_{minsig} found within a short-time segment may not correspond to the true GCI position. The true GCI time position corresponds to the time instant t_e of the glottal flow derivative. It is hidden in the speech signal since the glottal flow derivative is convolved with noise turbulences and the VTF. This leads to a possible time shift between the local minimum of the time waveform of the signal and the true GCI. A GCI estimation can therefore not be conducted by the straight-forward determination of local minima in the time domain signal. Different approaches to estimate the true GCI position are presented in section 3.4.

Step 7 - Time basis interpolation:
The R_d^{gci} estimate, established with the preceding steps 1 - 6, is used in *PSY* to synthesize a sequence $g(n)$ of glottal pulse derivatives $g_{R_d}^{gci}$ in the time domain at each corresponding GCI. It is required to estimate its spectral envelope sequence for the VTF extraction presented in chapter 6.2.2.2. Additionally it constitutes the basic voice descriptor to conduct a voice quality transformation as introduced in chapter 6.4.3.

Still, the R_d^{gci} contour does only provide information about the estimated R_d values per estimated GCI time instants T_{gci}. However, the energy modelling and the GMM-based parameter prediction of *PSY*, introduced in chapter 6.4.1 and respectively 6.4.2, require the R_d estimation to be available at the same STFT analysis/synthesis time steps as

with the F_0 and F_{VU} estimations. The R_d^{gci} curve is thus interpolated to the R_d curve on the STFT time grid.

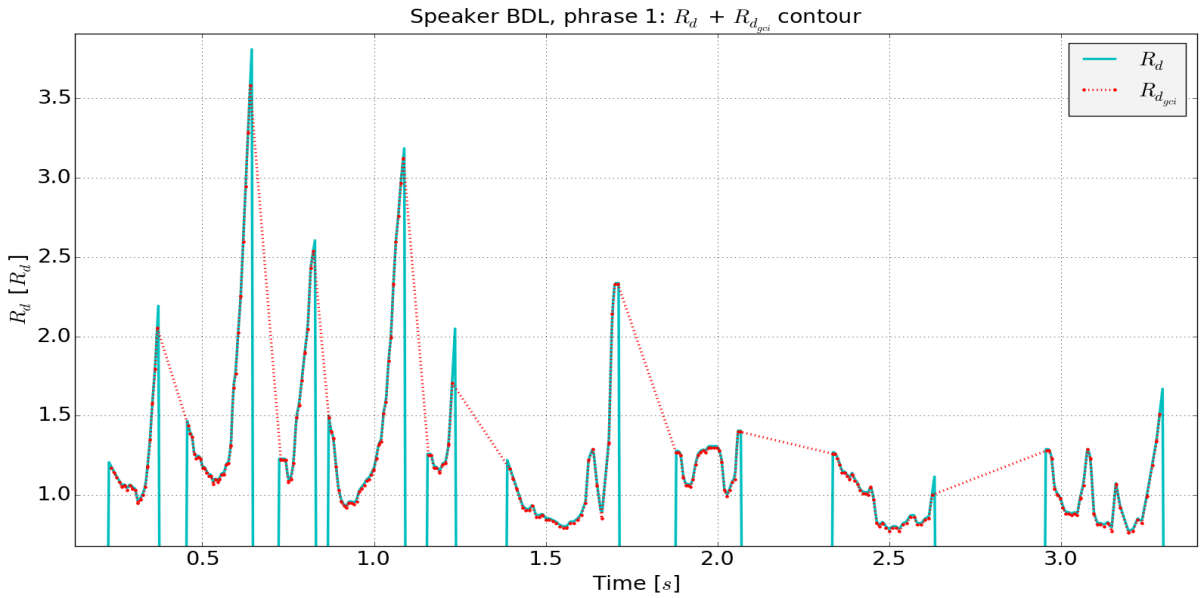


Figure 6.4: *Speaker BDL - Example of an interpolated R_d and its original R_d^{gci} contour*

Figure 6.4 shows the interpolated R_d and the R_d^{gci} contour for the first phrase of CMU Arctic speaker BDL. Please note that the denomination $R_{d_{gci}}$ given in the figure denotes R_d^{gci}.

6.2.2 Spectral envelopes

6.2.2.1 Spectral envelope of the input signal

A spectral envelope sequence $\mathcal{T}_{sig}$ is estimated on the input signal $s(n)$ using the True Envelope (TE) estimator described in chapter 2.6.4 and in [Villavicencio et al., 2006, Röbel et al., 2007]. The TE local order selection is adapted per STFT frame to an optimal TE order TE_{opt} given the previously estimated F_0 information. The automatic determination of the optimal TE order TE_{opt} can further be influenced in SuperVP by defining an additional scaling factor TE_{scale} to account for different audio signal types:

$$TE_{opt} = F_s/F_0 \cdot 0.5 \cdot TE_{scale}. \tag{6.5}$$

F_s is the sampling rate of the recording. The TE order can be reduced for comparably smooth spectral peak sequences by setting $TE_{scale} < 1.0$. A very detailed approximation of the spectral peak sequence is achieved with TE_{scale} values > 1.0. In *PSY*, the default TE_{scale} factor is set empirically to a relatively high order of 1.5. Additionally, the TE convergence criterion θ_{conv}^{TE} is set to a very small value of 0.01 dB in *PSY*. The iterative spectral envelope estimation with TE stops per default if a θ_{conv}^{TE} of 2 dB is reached. Lower values approximate a peak-to-peak like spectral envelope while higher values average over the spectral peaks. Both a high TE_{scale} and a low θ_{conv}^{TE} value allow to capture more specific spectral information such as narrow formant structures.

6.2.2.2 Spectral envelope of the glottal excitation source

The SVLN method of section 3.8.4 synthesizes per synthesis step m_k of the STFT one glottal pulse $G_{R_d}(\omega)$ in the spectral domain. SVLN determines by interpolation for each synthesis time instant the corresponding R_d value from any R_d or R_d^{gci} input contour. It does not matter if the given input sequence consists of R_d^{gci} values on a GCI time basis, or a sequence of R_d values on any analysis time grid [Degottex et al., 2013]. This technical issue is approached in a different manner in *PSY*.

Synthesis of glottal pulses $g_{R_d}^{gci}$ and the glottal pulse sequence $g(n)$:

Each differentiated glottal flow pulse $g_{R_d}^{gci}$ is synthesized at its corresponding glottal closure time instant t_{gci}. The LF model is parameterized by the corresponding R_d^{gci} and F_0 estimates, synthesized as $G_{R_d}^{gci}(\omega)$ in the spectral domain, convolved with the radiation filter $R(\omega)$, and transformed by the inverse Fourier transform to $g_{R_d}^{gci}$ in the

time domain. The LF amplitude parameter E_0 used to scale the LF model in amplitude at synthesis is set to 1.0. A precise explanation of the energy modelling in *PSY* is given in chapter 6.4.1. Here, each synthesized single glottal pulse $g_{R_d}^{gci}$ is scaled according to a linear Root Mean Square (RMS) energy measure. The local RMS energy E_{sig} of the original waveform $s(n)$ is measured with the operator F_{RMS} on a signal segment being additionally windowed with a Hanning window $w_h(n)$. The evaluated segment has a length of two fundamental periods $t_{seg} = 2 \cdot T_0 = 2/F_0$, with one period before and one period after the GCI t_{gci}.

$$E_{sig} = F_{RMS}(w_h(s[t_{gci} - 1/F_0 : t_{gci} + 1/F_0])) \tag{6.6}$$

$$E_{g_{R_d}} = F_{RMS}(g_{R_d}^{gci}) \tag{6.7}$$

$$g_{R_d}^{gci} * = E_{sig}/E_{g_{R_d}} \tag{6.8}$$

$$F_{RMS}(x,k) = \sqrt{1/K \cdot \Sigma^K (x(k)^2)} \tag{6.9}$$

$$w_h(k) = 0.5 \cdot (1 - \cos(2 \cdot \pi \cdot k/K)) \tag{6.10}$$

The glottal flow derivative sequence $g(n)$ constitutes the deterministic part of the glottal excitation source. It is constructed iteratively by overlap-adding each synthesized and energy scaled glottal pulse $g_{R_d}^{gci}$ to the signal $g(n)$. One example of a synthesized pulse sequence $g(n)$ is depicted in fig. 6.5 in blue solid line. The horizontal solid

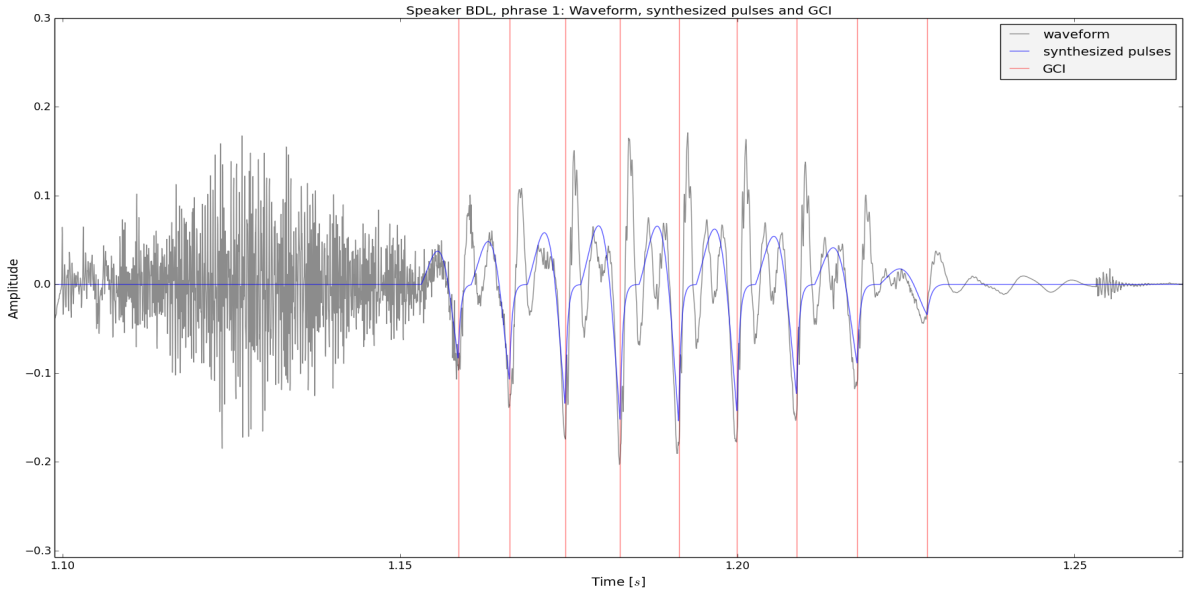


Figure 6.5: *Speaker BDL - Example of a synthesized glottal pulse sequence*

lines in red colour show the GCI time instants t_e of maximum formant excitation per pulse [d'Alessandro, 2006].

Estimation of the glottal source spectral envelope $\mathcal{T}_g$:
The TE spectral envelope $\mathcal{T}_g$ is estimated on the synthesized glottal flow derivative sequence $g_s(n) = \sum_i g(n, P_i) * \delta(n - P_i)$. Fig. 6.6 depicts the spectrum $G_s(\omega)$ of a windowed glottal pulse sequence $g_s(n)$ and its corresponding TE spectral envelope estimate $\mathcal{T}_g(\omega)$ for one frame of CMU Arctic speaker BDL. Please note that the shown R_d value of 1.04 does not reflect the spectral representation of the LF model being synthesized by the same R_d value. The shown glottal pulse spectrum is the STFT spectrum of the windowed glottal pulse sequence $g_s(n)$ synthesized in the time domain. $G_s(\omega)$ is thus influenced by the shape of the neighbouring glottal pulses.

Two ripple types exist in the spectral representation of $G_s(\omega)$. An example of both is given in fig. 6.7 with the sinusoidal content of a glottal pulse sequence in red and its corresponding spectral envelope in cyan colour.
a) Sinusoidal content (red colour):
A higher modulation in amplitude and a faster modulation in frequency leads to smaller and more densely sampled ripples. These ripples result from the sequence of quasi-harmonic sinusoids in the spectrum, convolved with the Hanning window $w_h(n)$ of the STFT. They origin from the quasi-periodic pulse sequence of the harmonic structure which generates the fundamental periodicity.
b) Spectral envelope (cyan colour):
A lower modulation in amplitude and a slower modulation in frequency leads to bigger and more widely sampled ripples. These ripples are the result of higher R_d values for which the LF model produces bigger curvatures

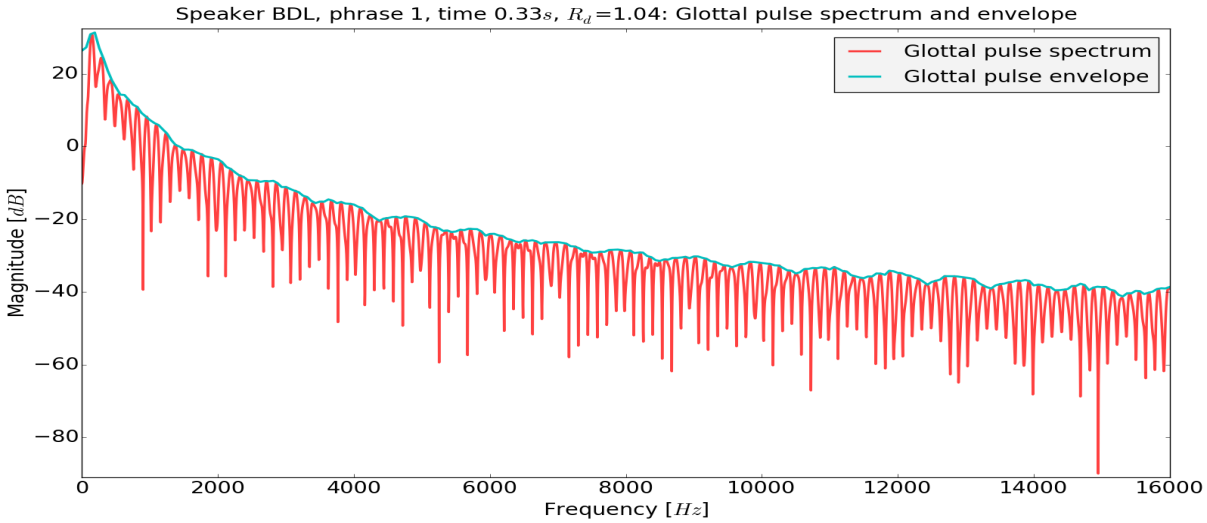


Figure 6.6: *Speaker BDL - Glottal pulse spectrum $G_s(\omega)$ and spectral envelope $\mathcal{T}_g(\omega)$ example*

[Fant, 1995]. The LF ripple effect increases with an increasing return phase t_a and an decreasing amplitude ratio A_e=E_e/E_i [Fant et al., 1985]. This effect is according to [Doval et al., 2003] caused by the truncation of the impulse response. It induces regularly spaced zeroes and ripples in the spectrum. The truncation originates from windowing the return phase signal being minimum phase. A truncated minimum phase filter impulse response may interfere the minimum or mixed phase characteristic [Bozkurt, 2005].

The mentioned LF ripples are not the type of ripple manifesting themselves as oscillations in the glottal source contour while the open phase [Fant and Ananthapadmanabha, 1982]. Peaks and valleys present while the open phase originate from formant oscillations which are induced by the interaction of the glottal excitation source with the VTF, as observed in [Fant and Lin, 1987, Childers and Wong, 1994, Båvegård and Fant, 1994, Titze, 2004, Titze et al., 2008, Zañartu et al., 2013].

The bigger ripples (lower amplitude modulation, slower frequency modulation) origin from a synthetic signal in the example shown in fig. 6.6. This type of ripples does not originate from a glottal source signal extracted by inverse filtering recordings of natural human speech [Fant and Lin, 1987]. The curvature pattern of the LF ripples depends not just on the R_d value but as well on the fundamental frequency F_0. Fig. 6.7 exemplifies the spectral contour of LF pulses for the two R_d values 0.5 and 3.0, each synthesized with the three F_0 values 80 Hz, 160 Hz, and 240 Hz. For R_d=0.5, almost no spectral ripples appear. The three LF pulses describe an approximately continuous spectral wave. Only for F_0=80 Hz a very subtle ripple can be observed. For R_d=3.0 however, all three LF pulses are interfered by a different spectral ripple structure. The latter being determined by F_0. Smaller F_0 values lead to finer ripples with a lower bandwidth. Higher F_0 values generate a bigger ripple curvature pattern. Fig. 6.8 illustrates the impact of the LF ripple pattern over the complete R_d range for two different frequencies. The LF spectra of the lower frequency F_0=80 Hz exhibits only minor undulations, predominantly present in higher frequency regions for higher R_d values in the upper R_d range R_d >2.7. The example with the higher frequency F_0=240 Hz depicts the interrelation of the smaller and bigger spectral ripples, being determined by F_0 and respectively by R_d. Fig. 6.9 illustrates the spectrograms of the synthesized glottal pulse derivative sequence $g_s(n)$ and the TE spectral envelope $\mathcal{T}_g$ for the complete phrase number one of CMU Arctic speaker BDL [Kominek and Black, 2004, Huber and Röbel, 2013].

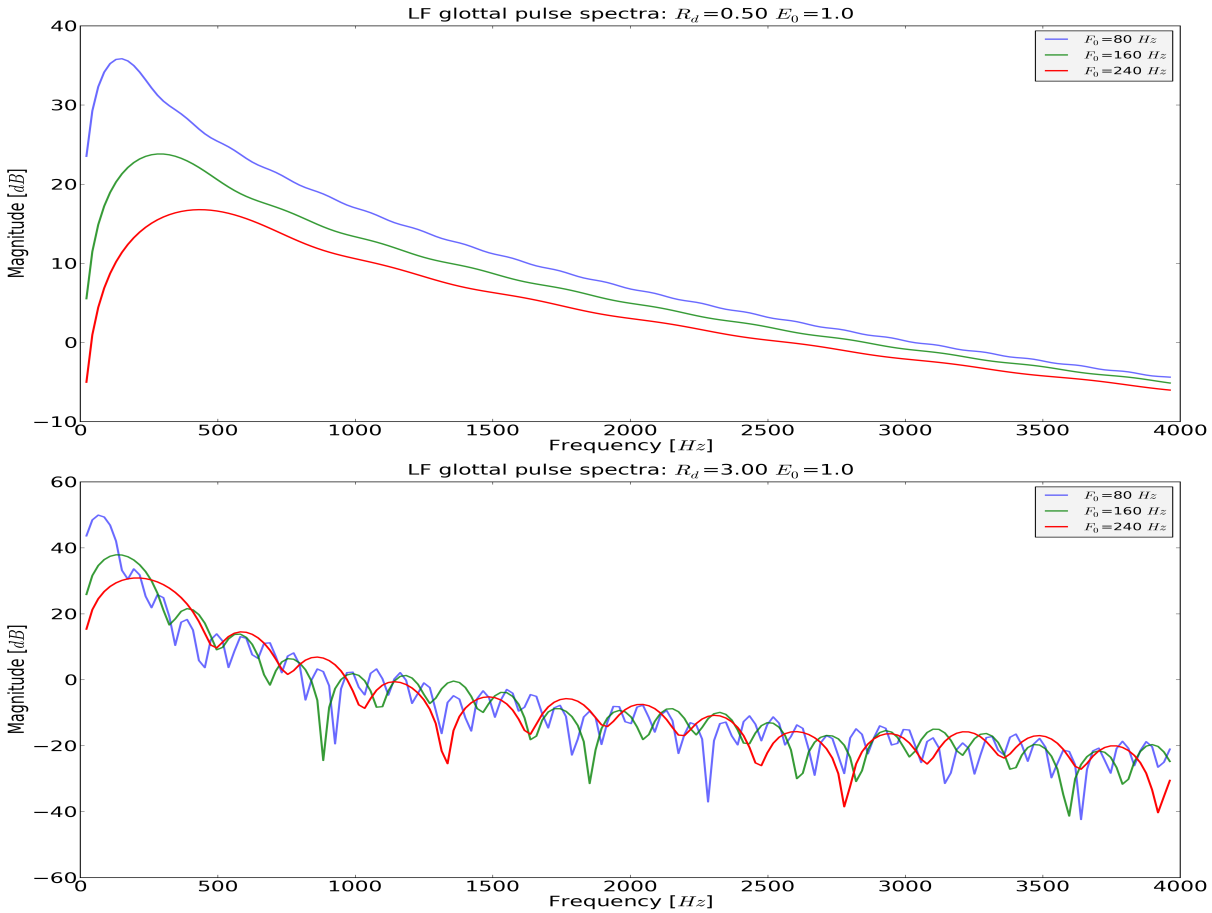


Figure 6.7: *Spectrum of three synthetic glottal pulses for different F_0 and R_d values*

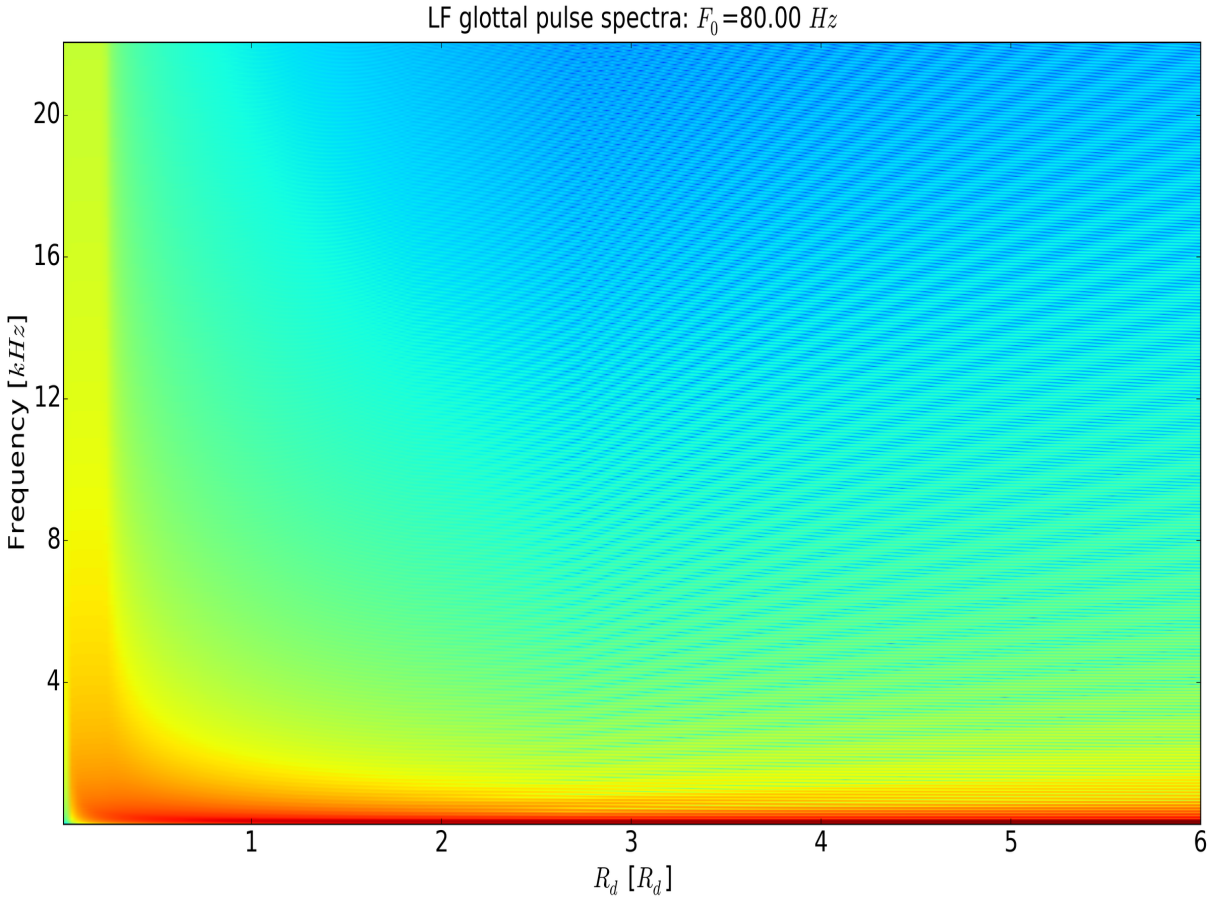
6.2.3 Vocal Tract Filter extraction

Two different versions of the Vocal Tract Filter $C(\omega)$ are implemented in *PSY*. The $C_{F_{VU}}(\omega)$, introduced in the following chapter 6.2.3.2, implies spectral splitting at the F_{VU}. It follows the assumptions introduced in chapter 2.3.2. The $C_{full}(\omega)$, introduced in chapter 6.2.3.4, does not reflect the splitting of the spectrum at the F_{VU}. If follows the signal interpretation outlined in chapter 6.2.3.3. Please note that the discussion presented in the following is restricted to voiced segments only. Such segments must contain valid F_0, F_{VU} and R_d^{gci} estimates.

6.2.3.1 Suppression of spectral ripples

SVLN divides the input signal $s(n)$ not by spectral envelope estimations of glottal pulses but by their direct signal representations. The latter is described below F_{VU} by a synthesized glottal pulse $G^{R_d}(\omega)$. Above F_{VU} it is described by a constant value of the energy level found at $|G^{R_d}(F_{VU})|$, in combination with the radiation filter $R(\omega)$. Defined in equ. 3.9, the cepstral spectral envelope $\mathcal{T}$ is estimated on the deterministic, and the real cepstrum envelope $\mathcal{P}$ on the stochastic part of the division. According to the discussion presented in section 6.2.2.2 such an approach is prone to produce ripples in the created VTF as a result of the LF ripples whose curvature pattern is determined by F_0 and R_d. Therefore, one should first estimate the spectral envelope of the implied signal components and then apply the division. Fig. 6.10 shows an example in which the glottal pulse signal $G_s(\omega)$, and not its corresponding spectral envelope $\mathcal{T}_g$, divides the spectral envelope $\mathcal{T}_{sig}$ of the input signal $s(n)$. Please note that the denomination $G_{R_{d_{gci}}}(\omega)$ given in the figure denotes $G_s(\omega)$.

As explained in the preceding chapter 6.2.2.2, smaller ripples from the spectral representation $\Delta(\omega)$ of the Dirac impulse sequence and bigger ripples generated with higher R_d values are present in the glottal excitation source $G_s(\omega)$. The $S(\omega)$ contains additionally the contribution of the VTF. As can be clearly visually inspected from fig. 6.10, the result of dividing $S(\omega)$ by $G_s(\omega)$, shown in solid line with magenta colour, results into a spectral curve not properly describing the intended VTF contour. Especially the smaller ripples of both signals used in the division mask the true VTF contour. Estimating after such division a spectral envelope $\mathcal{T}$ on the result does not properly reflect the formant structure described by the VTF. The estimation of the spectral envelope sequences $\mathcal{T}_g$ on the glottal pulse sequence $g(n)$ and $\mathcal{T}_{sig}$ on the input signal $s(n)$ is hence necessary. It suppresses at least for the influence of the smaller ripples present in the amplitude spectrum of $G_s(\omega)$. The impact is exemplified in fig. 6.11.



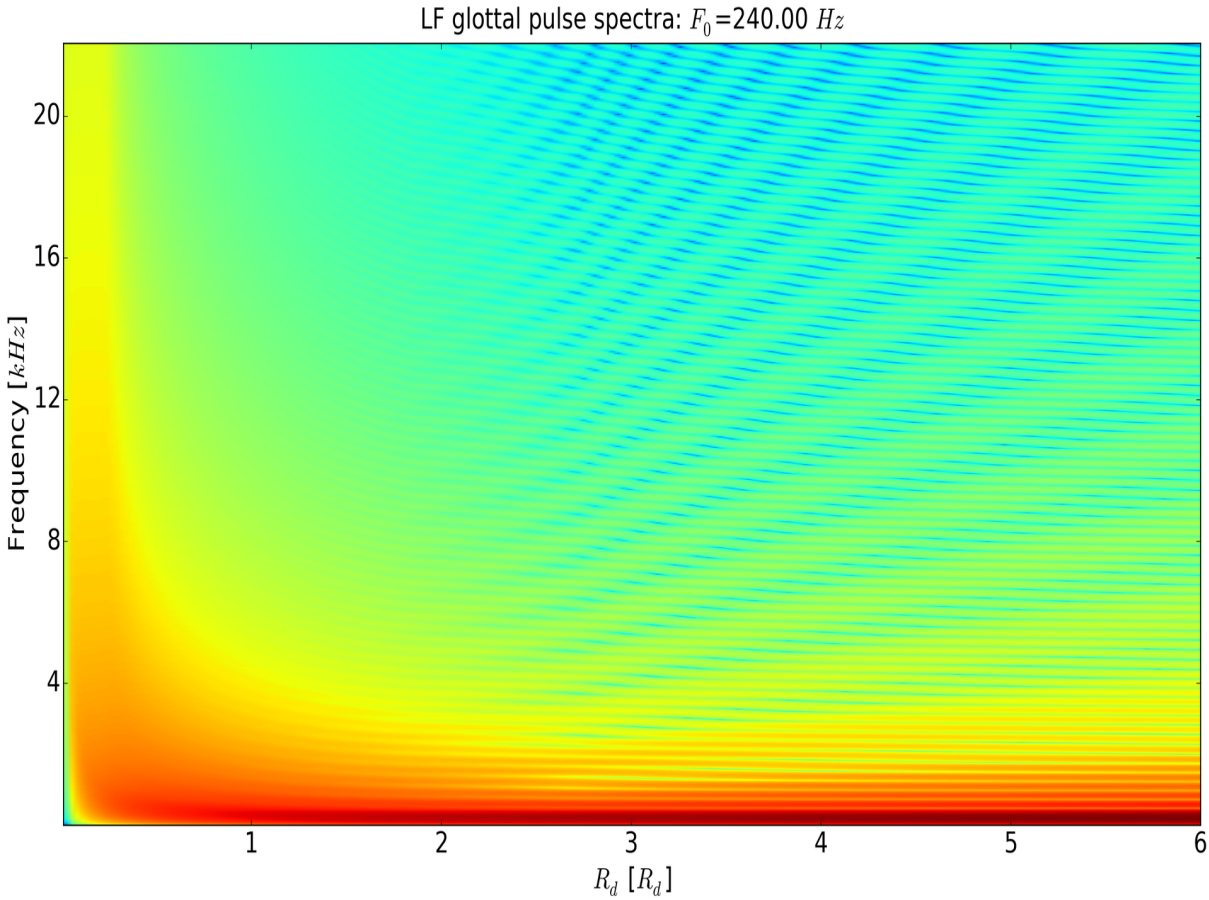


Figure 6.8: *Spectrogram of glottal LF pulses for two F_0 values over complete R_d range*

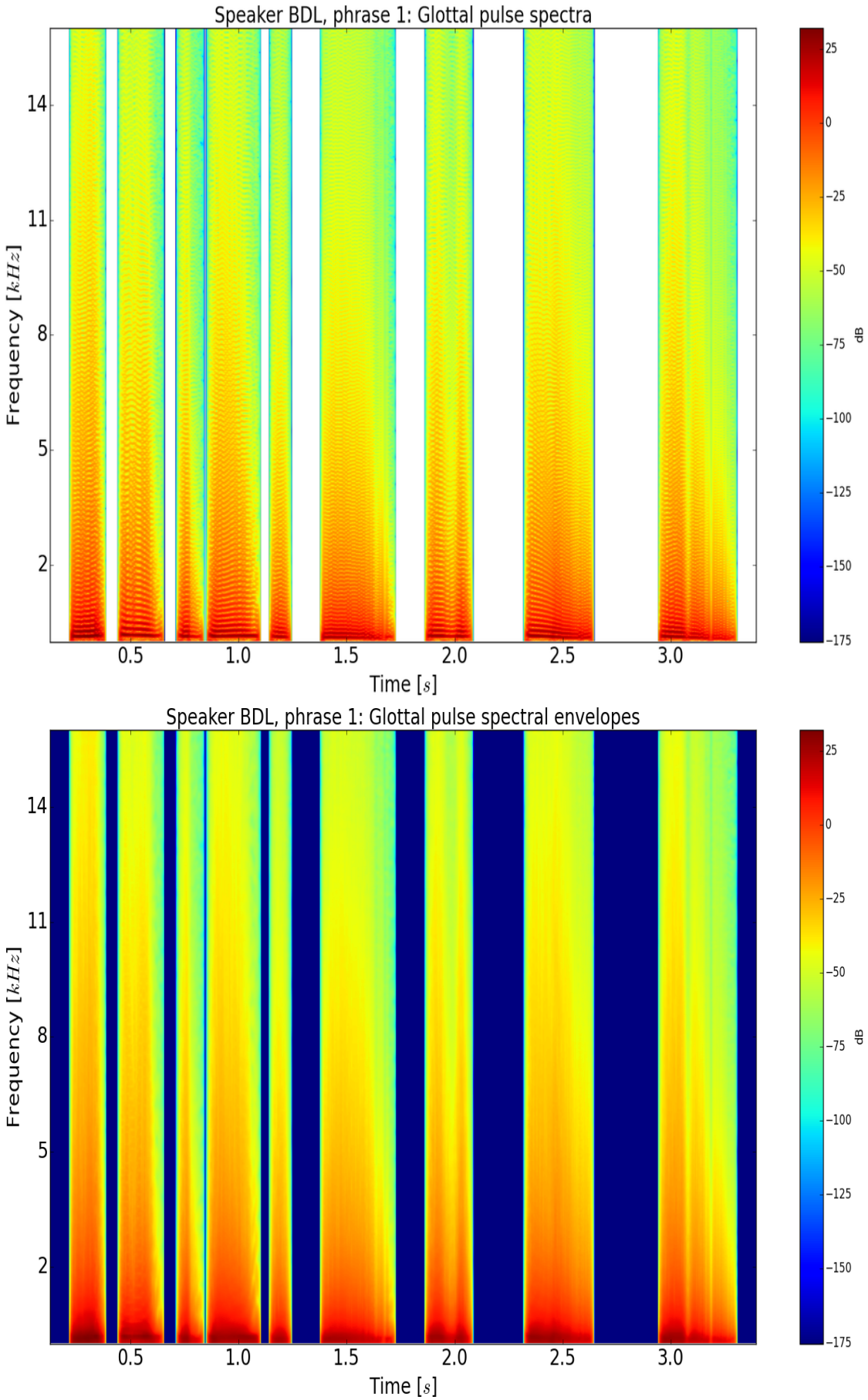


Figure 6.9: *Speaker BDL - Spectra of glottal pulse $g_s(n)$ and envelope $\mathcal{T}_g$ sequence*

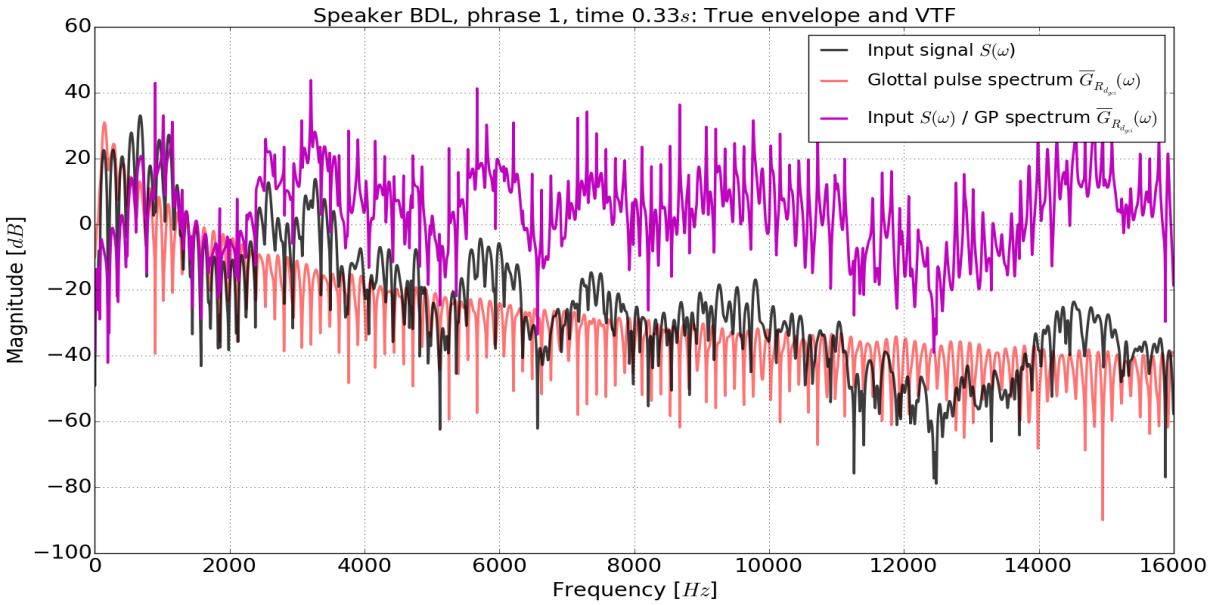


Figure 6.10: *Speaker BDL - Example of dividing $S(\omega)$ by glottal pulse $G_s(\omega)$, not $\mathcal{T}_{sig}$ by $\mathcal{T}_g$*

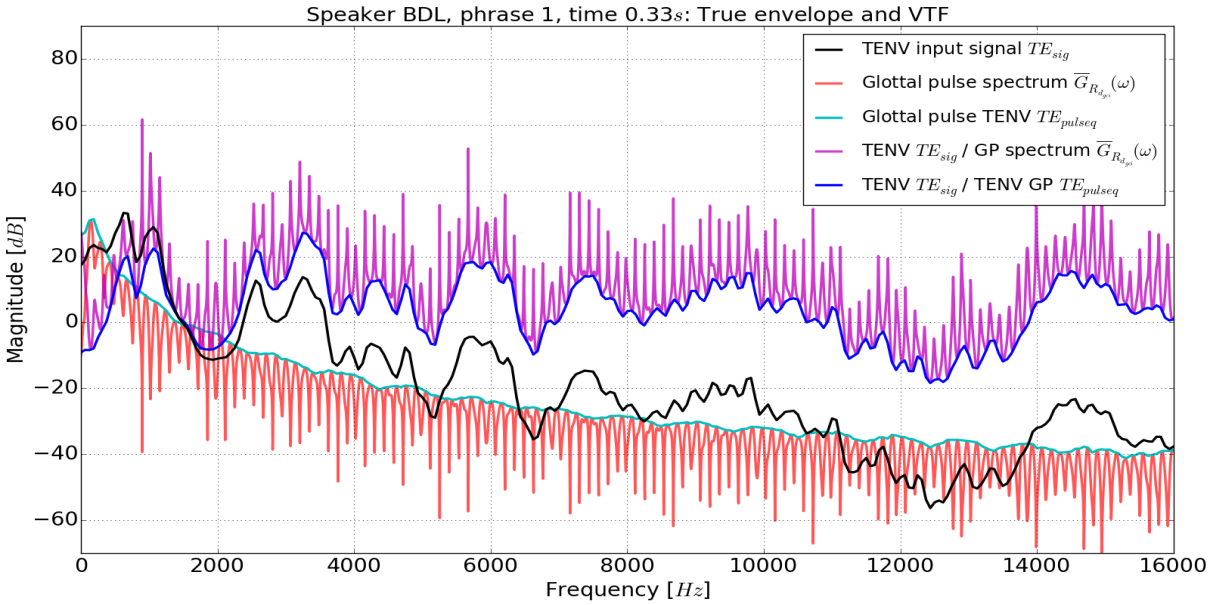


Figure 6.11: *Speaker BDL - Example of dividing $\mathcal{T}_{sig}$ by $|G_s(\omega)|$, not by $\mathcal{T}_g$*

Dividing $\mathcal{T}_{sig}$ by $G_s(\omega)$ transfers the smaller ripples to the VTF contour, shown in solid line with magenta colour. This approach mirrors the local minima of $G_s(\omega)$ on $\mathcal{T}_{sig}$. This results into local maxima at the opposite direction of $\mathcal{T}_{sig}$.

The proper manner to extract the VTF is by applying the spectral division on the estimated spectral envelopes of the glottal source and the speech signal. The solid line in blue colour of fig. 6.11 depicts an example of the VTF $C_{full}(\omega)$ which will be introduced in section 6.2.3.4. The spectral envelope $\mathcal{T}_{sig}$ is divided by the spectral envelope $\mathcal{T}_g$ over the complete spectrum. In *PSY*, both $C(\omega)$ versions are based on the division of the relevant spectral envelopes. No direct signal representation is involved. Please note that the denomination $\overline{G}_{R_{d_{gci}}}(\omega)$ given in fig. 6.11 denotes $G_s(\omega)$.

Please note that in SVLN, these smaller ripples originating from the quasi-harmonic sinusoids are not present. SVLN synthesizes per analysis and synthesis frame the LF model directly in the spectral domain, without employing a STFT. Therefore, only the bigger ripples for higher R_d values are present. However, SVLN still employs the input signal $s(n)$ in the spectral division. The VTF contour estimated by SVLN is not masked by the mentioned smaller ripples since $S(\omega)$ is used as dividend and divided by the glottal pulse being directly synthesized with the LF model. Therefore, only the LF ripples present for higher R_d values are transferred to the VTF estimation in SVLN. Please note that the original implementation of SVLN employed only the normal R_d range of [Fant, 1995, Fant, 1997] and not the extended one of [Huber et al., 2012, Huber and Röbel, 2013]. Thus only the LF ripples occurring below R_d<2.7 are interfering the VTF estimation of the SVLN version used in [Degottex et al., 2013].

6.2.3.2 $C_{F_{VU}}(\omega)$ - Split at F_{VU}

The $C_{F_{VU}}(\omega)$ in *PSY* follows the signal interpretation of SVLN to split the spectrum at the F_{VU}. The SVLN approach is introduced in section 3.8.4.2. The F_{VU} and the splitting of the spectrum into two bands is discussed in section 2.3.2. Equation 3.9 defines the VTF $C(\omega)$ of SVLN. It is based on the fact that at the Voiced / Unvoiced Frequency boundary F_{VU}, the spectrum is divided into a deterministic frequency band $\omega < \omega_{VU}$ below F_{VU}, and into a stochastic frequency band $\omega > \omega_{VU}$ above F_{VU}. The F_{VU} frequency is interpreted as angular frequency ω_{VU}. An example of both frequency bands is given in fig. 2.2. The idea behind the splitting theory is that
a) the deterministic and the stochastic signal part describe two different signal types and thus two different perceptual sensations, and that
b) the one signal part with higher energy perceptually masks in its frequency band the other signal part having lower energy,
with a) requesting and b) confirming a separate processing of both parts.

In *PSY*, the VTF $C_{F_{VU}}(\omega)$, being similar to the implementation of $C(\omega)$ in SVLN, is constructed by connecting the contributions of the deterministic band $\omega < \omega_{VU}$ in lower frequency regions with the stochastic frequency band $\omega > \omega_{VU}$ in higher frequency regions. In the deterministic band, the spectral envelope $\mathcal{T}_{sig}(\omega)$ of the signal $s(n)$ is divided per STFT frame by the spectral envelope $\mathcal{T}_g(\omega)$ of the glottal pulse sequence $g_s(n)$. In the stochastic band, the division is not applied. Instead, the spectral envelope $\mathcal{T}_{sig}$ is directly adopted in the higher frequency band. The linear operation $C_{F_{VU}}(F_{VU})$ - $\mathcal{T}_{sig}(F_{VU})$ accounts for the magnitude difference between the $C_{F_{VU}}$ and the spectral envelope $\mathcal{T}_{sig}$ of the signal found after the spectral division at the F_{VU}. The difference compensation assures a smooth magnitude continuation of $C_{F_{VU}}$ between the deterministic $\omega < \omega_{VU}$ and the stochastic $\omega > \omega_{VU}$ frequency band. An illustration of the difference compensation can be found in fig. 6.13 at the F_{VU}. Equation 6.11 defines mathematically the construction of the VTF $C_{F_{VU}}$ in *PSY*.

$$C_{F_{VU}}(\omega) = \begin{cases} \mathcal{T}_{sig}(\omega) \,/\, \mathcal{T}_g(\omega) & \forall\ \omega \leq \omega_{VU} \\ \mathcal{T}_{sig}(\omega) + C_{F_{VU}}(\omega_{VU}) - \mathcal{T}_{sig}(\omega_{VU}) & \forall\ \omega > \omega_{VU} \end{cases} \tag{6.11}$$

Fig. 6.12 illustrates the spectrogram of the constructed Vocal Tract Filter $C_{F_{VU}}$ for one whole phrase of speaker BDL. The F_{VU} contour is shown as blue solid line plotted on top of the spectrogram. Using a higher zoom level it can be inspected that on time positions where the F_{VU} estimate underlies jumps or is erroneous. The established $C_{F_{VU}}$ surface exhibits as well huge changes within short-time segments, e.g. at ~1.95, ~3.2, and ~3.3 seconds. Please note that the denomination V_{VUF} given in fig. 6.12 denotes $C_{F_{VU}}(\omega)$.

6.2.3.3 Drawbacks related to utilizing the F_{VU} boundary

In theory, the different handling of the spectral envelope $\mathcal{T}_{sig}$ at the F_{VU} to construct the VTFs $C(\omega)$ in SVLN or $C_{F_{VU}}(\omega)$ in *PSY* may be theoretically a good idea. However, in practice, the splitting of the spectrum into two

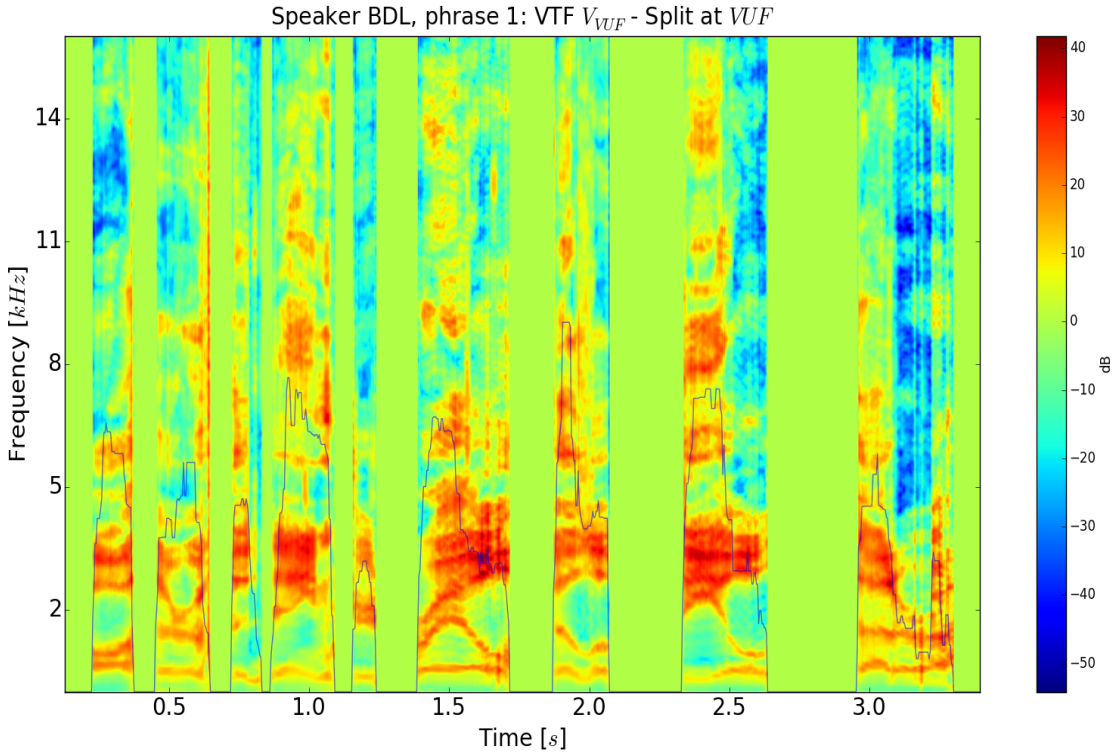


Figure 6.12: *Speaker BDL - VTF example $C_{F_{VU}}(\omega)$ reflecting the spectral split at the F_{VU}*

frequency bands, its distinct signal treatment, and the connection of the two differently processed frequency bands depends highly on the estimated F_{VU} contour.

Inaccurate F_{VU} estimate:
Erroneous F_{VU} estimations introduce misinterpretations of deterministic $V(\omega)$ or stochastic $U(\omega)$ signal types into the signal processing framework. Stable sinusoidal content is treated as a spurious noisy signal, and vice versa. High jumps of the F_{VU} contour in short-time segments may as well lead to difficulties in establishing a smooth and robust VTF contour over time.

Frequency dependent noise and signal level:
The deterministic signal content is described by quasi-harmonic sinusoids. The stochastic signal content is described by random noise. Both are generated by a convolution of the glottal excitation source $G(\omega)$ with the VTF $C(\omega)$. The glottal excitation source $G(\omega)$ consists of the deterministic sinusoidal-like and the stochastic noise-like parts. Both are filtered by the VTF while passing through the vocal tract. The filter $C(\omega)$ is physically determined by the pharyngeal, oral and nasal cavities of the vocal tract. In LTI system terms, the filter $C(\omega)$ is described by bandwidth, frequency and amplitude of all formants V_F comprising the vocal tract. The formants V_F characterize the physical resonances of the filter caused by the vocal tract cavities. The signal levels of both the deterministic $V(\omega)$ and the stochastic $U(\omega)$ component are thus
a) varying over frequency F at one quasi-stationary STFT frame k, with the formants V_F assumed to be fixed, and
b) varying over time t and consecutive STFT frames k, with the formants V_F assumed to change. The convolution of the vocal tract formants V_F with the deterministic and stochastic part of the glottal excitation source $G(\omega)$ introduces thus a energy variation over time t and frequency F for both signal parts.

Existence of several F_{VU} boundaries:
Due to this interference, one spectral frame $S(\omega)$ of the signal $s(n)$ is likely to contain several Voiced / Unvoiced Frequency boundaries F_{VU}. After the deterministic component $V(\omega)$ is masked by the stochastic component $U(\omega)$ at a first F_{VU} frequency, the deterministic part may again rise in energy above the noise energy level [Chan and Hui, 1996]. With this a second F_{VU} is found at the frequency where the deterministic component descends again into noise. The variation in energy level of both signal parts may originate several F_{VU} boundaries. This may lead to local instabilities of the F_{VU} estimate over consecutive signal frames. The F_{VU} estimator utilized in this work, implemented in SuperVP from section 2.5.2, is configured to determine the highest narrow frequency band having sinusoidal content as F_{VU} boundary. If the F_{VU} estimator has to select one out of several possible F_{VU} boundaries over time, the estimation is prone to jumps. If several F_{VU} boundaries coexist over time and the formant causing the highest F_{VU} diminishes into noise, the F_{VU} estimate will have to fallback to the second highest F_{VU}. This exemplifies one unavoidable cause of the mentioned jumps in the F_{VU} contour.

Due to the possibility of such local instabilities, the initial F_{VU} estimate is smoothed in *PSY* with a median filter covering 2.5 analysis windows, as indicated in section 6.2.1.2. In section 6.6 it will be shown with the tests on voice quality transformation that SVLN requires even more smoothing.

6.2.3.4 $C_{full}(\omega)$ - Full-band glottal source effect

The VTF $C_{full}(\omega)$ in *PSY* does not follow the signal interpretation to split the spectrum at the F_{VU}. It applies the simple division over the whole spectrum of the spectral envelope $\mathcal{T}_{sig}$ of the signal by the spectral envelope of the deterministic part of the glottal excitation source $\mathcal{T}_g$, defined in equ. 6.12:

$$C_{full}(\omega) = \mathcal{T}_{sig}(\omega) \,/\, \mathcal{T}_g(\omega) \quad \forall \quad 0 \leq \omega \leq \omega_{F_s/2}\,. \tag{6.12}$$

With this it is independent to the drawbacks of the F_{VU} estimate, outlined in the preceding section 6.2.3.3.

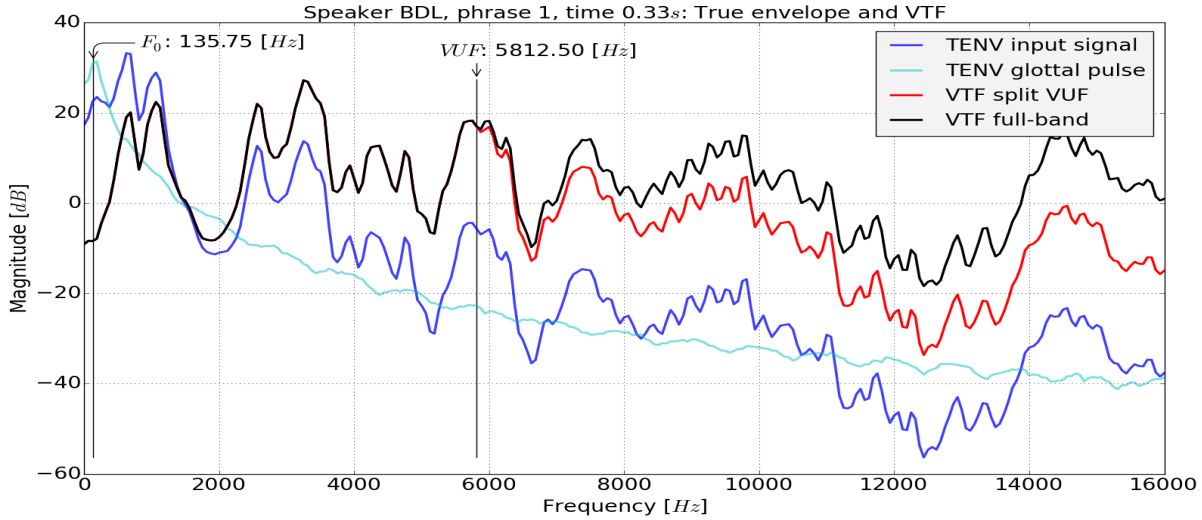


Figure 6.13: *Speaker BDL - Spectral division example to extract full-band $C_{full}(\omega)$ and split $C_{F_{VU}}(\omega)$*

Fig. 6.13 illustrates the creation of the full-band $C_{full}(\omega)$ and the $C_{F_{VU}}(\omega)$ with the split at the F_{VU}. The full-band $C_{full}(\omega)$ is described by higher magnitude values above the F_{VU} compared to the split version $C_{F_{VU}}(\omega)$. The latter applies the energy difference of $\mathcal{T}_{sig}(F_{VU})$ - $C_{F_{VU}}(F_{VU})$ on the stochastic band $\omega > \omega_{VU}$. The energy correction is defined in equ. 6.11. $C_{F_{VU}}(\omega)$ contains the spectral envelope $\mathcal{T}_{sig}(F_{VU})$ described by the stochastic part in the stochastic frequency band. It assumes that the contribution of the deterministic part of the glottal excitation source is masked above the F_{VU} by the contribution of its stochastic part. Please note that the denomination VUF given in fig. 6.13 denotes F_{VU}.

Contrariwise, the signal interpretation for the full-band $C_{full}(\omega)$ is based on the assumption that despite the masking by the noise part the contribution of the deterministic glottal excitation source part still influences the spectral slope in the stochastic frequency band $\omega > \omega_{VU}$. Additionally, another hypothesis not being validated here claims that the filter described by the vocal tract has zero spectral slope over the full frequency band. The LF model describes over its complete R_d parameterization range a descending slope from low to high frequencies. The energy contour in the stochastic band describes to a huge extent as well a descending spectral slope. However, this depends on the phonetic content type of the speech signal in voiced segments, with different phonemes having different spectral slopes. In most of the cases the slopes of the synthetic LF pulse and the slope described by the original signal cancel itself out such that a VTF contour having zero spectral slope remains.

Figure 6.14 shows the spectrogram of the constructed VTF $C_{full}(\omega)$ for speaker BDL. It exhibits throughout most of the spectrogram an amplitude contour having roughly a zero spectral slope. In voiced segments where the speech signal is comprised of harmonic sinusoids and random noise, the spectral slope descends. However, if the spectral slope of the stochastic component does not descend by an amount being higher than the spectral slope of the synthetic LF pulse, the division by a glottal pulse synthesized with the LF model results into a spectral slope f or the extracted VTF ascending with frequency. On the other hand, in purely unvoiced segments with no sinusoidal content, for example for the unvoiced fricative phonemes [f] or [s], the spectral energy rises from low in lower to high in higher frequency bands. But it descends when approaching the Nyquist frequency. A parabola-like shape can be observed for such phoneme types. If the spectral division to construct a VTF is executed in a transient region with nearly no deterministic but much more stochastic content present, the contour of $C_{full}(\omega)$ is prone to

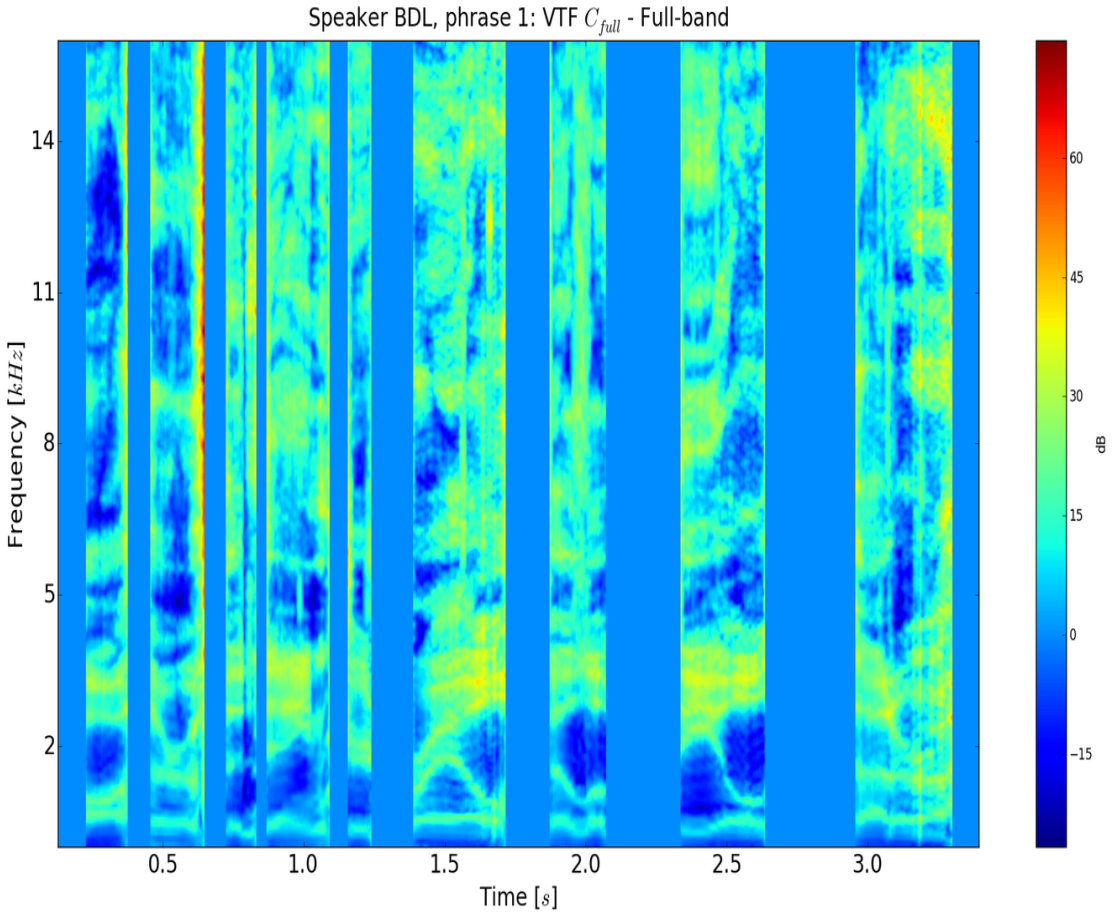


Figure 6.14: *Speaker BDL - Example of full-band VTF $C_{full}(\omega)$ without scaling to 0 dB mean*

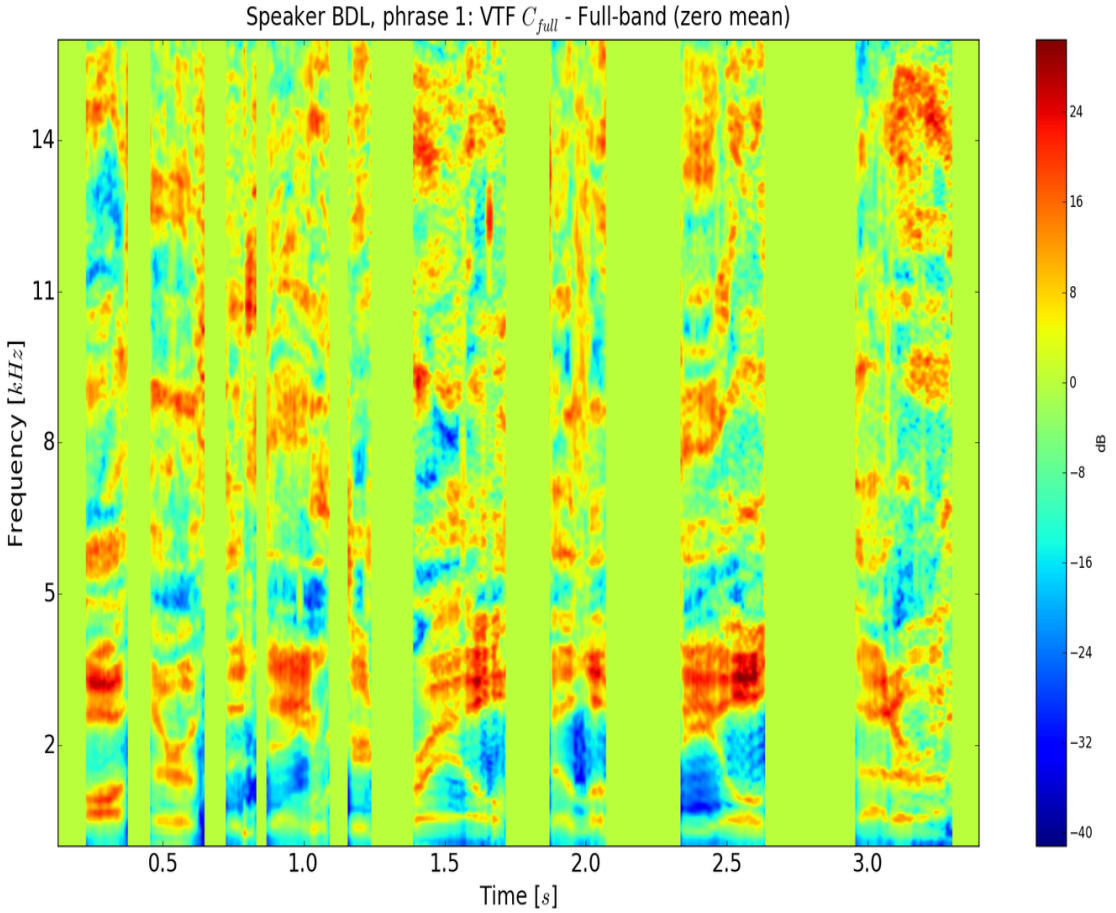


Figure 6.15: *Speaker BDL - Example of full-band VTF $C_{full}(\omega)$ with scaling to 0 dB mean*

rise with higher frequencies. This behaviour can be inspected in fig. 6.14 to a huge extent at ~0.65 seconds and slightly in the region around ~3.30 seconds.

Fig. 6.15 exemplifies the same $C_{full}(\omega)$ with each frame being scaled to have 0 dB mean. It better illustrates the formant structure contained in lower frequency regions present in $C_{full}(\omega)$. The preceding example of fig. 6.14 is biased by the very high amplitudes of the stochastic part at the transient around ~0.65 seconds. Please note that the Vocal Tract Filter example for $C_{F_{VU}}(\omega)$ shown in fig. 6.12 is not scaled to 0 dB mean. The only scaling present in *PSY* concerning the extraction of both VTF versions is the energy correction of each synthesized glottal pulse $g_{R_d}^{gci}$, defined in equ. 6.10.

It will be shown in section 6.5 that the convolution of $C_{full}(\omega)$ with a glottal pulse $G_s(\omega)$ at synthesis re-establishes a spectral slope contour over time reflecting the one of natural human speech. Obviously, changing the original R_d contour for the re-synthesis of a transformed glottal pulse sequence $g'_{R_d}(n)$ with $C_{full}(\omega)$ or $C_{F_{VU}}(\omega)$ results into a change of the spectral slope of the output signal $s'(n)$. The spectral fading synthesis variant of *PSY* introduced in section 6.5.5 is designed for the synthesis of transformed R'_d contours.

6.3 Analysis II - The unvoiced stochastic component

This section presents different means conducted for this work to estimate the stochastic signal $u(n)$ from a speech signal $s(n)$. The best performing approach of the presented methods utilized in *PSY* is the combination of:
1. The method "Re-Mixing with De-Modulation" of section 6.3.1.3 to delete the voiced part $v(n)$ generating the unvoiced residual signal $u_{res}(n)$
2. The posterior filter applying an high pass filter below F_{VU} of section 6.3.2.3, applied on $u_{res}(n)$ to synthesize $u(n)$.

The posterior filter is required since an erroneous sinusoidal detection usually provokes an increased energy level in $u_{res}(n)$. The deterministic part discussed in this section is represented by a sinusoidal parameter model. The separation of a speech signal $s(n)$ into the contributions of the voiced deterministic $v(n)$ and the unvoiced component $u(n)$ is based on the calculation of a residual of a sinusoidal model. Using the sinusoidal model has the advantage that pulse shape errors that are due to the rather limited coverage of the R_d parameterization of the LF model will not lead to an increase in the unvoiced component.

The separation of a speech signal into the contributions of a voiced deterministic $V(\omega)$ and an unvoiced stochastic $U(\omega)$ component represents one of the basic signal operation means performed in the presented *PSY* speech processing framework. An exact estimation of the quasi-harmonic sinusoidal and random noise signal types is indispensable for a high qualitative synthesis. The energy of a wrongly as noise classified sinusoid would be transformed into a random noise signal. This can be perceived as increased high frequency noise. On the other hand, wrongly as sinusoidal content classified spurious peaks, sub-harmonic peaks or a narrow-band noise region is prone to produce crackling or buzzing artefacts.

Section 6.3.1 presents the first part of estimating the stochastic signal $u(n)$. First, the classification of an input signal $s(n)$ into the voiced component $\hat{V}(\omega)$ and into the unvoiced component $\hat{U}(\omega)$ is discussed in section 6.3.1.1. An attempt to employ the Quasi-Harmonic Model of [Pantazis et al., 2008] for this task is shown in section 6.3.1.2. Section 6.3.1.3 introduces a novel method called "Re-Mixing with De-Modulation". It is based on the de-modulation of the spectral representation $\Delta(\omega)$ of the Dirac impulse sequence to optimize the estimation of amplitude $\hat{A}_k(n)$, instantaneous frequency $\hat{f}_k(n)$ and instantaneous phase $\hat{\phi}_k(n)$ for each sinusoid k contained in $\hat{V}(\omega)$. Informal listening test suggest that the "Re-Mixing with De-Modulation" method estimates most accurately the sinusoidal content $\hat{V}(\omega)$ for its spectral subtraction from $S(\omega)$, compared to the other approaches presented in the same section.

The first step of section 6.3.1 evaluates the residual waveform $U_{res}(\omega)$ as a result of the spectral subtraction performed by either the approach introduced in section 6.3.1.2 or the one of 6.3.1.3. A spectral envelope sequence $TE_{res}(\omega)$ is estimated on $U_{res}(\omega)$ and excited by white noise. The result is the unvoiced residual waveform $u_{res}^{noi}(n)$. Despite, not all sinusoidal content could be properly detected and deleted from $S(\omega)$ by these approaches. Sinusoidal energy is transformed into unvoiced energy and remains in $u_{res}^{noi}(n)$. This leads to a buzzy and metallic sounding unvoiced residual waveform $u_{res}^{noi}(n)$. Hence, section 6.3.2 presents two posterior filters to further suppress sinusoidal content, and to more accurately approximate the true stochastic signal $u(n)$. Section 6.3.2.2 presents a posterior filter which exclusively deletes remaining sinusoids at phoneme borders where the signal transitions either from purely unvoiced to a mixed voiced and unvoiced signal, or vice versa. Another posterior filter introduced in section 6.3.2.3 applies a high pass filter whose cutoff frequency is controlled by the F_{VU}.

6.3.1 Stochastic residual estimation

6.3.1.1 Signal classification into sinusoidal / noise peaks

The estimation of the random noise component $\hat{U}(\omega)$ is straightforward in the stochastic frequency band above the F_{VU} border. Its amplitude envelope $\mathcal{T}_{unv}$ is solely described by the frequency dependent noise floor in the stochastic frequency band $\omega > \omega_{VU}$. The problem to extract the true $U(\omega)$ from a speech recording is the superposition of the voiced component $V(\omega)$ in the deterministic frequency band $\omega < \omega_{VU}$ below the F_{VU} border. The estimation of the noise level is error-prone in this lower frequency area due to the strong or complete masking of the noise component $U(\omega)$ by the sinusoidal harmonic component $V(\omega)$. The latter has to be estimated precisely such that it can be completely cancelled from the input waveform $s(n)$ without further cancelling parts of the unvoiced waveform $u(n)$. The sinusoidal detection [Röbel et al., 2004] is more likely to fail if the sinusoidal peaks are close in amplitude to the peaks of the stochastic component. This occurs predominantly at frequency regions close to F_{VU}, as well as at time instants around phoneme boundaries and transient regions. A too long analysis window $w_h(n)$ may lead as well for such signal areas to an erroneous estimation of amplitude $\hat{A}_k(n)$, instantaneous frequency $\hat{f}_k(n)$ and instantaneous phase $\hat{\phi}_k(n)$ for all sinusoids K. Such erroneous sinusoidal estimation leads

in turn to the deformation of the spectral envelope $\mathcal{T}$ evaluated on the estimated sinusoids. At synthesis, these sinusoidal estimation errors lead to a wrong pulse form and pulse position.

An empiric investigation revealed problems with a standard STFT based method using an F_0-adaptive window size of four fundamental periods T_0. The drawbacks are related to the exact estimation of phase alignment and amplitude contour of sinusoids for certain signal segments, such as transients. At synthesis, not precisely estimated pulses are smeared since the voiced component $\hat{V}(\omega)$ could not properly be constructed.

The sinusoidal parameters of the deterministic component $\hat{V}(\omega)$ are estimated for every peak being classified as sinusoidal [Zivanovic et al., 2004]. The peak is re-synthesized according to the estimated parameters and subtracted from the spectrum $S(\omega)$. The aperiodic residual $U_{res}(\omega)$ is accordingly extracted by subtracting the estimated deterministic component $\hat{V}(\omega)$ from $S(\omega)$: $U_{res}(\omega) = S(\omega)$ - $\hat{V}(\omega)$.

Misclassified voiced $\hat{V}(\omega)$ and unvoiced $\hat{U}(\omega)$ components cause reconstruction errors in the re-synthesized waveform $\hat{s}(n)$ of the original waveform $s(n)$. A misclassification leads to an undesired energy shift between both parts [d'Alessandro et al., 1998]. Disturbing artefacts and undesired effects like reverberation may be induced.

The sinusoidal detection and deletion is evaluated as follows. A spectral envelope sequence $\mathcal{T}^{unv}_{adapt}(\omega)$ is estimated on $U_{res}(\omega)$. The unvoiced component $\hat{U}_{noi}(\omega)$ is generated by exciting $\mathcal{T}^{unv}_{adapt}(\omega)$ with white noise. The extracted aperiodic residual $u_{res}(n)$ and the re-synthesized unvoiced component $\hat{u}_{noi}(n)$ may still be comprised of sinusoidal content and sound therefore unnatural, unpleasant, buzzy and synthetic. This initial investigation proved to be not sufficient to estimate the unvoiced signal $u(n)$ with high quality.

6.3.1.2 Quasi-Harmonic Model (QHM)

The implemented QHM-based approach to extract the unvoiced residual waveform $U^{QHM}_{res}(\omega)$ and the unvoiced component $U^{QHM}_{noi}(\omega)$ is summarized by the pseudo-algorithm 2 defined below. The detailed steps are outlined as follows:

I. $V_{QHM}(\omega)$:
The Quasi-Harmonic Model presented in [Pantazis et al., 2008] and section 2.4.6 facilitates developing a harmonic sinusoidal estimator which requires an analysis window size of only two fundamental periods T_0. A shorter window size achieves a more robust sinusoidal analysis of $V_{QHM}(\omega)$ at especially problematic signal regions like phoneme boundaries and transient regions. An exact sinusoidal analysis facilitates its deletion from the signal to precisely construct the unvoiced residual $U_{res}(\omega)$.

II. $V^{fade}_{QHM}(\omega)$:
Even with a precise estimation of the voiced borders, sinusoidal content may remain around a voiced segment border in $U_{res}(\omega)$. Therefore, an optimization determining the best fade in and fade out times of the sinusoidal component $V^{fade}_{QHM}(\omega)$ is necessary. Sinusoids are synthesized using different fade times on a fixed time grid in the unvoiced region at each voiced segment border. The maximum length of the fade times in the evaluated time grid is 50 ms. The grid is partitioned into 1 ms steps. The subtraction of the synthesized sinusoids is applied for each fade length. The fade length with the lowest remaining energy is maintained as optimal fade in or fade out time for the best performing sinusoidal deletion.

III. $U^{QHM}_{res}(\omega)$:
The voiced component $V^{fade}_{QHM}(\omega)$ with optimal fade in/out times is subtracted from $S(\omega)$ to extract the unvoiced residual $U^{QHM}_{res}(\omega)$.

Algorithm 2 - QHM-based unvoiced stochastic residual $U^{QHM}_{noi}(\omega)$

I. $V_{QHM}(\omega)$: Compute the Quasi-Harmonic Model per GCI time in voiced segments
II. $V^{fade}_{QHM}(\omega)$: Estimate optimal fade in/out times on voiced segment borders
III. $U^{QHM}_{res}(\omega)$: Subtract $V^{fade}_{QHM}(\omega)$ from $S(\omega)$ to retrieve $U^{QHM}_{res}(\omega)$
IV. $U^{QHM}_{noi}(\omega)$: Excite with white noise a spectral envelope estimated on $U^{QHM}_{res}(\omega)$ to generate $U^{QHM}_{noi}(\omega)$

Synthesizing the unvoiced residual $U^{QHM}_{res}(\omega)$ gives the time domain waveform $u^{res}_{QHM}(n)$. Informal listening tests show that $u^{res}_{QHM}(n)$ suffers from a not perfect estimation of the voiced component $V^{fade}_{QHM}(\omega)$. The unvoiced residual waveform $u^{res}_{QHM}(n)$ still contains spurious sinusoidal content and artefacts, especially at voiced segment borders. The same is valid for unvoiced waveform $u^{QHM}_{noi}(n)$.

6.3.1.3 Re-Mixing with De-Modulation (ReMiDeMo)

The sinusoidal vs. noise classification using a four period F_0-adaptive windowing to perform the simple sinusoidal subtraction discussed in section 6.3.1.1 did not perform sufficiently precise. The following approach shown in section 6.3.1.2 using the Quasi-Harmonic Model with a F_0-adaptive windowing of only two fundamental periods T_0 and a dedicated optimization at voiced segment borders exhibits significant improvements. Since still the exact estimation of the voiced $\hat{V}(\omega)$ and unvoiced $\hat{U}(\omega)$ component is not achieved with both attempts, no perfect reconstruction of the re-synthesized waveform $\hat{s}(n)$ being close to the original waveform $s(n)$ can be computed.

Therefore, a novel methodology to extract the unvoiced residual $u_{res}(n)$ from an speech signal $s(n)$ is presented in this section. Generally, the "Re-Mixing with De-Modulation" (ReMiDeMo) approach aims to simplify the sinusoidal detection [Röbel et al., 2004]. Signal models to describe the deterministic part of an audio signal, of which a subset for speech is presented in section 2.4, struggle with the estimation precision if stronger modulations and fast transitions in amplitude and frequency occur [Pantazis et al., 2011]. Such AM–FM modulations invalidate the assumption of quasi-stationarity within one windowed STFT segment and are prone to produce an estimation bias [Röbel, 2008]. Different model extensions exist to reduce such estimation bias [Abe and Smith, 2004, Abe and Smith, 2005]. These extensions assume that the corresponding sinusoidal peaks can be observed in the spectrum. However, this is not applicable for stronger modulations. The proposed ReMiDeMo method applies therefore a dynamic re-sampling of the modulations such that their inversion can be conducted without suffering from significant signal distortions. The pseudo-algorithm 3 summarizes the procedure entitled as "Re-Mixing with De-Modulation" in the spectral domain.

Algorithm 3 - Re-Mixing with De-Modulation: Estimate the unvoiced stochastic residual $\boldsymbol{U_{noi}^{ReDe}(\omega)}$

- **I.** $\boldsymbol{S_{flat}^{FM}(\omega)}$: De-modulate the F_0 frequency contour with $F_0^{'}$ to compute $S_{flat}^{FM}(\omega)$
- **II.** $\boldsymbol{S_{flat}^{AMFM}(\omega)}$: De-modulate the smoothed amplitude contour $\mathcal{H}$ using the Hilbert transform
- **III.** $\boldsymbol{U_{ReDe}^{res}(\omega)}$: Subtracting the sinusoidal content of $S_{flat}^{AMFM}(\omega)$ gives the unvoiced residual $U_{ReDe}^{res}(\omega)$
- **IV.** $\boldsymbol{U_{ReDe}^{AM}(\omega)}$: Re-modulate the amplitude envelope $\mathcal{H}(S^{F_{flat}}(\omega))$
- **V.** $\boldsymbol{U_{ReDe}^{AMFM}(\omega)}$: Re-modulate the frequency contour F_0^{transp} to re-establish the original F_0 contour
- **VI.** $\boldsymbol{U_{noi}^{ReDe}(\omega)}$: Excite with white noise a spectral envelope estimated on $U_{ReDe}^{AMFM}(\omega)$ to generate $U_{noi}^{ReDe}(\omega)$

The following more detailed explanations are given in the time domain: "ReMiDeMo" de-modulates the F_0 contour and the smoothed Hilbert amplitude envelope $\mathcal{H}$ from the signal $s(n)$. The original F_0 contour of $s(n)$ is warped to become flat by means of dynamic time-varying re-sampling. It uses as target F_0 the mean of the original fundamental frequency contour. The re-sampling operation will locally and globally change the time duration of all signal features. The effect can be inverted after the extraction of the residual since the warping contour is stored in $F_0^{'}$. A flat F_0 contour is established by means of the $F_0^{'}$ transposition which de-modulates the original F_0 contour, depicted in fig. 6.16. Please note that the transposition contour has been divided by factor 10 for easier illustration purposes. The varying amplitude contour of $s(n)$ is demodulated by means of dividing the signal by its smoothed

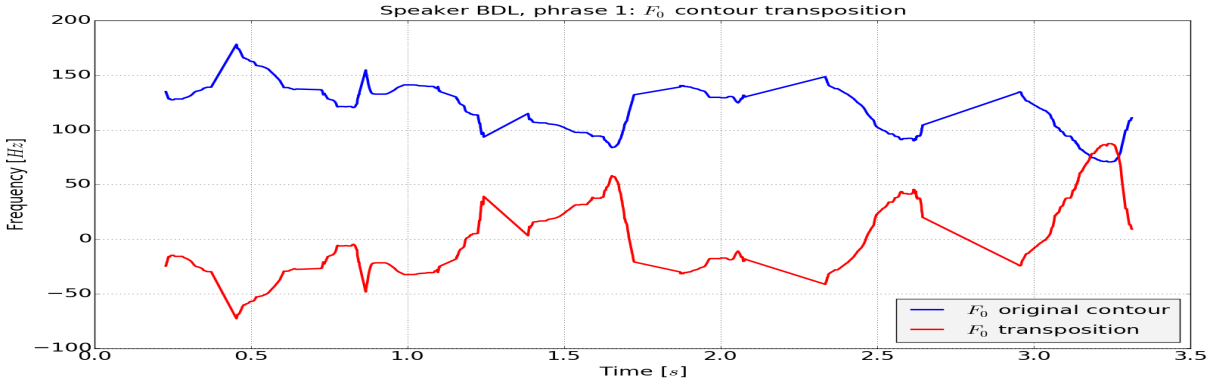


Figure 6.16: *Example of F_0 contour transposition for F_0 de-modulation*

Hilbert transform $\mathcal{H}(s(n))$, similar as in [Pantazis and Stylianou, 2008, Drugman and Dutoit, 2012]. The smoothing kernel is simply the Hanning window of duration exactly equal to $4/F_T$. This optimally removes all envelope fluctuations that are related to the deterministic components. The resulting signal $s_{flat}(n)$ is flat in amplitude en-

velope and fundamental frequency facilitating the detection of sinusoids following [Zivanovic and Röbel, 2008]. It avoids even for relatively high harmonic numbers the energy shift between voiced and unvoiced components [d'Alessandro et al., 1998]. The sinusoidal content is subtracted from $s_{flat}(n)$ and the demodulation steps are inverted. The original AM-FM modulation is recreated. This generates the unvoiced residual signal $u_{res}(n)$. A detailed description of each algorithmic step will be given in the following:

Index 1: The de-modulation of the F_0 time and frequency contour $F_0^{'}$ is applied as pitch transposition contour at $S(\omega)$ to generate $S_{flat}^{FM}(\omega)$. $S_{flat}^{FM}(\omega)$ has a constant and thus a flat fundamental frequency F_0 contour. The de-modulation curve $F_0^{'}$ is stored to facilitate a later inverse transposition for the re-modulation of the F_0 interference without time correction. The F_0 re-modulation re-establishes the original F_0 contour along with its time duration.

Index 2: The varying amplitude contour of $S(\omega)$ is as well normalized to be flat and constant. Similar as in [Pantazis and Stylianou, 2008, Drugman and Dutoit, 2012], this is achieved by means of dividing $\mathcal{H}$ from $S(\omega)$. The Hilbert transform $\mathcal{H}$ computes on $S_{flat}^{FM}(\omega)$ the analytic spectrum $\mathcal{H}(S_{flat}^{FM}(\omega))$ and its amplitude envelope $\mathcal{H}(S^{F_{flat}}(\omega))$. The latter is additionally smoothed by a time domain convolution with the Hanning window w_h covering four times the mean fundamental period $F_{0\mu}$ of $S_{flat}^{FM}(\omega)$. $S_{flat}^{FM}(\omega)$ is divided by the amplitude envelope $\mathcal{H}(S_{flat}^{FM}(\omega))$ to generate $S_{flat}^{AMFM}(\omega)$ having a flat amplitude and frequency contour.

Index 3: The procedure described above facilitates a comparably more robust and precise estimation of the sinusoidal content contained in $S_{flat}^{AMFM}(\omega)$. Likewise, the subtraction of the sinusoidal content is more robust and precise compared to the approaches described in the preceding sections. The sinusoidal deletion renders the unvoiced residual waveform $U_{ReDe}^{res}(\omega)$. The error tolerance parameter ρ determines for this operation how much detected spectral peaks are classified as sinusoidal content or as spurious noise [Zivanovic et al., 2004]. In *PSY*, ρ is set to 10 %.

Indices 4 and 5: A re-modulation of the beforehand removed amplitude envelope contour $\mathcal{H}(S^{F_{flat}}(\omega))$ and the frequency contour $F_0^{'}$ re-establishes the AM-FM modulation of $S(\omega)$ on the unvoiced residual $U_{ReDe}^{AMFM}(\omega)$.

Synthesizing the unvoiced residual $U_{ReDe}^{AMFM}(\omega)$ with the STFT generates the time domain signal $u_{ReDe}^{AMFM}(n)$. Informal listening tests suggest that with $u_{ReDe}^{AMFM}(n)$ a better cancellation of the sinusoidal content is achieved with the re-mixing of the de-modulated component $S_{flat}^{AMFM}(\omega)$ as with the QHM approach described in section 6.3.1.2. However, the unvoiced residual waveform $u_{ReDe}^{AMFM}(n)$ may be interfered by undesired remaining signal content resulting from a still not perfect estimation and cancellation of sinusoidal content.

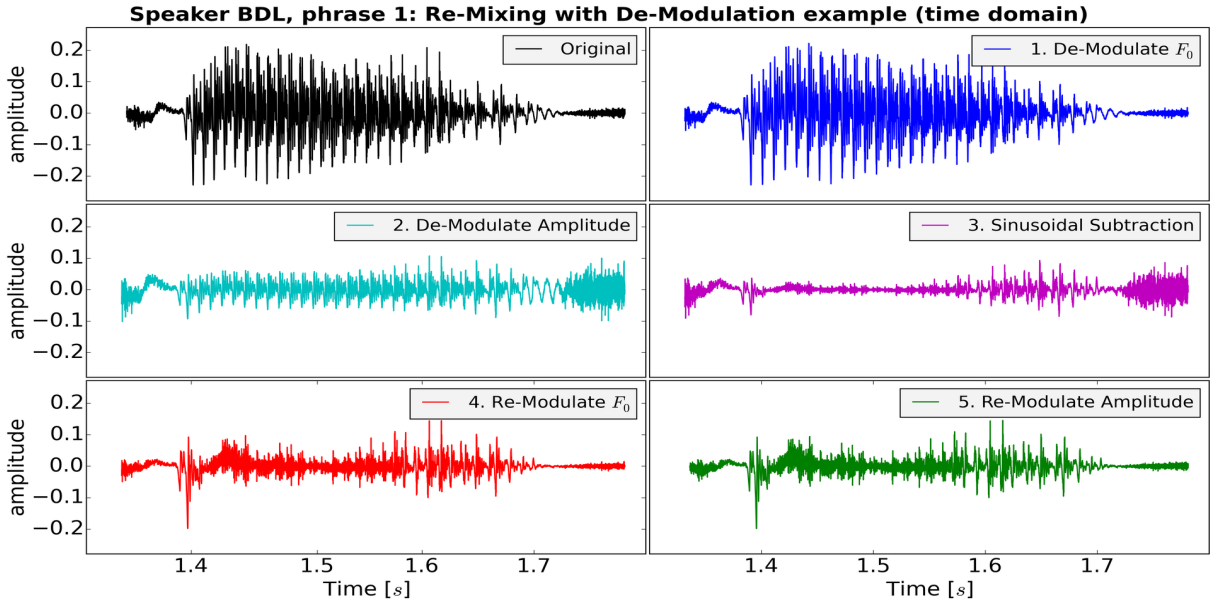


Figure 6.17: *Time domain example of "Re-Mixing with De-Modulation" to estimate* $u_{ReDe}^{AMFM}(n)$

Examples of the Re-Mixing using De-Modulation and Re-Modulation showing an audio segment for speaker BDL are given for the time domain in fig. 6.17 and for the spectral domain in fig. 6.18. Both figures exemplify the successive algorithmic steps introduced with the pseudo-algorithm 3. The indices from 1 to 5 given in both example figures correspond to the ones given in the pseudo-algorithm. One example of a non-perfect cancellation can be inspected for the transient region at ~1.40s in fig. 6.17. The shown failing sinusoidal deletion motivates the posterior filtering approaches introduced in the following. Still with the novel method presented in this section the

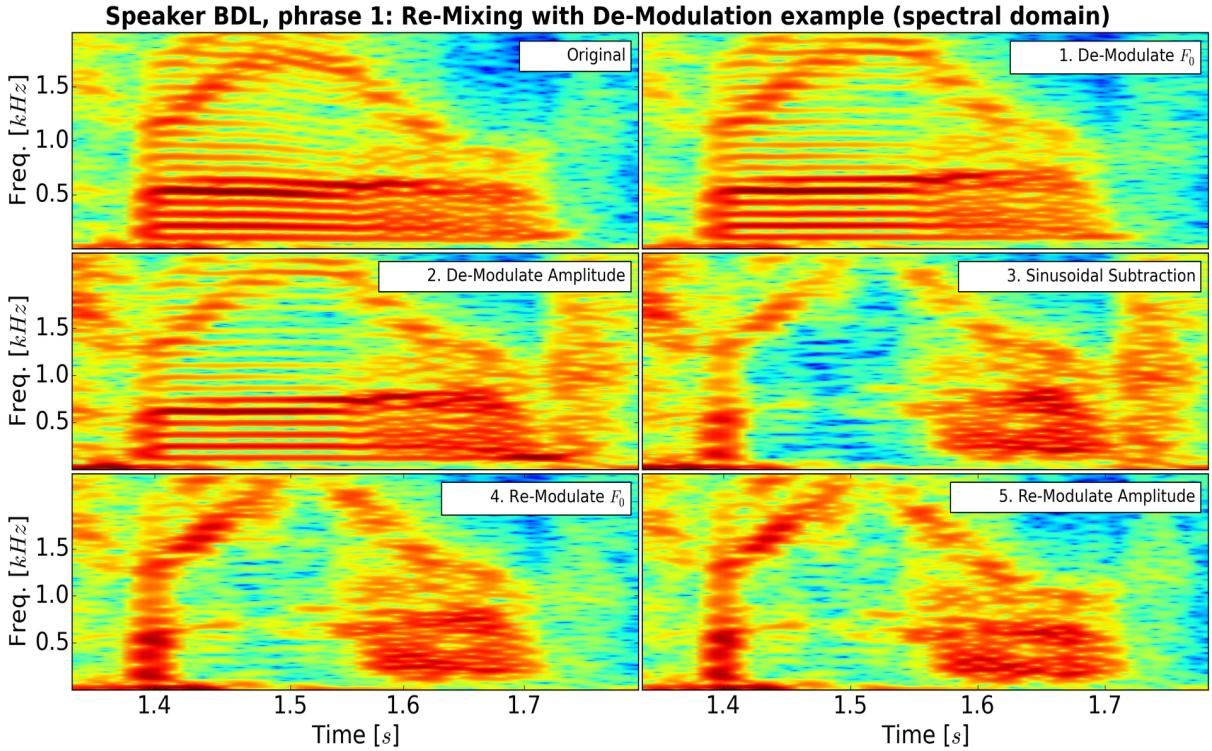


Figure 6.18: *Spectral example of "Re-Mixing with De-Modulation" to estimate* $U_{ReDe}^{AMFM}(\omega)$

created unvoiced residual proved to be not sufficient for a high quality speech system.

6.3.2 Posterior filtering

6.3.2.1 Noise excitation

The approach using the QHM along with the additional fade in/out at voiced borders, discussed in section 6.3.1.2, extracts the unvoiced residual waveform $u_{QHM}^{res}(n)$ from the input signal $s(n)$. The approach using the re-mixing of the de-modulated frequency and amplitude contour of sinusoidal content, discussed in section 6.3.1.3, estimates the unvoiced residual waveform $u_{ReDe}^{AMFM}(n)$. The non-perfect cancellation of sinusoidal content leaves both unvoiced residual waveforms $u_{QHM}^{res}(n)$ and $u_{ReDe}^{AMFM}(n)$ sounding unnatural, metallic and synthetic. Spurious remainders of sinusoidal content may additionally introduce artefacts.

Also the estimation of a spectral envelope sequence $\mathcal{T}$ on the residual waveforms and the excitation of $\mathcal{T}$ with white noise does not alleviate the mentioned problems. An erroneous sinusoidal detection leads to an undesired transformation of too much energy from the sinusoidal into the noise component. It is especially likely to fail at transients and voiced segment borders, e.g. depicted in fig. 6.17 at ~1.40s. The solutions presented in the following sections aim to correct such energy shifts.

6.3.2.2 Threshold onsets

The algorithm called "threshold onsets" aims to address the failing sinusoidal detection [Röbel et al., 2004, Zivanovic et al., 2004] at transients and voiced segment borders. It saturates the spectral envelope level of either $\mathcal{T}_{QHM}^{unv}(\omega)$ or $\mathcal{T}_{ReDe}^{unv}(\omega)$ to the energy level of the same spectral envelope found at a time distance of 1.5 fundamental periods T_0 before a voiced onset or respectively after a voiced offset. The scaling of the spectral envelope level before excitation with white noise suppresses too high unvoiced energies at the unvoiced part of a voiced / unvoiced segment border. The time domain waveforms of thresholding the onset regions are denominated $u_{QHM}^{thron}(n)$ or respectively $u_{ReDe}^{thron}(n)$.

However, informal tests have shown that signal clips are still present around transients. Too high unvoiced energies are found at the voiced part around a segment border. The reason may be that the "threshold onsets" approach treats only the unvoiced part around voiced/unvoiced segment borders. An extension of the proposed thresholding approach into the voiced part could minimize the observed problems in estimating $u(n)$. The informal tests used the

PSY energy model introduced in section 6.4.1. Misadjustments of the voiced and unvoiced energy parts occur if the unvoiced component is set to a wrong energy level. A re-synthesized waveform $\hat{s}(n)$ of the original waveform $s(n)$ will not be reconstructed precisely with erroneous unvoiced energies contained in $u^{thron}_{QHM}(n)$ or $u^{thron}_{ReDe}(n)$. Therefore, another posterior filter is introduced in the following section 6.3.2.3 to correct the unvoiced residual waveforms $u^{noi}_{QHM}(n)$ and $u^{noi}_{ReDe}(n)$ estimated by the methods of the preceding sections 6.3.1.2 and 6.3.1.3 by different means.

6.3.2.3 Below F_{VU} filter

The estimation of the Voiced / Unvoiced Frequency boundary F_{VU} is based on the assumption of separating the spectrum into two bands. The two bands consist of one deterministic part below the F_{VU} and one stochastic part above the F_{VU}, as discussed in the sections 2.3.2 and 6.2.3.3. Following such hypothesis for the deletion of the estimated voiced sinusoidal content $\hat{v}(n)$ from a speech signal $s(n)$ suggests that the sinusoidal deletion has to take effect exclusively in the deterministic frequency band $\omega < \omega_{VU}$. No sinusoidal content $\hat{V}(\omega)$ should remain due to a non-perfect sinusoidal cancellation in the unvoiced residual waveform $u^{res}_{QHM}(n)$ or $u^{AMFM}_{ReDe}(n)$.

A high pass filter is applied at the spectral envelope sequences $\mathcal{T}^{unv}_{QHM}(\omega)$ and $\mathcal{T}^{unv}_{ReDe}(\omega)$ of the unvoiced residual waveforms $u^{res}_{QHM}(n)$ or $u^{AMFM}_{ReDe}(n)$. The cutoff frequency f_c of the high pass filter is set to the F_{VU} estimation present for each frame k: $f_c(k) = F_{VU}(k)$. A gain of 1 is set in the filters passband equalling the stochastic frequency band $\omega > \omega_{VU}$. To further delete remaining sinusoidal content located below the F_{VU}, a linear ramp with a slope of m_{HP}=x dB per octave defines the high pass filtering. The filters stopband equals the deterministic frequency band $\omega < \omega_{VU}$. The high pass filter is defined by the following equation 6.13 as

$$\mathcal{T}^{unv}_{HP}(\omega) = \begin{cases} \mathcal{T}^{unv}(\omega) & \forall \quad \omega >= \omega_{VU} \\ \mathcal{T}^{unv}(\omega) \cdot m_{HP} \cdot log_2(F_{nyq}/F_{VU}) & \forall \quad \omega < \omega_{VU} \end{cases}, \tag{6.13}$$

with $F_{nyq} = F_s/2$ being the Nyquist frequency of half the sampling rate F_s. In *PSY*, the slope control factor m_{HP} is set to default -3 dB / octave. Informal re-synthesis tests conducted for establishing the listening tests presented in section 6.6 show that m_{HP} = -3 dB / octave achieves a good approximation to reconstruct the original signal $s(n)$.

If the F_{VU} estimation at time instant t is zero, no high pass filtering is applied: $F_{VU}(t) = 0$; $m_{HP}(t)$=0. The high pass filter is only active in voiced sections with values $F_{VU} > 0$. The filter is designed such that lower F_{VU} values lead to steeper filter ramps compared to higher F_{VU} values having a comparably more flat filter ramp in the filters stopband. If the F_{VU} estimate moves to higher frequencies because the evaluated signal segments contain more sinusoidal content, the filter ramp is automatically set less steep to not remove the complete unvoiced signal.

The spectral envelope sequence $\mathcal{T}^{unv}_{HP}$ reflects the high pass filtering in its contours. It is synthesized as the random noise component $u^{HP}_{QHM}(n)$ or $u^{HP}_{ReDe}(n)$ following the means explained in section 6.3.2.1.

6.3.2.4 Scale to $\mathcal{T}_{sig}$ level

The preceding sections introduced different means to establish the unvoiced signal $u(n)$. The unvoiced residual waveform $u^{QHM}_{res}(n)$ of section 6.3.1.2 using the Quasi-Harmonic Model can be further optimized with one of the two posterior filters of section 6.3.2.2 or 6.3.2.3. This produces the unvoiced waveforms $u^{thron}_{QHM}(n)$ or $u^{HP}_{QHM}(n)$. Respectively, the unvoiced residual waveform $u^{AMFM}_{ReDe}(n)$ of section 6.3.1.3 using the re-mixing of the de-modulated frequency and amplitude contour can be as well further processed with the two posterior filters to produce the unvoiced waveforms $u^{thron}_{ReDe}(n)$ or $u^{HP}_{ReDe}(n)$.

However, an erroneous sinusoidal detection may occur especially at some tricky signal segments such as transients [Röbel et al., 2004, Zivanovic et al., 2004]. A too strong or a too weak sinusoidal cancellation to produce the unvoiced residual waveforms $u^{QHM}_{res}(n)$ or $u^{AMFM}_{ReDe}(n)$ may introduce deviations of approximating the true unvoiced component $u(n)$ of the signal $s(n)$. Furthermore, the two posterior filters "threshold onsets" of section 6.3.2.2 and the below F_{VU} filter of section 6.3.2.3 can by their definition not assure the exact estimation of the unvoiced component $u(n)$.

The performance to robustly estimate the unvoiced signal $u(n)$ by the presented algorithmic combinations has been conducted by an analysis/synthesis approach on different speech phrase. Informal visual inspections and perceptual tests have shown that the reconstructed speech phrases may suffer for some unvoiced segments exhibiting a deviation in their energy level from the original signal. Therefore, a final scaling of a spectral envelope sequence

$\mathcal{T}_{unv}$ is conducted.

$$\eta = \frac{1}{k_{nyq} - k_{VU}} \sum_{k=k_{VU}}^{K=k_{nyq}} \left(\mathcal{T}_{sig}^{dB}(k) - \mathcal{T}_{unv}^{dB}(k) \right)$$
$$\mathcal{T}_{unv}^{w} = \mathcal{T}_{unv} \cdot (1 - k_{VU}/k_{nyq}) \cdot 10^{\eta/20} \tag{6.14}$$

k_{nyq} and k_{VU} are the DFT bins found closest to the Nyquist and the F_{VU} frequency. The scaling described in equ. 6.14 examines the difference between the spectral envelope sequence $\mathcal{T}_{sig}$ and $\mathcal{T}_{unv}$. The spectral envelope $\mathcal{T}_{unv}$ is estimated on any of the unvoiced waveforms $u_{ReDe}^{thron}(n)$, $u_{ReDe}^{HP}(n)$, $u_{QHM}^{thron}(n)$ or $u_{QHM}^{HP}(n)$ present in *PSY*. It approximately suppresses any appearing disparity between the observable stochastic component of the original signal $s(n)$ above the F_{VU} and any of the estimated stochastic signals $u(n)$. η equals the mean difference in dB between $\mathcal{T}_{sig}$ and the spectral envelope $\mathcal{T}_{unv}$. The scaling of $\mathcal{T}_{unv}$ is additionally weighted by the time-varying ratio of F_{VU} versus F_{nyq}. This regularization term is necessary to avoid an irregular envelope scaling, especially for high values of F_{VU}.

Algorithm 4 - Scale to $\mathcal{T}_{sig}$ level

1. $\boldsymbol{\eta = \mathcal{T}_{F_0 > F_{VU}}^{\delta}}$:
   Compute the mean difference $\eta = \mathcal{T}_{F_0 > F_{VU}}^{\delta}$ of $\mathcal{T}_{sig}(F_0 > F_{VU})$ and $\mathcal{T}_{unv}(F_0 > F_{VU})$ above the F_{VU}
2. $\boldsymbol{w_{F_{VU}}}$:
   Compute the weight $w_{F_{VU}} = (F_{nyq}\text{-}F_{VU})/F_{nyq}$ to reflect the implied voiced / unvoiced ratio
3. $\boldsymbol{\mathcal{T}_{unv}^{w}}$:
   Compute the weighted spectral envelope $\mathcal{T}_{unv}^{w} = \mathcal{T}_{unv} \cdot w_{F_{VU}} \cdot \mathcal{T}_{\delta_\mu}^{F_0 > F_{VU}}$

The pseudo-algorithm 4 summarizes the spectral envelope scaling. The additional weight $w_{F_{VU}}$ addresses the splitting at the F_{VU}. Setting the weight to a non-effective constant value of $w_{F_{VU}} = 1.0$ leads in general to a too high scaling and energy increase of the unvoiced component in both mixed voiced and unvoiced regions. The scaled spectral envelope $\mathcal{T}_{unv}^{w}$ is multiplied with a white noise signal to re-synthesize with the STFT any of the unvoiced waveforms $u_{ReDe}^{thron}(n)$, $u_{ReDe}^{HP}(n)$, $u_{QHM}^{thron}(n)$ or $u_{QHM}^{HP}(n)$ on which the original $\mathcal{T}_{unv}$ was estimated on.

6.3.3 Unvoiced stochastic component summary

Informal listening tests conducted throughout the work presented in the preceding section suggest that the sinusoidal cancellation of "Re-Mixing with De-Modulation" presented in section 6.3.1.3 combined with the below F_{VU} filter introduced in section 6.3.2.3 estimates the unvoiced waveform denominated as $u_{ReDe}^{HP}(n)$ with a higher sound quality compared to the other approaches. The unvoiced component $U(\omega)$ being utilized per default in *PSY* is therefore $u_{ReDe}^{HP}(n)$.

6.4 Transformation

The *PSY* analysis and synthesis framework is designed to perform advanced Voice Transformation or Voice Conversion tasks. The splitting into a voiced component $V(\omega)$ and an unvoiced component $U(\omega)$ enables means to process both parts separately. For Voice Transformation, a transformed glottal pulse shape contour R'_d enables to synthesize a new pulse sequence $g'_s(n)$ of the deterministic part of the glottal excitation source $G(\omega)$. For VC, a conversion of the VTF $C_{src}(\omega)$ of the source speaker into the target speakers $C'_{tar}(\omega)$ establishes the required exchange of the vocal tract filtering in a VC application. *PSY* enables the combination of a Voice Transformation and a Voice Conversion task. It can synthesize the combination of a glottal excitation pulse sequence $g'^{tar}(n)$ being transformed to the target speaker with a VTF sequence $C'_{tar}(\omega)$ being converted to the target speaker.

If a Voice Transformation and / or Voice Conversion from a source to a target speaker is performed, the synthesized waveform has to reflect the energy contour that the target speaker would have been used. If a Voice Transformation task changes the voice quality of a given speaker, as for example conducted in section 6.4.3, the energy contour of the synthesized waveform has to reflect the one of the given speaker. Such Voice Transformation or Voice Conversion tasks require to establish means to properly handle the energy of the synthesized waveform. This is reflected in the energy model of *PSY*. It will be introduced in the following section 6.4.1.

6.4.1 Energy modelling

The *PSY* energy model is based as well on the splitting of spectrum $S(\omega)$ into a voiced component $V(\omega)$ and an unvoiced component $U(\omega)$. The prosodic descriptor intensity [Pfitzinger, 2006] is expressed in *PSY* by a simple Root-Mean-Square (RMS) measure F_{RMS}. It evaluates the effective energy value E_{RMS} of an analyzed signal segment. Per analysis step k executed in *PSY*, the RMS energy measures E_{sig} on the signal $s(n)$, E_{voi} on the voiced component $V(\omega)$, and E_{unv} on the unvoiced component $U(\omega)$ are estimated. Each RMS energy measure E_{RMS} operates on the linear amplitude spectrum $A^y_{lin} = |Y(\omega)|$ of any arbitrary signal $y(n)$. Therefore, the energy modelling is dependent on the chosen STFT window size. However, the linear factor introduced by the Hanning window $w_h(n)$ can be neglected since it is present in each RMS measure on a signal segment. The energy interference of the window $w_h(n)$ applied at each analysis and synthesis of *PSY* cancels each other out.

$$F_{RMS}(A^y_{lin}, k) = \sqrt{1/K \cdot \Sigma_k \left(A^y_{lin}(k)^2\right)} \tag{6.15}$$

$$A^y_{lin} = |Y(\omega)| \tag{6.16}$$

$$E_{sig} = F_{RMS}(|S(\omega)|) \tag{6.17}$$

$$E_{unv} = F_{RMS}(|U(\omega)|) \tag{6.18}$$

$$E_{voi} = E_{sig} - E_{unv} \tag{6.19}$$

The equations 6.15 to 6.19 define the energy model of *PSY*. The RMS operator F_{RMS} measures the energies E_{sig} and E_{unv} of a signal segment $s(n)$ and of its corresponding unvoiced segment $u(n)$. Specifically, $u(n)$ refers to one of the four different unvoiced components explained in section 6.3.2.4. The energy measure E_{voi} is build on the simple energy difference between the speech signal $s(n)$ and its stochastic component $u(n)$, defined in equ. 6.19. An erroneous estimation in extracting the unvoiced component $U(\omega)$ leads to a misclassification of the energy distribution between the voiced and the unvoiced part of a re-synthesized signal $s'(n)$. The reason being that after a transformation or conversion and before synthesis, both the voiced component $V'(\omega)$ and the unvoiced component $U'(\omega)$ are scaled according to a chosen energy constraint. The operators $'$ and $''$ indicate that a transformation $'$ or a conversion $''$ has been applied to the variable identifying the same signal part. For example, the output signal $s'(n)$ is the result of a transformation applied to the input signal $s(n)$.

The possible energy constraints in *PSY* will be introduced in the following sections. The energy maintenance of section 6.4.1.2 simply re-scales the signal to the original energy of a synthesized phrase found before transformation. The GMM-based statistical energy model of section 6.4.2.3 models the energy behaviour of a speaker based on a dedicated voice descriptor set.

If the unvoiced component $U(\omega)$, estimated by the means shown in section 6.3, contains too much energy due to an erroneous sinusoidal detection, explained in section 6.3.1.1, the voiced deterministic component $V'(\omega)$ is scaled too much down in energy. Such erroneous sinusoidal detection may additionally lead to smeared and not well reproduced sinusoids. Especially the posterior filter of section 6.3.2.2 thresholding only the onsets in purely unvoiced speech segments is prone to a mislead reproduction of the voiced and unvoiced signal content around voiced segment borders. The posterior filter 'below F_{VU}' of section 6.3.2.3 requires the right adjustment of its slope parameter m_{HP} to achieve a good approximation of both signal types enabling a speech synthesis of high quality.

6.4.1.1 Energy behaviour of the LF model

The findings discussed in this chapter concern the open research question how to estimate and model vocal effort within the context of processing the glottal excitation source. A separated and more refined handling of vocal fold abduction / adduction expressed by R_d and vocal effort expressed by an energy metric is required for some expressive voices. Speaker are able to utter more loud with a relaxed voice quality, or to utter more silent with a a tense voice quality. This separation is partially addressed in *PSY*. The transformation of the R_d contour reflects the abduction and the adduction of the vocal folds. An alteration of the R waveshape parameters induced by an R_d value change requires additionally to alter the negative minimum E_e and the positive maximum E_i amplitude of the LF model accordingly. However, this alteration is not inherently handled by the LF model. Additionally, little information can be found in the literature concerning this behaviour [Strik and Boves, 1992, Gillett, 2003, Tooher et al., 2008, Degottex et al., 2013].

Please note that the discussion in this section intends by no means to criticise the works of Fant, one of the greatest researchers having contributed to the speech community. The presented energy behaviour of the LF model is just an interesting finding to justify the requirement of an energy model to conduct the transformation of the glottal excitation source. The properties of the LF model concerning its amplitude and energy characteristics will be evaluated over the R_d range and discussed in the following. The findings do not indicate a failure of the LF model. The discussed LF energy characteristic is a side effect being required to properly describe the shape of the glottal derivative pulse. It has to reflect the modelling established with the implied LF parameters.

The Liljencrants-Fant model LF, introduced in section 3.2.3, describes the deterministic part of the glottal excitation source. An efficient parameterization of the LF model is given with the LF shape parameter R_d, introduced in section 3.3.1. A robust estimation of R_d is outlined in chapter 5. The R_d regression parameterizes a subset of the three R waveshape parameters to describe the shape of the LF model, shown in section 5.2. However, the R_d estimation, introduced in section 3.7 and improved in chapter 5, does not offer means to characterize two additional parameters required to synthesize the LF model: The fundamental period T_0 defining the time length and the LF amplitude parameter E_e defining the amplitude of the synthesized LF model. While the fundamental period T_0 is already estimated in *PSY*, shown in section 6.2.1.1, the LF amplitude parameter E_e is not.

The LF model was developed by Liljencrants and Fant in [Fant et al., 1985, Fant, 1995] within the context of inverse filtering, similar to [Alku, 1992]. With this the LF amplitude parameter E_e can be measured as direct amplitude value E_e at the time instant t_e of maximum (negative) excitation, commonly known as the Glottal Closure Instant GCI [d'Alessandro et al., 2007]. Prominent methods to estimate E_e on a glottal source waveform are detailed in [Strik et al., 1993, Strik, 1998]. However, little information can be found in the literature how to estimate the amplitude E_e if no iterative inverse filtering is conducted [Degottex, 2010]. Moreover, the E_e measure may be erroneous if the estimation of one or several formants to inverse filter out its influence is erroneous. The E_e should be set to a certain amplitude level such that a glottal pulse synthesized with LF properly reflects the energy of the underlying deterministic part of the glottal excitation source [Degottex et al., 2011b, Degottex et al., 2013].

However, the equations defining the LF model do not allow to determine the amplitude value of E_e for synthesis directly. Instead, the factor E_0 allows to scale a synthesized LF model proportionally in amplitude. E_0 is defined in equ. 5 in [Fant et al., 1985]. This definition this not allow for a straightforward linear mapping of E_0 on E_e, or vice versa. The chosen energy modelling to handle voice quality transformations in terms of an R_d contour transformation is thus based on a linear RMS metric in *PSY*, explained in the following.

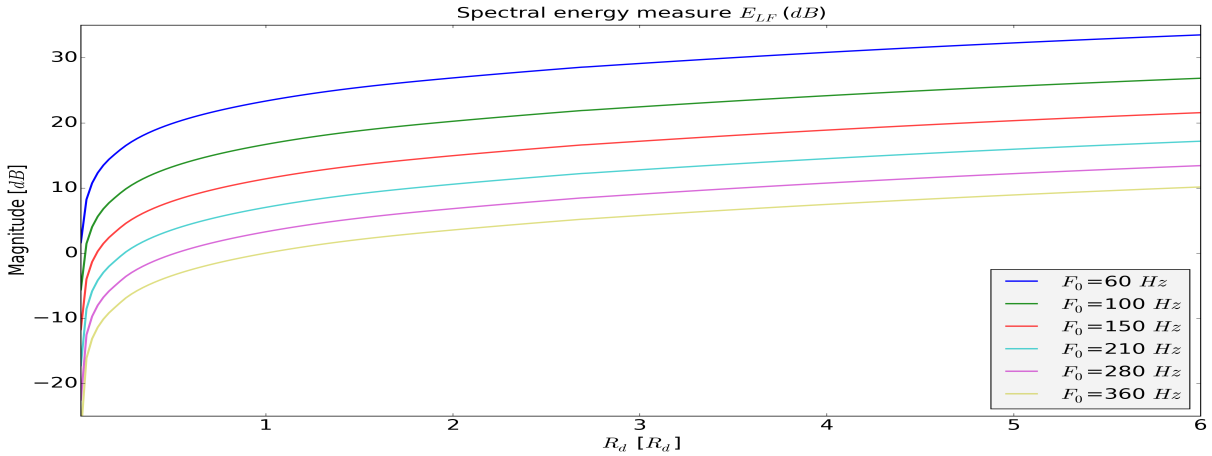


Figure 6.19: *RMS-based energy contours E_{LF} of synthesized LF glottal pulses in dB*

An RMS measure E_{LF} estimated on synthesized LF models is shown for different F_0 values in fig. 6.19. Each glottal pulse is synthesized with E_0=1.0 using different R_d values along the R_d range shown on the x-axis. The E_{LF} energy measure follows the definition given in 6.19. It is expressed in dB on the y-axis. The linear amplitude spectrum A_{lin} of the synthesized glottal sequence $G_s(\omega)$ is measured with the RMS operator $F_{RMS}(A_{lin})$. Lower R_d values are associated with a more tense voice quality, which can be interpreted as speech having a higher loudness level [Childers and Lee, 1991, Fant and Kruckenberg, 2005, Liénard and Barras, 2013]. Consequently, the hypothesis on any energy measure on the glottal source $g(n)$, the voiced $v(n)$ or the whole signal $s(n)$ part is that higher energy values should be observed at lower R_d values. However, fig. 6.19 exhibits the contrary of this expectation. Lower energy values are measured for lower R_d values on the synthetically generated glottal source $G(\omega)$. This violates the R_d border hypothesis discussed with step 4 of section 6.2.1.3. The hypothesis expects higher energy values for lower R_d values. As well contrary to such assumption is the study of [Fant and Kruckenberg, 1996] stating that an increase of 1 dB in E_e is associated with 0.5 dB increase in R_d for a constant F_0 value. The co-variation of F_0, R_d and E_e observed in [Fant and Kruckenberg, 1996] leads as well to a dependency of the positive amplitude peak U_0 of the glottal flow. R_d is according to equ. 3.4 determined by the ratio of U_0/E_e and the ratio $F_0/110$. The latter is the fundamental frequency of the current signal frame, normalized by the frequency 110 Hz being typical for a male speaker [Fant, 1995].

Increasing R_d at a constant level of E_e and F_0 requires that U_0 has to rise as well. The positive amplitude E_i of the glottal derivative pulse increases with R_d. With E_0 set to a constant amplitude, the amplitude level E_e of the synthesized LF model changes with F_0, independent of R_d. This property of the LF model leads to a bigger surface of the signal described by a synthesized LF pulse. This leads in turn to a higher measured RMS energy. The behaviour is required to properly approximate the contour of $E(t)$ over the whole period.

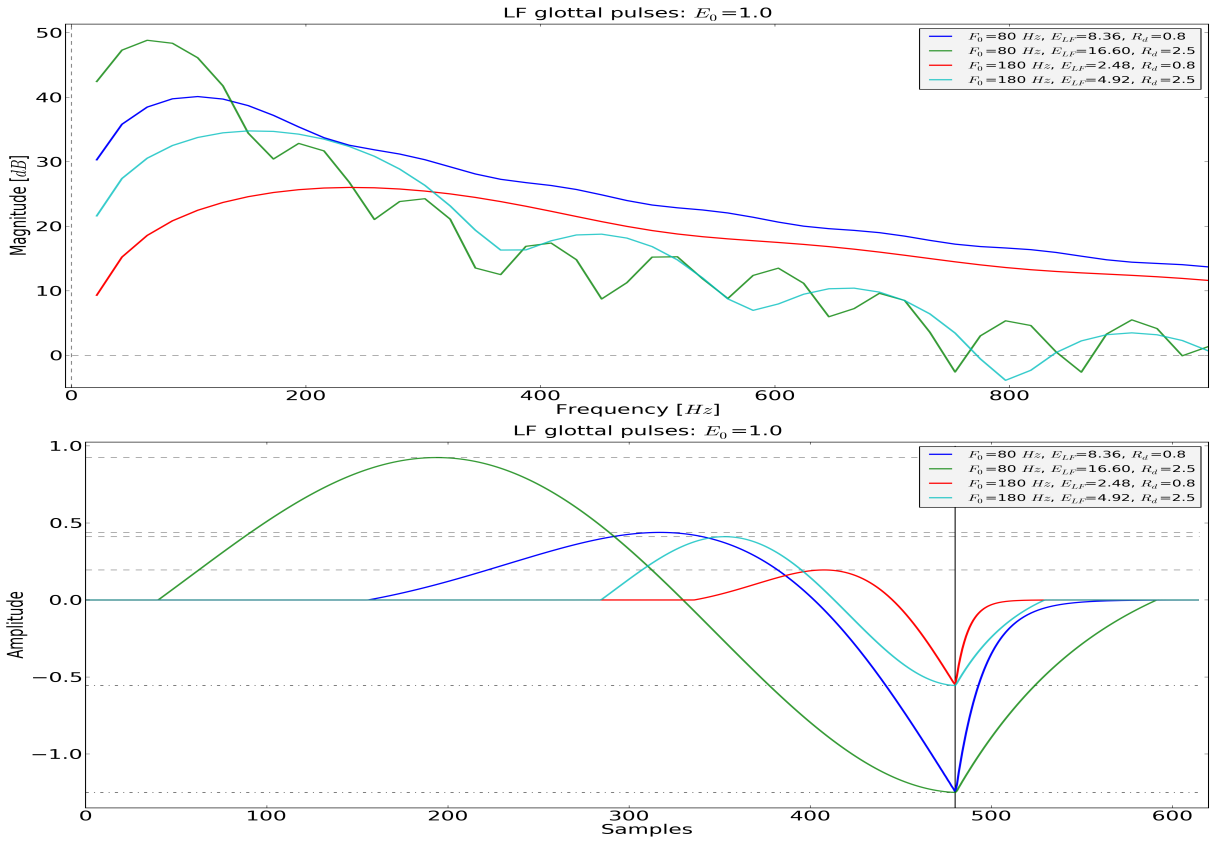


Figure 6.20: *Synthesized glottal pulse examples in the spectral (above) and the time (below) domain*

Figure 6.20 depicts four different glottal pulses, synthesized with the LF model in time and spectral domain. All four glottal pulses are synthesized with a LF amplitude scaling factor of E_0=1.0. The two frequencies 80 and 180 Hz are utilized along with the two R_d values 0.8 and 2.5. The grey dashed horizontal lines shown in the positive amplitude range in the time domain representation refer to the LF amplitude E_i. The grey dashed-dotted horizontal lines in the negative amplitude range in the time domain representation refer to the LF amplitude E_e. The grey solid vertical line in the time domain representation refers to GCI time instant t_e. All glottal pulses are aligned in time at t_e for comparison purposes. The corresponding RMS energy measures E_{LF} are shown in fig. 6.20 for each of the four glottal pulses. Each E_{LF} measure is conducted on the same DFT spectrum size N=4096. Naturally, lower

F_0 frequencies lead to longer T_0 periods. This causes higher values in the energy measure E_{LF}. The drawback is that the energy measures for the higher R_d values are higher than its corresponding lower R_d values for the same frequency. However, higher R_d values should be associated with lower energy values.

Please note that the discussion presented here treats the theoretic analysis of synthesizing the LF model over the R_d regression parameter space. A continuative investigation discusses the findings of RMS-based energy values over the whole R_d range measured on natural human speech in section 6.6.3. These findings on natural speech differ from the theoretical analysis presented in this section. Further interesting analyses on natural human speech in [Fant and Kruckenberg, 1996, Fant et al., 2000, Fant and Kruckenberg, 2005] use the LF model in the context of supra- and sub-glottal pressure to evaluate the co-variation of intensity, sound pressure level (SPL), F_0, E_e and U_0. Some empiric interrelations on the analysed data set could be determined. However, the studies conclude that its findings may be biased e.g. on speaker dependency and speaking style. Another paper in [Alku et al., 2002b] studies in the same context the correlation of vocal intensity and F_0. A loud, flow and weak voice quality or phonation type can be associated with high, normal or low sub-glottal pressure [d'Alessandro, 2006].

Trying to establish a formula to reflect the required changes in amplitude or energy which generalizes over different speakers and speaking styles appears to be almost impossible. Too many dependencies on vocal effort and intonation, interferences from articulatory processes, differences in speaker identity and other influences just permit a data modelling based solution to better approximate the required energy changes implied with common Voice Transformation and Voice Conversion tasks. Moreover, the energy behaviour of the underlying LF model is contradictory to the R_d border hypothesis of section 6.2.1.3.

Therefore, two energy models are implemented in *PSY* to control energies changes. The following section 6.4.1.2 introduces a simple and robust method to maintain the energy required to properly execute Voice Transformation and Voice Conversion applications. The advanced GMM-based methodology of section 6.4.2.3 seeks to reflect the underlying energy data measured for each analyzed speaker.

6.4.1.2 Energy maintenance

A very simple and robust procedure to correct undesired energy changes introduced by transforming or converting a given speech phrase is to re-establish the energy contour of the original speech signal. In general, this implies in *PSY* to either re-scale the complete re-constructed signal $s'(n)$ with the energy E'_{sig} to the original energy E_{sig} measured at the input signal $s(n)$. Or the energies E'_{voi} and E'_{unv} measured on the transformed voiced $V'(\omega)$ and transformed unvoiced $U'(\omega)$ components are used for re-scaling. Please note that also a conversion denoted with operator $''$ can be handled in the same manner. It is dropped here for more efficient illustration purposes. The equations of 6.20 define the energy re-scaling for the whole signal:

$$\begin{aligned} E_{sig} &= F_{RMS}(|S(\omega)|) \\ E'_{sig} &= F_{RMS}(|S'(\omega)|) \\ S'(\omega) &= (E_{sig}/E'_{sig}) \cdot S'(\omega) \end{aligned} \tag{6.20}$$

The equations of 6.21 for the voiced component:

$$\begin{aligned} E_{voi} &= F_{RMS}(|V(\omega)|) \\ E'_{voi} &= F_{RMS}(|V'(\omega)|) \\ V'(\omega) &= (E_{voi}/E'_{voi}) \cdot V'(\omega) \end{aligned} \tag{6.21}$$

The equations of 6.22 for the unvoiced component:

$$\begin{aligned} E_{unv} &= F_{RMS}(|U(\omega)|) \\ E'_{unv} &= F_{RMS}(|U'(\omega)|) \\ U'(\omega) &= (E_{unv}/E'_{unv}) \cdot U'(\omega) \end{aligned} \tag{6.22}$$

The different energy re-scaling methods maintaining the original energy of the input signal applied in *PSY* are explained in chapter 6.5 for different synthesis schemes.

6.4.2 GMM-based contour prediction

The iterative approximation of the Expectation Maximization (EM) algorithm as an open form solution to train Gaussian Mixture Models (GMM) facilitates the modelling of huge data sets such as one or several speech corpora.

The *PSY* speech system employs different GMMs being introduced in the following sub-chapters. Each of the GMMs is based on the data prediction model introduced in section 5.5.2. A similar or same set of voice descriptors D as presented in section 5.5.1 is used for the following models. The models are based on the assumption that the features describing a speech signal underlie a certain co-variation among each other. The co-variation is a result of the fact that each of the selected voice descriptors relates to the same physiological process to generate speech according to the model of human voice production. The selection of a voice descriptor set D is based on the correlation measured between the contours of each descriptor. Only voice descriptors which exhibit a relatively high correlation with the reference value to be predicted are considered. The descriptor contours are collected over the complete corpus data for one speaker. This establishes a model exploiting the correlation present between sufficiently enough co-varying voice descriptors.

6.4.2.1 Generic GMM-based contour modelling

The GMM implementation to predict any specific voice descriptor curve in *PSY* will be presented in the following sections. All follow the same generic principle explained in this section. The basic GMM modelling follows the approach already detailed in section 5.5.2 and in [Lanchantin and Rodet, 2010, Lanchantin and Rodet, 2011]. A GMM model $\mathcal{M}$ is trained on a set of chosen voice descriptor features D and one reference feature R. The GMM training is defined by equ. 5.22. A prediction function F defined by equ. 5.23 is derived from a trained GMM model $\mathcal{M}$. A GMM-predicted voice descriptor contour R^p is computed from a prediction function $F(d)$ using a voice descriptor feature set D as input.

Machine learning and data prediction algorithms may suffer from common statistical modelling drawbacks. In general this can be expressed by the modelling error ϵ_M. This error can be decomposed into a bias B and a variance V. The following paragraph discusses shortly the bias-variance decomposition trade-off [Geman et al., 1992]. Any prediction of a true target contour T by any model $\mathcal{M}$ reflects an approximation of the underlying data pattern. The model predicts a target contour $\hat{T}$ and is influenced by bias B and variance V.

The bias B:
It constitutes a systematic error due to the model deficiency and is independent of the data. The model $\mathcal{M}$ can be biased because its structure does not allow to fit the data D. Any statistical model not being properly designed for the underlying data pattern may introduce another source of approximation error. It prevents a perfect re-construction of the data D by a model $\mathcal{M}$. The error ϵ_M may thus originate from modelling mistakes introducing a systematic bias B, like a non-perfectly chosen amount Q of Gaussian densities [Lanchantin and Rodet, 2011], an unfortunate initialisation of the Gaussian component priors [Dognin et al., 2009], the GMM approximation of the underlying data is stuck in a local instead of the global optimum [Paalanen et al., 2005], or the estimated mixture components overlap and do not properly reflect the underlying data [Ververidis and Kotropoulos, 2008].

The variance V:
It relates to noise in the data, sparsely structured data, data of uncorrelated nature etc. The precise re-production of the data pattern D from the model $\mathcal{M}$ is cumbersome and error prone. For example, the probability density function of each EM-fitted Gaussian normal distribution may not be able to properly match the underlying data. Examples of sparse data areas can be inspected by the data examination of the voice descriptors estimation of three different speech corpora, presented in section 6.6.3.

Approximation and prediction errors are related to the over-fitting of the model $\mathcal{M}$ on the data D creating bias B, and / or the under-fitting of the data D to the the model $\mathcal{M}$ introducing the variance V. The bias-variance problem leads to the drawback that the statistical model does not generalize well over the data.

An error GMM $\mathcal{M}_{err}$ is trained on ϵ_M to minimize possible prediction errors introduced by a modelling error ϵ_M. The modelling error ϵ_M is measured by means of a squared error operation as

$$\epsilon_M = \sqrt[2]{(R - R^p)^2}. \tag{6.23}$$

A voice descriptor set D is employed as input together with a reference feature R to train a standard model $\mathcal{M}$. The prediction function $F(d)$ uses the same input voice descriptor set D to compute its predicted voice descriptor contour R^p_μ, described in equ. 6.25. The latter serves as measure to evaluate the modelling error ϵ_M. The original reference feature R is replaced by ϵ_M to form together with the still same input voice descriptor set D the conjoint data set Z_{err} as

$$Z_{err} = \{z_k\}, z_k = [d_k^T \epsilon_k^T]^T, \tag{6.24}$$

with K expressing the data set length. An error GMM $\mathcal{M}_{err}$ is trained on the conjoint data set Z_{err} to derive an error prediction function $F_{err}(d)$ according to equ. 5.23. The GMM-based modelling to predict a transformed voice

descriptor contour R'^p from a transformed voice descriptor set D' is designed in *PSY* as follows:

$$R_\mu^p = \mathcal{M}(F(d)) \tag{6.25}$$

$$R_\mu^{'p} = \mathcal{M}(F(d')) \tag{6.26}$$

$$R_\sigma^p = \mathcal{M}_{err}(F_{err}(d)) \tag{6.27}$$

$$R_\sigma^{'p} = \mathcal{M}_{err}(F_{err}(d')) \tag{6.28}$$

$$r' = R_\mu^{'p} + (r - R_\mu^p) \cdot \frac{R_\sigma^{'p}}{R_\sigma^p} \tag{6.29}$$

Each trained model pair $\mathcal{M}$ and $\mathcal{M}_{err}$ is utilized to predict via their derived prediction functions F and F_{err} the "mean" prediction value R_μ^p and the predicted "standard deviation" R_σ^p from the "true" prediction value introduced by the model error ϵ_M. The "true" prediction value would equal R_μ^p if no model error occurs: ϵ_M=0. The calculation of a transformed reference value r' from a transformed and the original voice descriptor set d' and respectively d is defined by equ. 6.29. It evaluates the difference between the original reference feature value r and the from its corresponding original voice descriptor set d predicted mean value R_μ^p, normalized by the ratio of the original and transformed standard deviations R_σ^p and $R_\sigma^{'p}$, and corrected by the transformed predicted mean value $R_\mu^{'p}$.

The energy contours E_{voi}^p and E_{unv}^p of a speech phrase shown in section 6.4.2.3, the F_{VU}^p contour described in section 6.4.2.4, or the R_d^p contour presented in section 6.4.2.5 are predicted by the means defining equation 6.29. Please note that the terminology "up to all" used for the explanation of each GMM model in the following refers to the fact that each listed voice descriptor is evaluated on its correlation with the reference value R for the given speaker. If the voice descriptor exhibits a comparably low correlation with the reference value R to be predicted, it is not used in the GMM training and later prediction for the corresponding test case. This will be presented in the evaluation section 6.6.

6.4.2.2 Voice descriptor selection

The following three sections explain the algorithmic setup how to design GMM models for the prediction of the RMS energy in section 6.4.2.3, the F_{VU} in 6.4.2.4 and R_d in section 6.4.2.5. Each model employs a voice descriptor set of original D and transformed D' features. Each voice descriptor presented in the following chapters is selected as feature for GMM modelling since it proved to be sufficiently high correlated with the to be predicted reference feature for some speakers. The correlation is evaluated by the Pearson r coefficient [Pearson, 1900] between both features over the whole corpus per speaker. A high correlation is given by high absolute values of r since r=-1 expresses a perfect anti-correlation between two feature contours. However, in practice one feature pair may not exhibit a strong correlation for one speaker. The utilization of this voice descriptor should be disabled in this case. It is up to the user of *PSY* to carefully select a voice descriptor set of higher inter-correlation given the underlying data set of a speaker.

The theoretical foundation for the voice descriptors $H1$-$H2$, F_0, F_{VU} and R_d has been discussed in section 5.5.1. The correlation of energy related descriptors and R_d is discussed in section 6.4.1.1. This paragraph presents further justification for the usage of the voiced E_{voi} and unvoiced E_{unv} RMS energy as co-varying feature to the other voice descriptors, without the relation to R_d. The sound level and the loudness of phonation are found in [Gramming et al., 1988] to accompany fundamental frequency changes in singing. The interdependencies of sub- and supra-glottal pressure, intensity, glottal source parameters and F_0 is studied in [Fant et al., 2000]. The co-variation of F_0 with speech intensity in terms of sub-glottal pressure and the Sound Pressure Level (SPL) is discussed in [Fant and Kruckenberg, 2007]. Intensity is correlated to OQ, $H1$-$H2$ and the spectral tilt, vocal effort to F_0 [Liénard and Barras, 2013]. A linear regression of F_0 in logarithmized Hz and energy in dB in [Sorin et al., 2015] exhibits a high correlation between both features. It proofs the fundamental relationship between instantaneous pitch and instantaneous energy of a signal. The pitch-energy relationship is used to modify the instantaneous energy coherently with changes applied to the instantaneous frequency. The study was conducted within the context of a Voice Transformation to create an expressive TTS system.

6.4.2.3 GMM energy models

Three GMM-based energy models exist in *PSY* according to the three energy maintenance methods defined with equations 6.20, 6.21, and respectively 6.22 in section 6.4.1.2. An original voice descriptor set D_E for the energy modelling in *PSY* may consist of up to all considered voice descriptors: D_E=[R_d, F_0, F_{VU}, $H1$-$H2$]. A transformed voice descriptor set D'_E, denoted by the operator $'$, may contain the original F_0 contour but transformed values

for up to all the remaining voice descriptors: $D'_E = [R'_d, F_0, F'_{VU}, H'1\text{-}H'2]$. The voice quality transformation of an original R_d to a transformed R'_d contour according to the means explained in section 6.4.3 causes in the output signal $s'(n)$ transformed F'_{VU} and $H'1\text{-}H'2$ curves. The transformed contour of $H'1\text{-}H'2$ is measured in the magnitude spectrum of $|S'(\omega)|$ in dB. The predicted F'_{VU} value is retrieved from the signal $s'(n)$ and the GMM model introduced in the following section 6.4.2.4. The F_0 voice descriptor is added to the energy modelling due to its high co-variation with the other selected voice descriptors. However, no means have yet been implemented in *PSY* to transform the F_0 contour of the original speech recording. The manually transformed R'_d, the re-estimated $H'1\text{-}H'2$, the predicted F'_{VU} and the original F_0 contour define the transformed voice descriptor set D'_E. Each energy model receives for training its corresponding reference feature R defined in equ. 6.19. The reference feature R can be either E_{sig}, E_{voi} or E_{unv}.

The energy models $\mathcal{M}^{voi}$ and $\mathcal{M}^{unv}$ are used via their prediction functions F^{voi} and F^{unv}, along with their corresponding error models $\mathcal{M}^{voi}_{err}$ and $\mathcal{M}^{unv}_{err}$ and error functions F^{voi}_{err} and F^{unv}_{err}, to predict the RMS-based energy measures E^p_{voi} and E^p_{unv}, according to equ. 6.29. Examples of different original and transformed voice descriptor sets D_E and D'_E to utilize the RMS-based energy model of *PSY* to transform the voice quality of natural human speech recordings are given in the evaluation section 6.6.3.

6.4.2.4 GMM F_{VU} model

Any F_{VU} estimation is performed in *PSY* by means of calling offline the SuperVP system of section 2.5.2 for a complete speech phrase. However, no means to conduct a robust frame-based F_{VU} estimation on single STFT frames of $S'(\omega)$ is available in *PSY*. The GMM models $\mathcal{M}^{F_{VU}}$ and $\mathcal{M}^{F_{VU}}_{err}$ are therefore established in *PSY* to predict a transformed F'_{VU} contour by frame-based means. The F'_{VU} contour prediction uses the voice descriptor sets $D_{F_{VU}}$ describing the input signal $s(n)$ and $D'_{F_{VU}}$ describing the transformed signal $s'(n)$. It is based on the same means described in the preceding sections 6.4.2.1 and 6.4.2.3. An original voice descriptor set $D_{F_{VU}}$ for the F_{VU} prediction may consist in *PSY* of up to all voice descriptors listed in the following: $D_{F_{VU}}$=$[R_d, F_0, H1\text{-}H2, E_{voi}, E_{unv}]$. Its transformed voice descriptor counterpart $D'_{F_{VU}}$ contains again the original F_0 contour but transformed values for the remaining voice descriptors: $D'_{F_{VU}}$=$[R'_d, F_0, H'1\text{-}H'2, E'_{voi}, E'_{unv}]$. The transformed $H'1\text{-}H'2$ values are measured on the transformed signal $s'(n)$. Please note that the RMS energy values for the transformed voiced component E'_{voi} and the transformed unvoiced component E'_{unv} are re-measured on the transformed signal $s'(n)$. *PSY* could on the other hand handle the prediction of E'^{p}_{voi} and E'^{p}_{unv} using the energy models described in the preceding section 6.4.2.3 and use the predicted energies instead. Examples of different original and transformed voice descriptor sets $D_{F_{VU}}$ and $D'_{F_{VU}}$ to predict a transformed F'_{VU} contour are again given in section 6.6.3 on three speaker examples of natural human speech. The F'_{VU} contour prediction is used in *PSY* predominantly for the spectral fading synthesis variant described in section 6.5.5.

6.4.2.5 GMM R_d model

The GMM modelling to predict a R_d contour according to a given data set is implemented in *PSY* solely to reflect the characteristic R_d contour of the target speaker in the VC context. It does not for the time being utilize the additional error correction of the GMM model $\mathcal{M}_{err}$ described by equ. 6.28 in section 6.4.2.1. Only the basic GMM modelling means defined by equ. 6.25 are utilized. The GMM model $\mathcal{M}^{R_d}$ is trained on the target speakers database. It represents the acoustic space of the voice descriptor set D_{R_d} of the target speaker. The voice descriptor set D_{R_d} considers the following voice descriptors extracted from the source speakers speech phrase chosen for conversion: D_{R_d}=$[F_0, F_{VU}, H1\text{-}H2, E_{voi}, E_{unv}]$. The predicted R'_d contour reflects the conditioning of the source speakers data set D_{R_d}, given the target speakers GMM model $\mathcal{M}^{R_d}$.

6.4.3 Voice quality transformation

The extended set of features tackled in the VC process requires some extensions of the existing Voice Transformation algorithms. Recent algorithms have established means for real time transformation of the perceived gender and age of a speaker [Lanchantin et al., 2011a], or perceived fine voice quality details such as jitter and shimmer reflecting time and amplitude modulations [Bonada, 2004]. The fine control of the voice quality features that is required for transformation into a well defined target speaker requires extended means for coherent transformation of the excitation source characteristics. The results will not only be beneficial for the VC problem, but will in general increase the potential transformations that are available for Voice Transformation and Voice Conversion.

The theoretical basis of voice quality has been discussed in section 3.5. Open questions when transforming the

voice quality of natural human speech are
a) to what detail listeners do perceive shape differences of the deterministic part, and energy differences of the stochastic part of the glottal excitation source,
b) in what technical manner the glottal excitation source shall be transformed such that it reflects the natural behaviour of a human speaker, and
c) how such transformation shall be treated such that a separated processing of the voiced deterministic and the unvoiced stochastic component results into a coherently transformed glottal excitation source.

The construction of a speech phrase within the context of VC shall as much as possible be described by voice descriptors of the target speaker. This could include the adaptation of the R_d mean difference between source and target speaker. Concerning **a)**, if this R_d mean difference is below the Just Noticable Difference (JND) of human auditory perception it won't be perceivable. The same is valid when altering the R_d of the same speech phrase in the context of Voice Transformation. Care has to be taken that the applied R_d contour modification is perceptible. Following the discussion about JND in section 3.5.3, it is practically very cumbersome to establish a well-defined grid of JND measurements covering all involved synthesis parameters and its required changes in different step sizes. As well the evaluation of such a huge perceptual tests may lead to the conclusion that any JND concerning any synthesis parameter is listener dependent. This assumption was already indicated in the study of [Henrich et al., 2003] discussed in section 3.5.3. It reports different results between the trained and the untrained listener group. This indicates a speaker dependency for each JND. The listening tests conducted in sections 6.6.4 and 6.6.5 examine if the participants could perceive differences among the different re-synthesized speech phrases for which the original R_d^{gci} contour has been transformed.

The studies in [Henrich et al., 2003, van Dinther et al., 2004] changed per test stimuli exclusively either OQ or α_m. In contrast, the alteration of the original R_d contour or any R_d value implies a change effectuated in the corresponding values of OQ, α_m and ta. The relative JND $\Delta R_d / R_d$ values required to excite the perceptual sensation of a voice quality change lie therefore lower for R_d compared to the single value changes required for either OQ or α_m.

The method reported in [van Dinther et al., 2004] provides means to systematically quantify the relation between the R waveshape parameter space of the LF model and different voice quality characteristics. The following two sections describe different means to transform the voice quality of a speech phrase within *PSY*. It concerns only the voiced component of a speech signal by means of altering the original R_d^{gci} contour. Both follow a simple empiric strategy considering the questions **a)** and **b)**. The tests presented in the sections 6.6.4.2 and 6.6.4.3 address with the additional GMM-based energy prediction for the voiced and the unvoiced parts question **c)**.

6.4.3.1 Simple R_d^{gci} offsets

An intended change in voice quality can be user driven. Care has to be taken that any R_d^{gci} contour transformation lies for each R_d value above the given JND present in its current parameter region. Too small R_d value changes would not result into the desired perceptual sensation of a change in voice quality. A simple empiric solution is to add to or subtract from the original R_d^{gci} contour a reasonably high R_d offset above the JNDs. The listening test presented in section 6.6.4 examines as a proof-of-concept this attempt on natural human speech recordings of two French speakers.

6.4.3.2 R_d^{gci} contour transformation

The study on JND thresholds of [Henrich et al., 2003] discussed in section 3.5.3 shows that smaller (higher) changes of ΔOQ ($\Delta\alpha_m$) are required) in lower Open Quotient OQ (higher asymmetry coefficient α_m) value regions. Likewise, higher changes (lower changes) are required in higher (lower) value regions of OQ (α_m). In the following, these experimental findings will be entitled as JND threshold objective. A transformation of the original R_d^{gci} contour into several $R_d^{'gci}$ contours in positive R_d value direction towards a more relaxed and in negative R_d value direction towards a more tense voice quality to cover the complete R_d range is implemented in *PSY*. The partitioning over the R_d range adds to or subtracts from the original R_d^{gci} contour mean offsets, and expands or compresses each transformed $R_d^{'gci}$ contour in its variance. The semi-automatic $R_d^{'gci}$ contour transformation is summarized by pseudo-algorithm 5.

The amount of K offsets span over the OQ and R_d range is defined by n positive and m negative OQ offsets: $K = n + m$. The logarithmic offset generation on the OQ scale at step III. causes the desired non-linear spacing in the generated $R_d^{'gci_K}$ matrix on the R_d scale. The difference between each $k \cdot R_d$ offset subtracted from and added to the original R_d^{gci} contour increases with k. It results into progressively lower mean R_d^{μ} value differences between

Algorithm 5 - Semi-automatic R_d^{gci} contour transformation

I. $\boldsymbol{R_d^{gci}}$**->**$\boldsymbol{OQ^{gci}}$:
Convert the original R_d^{gci} contour to OQ^{gci} on the OQ scale
II. $\boldsymbol{OQ_\mu}$:
Compute the mean OQ_μ of the original OQ^{gci} contour
III. $\boldsymbol{OQ_\mu}$ **offsets**:
Compute n (m) positive (negative) logarithmically spaced OQ_μ offsets
in range $[OQ_\mu\ OQ_{max}]$ ($[OQ_{min}\ OQ_\mu]$)
IV.
for $\boldsymbol{OQ_\mu^k}$ in ($\boldsymbol{OQ_\mu^K}$ **offsets**): **do**
$\boldsymbol{OQ_k^{'gci}}$ **offsets**: Add (subtract) the OQ_μ^k offset to (from) OQ_k^{gci}
$\boldsymbol{OQ_k^{'gci}}$ **soft saturate**: Restrict each $OQ_k^{'gci}$ contour to OQ range $[OQ_{min}\ OQ_{max}]$ via soft saturation
$\boldsymbol{OQ_k^{'gci}}$**->**$\boldsymbol{R_d^{'gci_k}}$: Convert the soft saturated $OQ_k^{'gci}$ contour to $R_d^{'gci_k}$ on the R_d scale
end for

each generated $R_d^{'gci}$ in lower R_d regions. Contrariwise, higher mean R_d^μ value differences between each $R_d^{'gci}$ curve are created in higher R_d regions. The same behaviour applies to the variance R_d^σ described by each $R_d^{'gci}$ contour. R_d^{gci} contour transformation works according to the JND threshold objective.

The introduced $R_d^{'gci}$ contour transformation to span over the R_d range is semi-automatic since care has to be taken concerning mean and variance of the original R_d^{gci} contour. If it lies close to one R_d range border either the number of the n positive or of the m negative OQ offsets has to be adapted such that the transformed $R_d^{'gci}$ contours are well spread over the R_d range.

It is not assured that the transformed $R_d^{'gci}$ contours remains within the R_d range. A hard saturation on the R_d range borders would destroy the natural variation of the $R_d^{'gci}$ contours. A soft saturation restricts any OQ or R_d contour to its range in PSY by applying a soft compression of the parameter contour inside and outside its range borders. The sigmoid function $F_{sig}(t)$ defined by equ. 6.30 can be parameterized with the strength factor p_{sig} defining different soft compression levels.

$$F_{sig}(t) = \frac{1}{1 + e^{-t \cdot p_{sig}}} \tag{6.30}$$

Higher compression levels of the sigmoid function would result in a strong deformation of a soft saturated contour especially in the middle of the OQ or R_d range, not being shown in fig. 6.21. Therefore a linear function substitutes the curve described by the sigmoid function in the inner valid range to minimize the deformation. An example of different soft saturation contours warping a transformed OQ' contour into its range is given in fig. 6.21. The dashed

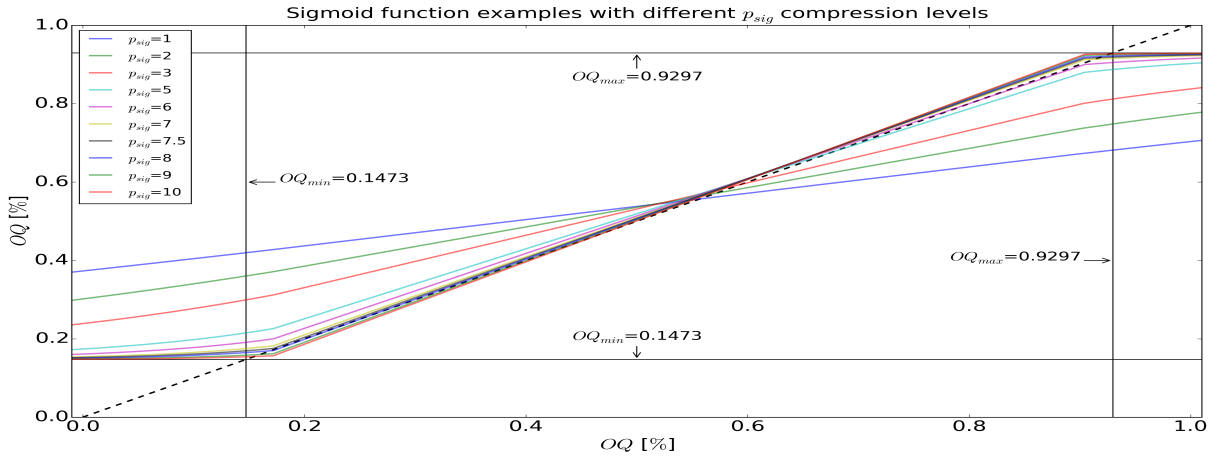


Figure 6.21: *Example of different soft saturation contours*

black line exemplifies an ideal linear line without any deformation. The connection points of the substituted linear lines with the sigmoid functions can be inspected short after the vertical line of OQ_{min} and short before the vertical

line of OQ_{max}. The default soft saturation is set to p_{sig}=7.

Different examples of the computation of transformed $R_d^{'gci}$ contours over the R_d range will be given in the evaluation chapter 6.6.5. Lower $R_d^{'gci}$ contours correspond to a more tense voice quality described by lower R_d^{μ} and R_d^{σ} values. Higher $R_d^{'gci}$ contours correspond to a more relaxed voice quality described by higher R_d^{μ} and R_d^{σ} values. The tables 6.10, 6.14, and 6.17 exemplify the relation between the different voice qualities and the different R_d^{μ} and R_d^{σ} values for the three speakers under evaluation.

6.5 Synthesis

Several synthesis approaches have been implemented and studied in *PSY* to execute advanced Voice Transformation and Voice Conversion tasks. Each synthesis variant is based on the extended source-filter model and the DSM-based unvoiced component estimation. The separate processing of the voiced component $V(\omega)$ and the unvoiced component $U(\omega)$ operates on the multiplication of a smooth spectral envelope sequence with an excitation signal.

6.5.1 Short-Time Fourier Transform and Overlap-Add

Both synthesis and some parts of the analysis in *PSY* are based on the well-known Short-Time Fourier Transform (STFT) [Griffin and Lim, 1983, Griffin and Lim, 1984]. The synthesis is realized on the simple Overlap-Add technique of single STFT buffers in the time domain [Allen, 1977]. The influence of the Hanning window applied in the STFT analysis and synthesis steps is normalized before the final synthesis. *PSY* operates for the time being on a constant time basis with a constant STFT step and window size in analysis and synthesis. The reason being that the current energy model in *PSY* introduced in section 6.4.1 cannot handle pitch-adaptive window lengths. Future work in *PSY* on an improved energy handling indicated in section 8.2.2 and pitch-adaptive processing indicated in section 8.2.4 should improve the flexibility and synthesis quality of *PSY*.

6.5.2 Noise excitation

Any unvoiced component $U(\omega)$ is synthesized by the means presented in 6.3.2.1. The multiplication of the STFT of a white noise signal with any spectral envelope sequence $\mathcal{T}_{unv}$ generates the unvoiced waveform $u(n)$ in the time domain. $\mathcal{T}_{unv}$ may be estimated on any of the four in *PSY* possible unvoiced signals $u^{thron}_{ReDe}(n)$, $u^{HP}_{ReDe}(n)$, $u^{thron}_{QHM}(n)$ or $u^{HP}_{QHM}(n)$, being close to the signal level as described in section 6.3.2.4.

6.5.3 Pulse excitation

An initial version $V^{dit}_{\delta}(\omega)$ of $V(\omega)$ is implemented in *PSY* for comparison purposes. The spectral envelope $\mathcal{T}_{sig}$ of the input signal $s(n)$ is excited by the flat amplitude spectrum $\Delta_{F_0}(\omega)$ of a Dirac delta function $\delta_{F_0}(n)$, parameterized by F_0. This well-known synthesis method exhibits problems in properly modelling the phases of the pulses since the glottal pulse shape is contained in $\mathcal{T}_{sig}$. Each estimated GCI location of $\delta_{F_0}(n)$ is shifted to the closest sample bin in the time domain which neglects a possible sub-sample offset between the GCI estimated in seconds and the sample bin. The same applies to each GCI estimation of each synthesized glottal flow derivative $g^{gci}_{R_d}$. Two versions of the purely voiced component $V(\omega)$ exist according to the two VTF versions $C_{F_{VU}}(\omega)$ and $C_{full}(\omega)$ introduced in section 6.2.3 in *PSY*. Both $V(\omega)$ versions will be introduced in the following.

6.5.3.1 Full-band excitation

The voiced version $V_{full}(\omega)$ takes the Vocal Tract Filter $C_{full}(\omega)$ and excites it by the spectral representation $G_s(\omega)$ of a few consecutive glottal pulses $g_s(n)$ being windowed by the STFT. The latter presents a sequence of glottal flow derivatives $g^{gci}_{R_d}$ synthesized at each estimated GCI time instant in the time domain. Since no splitting at the F_{VU} frequency was implied to extract $C_{full}(\omega)$ from the speech signal $s(n)$, the multiplication of excitation and filter part in the spectrum operates full-band without any restriction for $V_{full}(\omega)$.

Both VTF versions, $C_{full}(\omega)$ and $C_{F_{VU}}(\omega)$, describe a filter whose frequency response operates exclusively in the linear amplitude spectrum $|A_{lin}(\omega)|$. Their frequency response $H(e^{j\omega T})$, evaluated on the unit circle in the complex z-plane by $z = e^{j\omega T}$, is decomposed into the magnitude response $H|(e^{j\omega T})|$ and phase response $\angle H(e^{j\omega T})$, with the latter being set to zero: $\angle H(e^{j\omega T}) = 0 \; \forall \; \omega$. The pseudo-algorithm 6 explains the two simple steps to compute $V_{full}(\omega)$. Please note that the windowing of a few consecutive glottal pulses $g_s(n)$ synthesized in the time domain and transformed to the spectral domain by means of the STFT reproduces in $G_s(\omega)$ the underlying harmonic periodicity. The latter is given by the sequence of fundamental periods generated by each synthesized glottal pulse at the GCIs in the time domain.

Algorithm 6 - Voiced deterministic full-band version $V_{full}(\omega)$ without spectral splitting

I. Zero to minimum phase: $C_{full}(\omega) \rightarrow C^{-}_{full}(\omega)$
Convert the zero phase spectrum $C_{full}(\omega)$ to its minimum phase equivalent $C^{-}_{full}(\omega)$ following the means explained in section 2.2.3.
II. Excitation: $V_{full}(\omega) = C^{-}_{full}(\omega) \cdot G_s(\omega)$
Excitation in the spectral domain of $C^{-}_{full}(\omega)$ with $G_s(\omega)$ by a simple multiplication.

6.5.3.2 Excitation split at F_{VU}

The re-construction of the purely voiced component $V_{F_{VU}}(\omega)$ from the VTF $C_{F_{VU}}(\omega)$ and the contribution of the glottal excitation source $G_s(\omega)$ implies the consideration of the spectral splitting at the F_{VU} frequency. The excitation by $G_s(\omega)$ as with $V_{full}(\omega)$ cannot be applied over the full spectral band and is only valid in the lower frequency band below F_{VU}. The underlying harmonic periodicity above F_{VU} is therefore reproduced by exciting $C_{F_{VU}}(\omega)$ with the spectral representation $\Delta_{F_0}(\omega)$ of a Dirac delta impulse sequence $\delta_{F_0}(n)$ in the higher frequency band above F_{VU}.

Algorithm 7 - Voiced deterministic version $V_{F_{VU}}(\omega)$ with spectral splitting at F_{VU}

I. Zero to minimum phase: $C_{F_{VU}}(\omega) \rightarrow C^{-}_{F_{VU}}(\omega)$
Convert the zero phase spectrum $C_{F_{VU}}(\omega)$ to its minimum phase equivalent $C^{-}_{F_{VU}}(\omega)$
II. a) Excitation 1: $V^{lo}_{F_{VU}}(\omega{<=}\omega_{VU}) = C^{-}_{F_{VU}}(\omega{<=}\omega_{VU}) \cdot G(\omega{<=}\omega_{VU})$
Spectral excitation of VTF $C^{-}_{F_{VU}}(\omega{<=}\omega_{VU})$ with glottal source $G(\omega{<=}\omega_{VU})$
II. b) Excitation 2: $V^{hi}_{F_{VU}}(\omega{>}\omega_{VU}) = C^{-}_{F_{VU}}(\omega{>}\omega_{VU}) \cdot \Delta^{F_0}_{dit}(\omega{>}\omega_{VU})$
Spectral excitation of VTF $C^{-}_{F_{VU}}(\omega{>}\omega_{VU})$ with dirac impulse $\Delta_{F_0}(\omega{>}\omega_{VU})$
III. Spectral concatenation: $V_{F_{VU}}(\omega) = [V^{lo}_{F_{VU}}(\omega{<=}\omega_{VU}) \;\; (V^{hi}_{F_{VU}}(\omega{>}\omega_{VU}) + A_{diff})]$
Account for amplitude difference A_{diff} measured between $|V^{lo}_{F_{VU}}(\sim F_{VU})|$ and $|V^{hi}_{F_{VU}}(\sim F_{VU})|$ above F_{VU}, concatenate both parts
IV. Spectral interpolation:
Interpolate magnitudes and phases separately around the concatenation point F_{VU} to reduce possible discontinuities and unsteadiness

The voiced component $V_{F_{VU}}(\omega)$ is computed according to the pseudo-algorithm 7. The simple concatenation of algorithmic step III. requires to adjust for a possible amplitude difference A_{diff} found between the voiced components $|V^{lo}_{F_{VU}}(\sim F_{VU})|$ and $|V^{hi}_{F_{VU}}(\sim F_{VU})|$ in the lower and higher frequency band. The amplitude correction does not measure the amplitude difference A_{diff} at the DFT bin being closest to the F_{VU} frequency. Instead it searches for a sinusoidal maximum peak being closest to F_{VU} in both voiced components. A quadratic interpolation at both maximum peaks evaluates both amplitudes whose difference defines A_{diff}. The amplitude difference A_{diff} is added to the amplitude spectrum of $V^{hi}_{F_{VU}}(\omega{>}\omega_{VU})$ before the spectral concatenation.

6.5.4 Time domain: Simple mixing

The straight-forward mixing in the time domain adds one of the synthesized unvoiced waveform $u(n)$, listed in section 6.5.2 as $U(\omega)$, to one of the synthesized voiced deterministic waveform $v(n)$, listed in section 6.5.3 as $V(\omega)$. The time domain mixing operates thus full-band without any restriction. However, as will be shown in a first listening test shown in section 6.6.4, it is prone to create artefacts caused by the voiced component $V(\omega)$ for the case of a voice quality transformation synthesizing a tense voice character. Therefore, another synthesis variant operating in the spectral domain is presented in the following section 6.5.5.

6.5.5 Spectral domain: Fade unvoiced in - fade voiced out

The multiplication of the glottal pulse derivative $G(\omega)$ with the VTF $C(\omega)$ defines the voiced component $V(\omega)$. The transformation of the original R^{gci}_{d} contour used to extract $C(\omega)$ introduces an energy variation in the re-synthesis of a transformed $V'(\omega)$. However, even with the energy maintenance of section 6.4.1.2 the alteration of a modal to a "very tense" voice quality may lead to sinusoidal content being of higher energy than the noise part at the Nyquist

frequency $F_{nyq} = F_s/2$. This sets $F'_{VU} = F_{nyq}$ and causes audible artefacts. The spectral fading synthesis variant of *PSY* presented in this section is therefore designed to suppress such artefacts which possibly occur for glottal source shape parameter transformations.

Glottal source shape vs. spectral slope:
A short summary discusses here the strong correlation of the glottal source shape parameter R_d with the spectral slope. The discussion is required to understand the motivation for the spectral fading synthesis variant presented in this section. References to an extensive analysis of the spectral correlates of glottal excitation source parameters can be found in section 3.3.1 which introduces the R_d parameter. A more relaxed voice quality is reflected by higher R_d values and is related to a glottal flow derivative having a sinusoidal-like shape which generates steep spectral slopes. A more tense voice quality is parameterized by lower R_d values and is related to a glottal flow derivative having an impulse-like shape which originates flat spectral slopes. A flat (steep) spectral slope indicates that more (less) sinusoidal content can be observed in higher frequency regions. The voice quality transformation to change an original speech recording having a modal voice quality to a more tense voice character has to alter the spectral slope by means of extending the quasi-harmonic sequence of sinusoids above the F_{VU}. Contrariwise, a transformation to a more relaxed voice quality needs to reduce the sequence of quasi-harmonic sinusoids. Therefore, a modification of the glottal excitation source required for voice quality transformations implies an alteration of the Voiced / Unvoiced Frequency boundary F_{VU} to modify the spectral slope. The altered F_{VU} frequency has to be naturally represented by properly joining the voiced $V(\omega)$ and unvoiced $U(\omega)$ signal components.

A simple mixing in the spectral domain of any voiced $V(\omega)$ and unvoiced $U(\omega)$ component results in the same signal output as with the time domain mixing of the preceding section 6.5.4. A simple spectral concatenation of any voiced component $V(\omega<=\omega_{VU})$ in the lower deterministic band with any unvoiced component $U(\omega>\omega_{VU})$ in the higher stochastic band sounds reasonably well for the direct re-synthesis of the original speech recording without any further transformation or conversion. However, even for this re-synthesis case without modification the abrupt switchover at the F_{VU} does not reflect the signal characteristic observed with natural human speech. The spectral band around F_{VU} is comprised of a mix of both voiced deterministic $V(\omega)$ and unvoiced stochastic $U(\omega)$ parts. As already discussed in sections 2.3.2 and 6.2.3.2, the voiced component $V(\omega)$ has higher energy and perceptually masks therefore the unvoiced component $U(\omega)$ in the lower deterministic frequency band below F_{VU}. The inverse condition applies in the higher stochastic frequency band above F_{VU} where $U(\omega)$ has a higher energy and masks $V(\omega)$.

As already indicated in section 6.4.2.4, no robust frame-wise F_{VU} estimator is available in *PSY*. The GMM models $\mathcal{M}^{F_{VU}}$ and $\mathcal{M}^{F_{VU}}_{err}$ are therefore employed to predict a transformed F'_{VU} contour reflecting the changes implied with the voice quality transformation and VC tasks. The transformed F'_{VU} contour defines the cutoff frequency f_c of two linear filters. A low pass filter P_L fades out the voiced component $V(\omega)$ and a high pass filter P_H fades in the unvoiced component $U(\omega)$ with increasing frequency. The linear ramps with a slope of m_{LP}=-96 dB and m_{HP}=-48 dB per octave define the steepness of the low pass P_L and respectively the high pass P_H filter. A higher value is chosen for m_{LP} since the F'_{VU} prediction may be very high for "very tense" voice qualities. A less steep fade out filter would for such cases not be effective enough to suppress possibly occurring artefacts of the sinusoidal content at F_{nyq}.

A listening test presented in section 6.6.5 evaluates the spectral fading synthesis variant within the context of a voice quality transformation on natural human voice recordings of three different speakers.

6.5.6 *PSY* synthesis system layout

Fig. 6.22 illustrates the different system blocks of the synthesis part of the speech framework *PSY*. Most system blocks are required and some are optional to synthesize a *PSY* voice descriptor set into an audio waveform. A voice descriptor set D consists of the estimated features F_0, F_{VU}, R_d^{gci}, E_{voi} and E_{unv} for one speech phrase. It is required for synthesis in *PSY*, along with a sequence of Vocal Tract Filters $C(\omega)$ and spectral envelopes $T_{unv}(\omega)$ of the unvoiced component $U(\omega)$. The synthesis module can transparently handle a set of (partially) transformed voice descriptors D', and / or a converted VTF $C''(\omega)$ as well as a converted spectral envelope $T''_{unv}(\omega)$ sequence. The conversion resulting into $C''(\omega)$ and / or $T''_{unv}(\omega)$ will be described in chapter 7. Each of the transformed voice descriptors F'_{VU}, $R_d^{'gci}$, E'_{voi} and E'_{unv} can be predicted from its corresponding prediction model $\mathcal{M}^{F_{VU}}$, $\mathcal{M}^{R_d}$, $\mathcal{M}^{voi}$ and $\mathcal{M}^{unv}$, as explained in section 6.4.2. A transformed $R_d^{'gci}$ contour can additionally result from a voice quality transformation as introduced in section 6.4.3.2. The transformation of the F_0 contour along with the corresponding alteration of the GCI time instants is not yet implemented in *PSY*. The possible but not imperative utilization of a voice descriptor contour transformation is indicated by the dashed lines in fig. 6.22.

Block diagram illustrating the different synthesis parts in *PSY*

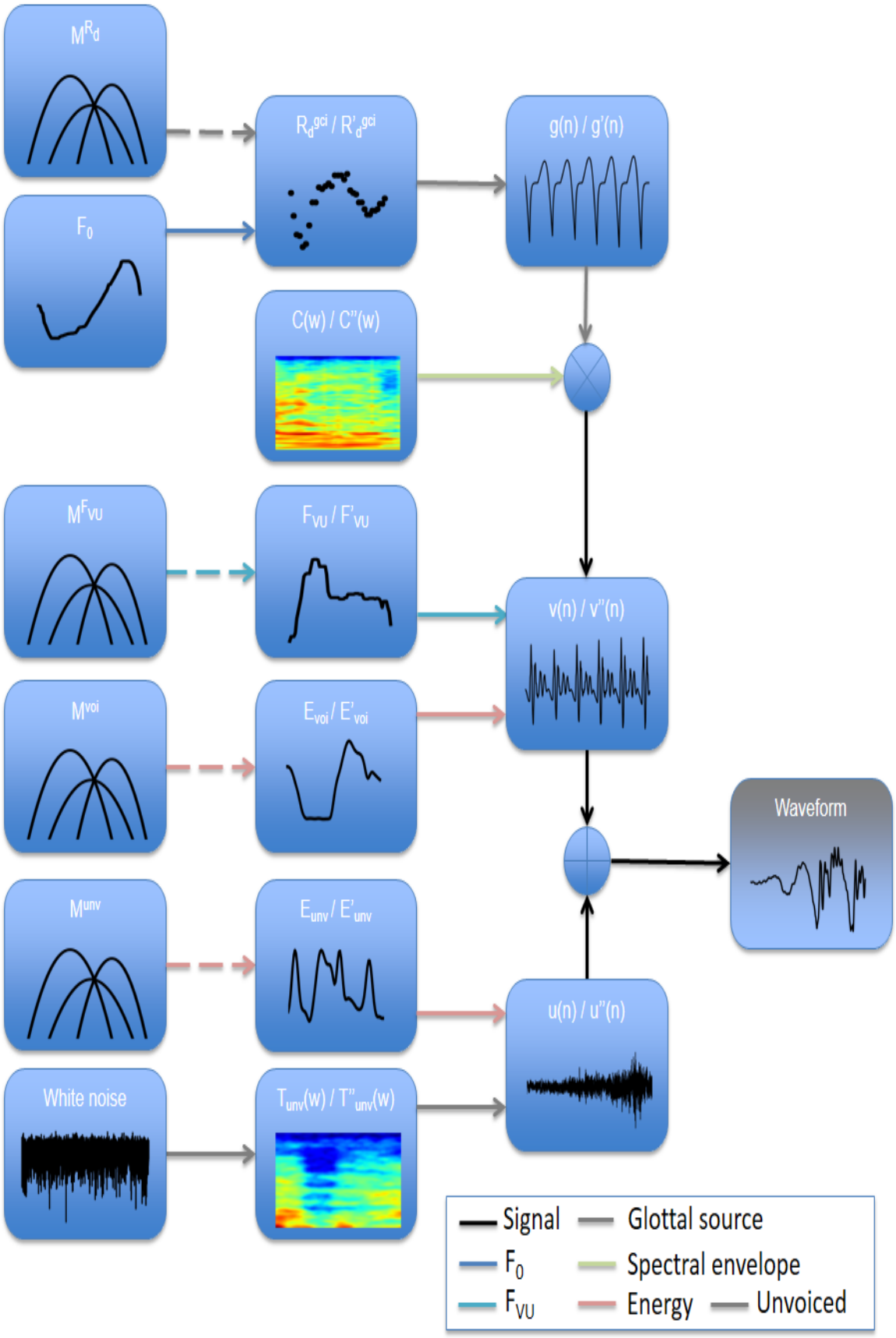


Figure 6.22: *System overview of the synthesis stage in PSY*

PSY processes the voiced $v(n)$ and unvoiced $u(n)$ parts separately. Each part is first synthesized and then mixed together in the time domain.

Voiced:
A sequence of glottal pulses is described by an R_d^{gci} / $R_d^{'gci}$ contour. It is used to synthesize a glottal pulse waveform $g(n)$ / $g'(n)$ in the time domain. Transformed to the spectral domain per STFT frame, it is convolved with one spectral representation of VTF $C(\omega)$ / $C''(\omega)$. The application of the spectral fading mechanism of section 6.5.5 is required if the voice quality contained in the original recording is altered before re-synthesizing the speech phrase. This is achieved by means of transforming R_d^{gci} into $R_d^{'gci}$. The employment of E_{voi} assures the maintenance of the original energy contour for the voiced part, as explained in section 6.4.1.2. Utilizing E'_{voi} instead of E_{voi} shall impose an energy contour in accordance with a corresponding voice quality transformation. The explained steps allow synthesizing an original $v(n)$ or a transformed and / or converted voiced part $v''(n)$.

Unoiced:
A sequence of original $T_{unv}(\omega)$ or converted $T''_{unv}(\omega)$ spectral envelopes is convolved with white gaussian noise per STFT frame in the spectral domain. Applying E_{unv} assures energy maintenance. E''_{unv} shall accordingly reflect the alteration of the unvoiced energy if the voice quality is transformed. An original $u(n)$ or a transformed and / or converted unvoiced part $u''(n)$ is synthesized in the time domain.

Mixing:
A simple addition of $v(n)$ or $v''(n)$ with $u(n)$ or $u''(n)$ mixes the voiced and unvoiced parts together. Its result is the re-synthesized signal waveform $s(n)$, or the transformed and / or converted signal waveform $s''(n)$.

6.6 Evaluation on voice quality transformation

Recordings of three different speakers are employed to evaluate the analysis robustness and synthesis quality of *PSY* on natural human speech. The baseline method SVLN of section 3.8.4 is utilized for comparison. The three chosen recordings of natural human speech are:

a) The English phrase *"Author of the danger trail, Philip Steels, et cetera."* as the first phrase of the English male speaker "BDL" of the CMU Arctic speech corpus proposed in [Kominek and Black, 2004].
c) The French phrase *"Que faire en cas de conflit avec sa banque?"* by the French female speaker "Margaux" taken from one speech corpus kindly provided by the Acapela group [1]. Please find the voice of "Margaux" via the type and talk demo found via the given link.
b) The phrase *"Il se garantira du froid avec un bon capuchon."* in French language spoken by the Hispanic male speaker "Fernando" recorded at IRCAM [Lanchantin et al., 2008].

Section 6.6.4 presents the utilization of the simple manual R_d offsets modification introduced in section 6.4.3.1 to perform a voice quality transformation. The test utilizes the time domain mixing synthesis variant of section 6.5.4. The corresponding listening test was conducted internally in the laboratory as a preliminary investigation on the two speech phrases b) and c) in French language. The test examines as well the GMM-based energy prediction of section 6.4.2.3.

The conclusions drawn from the test results motivates the subsequent test on voice quality transformation of section 6.6.5. It discusses the results of a second listening test spread to a bigger audience, including a post to the Auditory list [2] of the McGill University located in Montreal, Quebec, Canada. The spectral fade in/out synthesis variant of section 6.5.5 is evaluated since it is designed to assure a proper handling of voice quality transformations. The listening test covers all three phrases mentioned above to evaluate the voice quality transformation of section 6.4.3 on natural human voices. It examines the application of the generated R_d mean offset and R_d variance compression / expansion explained in section 6.4.3.2. The transformed $R_d^{'gci}$ contours and the original R_d^{gci} contour were employed by *PSY* and SVLN for synthesis. Both systems synthesized R_d contours which covered the complete R_d range to examine with which quality they are able to represent the transition in voice quality from a tense to a modal to a relaxed voice quality characteristic. *PSY* and SVLN received the same voice descriptors R_d, F_0 and F_{VU} as pre-estimated input to analyse $C(\omega)$. Both tests are based on the following two evaluation metrics.

Table 6.2: *Voice quality rating indices and suggested characteristics*

Index	Voice quality characteristic
-3	Very tense
-2	Tense / pressed
-1	Tense / pressed to modal / normal
+0	Modal / normal
+1	Modal / normal to relaxed / breathy
+2	Relaxed / breathy
+3	Very relaxed / breathy

Since three more tense and three more relaxed voice qualities are generated from the starting point of the R_d^{gci} contour of the original phrase, the listener is presented with the different voice qualities listed in table 6.2 to choose from. The novel speech analysis and synthesis framework *PSY* presented in this chapter is compared to the SVLN baseline method of section 3.8.4. The voice quality assessment examines how well both synthesis systems are able to produce different voice quality variations in terms of a tense, modal or relaxed voice quality characteristic. Ideally each test participant is able to perceptually associate each synthesized voice quality example to its corresponding voice quality characteristic, shown in table 6.2.

Table 6.3: *Synthesis quality rating according to the Mean Opinion Score (MOS) scale*

Index	Mean Opinion Score (MOS)
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

[1] Acapela group: www.acapela-group.com
[2] Auditory list: www.auditory.org

A second evaluation metric examines the synthesis quality on the Mean Opinion Score (MOS) generated by both synthesis systems. The listeners are presented with the list of possible sound quality ratings shown in table 6.3. Please note that the presented speech examples are randomized in their numerical order for each listening test such that the listeners are not able to conclude from each test phrase index the underlying voice quality characteristic.

6.6.1 SVLN voice descriptor smoothing

SVLN requires to smooth the input voice descriptors F_0, F_{VU} and R_d to avoid possible artefacts. These may origin from the energy modelling of SVLN which measures the energy level at the F_{VU}. If the F_{VU} contour contains a comparably high value change within a short-time segment, the connected energy measure may as well result in a huge value difference. This in turn leads to an erroneous energy modelling which results in audible artefacts. The artefacts originate from too high signal differences between consecutive synthesis steps, caused by too high changes of the implied energy measure. For similar reasons, SVLN requires to smooth as well the R_d and F_0 curve. The assumption being that the voice quality does not change within one phoneme [Degottex, 2010]. Thus, an additional median smoothing filter covering 100 *ms* is applied in SVLN to the input voice descriptors F_0, F_{VU} and R_d.

PSY seeks to avoid the smoothing of the input contours in order to maintain a certain naturalness in the synthesis. Only the windowing of a few glottal pulses $g_s(n)$ introduces a certain smoothing of the original R_d^{gci} contour. The windowing length corresponds to the standard STFT size defined in *PSY*. The RMS energy constraint of section 6.4.1 demands a constant STFT window size such that the energy interference of the window introduced to the energy measure of the analysis cancels out in the synthesis. It prohibits a shorter window length covering only one or two single $g_{R_d}^{gci}$ glottal pulses by means of pitch-adaptive windowing.

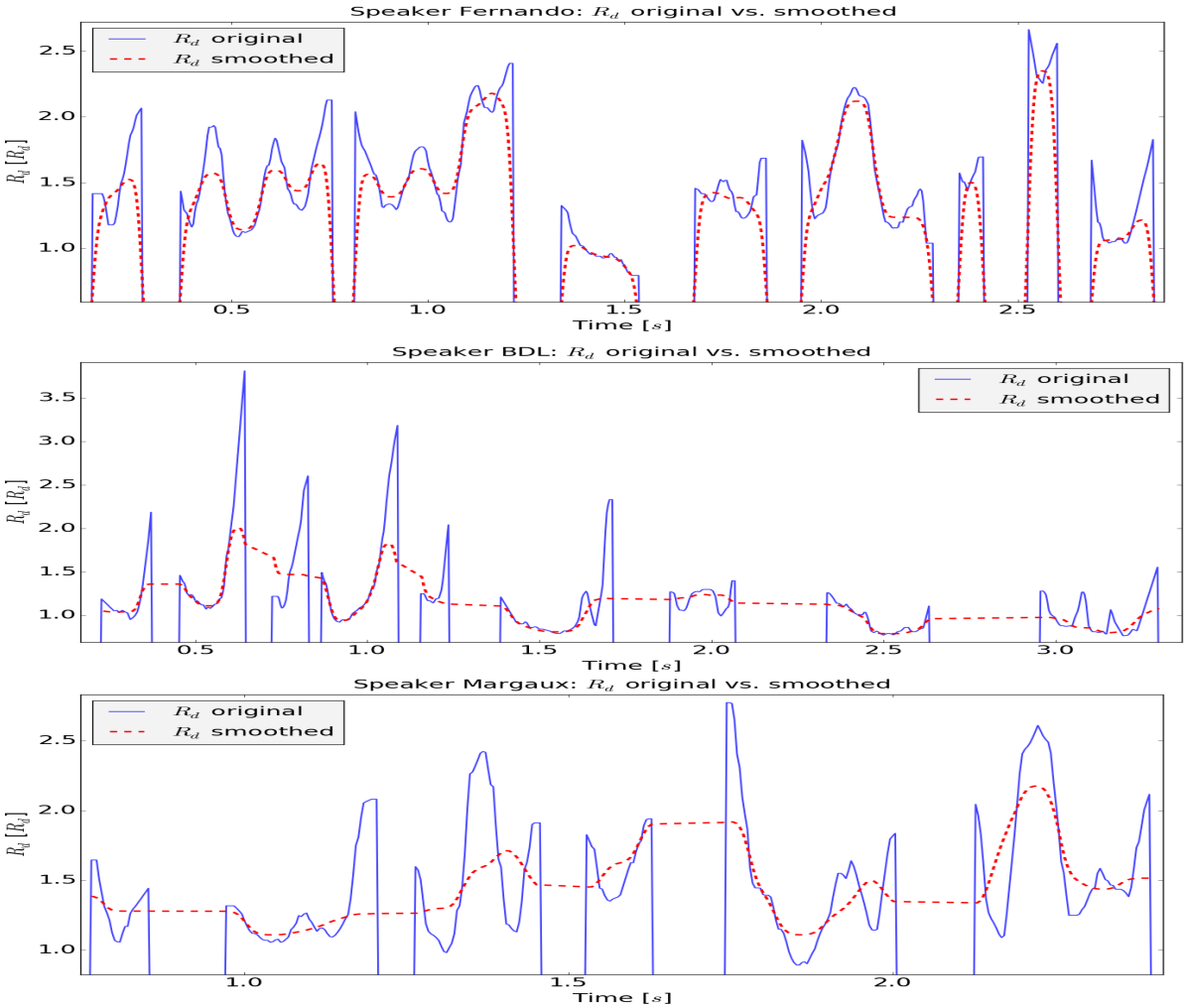


Figure 6.23: *R_d contour smoothing examples - Original for PSY, smoothed for SVLN*

The three illustrations shown in fig. 6.23 exemplify the smoothing of each R_d contour per speaker required by

SVLN, while the original R_d contour is processed by *PSY*. The smoothing reduces especially at voiced borders the original R_d contour towards lower instead of higher R_d values. This violates the hypothesis given for step 4 of section 6.2.1.3 to fade in/out the R_d contour at voiced borders. While the R_d smoothing appears for the most part tolerable for the speakers Fernando and BDL, it appears more like an intolerable cut around 1.2 - 1.5 seconds for speaker Margaux.

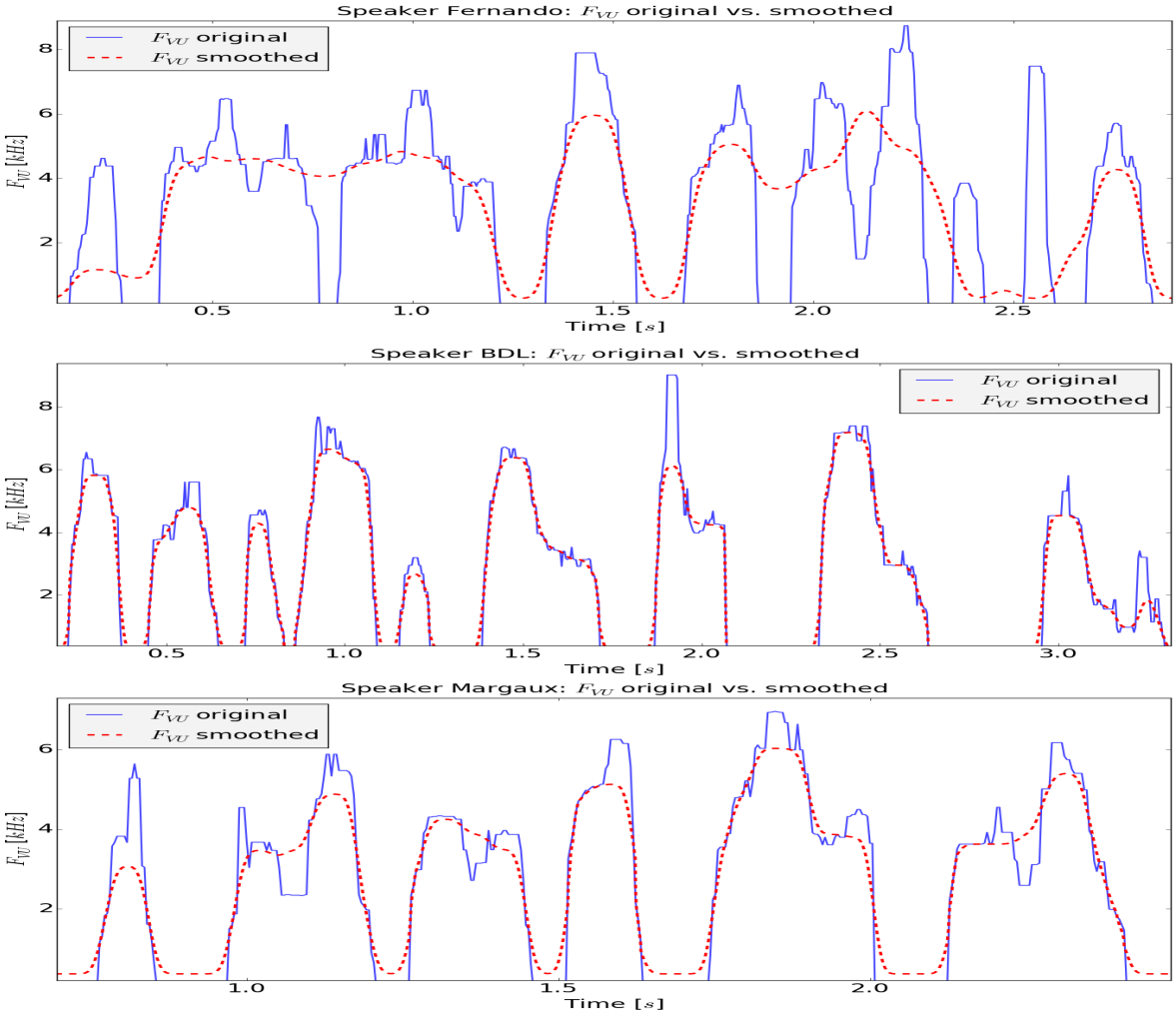


Figure 6.24: *F_{VU} contour smoothing examples - Original for PSY, smoothed for SVLN*

The smoothing of the F_{VU} contour for SVLN appears in general less dramatic. Examples are given for the three speakers Fernando, BDL and Margaux in fig. 6.24. The smoothing introduces only around seconds 2.2 of speaker Fernando a contradictory F_{VU} measure compared to the original F_{VU} contour. More severe is the unfortunate strong F_{VU} reduction for the voiced segments before and after 2.5 seconds.

6.6.2 STFT setup

The RMS energy constraint of section 6.4.1 requires a constant window size. It has to be set in such a way that the STFT window size leads to a proper representation of single sinusoidal peaks [Oppenheim and Schafer, 1975]. Windows in the time domain have to be long enough to distinctively present sinusoidal content in the spectral domain. However, too long window sizes invalidate the quasi-stationary assumption that the signal content remains constant in amplitude and frequency within the evaluated signal segment [Oppenheim and Schafer, 1989]. Two window sizes and its corresponding STFT step sizes were determined heuristically for the voice quality transformation tests in *PSY*. The standard window size used for analysis and synthesis of voiced signal content and voice descriptors such as F_0, F_{VU} or $\mathcal{T}_{sig}$ is set to a comparably small value of 32 *ms* for male and 22 *ms* for female speakers having higher pitched voices. The relatively small window size was heuristically determined by informal re-synthesis tests on the evaluated three speakers. It is accompanied with a STFT step size of 4 *ms*. A smaller STFT step size of only 1 *ms* is required for the analysis of R_d^{gci} and the estimation of the unvoiced residual $U(\omega)$.

A smaller STFT window size of only 8 *ms* is set for the estimation of $U(\omega)$ to better handle transient regions. Additionally, the random noise component does not require longer window sizes to properly resolve sinusoidal peaks. An informal empiric investigation determined 8 *ms* for the unvoiced synthesis as most appropriate. Smaller window sizes introduce muffling and ticking effects in the stochastic part. Higher window sizes may sound clearer and less muffled but suffer from increased reverberation effects. Before synthesis, the interference of applying the window $w_h(n)$ at the analysis and synthesis steps of the STFT to the corresponding signal parts is normalized following [Griffin and Lim, 1984].

6.6.3 Voice descriptor data analysis

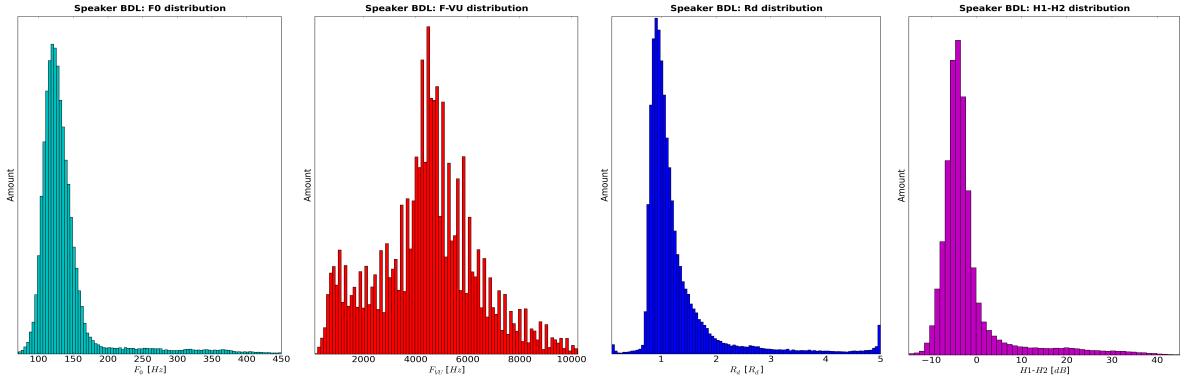
This section gives an overview on the value distribution of the complete voice descriptor set analyzed over the whole corpus of each evaluated speaker. The specific interrelation between the mean and variance distribution of the RMS-based voiced E_{voi} and unvoiced E_{unv} energies, as well as selected voice descriptors such as R_d and F_0 are shown and discussed. The data analysis provides another perspective on the correlation between energy and different shapes of the deterministic part of the glottal excitation source. This has been already studied theoretically in section 6.4.1.1 concerning the energy behaviour of the LF model. Here the analysis is based on real data of natural human speech of three different speech corpora. The analyzed data sets are employed as training set for the different GMM models presented in section 6.4.2.

The figures represent per speaker for each selected voice descriptor or RMS energy measure (despite the last SNR) their distribution over its own range. Each range is sliced into 100 parts. The y-axis parameterization represents thus the amount of the data occurrence per slice of the employed histogram. Differing amounts between different plots are removed to concentrate the illustration on the shapes of the data distributions itself, and not on its less interesting number of occurrences per descriptor.

Each shown RMS energy measure E_{unv} is evaluated on the synthesized unvoiced signal $u(n)$. The latter is constructed as a combination of the sinusoidal cancellation method "Re-Mixing with De-Modulation" of section 6.3.1.3 with the posterior below F_{VU} filter of section 6.3.2.3. This combination, denominated as $u^{HP}_{ReDe}(n)$, is the default methodology to represent the unvoiced signal $u(n)$ in *PSY*. It proved by informal listening tests conducted throughout this thesis work to achieve the best synthesis quality for $u(n)$.

6.6.3.1 Signal measures - Speaker BDL

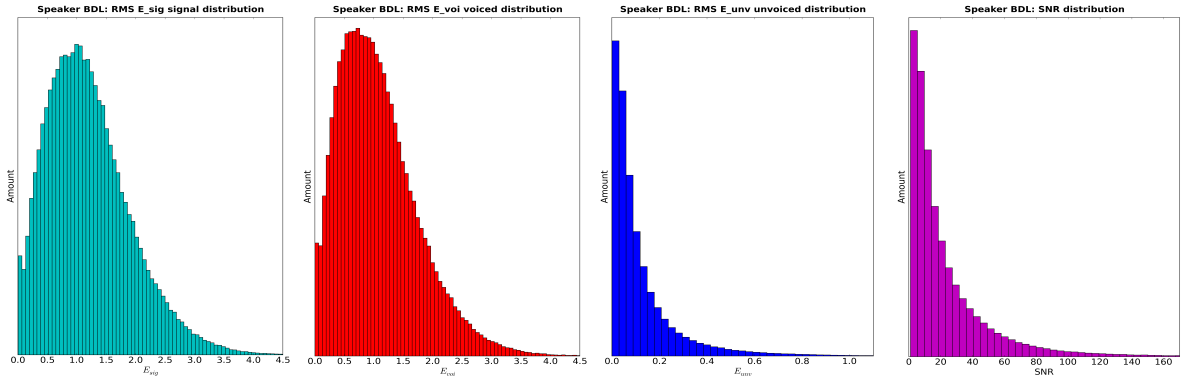
The voice descriptor distributions of speaker BDL, shown in fig. 6.25(a) for F_0 and fig. 6.25(d) for $H1$-$H2$, describe a symmetric Gaussian shape around its mean in lower value regions. Higher values descent in its amount without describing a specific statistical shape. R_d exhibits in fig. 6.25(c) an exponential Gaussian shape being skewed to the left in lower value regions. An obviously higher amount of R_d estimations is visible at the low R_d^{min} and high R_d^{max} value range border compared to neighbouring regions. This indicates partially erroneous results with a saturating R_d estimator. One reason could be the higher desire of speaker BDL to use more often the creaky voice quality mentioned in [Drugman et al., 2013]. A further discussion concerning the here employed R_d estimator in the context of creaky voice speech segments can be found in section 6.6.5.2.



(a) BDL F_0 distribution (b) BDL F_{VU} distribution (c) BDL R_d distribution (d) BDL $H1$-$H2$ distribution

Figure 6.25: *Speaker BDL - Value distribution of voice descriptor set*

The RMS energy distributions for the signal E_{sig} shown in fig. 6.26(a) and the voiced part E_{voi} in fig. 6.26(b)



(a) BDL E_{sig} distribution (b) BDL E_{voi} distribution (c) BDL E_{unv} distribution (d) BDL SNR distribution

Figure 6.26: *Speaker BDL - Value distribution of energy descriptor set*

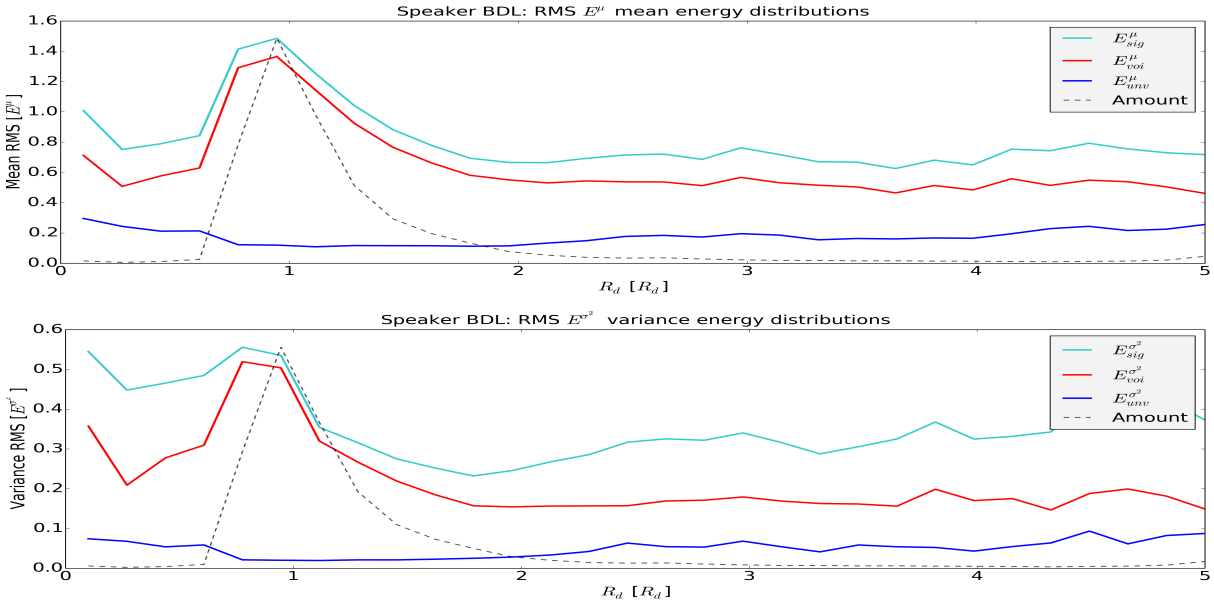


Figure 6.27: *Speaker BDL - RMS E^{μ,σ^2}_{sig}, E^{μ,σ^2}_{voi} and E^{μ,σ^2}_{unv} mean and variance distributions*

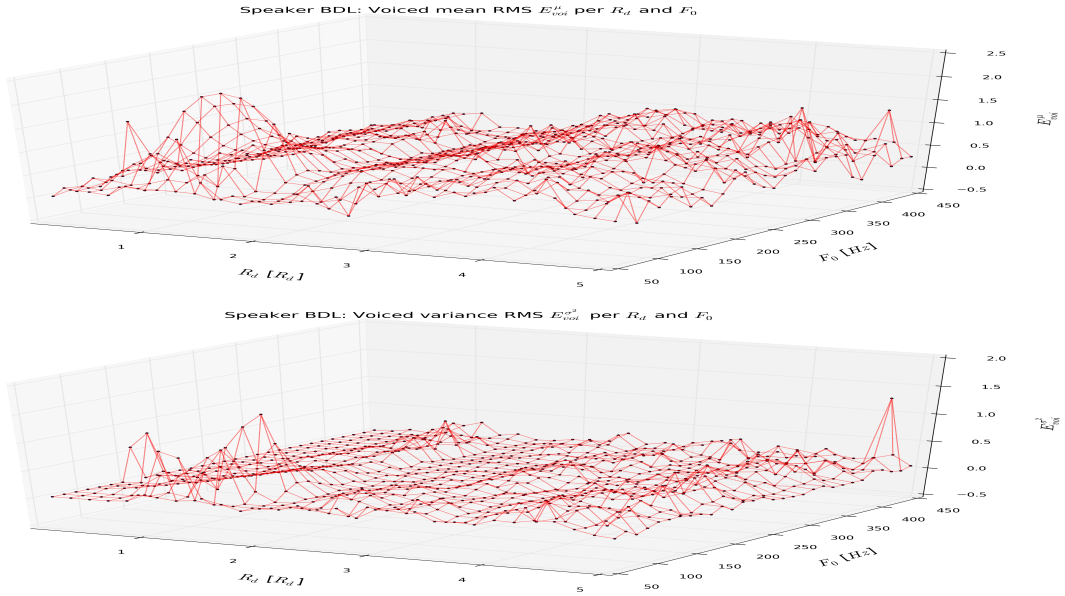


Figure 6.28: *Speaker BDL - Voiced RMS energies E^{μ}_{voi} and $E^{\sigma^2}_{voi}$ per R_d and F_0*

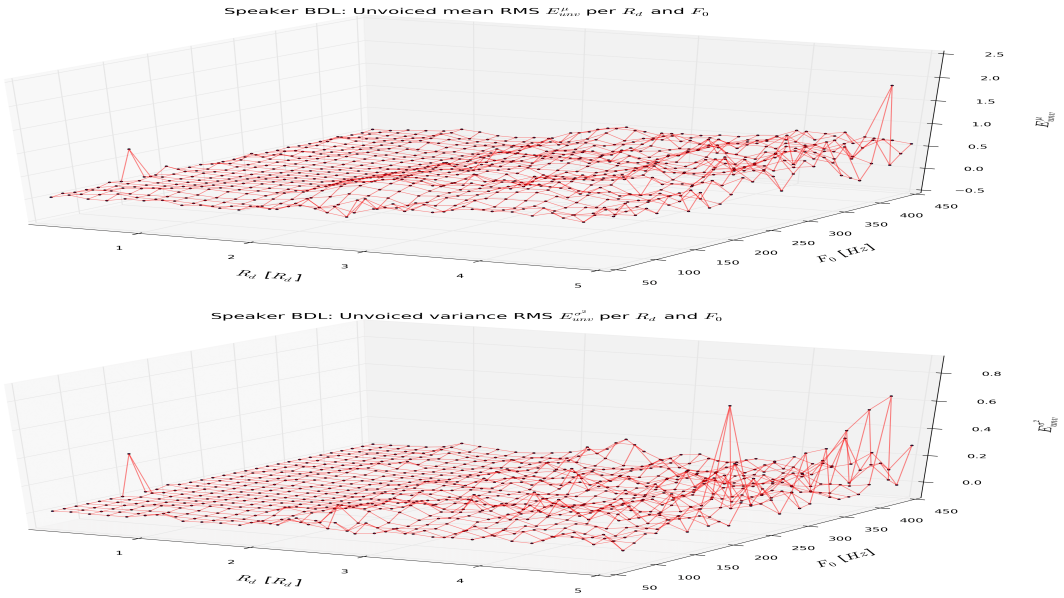


Figure 6.29: *Speaker BDL - Unvoiced RMS energies E^{μ}_{unv} and $E^{\sigma^2}_{unv}$ per R_d and F_0*

describe an exponential Gaussian shape being skewed to the left. The unvoiced energy distribution E_{unv} in fig. 6.26(c) and the SNR energy in fig. 6.26(d) follow an exponential decay.

An interesting finding are the RMS-based energy distributions sampled on a grid of 30 R_d values over the complete R_d range. Figure 6.27 shows the RMS mean E_{μ} and the RMS standard deviation E_{σ} values per R_d. The grey dashed lines correspond to the amount of voice descriptor entries observed per R_d value on the grid. The amount is normalized to the highest value present in each plot to facilitate the illustration. The hypothesis of higher voiced energies for lower R_d values and vice versa, explained in section 6.4.1.1, is not validated by the RMS mean energy measures depicted in figure 6.27. The expectation assumes that the maximum energy concentration is found on average for the lowest R_d value, to descent with increasing R_d values, and to find the lowest energy at the highest R_d value. However, in the range below R_d<1.0 the RMS-based energy does not further steadily increase with lower R_d values. At least the unvoiced energy E^{μ}_{unv} which is supposed to behave conversely exhibits a certain steady increase of its energy distributions for higher R_d>1.5 values. On the other hand, the presented observation may be valid considering the creaky voice quality of speaker BDL. Speech segments containing creak can be found predominantly at word and syllable endings having comparably lower energies. The creaky voice quality is comprised of impulse-like glottal pulses being similar to a tense voice quality. This may explain the energy findings illustrated in fig. 6.27 for R_d<1.0 for the voiced component. The observation is biased by the low amount of signal measures present in the lower R_d value region R_d<~0.6 and higher R_d value region R_d>~3.0. However, the figure 6.27 does only reveal certain aspects of the underlying data interrelationship.

The 3D plot illustrated in fig. 6.28 for the voiced energy distributions E^{μ}_{voi} and $E^{\sigma^2}_{voi}$ shown on the z-axis identifies that the strong energy concentration around R_d=1.0 occurs only for lower F_0 values. The 3D plot depicted in fig. 6.29 confirms for the unvoiced energy distributions E^{μ}_{unv} and $E^{\sigma^2}_{unv}$ shown on the z-axis the overall trend that the unvoiced energy rises with higher values of R_d.

Fig. 6.30 summarizes the data analysis on a 2D plot of mean and standard deviation found for the voiced E_{voi} and unvoiced E_{unv} energy on a grid of 30 R_d and F_0 values for each respective value range. Two distinct patterns can be observed for both the voiced and unvoiced energies: Below 200 Hz the data distribution appears more structured, while above 200 Hz mean and standard deviation behave more randomly and chaotic. Reflecting the amount of F_0 measures per F_0 value shown in fig. 6.25(a) reveals that above 200 Hz the data becomes sparse. Training a GMM to predict F_0 values in the context of a pitch transposition application may lead to modelling errors for upwards transpositions into the sparse data range.

6.6.3.2 Signal measures - Speaker Fernando

The distributions of the voice descriptors F_0, R_d and $H1$-$H2$ shown in the figures 6.31(a), 6.31(c) and respectively 6.31(d) appear to describe a Lorentzian bell shape. Contrary to speaker BDL, no descending value distribution can be observed for higher value regions. The distribution of estimated R_d values exhibits in fig. 6.31(c) only a slightly higher amount of saturated measures on the lower R_d^{min} border. This indicates that the R_d estimator could robustly estimate the R_d contours most of the times. The signal E_{sig} and voiced E_{voi} energy distributions depicted

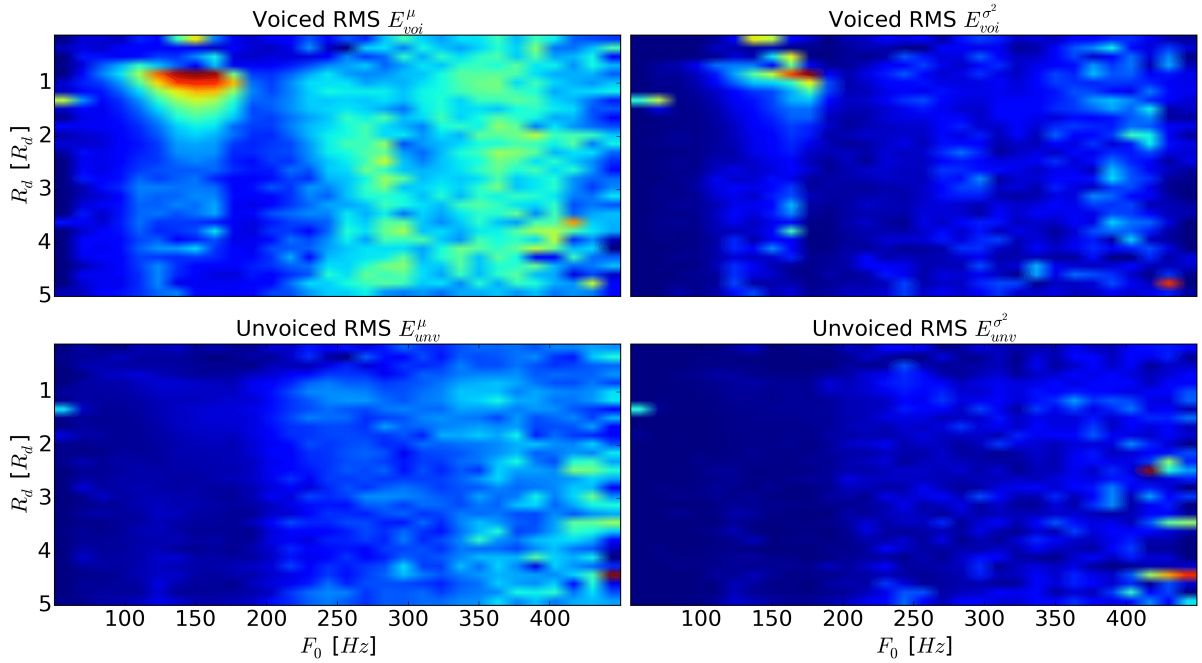
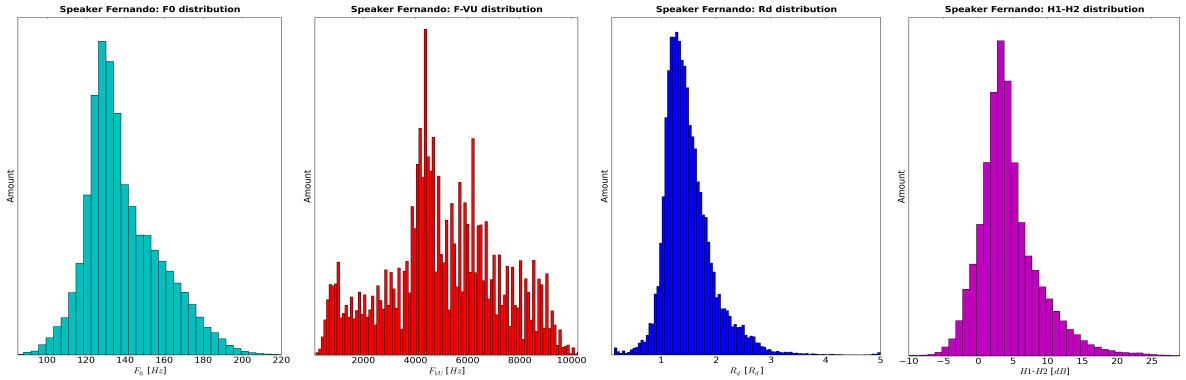


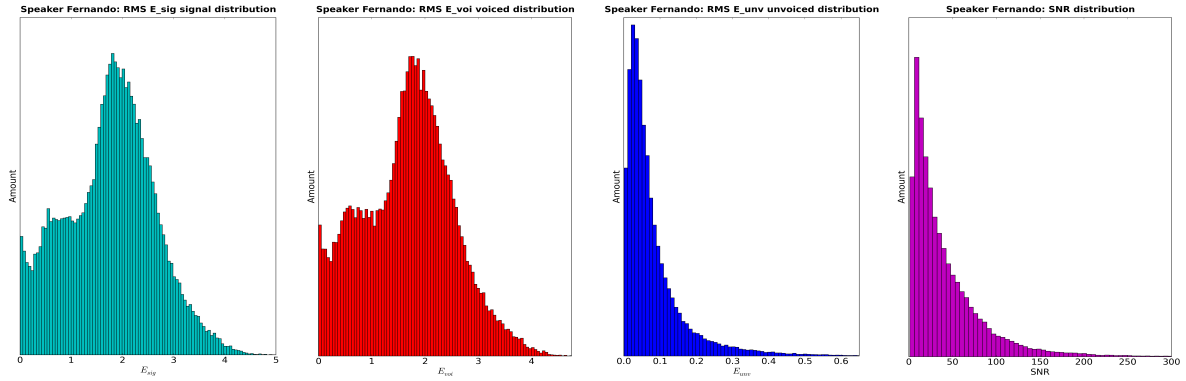
Figure 6.30: *Speaker BDL - RMS energies E^{μ,σ^2}_{voi} and E^{μ,σ^2}_{unv} per R_d and F_0 in 2D*

in fig. 6.32(a) and fig. 6.32(b) are not following a known statistical data shape. The data distributions observed for speaker Fernando are more complex than the multi-modal distributions exhibited by speaker BDL.



(a) Fernando F_0 distribution (b) Fernando F_{VU} distribution (c) Fernando R_d distribution (d) Fernando $H1$-$H2$ distr.

Figure 6.31: *Speaker Fernando - Value distribution of voice descriptor set*



(a) Fernando E_{sig} distribution (b) Fernando E_{voi} distribution (c) Fernando E_{unv} distribution (d) Fernando SNR distribution

Figure 6.32: *Speaker Fernando - Value distribution of energy descriptor set*

The unvoiced energy distribution E_{unv} in fig. 6.32(c) and the SNR energy in fig. 6.32(d) follow as for speaker BDL an exponential decay. The RMS-based energy distributions for the voiced and unvoiced signal parts are again

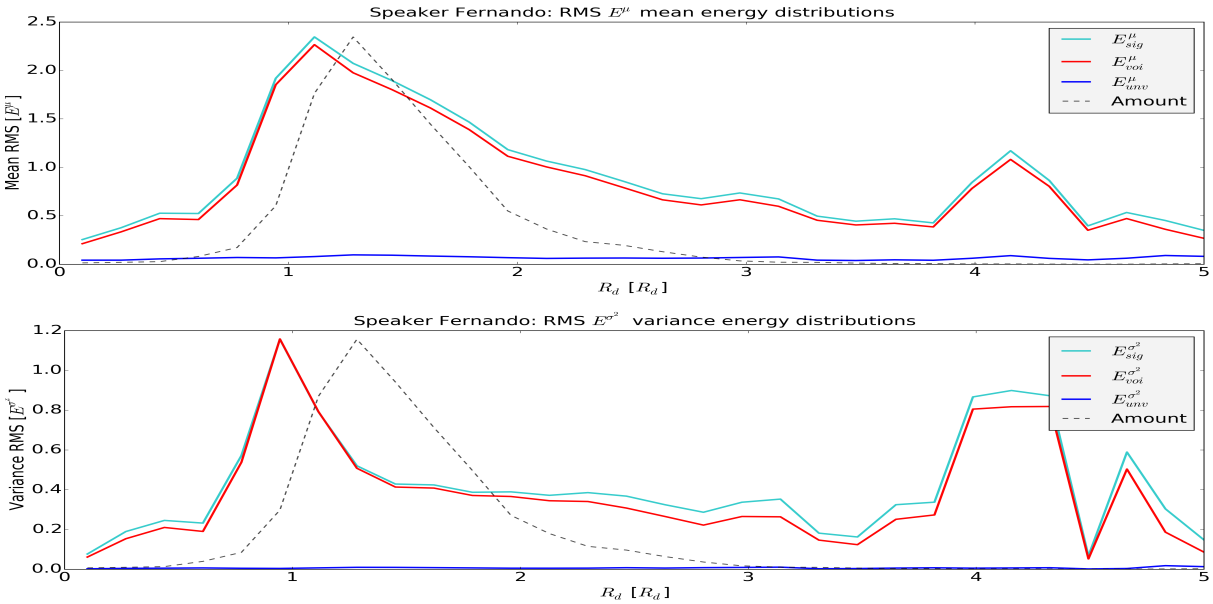


Figure 6.33: *Speaker Fernando - RMS E^{μ,σ^2}_{sig}, E^{μ,σ^2}_{voi} and E^{μ,σ^2}_{unv} mean and standard deviation distributions*

evaluated over the complete R_d range. The RMS mean and standard deviation energy is illustrated in fig. 6.33. As with speaker BDL, a peak amount of R_d estimation is observed in the normal R_d range. Again, the hypothesis of higher voiced energies for lower R_d values and vice versa of section 6.4.1.1 is not completely validated. A rough trend within the R_d range [1.0, 5.0] validates the hypothesis. However, below R_d<1.0 the voiced and signal RMS energies E_{voi} and E_{sig} decrease. One explanation for this behaviour can be inspected in fig. 6.53 around ~1.50 seconds. The estimated R^{gci}_d contour approaches continuously lower R_d values at the end of an utterance with continuously lower energies while the R^{gci}_d estimation should actually increase. This can be either a failure of the R^{gci}_d estimator or a speaker dependent behaviour observed with speaker Fernando. As with the other analyzed speakers, the observation is biased by the low amount of signal measures present in the lower R_d value region R_d<~0.5 and higher R_d value region R_d>~3.0.

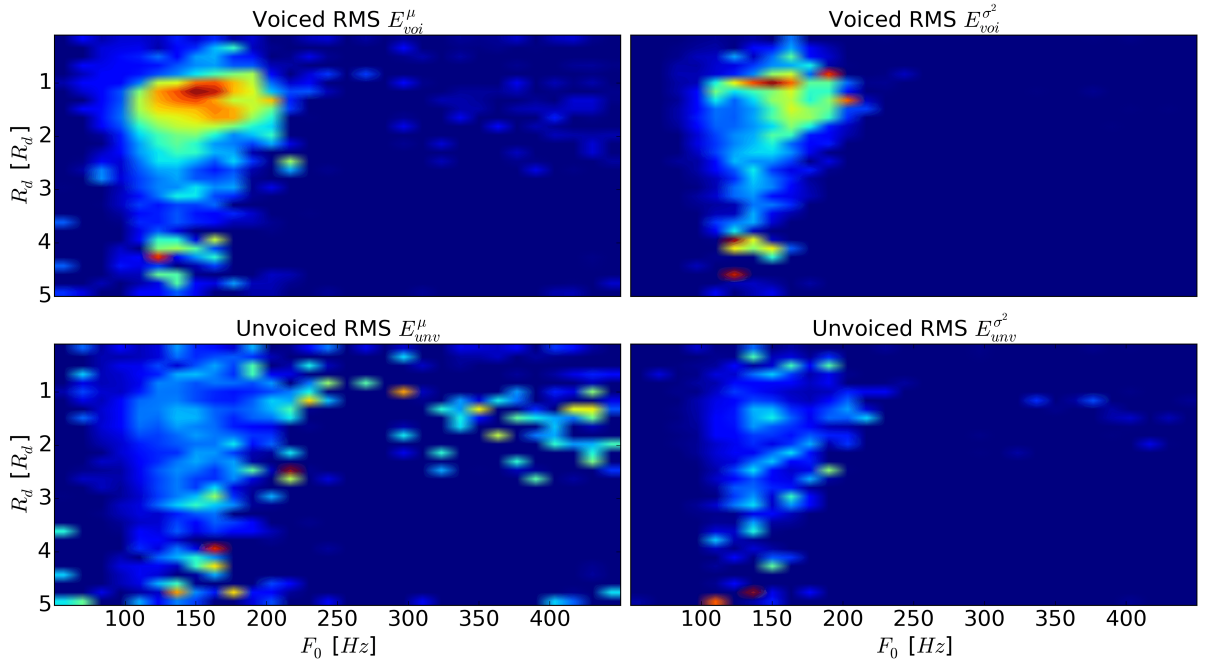


Figure 6.34: *Speaker Fernando - RMS energies E^{μ,σ^2}_{voi} and E^{μ,σ^2}_{unv} per R_d and F_0 in 2D*

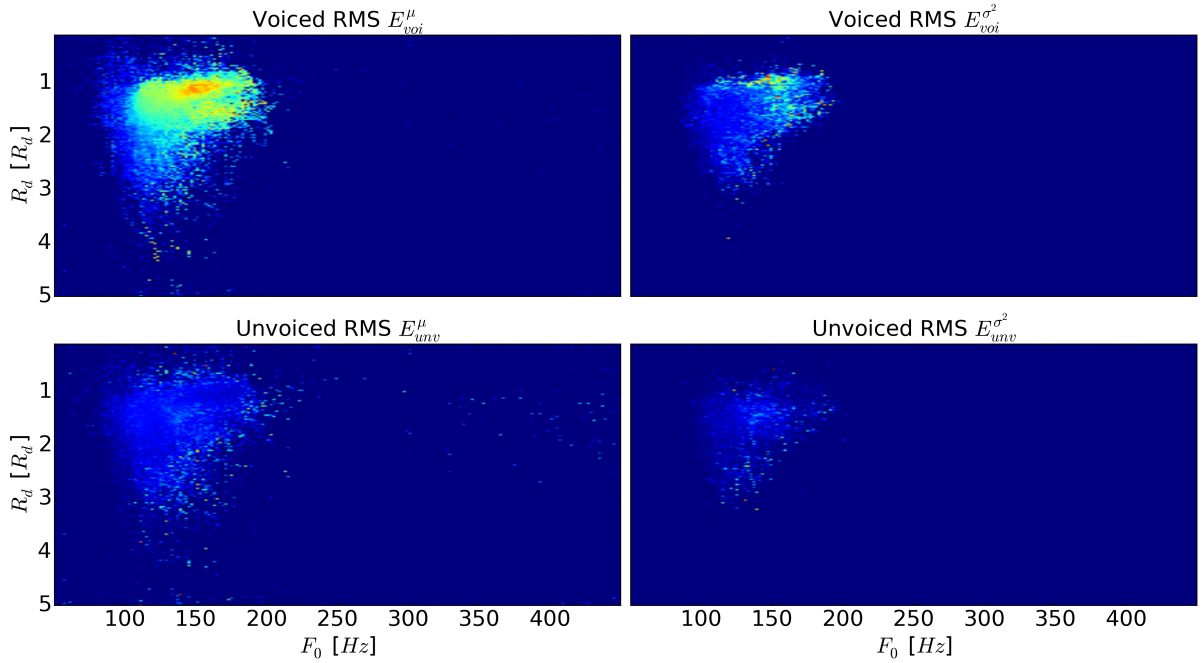
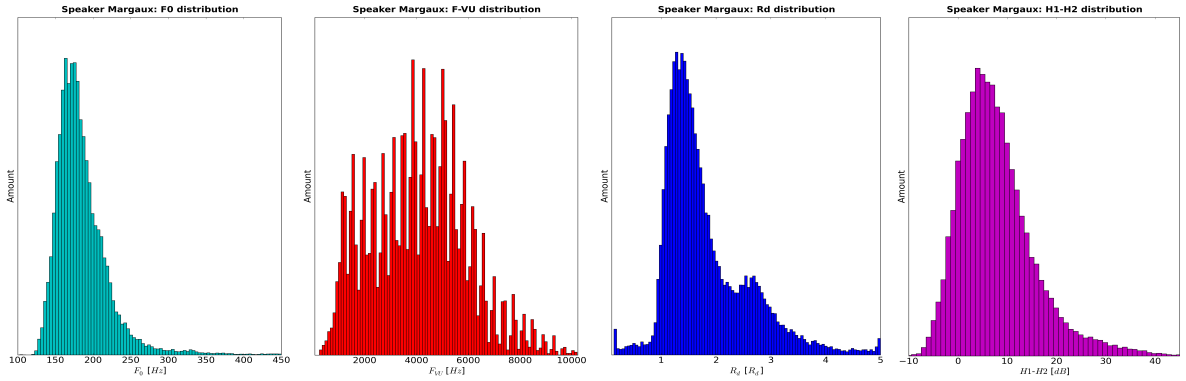


Figure 6.35: *Speaker Fernando - RMS energies E^{μ,σ^2}_{voi} and E^{μ,σ^2}_{unv} per R_d and F_0 in 2D (High Res.)*

Fig. 6.34 summarizes the data analysis findings on a 2D plot of mean and variance found for the voiced E_{voi} and unvoiced E_{unv} energies on a grid of 30 values over the complete R_d and F_0 range. The mean and standard deviation energy distribution for voiced and unvoiced becomes very sparse above 200 Hz. Higher R_d value regions for F_0 values below 100 Hz exhibit a sparse data distribution. The non-sparse data area for lower R_d and F_0 values resembles roughly a conical shape. The application of a speaker-dependent energy model, a glottal excitation source model or a pitch transposition model is error prone or at least cumbersome if the areas with sparse data are covered. Fig. 6.35 illustrates with a higher resolution grid of 200 instead of 30 R_d and F_0 values more detailed the same data distribution.

6.6.3.3 Signal measures - Speaker Margaux

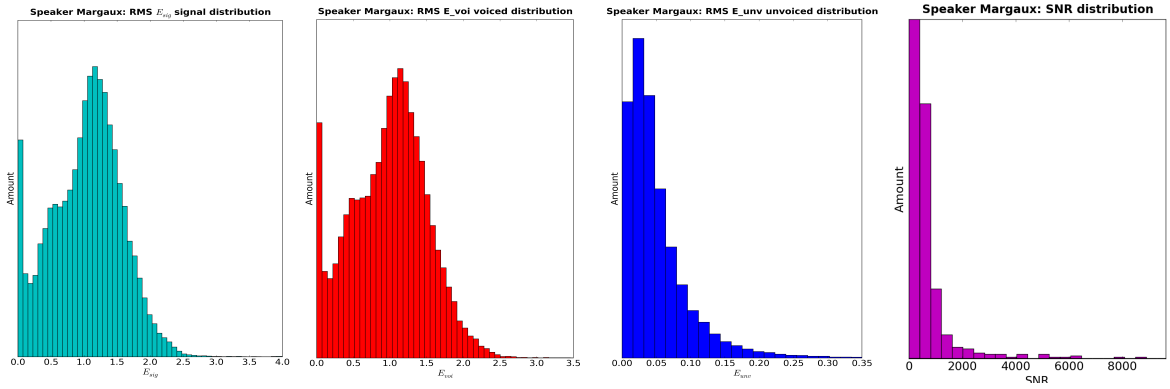
The French female speaker Margaux exhibits higher variance distributions and mean values compared to the male speakers for the voice descriptors F_0, R_d and $H1$-$H2$, illustrated in the figures 6.36(a), 6.36(c), and 6.36(d). R_d and $H1$-$H2$ share a negative correlation with F_{VU} [Huber and Röbel, 2013]. The F_{VU} data distribution for Margaux is accordingly skewed to comparably lower value regions. The same conclusion as with speaker BDL can be



(a) Margaux F_0 distribution (b) Margaux F_{VU} distribution (c) Margaux R_d distribution (d) Margaux $H1$-$H2$ distr.

Figure 6.36: *Speaker Margaux - Value distribution of voice descriptor set*

observed for speaker Margaux concerning the R_d estimation on her complete corpus, depicted in fig. 6.36(c): An obviously higher amount of R_d estimations is visible at the low R_d^{min} and high R_d^{max} value range border compared to neighbouring regions. The saturation at the R_d range borders indicates partially erroneous R_d estimation results.



(a) Margaux E_{sig} distribution (b) Margaux E_{voi} distribution (c) Margaux E_{unv} distribution (d) Margaux SNR distribution

Figure 6.37: *Speaker Margaux - Value distribution of energy descriptor set*

For speaker Margaux, the reason is mainly the lack of robustness to estimate the pulse shape of the glottal excitation source for voices with higher F_0 values [Drugman et al., 2008] and a more relaxed (breathy) voice quality [Huber and Röbel, 2013].

The RMS energy distribution of speaker Margaux resembles roughly the one of speaker Fernando. The signal E_{sig} and voiced E_{voi} energy distributions depicted in fig. 6.37(a) and fig. 6.37(b) are not following a known statistical data shape. The unvoiced energy distribution E_{unv} in fig. 6.37(c) and the SNR energy in fig. 6.37(d) describe an exponential decay. As with speaker Fernando, the mean RMS-based energy distributions for the signal E^{μ}_{sig} and the

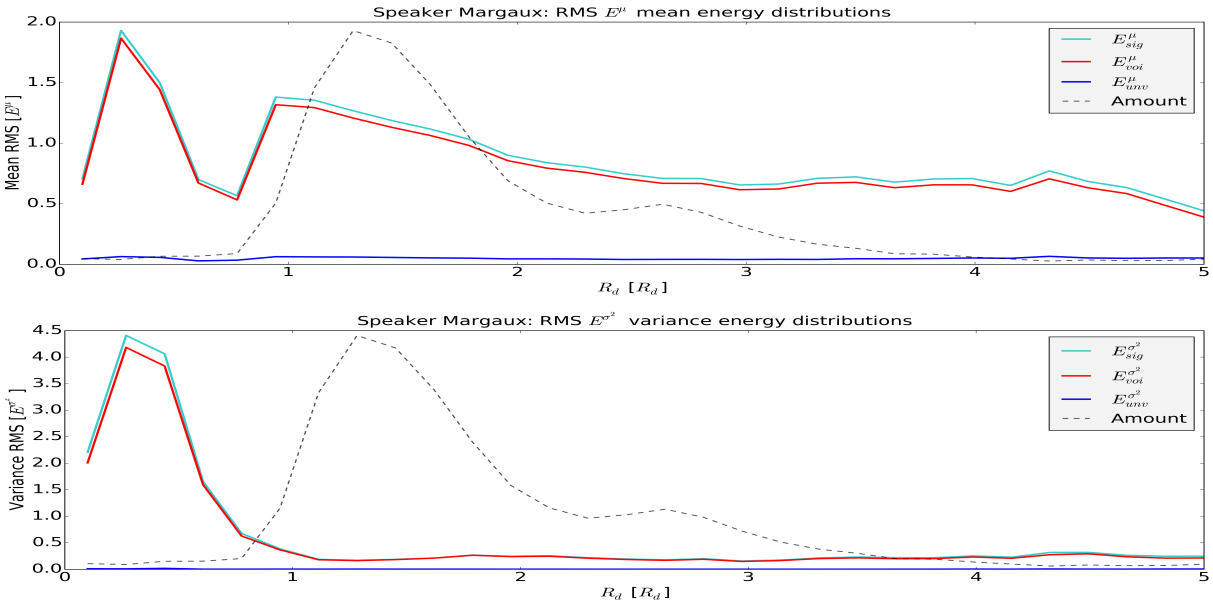


Figure 6.38: *Speaker Margaux - RMS E^{μ,σ^2}_{sig}, E^{μ,σ^2}_{voi} and E^{μ,σ^2}_{unv} mean and variance distributions*

voiced part E^{μ}_{voi} follow a roughly but not steadily decrease from lower to higher R_d values only within the R_d range [1.0, 5.0], illustrated in fig. 6.38. Below R_d<1.0 the voiced and signal RMS energies E_{voi} and E_{sig} do not confirm the hypothesis of section 6.4.1.1 that higher voiced energies E_{voi} are observed for lower R_d values, and vice versa. As with the other analyzed speakers, the findings are not very significant and may be biased by the low amount of signal measures present in the lower R_d value region R_d<~0.75 and higher R_d value region R_d>~4.0.

Another interesting finding is the voiced E^{μ,σ^2}_{voi} and unvoiced E^{μ,σ^2}_{unv} RMS energy distribution found on a 2D grid of 30 R_d and F_0 values for speaker Margaux, illustrated in fig. 6.39. The voice descriptor measures start for the female speaker with a minimum F_0 border of F_0 ~=120 Hz. A high concentration of the voiced mean energy E^{μ}_{voi} can be observed in the F_0 range [200, 300] Hz for R_d values below R_d<1.0. Contrary to the two-dimensional plots shown for speaker BDL in fig. 6.30 and for speaker Fernando in 6.34, no sparsely distributed voice descriptor value regions can be observed for the voiced mean E^{μ}_{voi} and variance $E^{\sigma^2}_{voi}$ measures. The unvoiced mean E^{μ}_{unv} and variance $E^{\sigma^2}_{unv}$ measures exhibit a more unstable and chaotic data pattern.

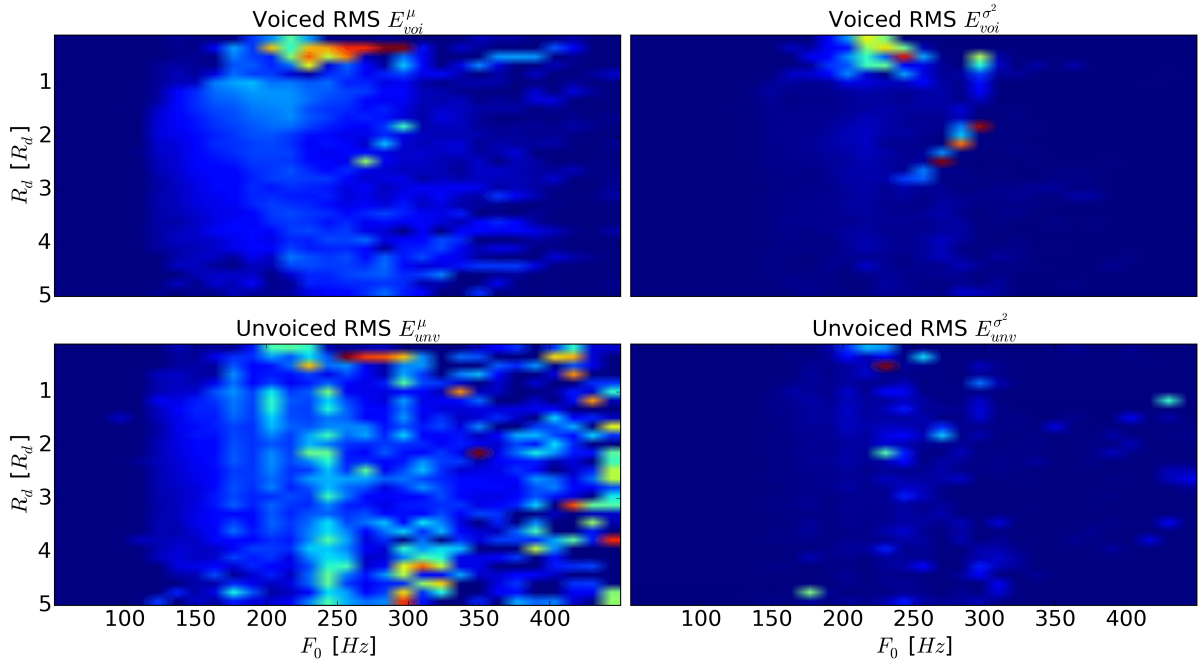


Figure 6.39: *Speaker Margaux - RMS energies E^{μ,σ^2}_{voi} and E^{μ,σ^2}_{unv} per R_d and F_0 in 2D*

6.6.4 Voice Quality (VQ) Test 1: Time domain mixing and R_d shifting

The simple R_d offset introduced in section 6.4.3.1 is examined in section 6.6.4.2 on the Hispanic male speaker Fernando speaking French, and in section 6.6.4.3 on the French female speaker Margaux. The test examines the *PSY* synthesis variant called "time domain mixing" introduced in section 6.5.4.

6.6.4.1 GMM-based energy scaling drawbacks

The test evaluates in particular if listening experts are able to perceive differences between phrases synthesized with either the standard energy maintenance of section 6.4.1.2 or with the advanced GMM-based energy modelling of section 6.4.2.3. The latter is intended to properly predict the energy of the voiced and unvoiced components separately. It shall reflect a speakers behaviour when uttering different voice quality characteristics more precisely than simple energy maintenance. The GMM energy model, trained on a voice descriptor set D of the given speaker, shall predict the proper energies from a new voice descriptor set D'. The latter is computed on a transformed signal $s'(n)$ being generated by a voice quality transformation. However, several problems may arise with the current status of the implemented GMM-based energy model in *PSY*:

I. Energy scaling impact:
A voice quality transformation covering bigger R_d changes creates too huge changes in the predicted energy contour for usually the voiced and possibly the unvoiced part. It results especially for a transformation from a modal to a tense voice quality into a predicted RMS energy contour which is prone to result in amplitude values in the time domain being outside the valid range [-1 1]. A synthesized phrase is prone to clip without a preceding amplitude normalization. Normalizing the amplitude of one phrase requires to apply the same normalization factor to all other phrases of the test such that their energy relations are properly reflected. However, the predicted energy contours for more tense voice qualities may lead to huge energy jumps within short-time segments, or even just single frames. Reflecting the energy interrelation to all other phrases would scale the corresponding relaxed voice quality phrases too much down such that they would not anymore be audible. Moreover, the energy prediction and its corresponding amplitude scaling from modal towards relaxed can be as well relatively huge. The synthesized speech phrases being transformed to a more relaxed voice quality with the GMM energy prediction are already of comparably low energy. The tense voice quality phrases which produce clipping are removed from the listening test. Their required normalization is avoided and the remaining speech phrases are examined directly without normalization.

II. Unstable data modelling and prediction:
The predicted energy contours appear to a huge extent reasonably for the voiced component. The energy scaling for the unvoiced component is more prone to errors or apparent mispredictions. Its resulting synthesis may suffer from unnatural sounding energy changes, especially for voice quality transformations having a higher impact.

One reason for that unfortunate behaviour is that the measured RMS unvoiced energy turned out to be the most uncorrelated one to all other employed voice descriptors for the two tested speech phrases.

III. Data approximation:
The data excerpts of section 6.6.3 illustrate partially the underlying energy data basis to be modelled. The trained GMM components and the complete GMM modelling may not be able to approximate well and to generalize on the given data. This applies especially for sparse as well as for highly fluctuating and thus tricky value regions.

IV. RMS energy basis:
The RMS energy measures E_{sig} on the signal $s(n)$ and E_{unv} on the synthesized unvoiced component $U(\omega)$ along with the energy scaling of the transformed voiced $V(\omega)$ and unvoiced $U(\omega)$ components before synthesis reflect the basis of the underlying energy metric defined in the equations from 6.19 to 6.15. The RMS measure evaluates the effective energy of a short-time audio signal segment. It does not consider the mechanism of human auditory perception. The future work section 8.2.2 lists some methods approximating human perception. A perceptually based energy measure may solve the problems with the too huge GMM-based energy prediction explained in this section as step I.

V. Erroneous R_d estimations:
The data analysis on the estimated voice descriptor values over the whole corpus of the three evaluated speakers presented in section 6.6.3 could not validate the hypothesis introduced in section 6.4.1.1 that lower R_d values are associated with a more tense voice quality. However, this does not signify that the given hypothesis is wrong. Informal tests on estimated R_d contours show that the R_d curve tends sometimes towards lower instead of higher R_d values in abducted speech segments when approaching a voiced border. One reason can be a violated periodicity assumption occurring predominantly for creaky voice segments [Drugman et al., 2014]. The algorithmic step 4 of section 6.2.1.3 to fade in and out the R_d contour at voiced segment borders constitutes an initial approach to overcome this false R_d estimation if it doesn't follow the hypothesis. However, apparently more investigations are required to further improve the R_d estimation for theses cases. Additionally, a creaky voice modelling is required to further analyse this aspect since the creaky voice quality contains glottal pulses being similar to the glottal source shapes of a tense voice quality. Simultaneously, creak is observed at the end of utterance having lower energies.

6.6.4.2 VQ Test 1 Results - French male speaker

A preliminary listening test on speaker Fernando was conducted by 6 sound processing experts internally in the laboratory. Fig. 6.40 depicts six R_d^{gci} contours with a manually defined offset from the original R_d^{gci} contour in the

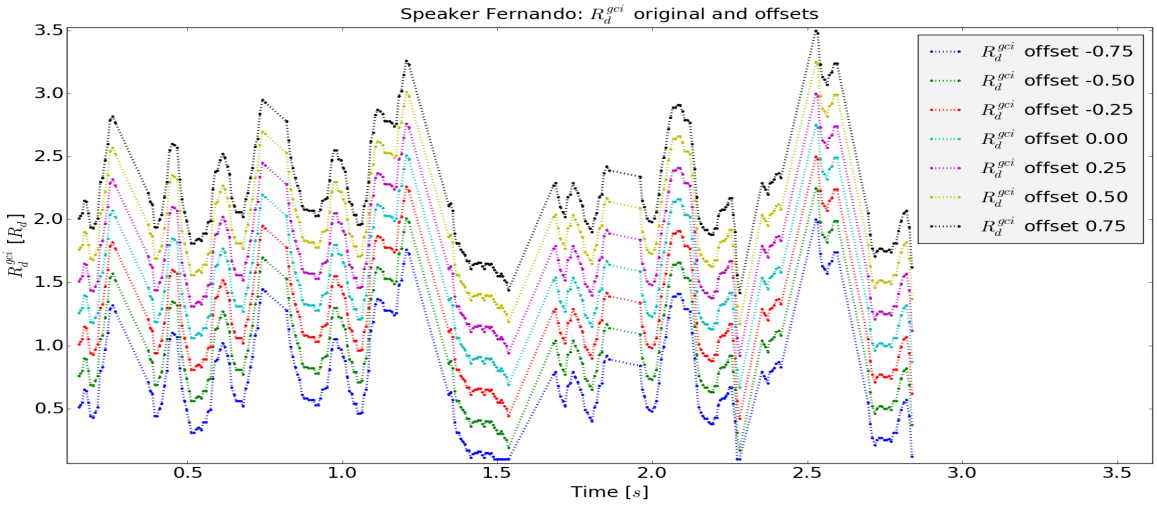


Figure 6.40: *VQ Test 1 Speaker Fernando - Manually set R_d mean offsets with step size R_d±0.25*

middle shown in cyan colour. Each positive and negative offset constitutes an empirically determined mean R_d offset of R_d±0.25 to the previous offset in its respective direction. The offset amount was chosen such that one of the R_d^{gci} offset contours reaches a border of the R_d range [0.1, 5.0]. The R_d^{gci} offset -0.75 saturates in this example around ~1.50 seconds on the lower R_d border. Please note that the algorithm to fade an R_d contour in or out at voiced borders has not yet been implemented in *PSY* at the time of this test. As well the soft saturation algorithm at R_d range border was developed later in time. SVLN applies to each R_d^{gci} offset contour the time basis interpolation and median smoothing explained in section 6.6.1 in the same sense as illustrated in fig. 6.23.

Table 6.4: *VQ Test 1 Speaker Fernando - Test indices per synthesis method and R_d offset*

Method	-0.75	-0.50	-0.25	0.0	+0.25	+0.50	+0.75
PSY	06	11	07	12	02	14	15
PSY (GMM)			04		13	03	19
SVLN	01	18	17	10	05	08	09

The synthesized speech phrases can be found online via the link for Speaker Fernando [3]. Table 6.4 lists per synthesis method and R_d offset the indices which are given for the listening test online such that readers can listen to each sound example to transparently follow the work here presented. Please note that the GMM energy model is not used for the direct re-synthesis with R_d offset 0.0 without any glottal source transformation since it is designed to predict the respective energies on transformed R_d values. The too huge energy scaling impact mentioned at step I. in the preceding section prevents its usage for the voice quality transformations with R_d offset -0.50 and -0.75. The hidden original speech phrase is placed at test index 16.

Voice quality rating results:

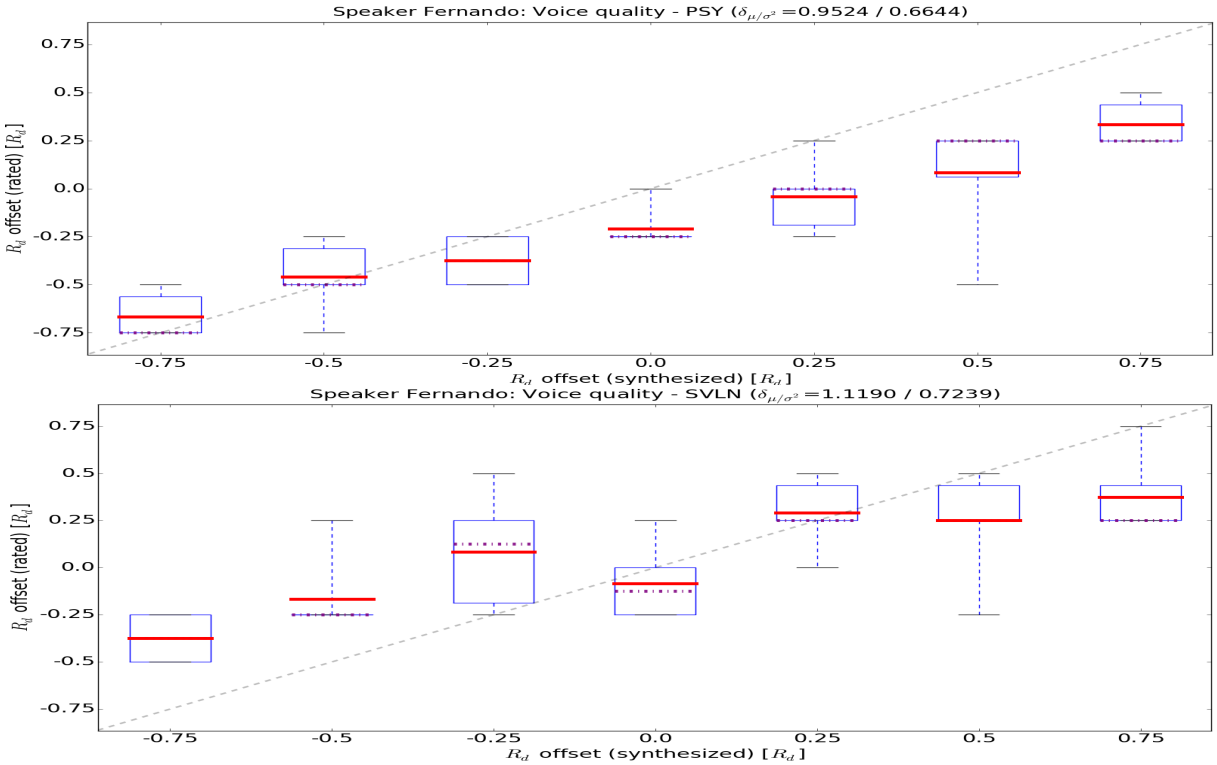


Figure 6.41: *VQ Test 1 Speaker Fernando - Voice quality rating results*

Fig. 6.41 depicts the voice quality rating results of VQ test 1 for the novel speech framework *PSY* and the baseline SVLN on speaker Fernando. The small horizontal grey lines at both ends (whiskers) per boxplot are set to show the minimum and maximum value for each evaluation. The horizontal red (violet) lines reflect the mean (median) voice quality ratings of all participants per test phrase with the same indices as in table 6.2. The dialog grey dashed line exemplifies their ideal placement if each test participant would have been able to associate perceptually each synthesized voice quality example to its corresponding voice quality characteristic. The mean deviation value δ_μ=0.95 for *PSY* expresses the disagreement of the listeners, being ideally δ_μ=0.00. *PSY* received very low mean deviation δ_μ values for more tense voice qualities. However, the stronger the original modal voice quality is transformed towards a more relaxed voice quality the less well could the participants identify its perceptual sensation. Drawing a regression line through each mean value shown in red horizontal lines per rated R_d offset would result in a less steep line for *PSY* than the ideal one depicted as grey dash line. A higher mean deviation value δ_μ=1.12 as compared to *PSY* is shown for the baseline method SVLN in fig. 6.41. It indicates that the listeners could less well capture the different synthesized voice qualities and associate them with the corresponding offset indices. A roughly matching trend of the voice quality associations can be concluded for both systems.

[3] Speaker Fernando: http://stefan.huber.rocks/phd/tests/RdMisterF/

MOS synthesis quality rating results:

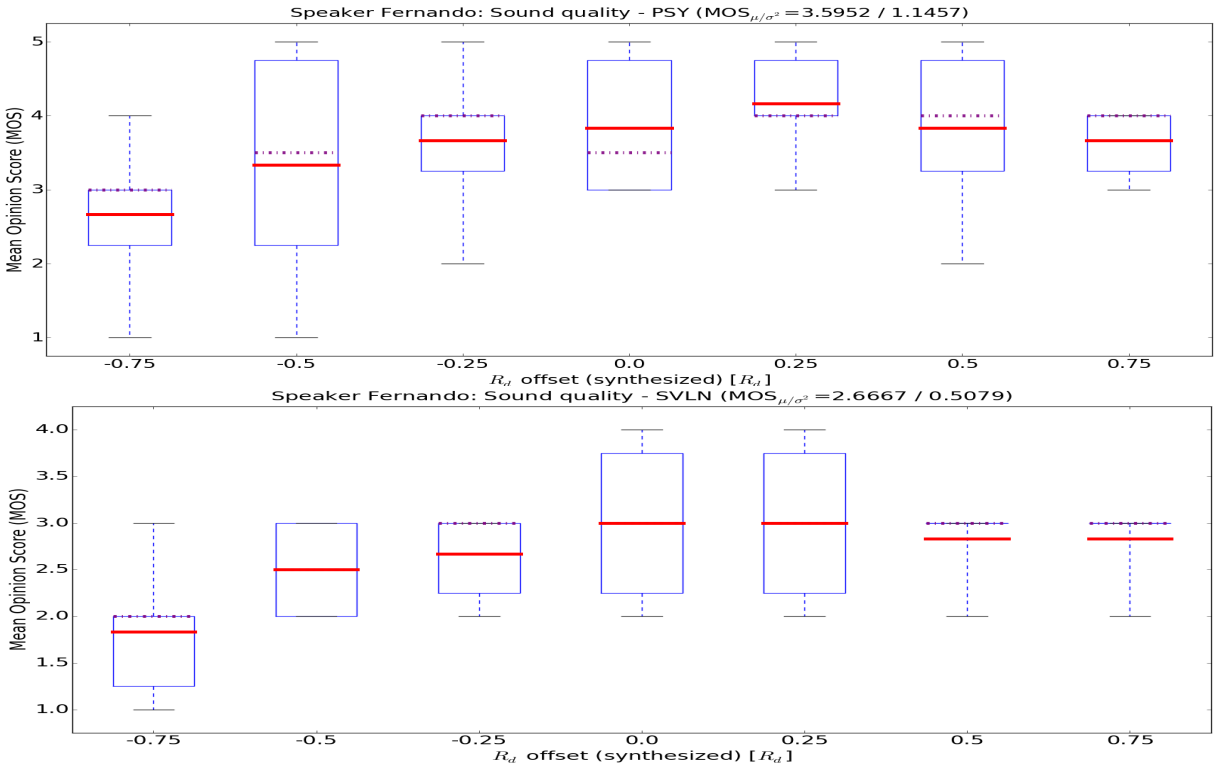


Figure 6.42: *VQ Test 1 Speaker Fernando - MOS synthesis quality rating results*

The evaluation results on the MOS synthesis quality are shown in fig. 6.42 for SVLN and *PSY*. The latter exhibits partially highest ratings up to an excellent synthesis quality of 5 for all but the "very tense" and "very relaxed" voice quality characteristics with the R_d offsets ±0.75. Contrariwise, the voice qualities "very tense" and "tense" are partially rated with the lowest MOS synthesis quality "poor". The evaluated mean synthesis quality MOS$_\mu$=2.67 of SVLN is comparably lower than MOS$_\mu$=3.60 for *PSY*. However, *PSY* has a higher MOS variance. The "very tense" voice quality of SVLN received comparably lower MOS ratings than its other synthesized R_d offsets. Stronger voice quality changes are assessed with less good MOS synthesis qualities for both systems. In general, *PSY* received a lower deviation from the true voice quality rating and a higher MOS synthesis quality compared to SVLN.

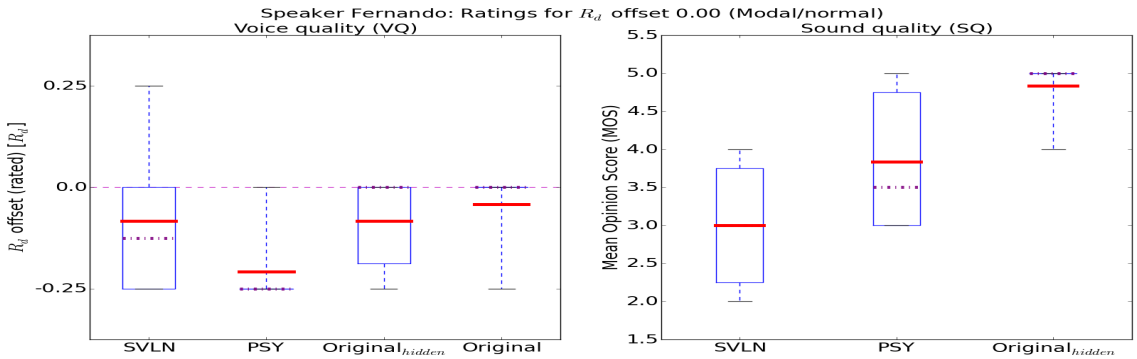


Figure 6.43: *VQ Test 1 Speaker Fernando - Results for R_d offset 0.0 with RMS energy scaling*

SVLN and *PSY* received similar voice quality ratings shown in 6.43 than the hidden original and the non-hidden original reference. A discussion concerning why the participants were asked to rate the voice quality of the non-hidden original reference is given in section 6.6.6.

GMM-based energy scaling results (in VQ and MOS):

The following paragraph discusses the results of synthesizing the different R_d offset contours in *PSY* with an additional energy scaling of the voiced $V(\omega)$ and unvoiced $U(\omega)$ component. The respective RMS energies are

predicted by dedicated GMM energy models for each part. Table 6.5 lists the Pearson product moment correlation

Table 6.5: *VQ Test 1 Speaker Fernando - Voice descriptor correlations (Pearson r)*

RMS measure	R_d	F_0	F_{VU}	$H1$-$H2$
RMS E_{voi} voiced	-0.59	0.42	0.69	-0.67
RMS E_{unv} unvoiced	-0.12	-0.02	-0.24	-0.07

coefficient r [Pearson, 1900]. It was measured over the whole corpus of speaker Fernando for the voice descriptors R_d, F_0, F_{VU} and $H1$-$H2$ against the reference values RMS E_{voi} voiced and RMS E_{unv} unvoiced. E_{voi} exhibits for all four chosen voice descriptors a reasonably high r-correlation. The descriptor set employed for this test consisted of D_E=[R_d, F_0, F_{VU}, $H1$-$H2$]. It was used to train on and predict from the voiced GMM energy models $\mathcal{M}^{voi}$ and $\mathcal{M}^{voi}_{err}$ the corresponding RMS-based energies for the voiced and unvoiced components. The fundamental frequency F_0 was omitted for the unvoiced GMM energy models $\mathcal{M}^{unv}$ and $\mathcal{M}^{unv}_{err}$ to define D'_E = [R'_d, F'_{VU}, $H'1$-$H'2$]. The reason being that F_0 exhibited a very low correlation of r=-0.02 versus the unvoiced RMS-based energy E_{unv}. 15 GMM components have been employed to train each GMM energy model for speaker Fernando.

In general it can be observed from fig. 6.44 that the GMM-based energy scaling of *PSY* received roughly similar voice transformation and synthesis quality ratings as the standard *PSY* method. This suggests that the RMS energy contours predicted by the GMM models for the voiced $V(\omega)$ and unvoiced $U(\omega)$ parts do neither increase nor decrease the synthesis quality and the voice quality characteristic to a significant extent.

Table 6.6: *VQ Test 1 Speaker Fernando - VQ voice and MOS sound quality summary*

Method	Δ VQ$_\mu$	Δ VQ$_{\sigma^2}$	MOS$_\mu$	MOS$_{\sigma^2}$
PSY	0.9524	0.6644	3.5952	1.1457
PSY (GMM)	0.6667	0.4722	3.8333	0.8056
SVLN	1.1190	0.7239	2.6667	0.5079

Table 6.6 summarizes the deviation mean Δ VQ$_\mu$ and the deviation variance Δ VQ$_{\sigma^2}$ from the optimal voice quality rating for each method in the first two columns. The corresponding mean and variance of the MOS sound quality ratings are listed in the last two columns to the right. Please note that the lower VQ and higher MOS values for *PSY* (GMM) are partially a result of having omitted the two voice quality transformations towards a tense and "very tense" voice quality. The expectation for these two omitted test cases is that they would have decreased the good test results for *PSY* (GMM).

6.6.4.3 VQ Test 1 Results - French female speaker

Another preliminary listening test on speaker Margaux was conducted by the same 6 sound processing experts of the laboratory. Fig. 6.45 depicts the six R_d^{gci} contours utilized for the test of speaker Margaux with a manually defined offset from the original R_d contour in the middle shown in cyan colour. The descriptor smoothing of SVLN is executed in the same manner as explained for speaker Fernando. The same amount of positive and negative R_d^{gci} offsets of R_d±0.25 is chosen to have the same comparison basis.

Table 6.7: *VQ Test 1 Speaker Margaux - Test indices per synthesis method and R_d offset*

Method	-0.75	-0.50	-0.25	0.0	+0.25	+0.50	+0.75
PSY	03	02	08	04	17	05	15
PSY (GMM)			01		09	11	16
SVLN	19	12	10	18	14	13	06

The synthesized speech phrases can be found online via the link for Speaker Margaux [4]. Table 6.7 lists per synthesis method and R_d offset the indices which are given for the listening test of speaker Margaux online. Readers can listen to each sound example such that the work here presented can be perceptually comprehended. Again, the lower R_d values for the voice qualities "very tense" and "tense" cause too high GMM-predicted RMS energies and

[4] Speaker Margaux: http://stefan.huber.rocks/phd/tests/RdMissM/

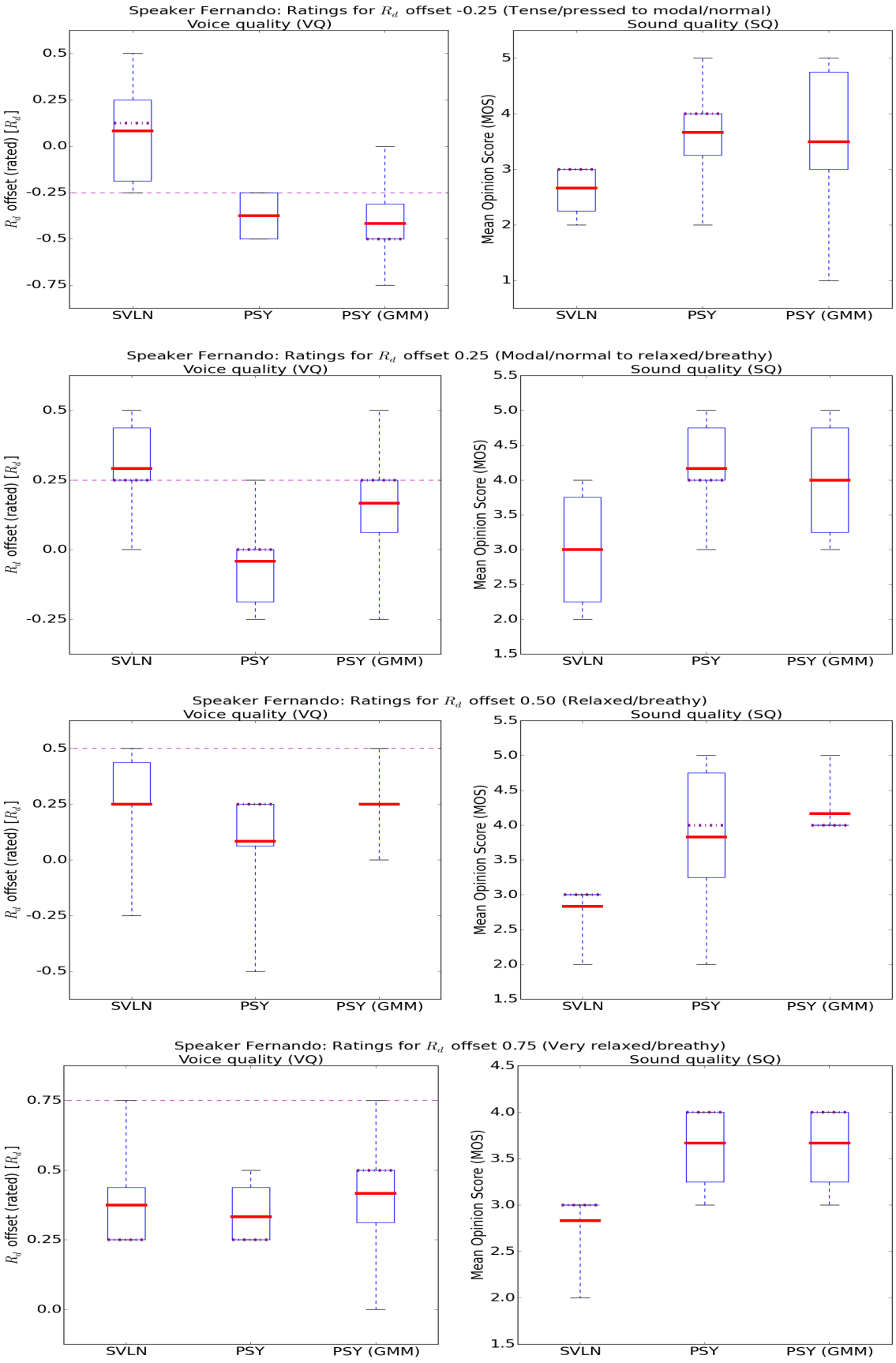


Figure 6.44: *VQ Test 1 Speaker Fernando - Results with GMM-based energy scaling*

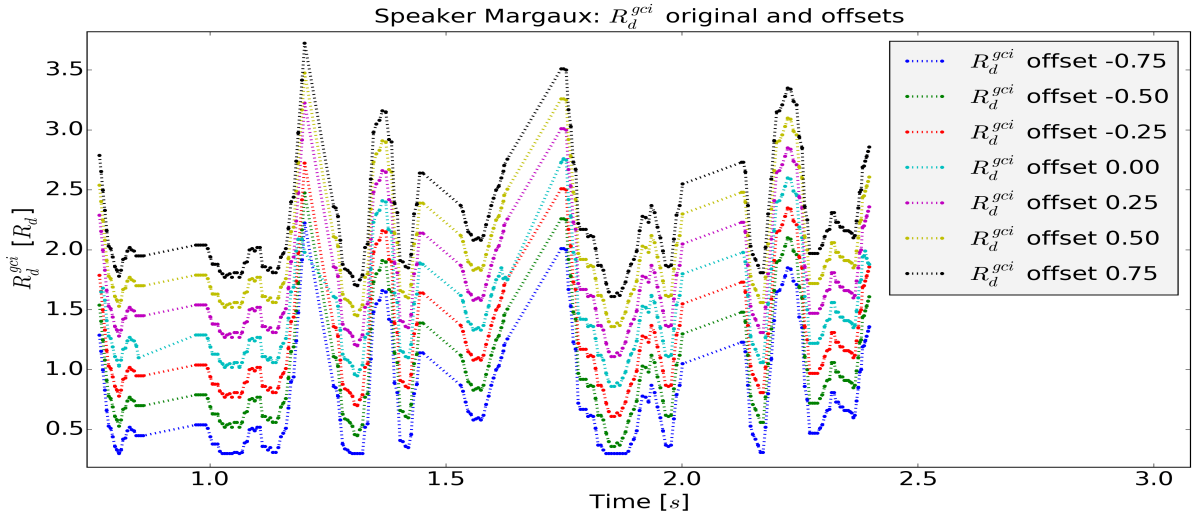


Figure 6.45: *VQ Test 1 Speaker Margaux - Manually set R_d mean offsets with step size R_d±0.25*

are excluded from the test. The hidden original speech phrase is placed at test index 07.

Voice quality (VQ) rating results:

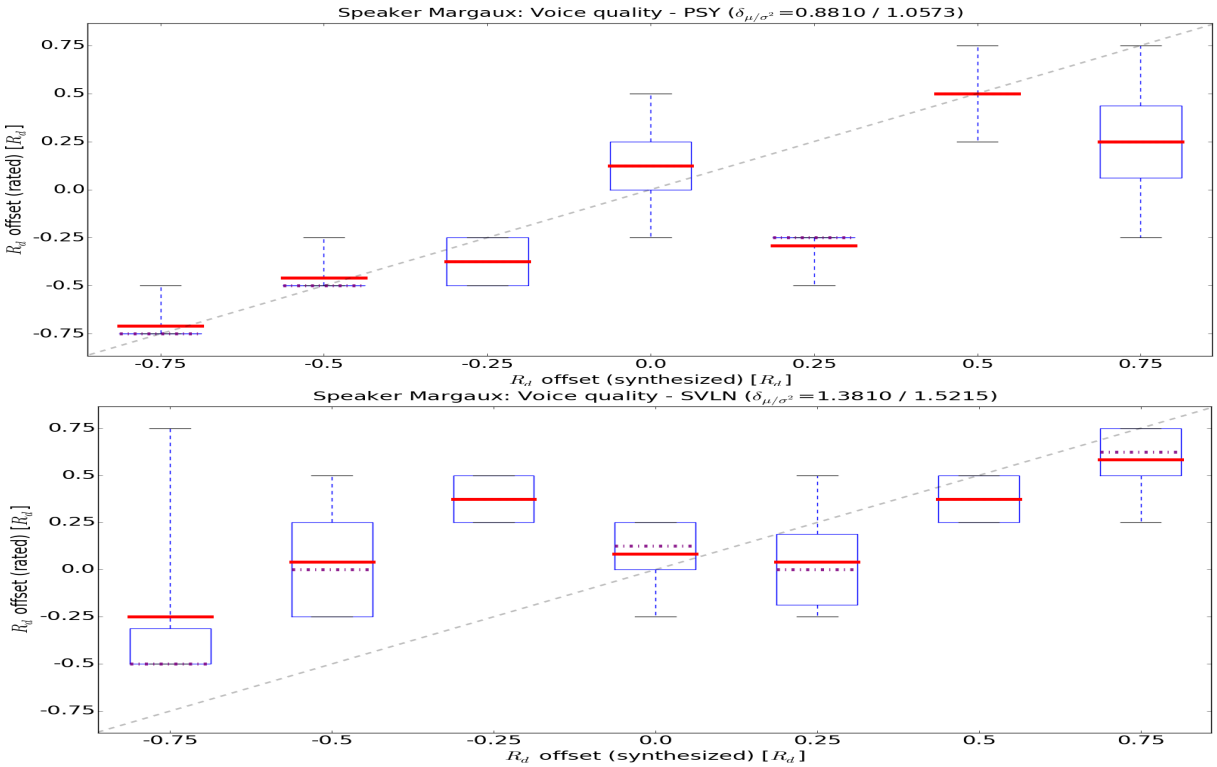


Figure 6.46: *VQ Test 1 Speaker Margaux - Voice quality rating results*

Fig. 6.46 depicts the voice quality ratings for *PSY* and SVLN. The female speaker Margaux has a lower mean deviation value δ_μ=0.88 than the male speaker Fernando with δ_μ=0.95. Especially the more tense voice qualities could be recognized well. Contrariwise, comparably higher deviations can be inspected for the R_d offsets 0.25 and 0.75 but strangely not for 0.50. SVLN has a higher mean deviation value of δ_μ=1.38. Especially the more tense voice qualities could be recognized less well. They exhibit a very high deviation from the ideal grey dashed line. The voice quality associations are for SVLN less good for the female speaker with δ_μ=1.38 compared to δ_μ=1.12 for the male speaker.

MOS synthesis quality rating results:

The MOS synthesis quality evaluation for *PSY* and SVLN are shown in fig. 6.47. *PSY* exhibits good ratings for more relaxed voice qualities. The voice quality "very tense" is mostly rated with the lowest MOS synthesis quality

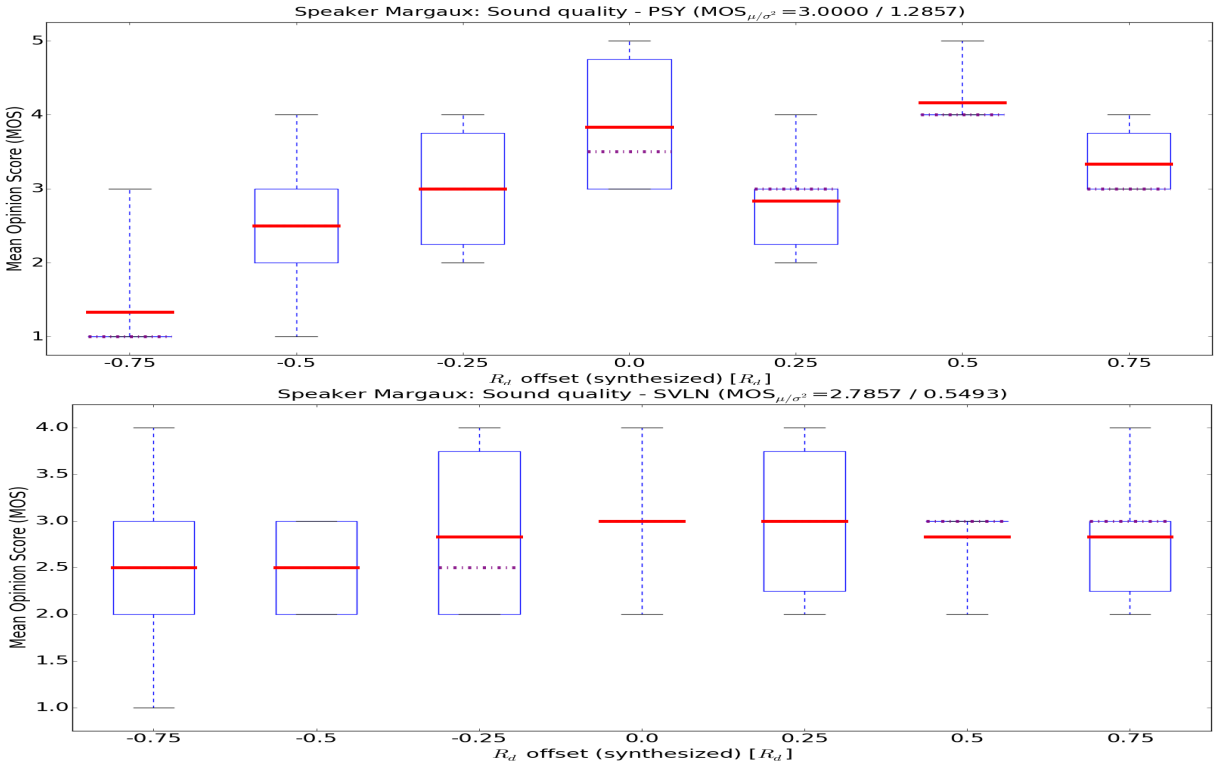


Figure 6.47: *VQ Test 1 Speaker Margaux - MOS synthesis quality rating results*

"poor". This indicates that transformations towards more tense voice qualities pose problems to the time domain mixing synthesis variant of *PSY* and may introduce degradations into the synthesized speech phrase. SVLN is rated for all presented voice qualities with a roughly same MOS quality within the range 2.5 to 3.0. But it received a comparably lower variance. This may be a result of smoothing the utilized voice descriptors. It reduces the voice quality transformation effect caused by each applied R_d offset. In general, *PSY* received a lower deviation from the true voice quality rating and a higher MOS synthesis quality compared to SVLN.

GMM-based energy scaling results (in VQ and MOS):
The rating results on the GMM-predicted RMS energies used to alter the energies of the voiced $V(\omega)$ and unvoiced $U(\omega)$ parts are examined as well for speaker Margaux. Table 6.8 lists again the Pearson r-correlations for the voice

Table 6.8: *VQ Test 1 Speaker Margaux - Voice descriptor correlations (Pearson r)*

RMS measure	R_d	F_0	F_{VU}	$H1$-$H2$
RMS E_{voi} voiced	0.10	-0.05	0.49	-0.51
RMS E_{unv} unvoiced	-0.10	0.30	-0.12	-0.27

descriptor set measured against the reference values E_{voi} and E_{unv}. The r-correlations for the energy references are in general lower for the female compared to the male speaker. The fundamental frequency F_0 exhibits a very low correlation with the voiced RMS energy and is not included into the GMM energy models $\mathcal{M}^{voi}$ and $\mathcal{M}^{voi}_{err}$ for speaker Margaux: D_E=[$R_d, F_{VU}, H1$-$H2$]. The unvoiced GMM energy models $\mathcal{M}^{unv}$ and $\mathcal{M}^{unv}_{err}$ received all four voice descriptors for training and prediction: $D'_E = [R'_d, F'_0, F'_{VU}, H'1$-$H'2]$. 10 GMM components have been employed to train each GMM energy model. SVLN and *PSY* received similar voice quality ratings than the hidden original and the non-hidden original reference, shown in 6.48. The direct re-synthesis of both methods as well as the original and the hidden original references are perceived in general as slightly more relaxed. The listeners could well recognize the original hidden reference as highest performing synthesis quality, whereas SVLN received the lowest MOS ratings. In general it can be observed that the GMM-based energy scaling of *PSY* received roughly similar voice and synthesis quality ratings as the standard *PSY* method, depicted in fig. 6.49. This suggests that the GMM predicted energy contours for the voiced $V(\omega)$ and unvoiced $U(\omega)$ parts do neither increase nor decrease the synthesis quality and the voice quality characteristic to a significant extent.

Table 6.9 summarizes again the mean deviation Δ VQ$_\mu$ and its variance Δ VQ$_{\sigma^2}$ from the optimal voice quality rating, as well as its corresponding mean and variance MOS quality rating for each method. Please note again that the lower VQ and higher MOS values for *PSY* (GMM) are partially a result of omitting the tense voice quality

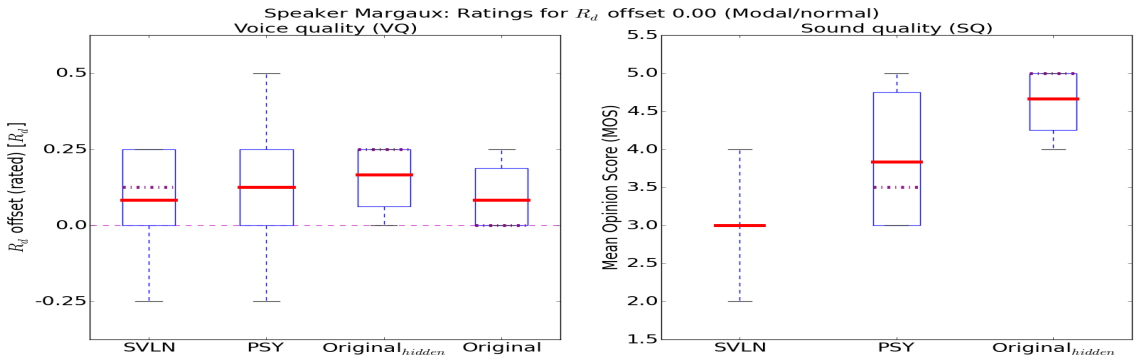


Figure 6.48: *VQ Test 1 Speaker Margaux - Results for R_d offset 0.0*

Table 6.9: *VQ Test 1 Speaker Margaux - VQ voice and MOS sound quality summary*

Method	Δ VQ$_\mu$	Δ VQ$_{\sigma^2}$	MOS$_\mu$	MOS$_{\sigma^2}$
PSY	0.8810	1.0573	3.0000	1.2857
PSY (GMM)	0.7083	0.5399	3.1667	0.9722
SVLN	1.3810	1.5215	2.7857	0.5493

transformations.

6.6.4.4 Conclusions

The sound quality assessments of *PSY* exhibits for more tense voice qualities in comparison with SVLN lower MOS ratings, illustrated in figures 6.42 and 6.47. This indicates that the evaluated time domain mixing synthesis variant of *PSY* is not able to produce such voice quality characteristics with a sufficiently high synthesis quality. The reason is already explained in section 6.5.5. Audible artefacts may be introduced in the synthesis if the voiced part describes higher amplitudes up to F_{nyq} as the unvoiced counterpart. This occurs for lower R_d values for transformations towards a tense voice quality. This motivates the spectral fading synthesis variant of section 6.5.5. It will be evaluated in the following section. Please note that the GMM-based energy prediction evaluated in this section will not be examined further. It proved to not achieve any improvements due to the partially erroneous energy predictions explained by step I in section 6.6.4.1.

6.6.5 Voice Quality (VQ) Test 2: Spectral fading and R_d transformation

The following sections present the results of a listening test conducted on natural human speech for the three speakers BDL, Fernando and Margaux. The *PSY* synthesis variant of section 6.5.5 to fade in the unvoiced and fade out the voiced part around the F_{VU} frequency is examined on its ability to establish different voice quality characteristics from an original speech recording. The voiced component $V(\omega)$ is modified according to the generation of different $R_d^{'gci}$ contours span over the complete R_d range. The contour spanning is explained in section 6.4.3.2. The unvoiced component $U(\omega)$ remains unmodified since the GMM-based energy modification, evaluated in the preceding section, does not predict energies within the desired amplitude range and may lead to artefacts. No algorithmic changes are applied to the SVLN baseline method. It receives contrary to VQ Test 1 the transformed $R_d^{'gci}$ contours and not the simple R_d mean offset contours.

6.6.5.1 VQ Test 2 Results - French male speaker

Transformation of the R_d^{gci} contour:
Fig. 6.50 illustrates the 3 positive and the 3 negative OQ_μ offsets for speaker Fernando, span logarithmically over the OQ range in higher and lower OQ value direction. The horizontal dashed black line depicts the upper OQ border at OQ_{max}=0.9297. This corresponds to R_d^{max}=5.0. The transformation to a "very relaxed" voice quality with OQ offset +3 in blue results in OQ values lying above OQ_{max}. Fig. 6.51 exemplifies the transformation back to the R_d range. It applies a hard saturation at the R_d range borders. The natural variation of the glottal pulse shape contour is lost for the R_d offset +3 contour in blue dots, e.g. around 1.2, 2.1 and 2.6 seconds. Fig. 6.52 shows the effect of applying the soft saturation algorithm, introduced in section 6.4.3.2, to the different $OQ^{'gci}$ contour of fig. 6.50. Transforming back the soft saturated $OQ^{'gci}$ contours to the R_d range gives the transformed $R_d^{'gci}$ contours

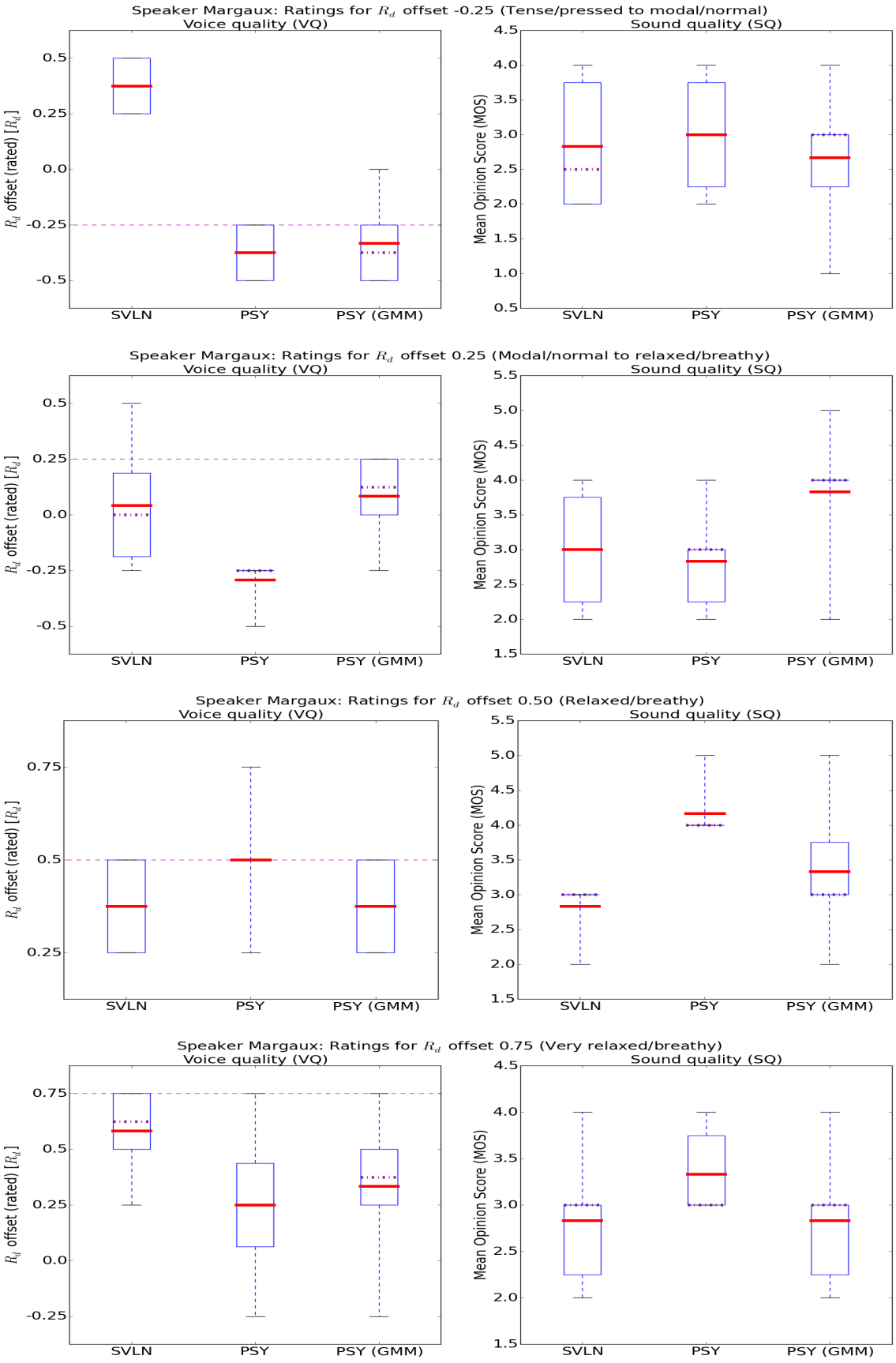


Figure 6.49: *VQ Test 1 Speaker Margaux - Results with GMM-based energy scaling*

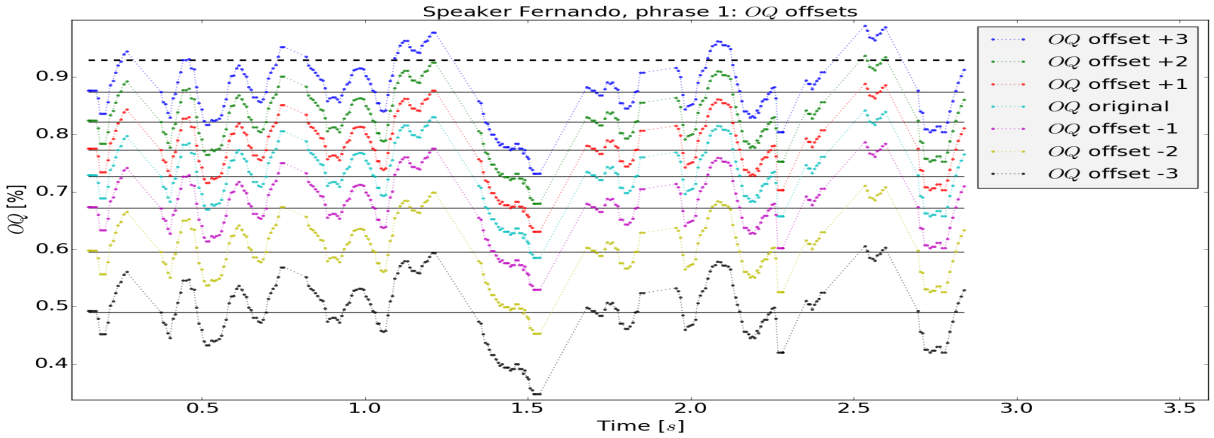


Figure 6.50: *VQ Test 2 Speaker Fernando - Example of $OQ^{'gci}$ contour generation with hard saturation*

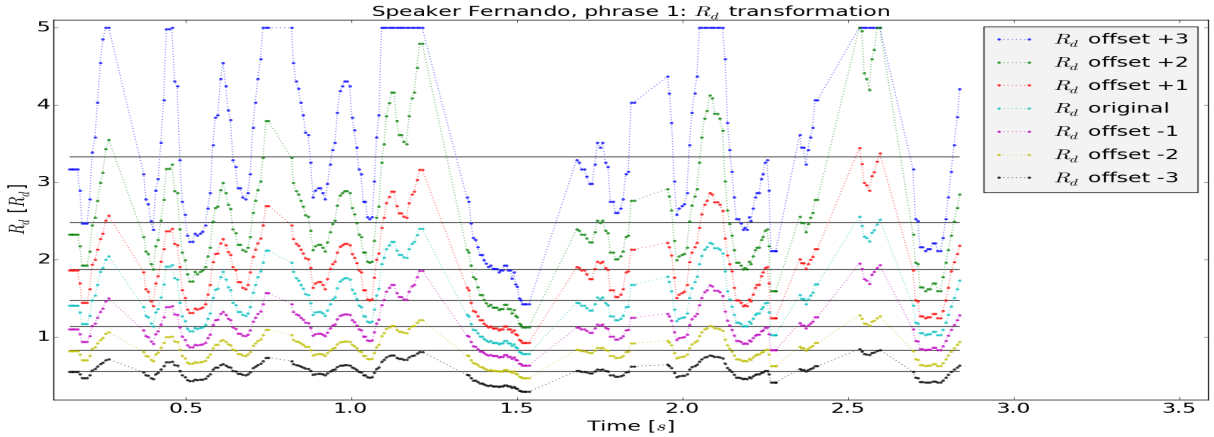


Figure 6.51: *VQ Test 2 Speaker Fernando - Example of $R_d^{'gci}$ contour generation with hard saturation*

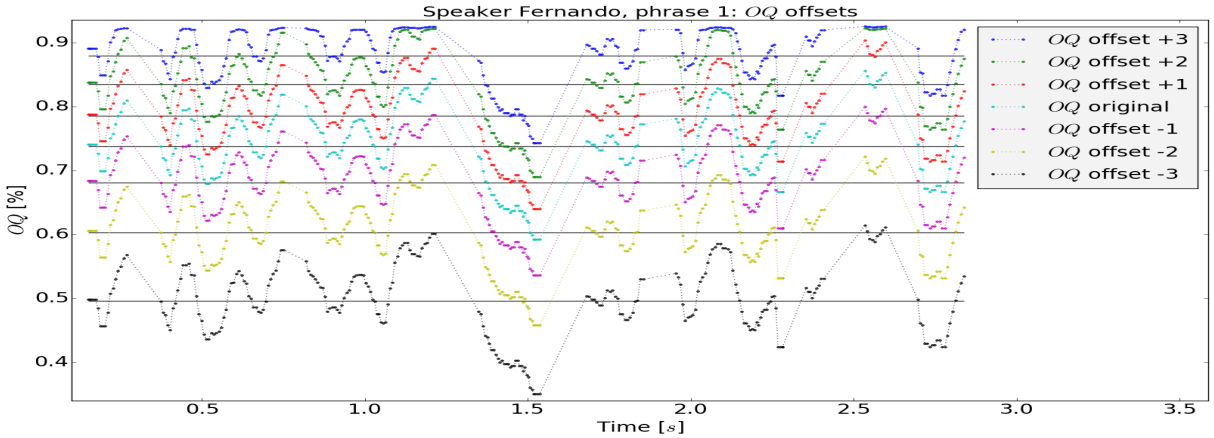


Figure 6.52: *VQ Test 2 Speaker Fernando - Example of $OQ^{'gci}$ contour generation with soft saturation*

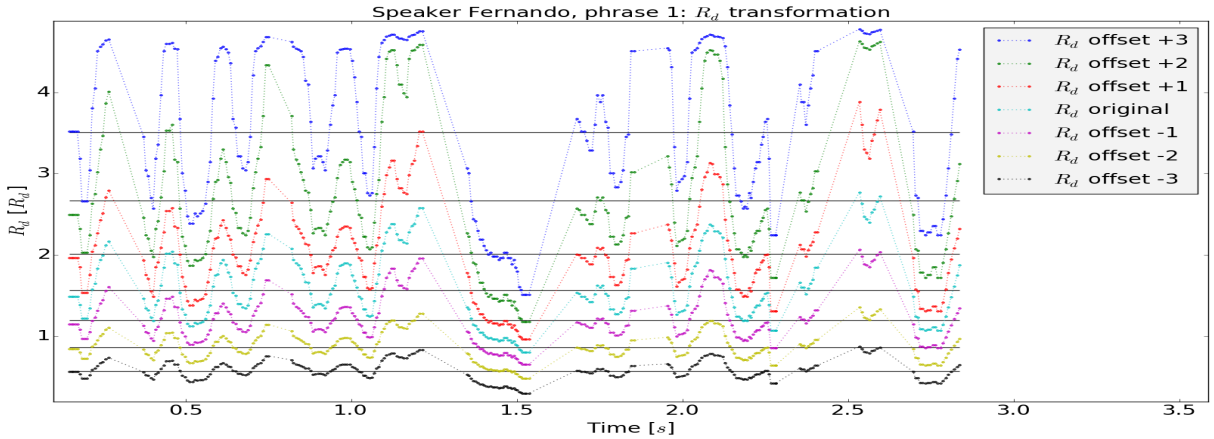


Figure 6.53: *VQ Test 2 Speaker Fernando - Example of $R_d^{'gci}$ contour generation with soft saturation*

depicted in fig. 6.53. The $R_d^{'gci}$ contours with soft saturation maintain a certain variation, remain within the R_d range and are therefore utilized for this listening test of speaker Fernando. Table 6.10 shows the OQ_μ and R_d^μ mean

Table 6.10: *VQ Test 2 Speaker Fernando - Example OQ and R_d values for voice quality transformation*

Voice quality (index)	OQ_μ	R_d^μ	$R_d^{\sigma^2}$	ΔR_d^μ
Very relaxed (+3)	0.8793	3.5109	0.9031	-0.8397
Relaxed (+2)	0.8344	2.6711	0.7825	-0.6597
Modal to relaxed (+1)	0.7853	2.0114	0.3631	-0.4442
Modal (original) (0)	0.7382	1.5673	0.1937	
Tense to modal (-1)	0.6814	1.1936	0.0941	-0.3737
Tense (-2)	0.6034	0.8601	0.0341	-0.3335
Very tense (-3)	0.4960	0.5704	0.0154	-0.2898

values of the original unmodified R_d^{gci} contour with index 0, and respectively 3 positive and 3 negative μ values for each voice quality change identified by the indices in parentheses. $R_d^{\sigma^2}$ lists their variance σ^2 around the mean. It increases with increasing R_d to reflect the JND threshold objective of having to apply higher ΔR_d steps with higher R_d values, discussed in section 6.4.3.2. The R_d^μ (diff) column reflects the ΔR_d steps measured between each row index on the R_d^μ values. The μ difference increases with increasing R_d^μ as well. Please note that the OQ steps in lower value regions are due to the non-linear warping between the OQ and the R_d range higher. This behaviour does not follow the assumption observed with the study on perceptual JND differences in [Henrich et al., 2003]. However, since a change in the R_d value implies a change of OQ, α_m and t_a, the JND threshold objective is still valid.

Listening test setup:
11 participants rated each speech phrase according to the voice quality characteristics given in the first column of table 6.10. The speech phrases are available online via the following link for Speaker Fernando [5]. Table 6.11 lists

Table 6.11: *VQ Test 2 Speaker Fernando - Test indices per synthesis method and transformation index*

Method	-3	-2	-1	0	+1	+2	+3
PSY	10	03	14	06	04	13	07
SVLN	01	12	05	08	11	02	09

for SVLN and *PSY* the listening test indices for each voice quality transformation offset. This facilitates to listen to each sound example such that the work presented here can be transparently followed. The hidden original speech phrase is placed at test index 15.

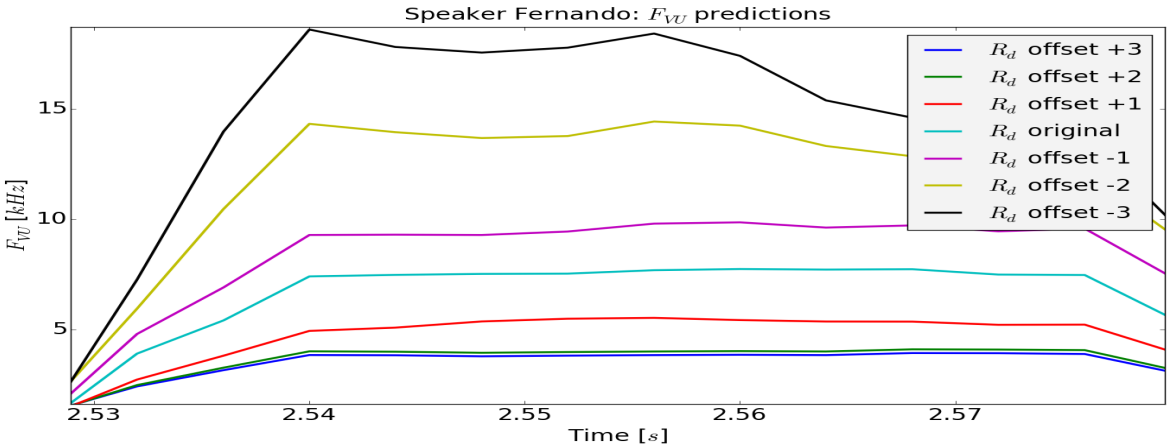


Figure 6.54: *VQ Test 2 Speaker Fernando - F_{VU} prediction excerpt in PSY*

The spectral fading synthesis variant of *PSY*, presented in section 6.5.5, requires the F_{VU} prediction of section 6.4.2.4. A short speech segment example to predict F_{VU} for speaker Fernando is depicted in fig. 6.54.

[5] Speaker Fernando: http://stefan.huber.rocks/phd/tests/vqMisterF/

Listening test results:

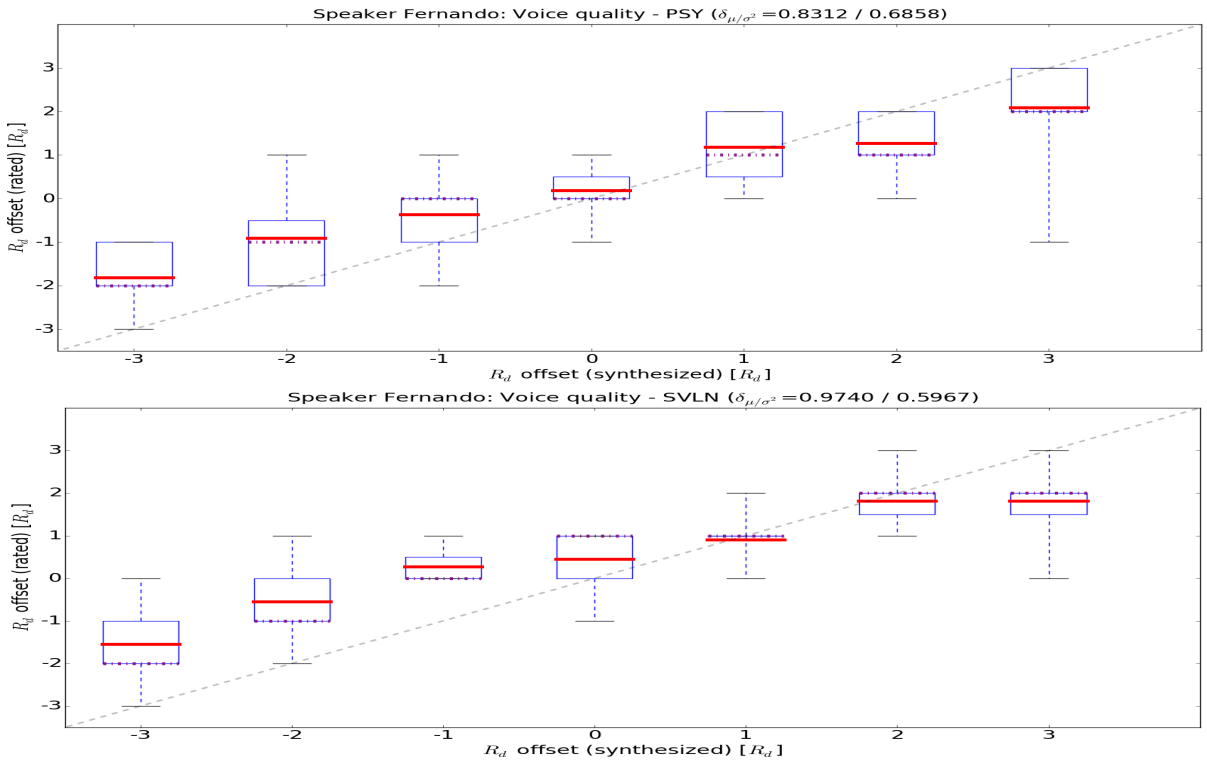


Figure 6.55: *VQ Test 2 Speaker Fernando - Voice quality rating results*

Fig. 6.55 depicts the voice quality ratings for the methods *PSY* and SVLN. The horizontal red (violet) lines reflect the mean (median) voice quality ratings of all participants per test phrase with the same indices as in table 6.10. The dialog grey dashed line exemplifies their ideal placement if each test participant would have been able to associate perceptually each synthesized voice quality example to its corresponding voice quality characteristic. The mean deviation value δ_μ=0.83 for *PSY* expresses the disagreement of the listeners, being ideally δ_μ=0.00. A higher mean deviation value δ_μ=0.97 indicates for the baseline method SVLN that the listeners could less well capture the different synthesized voice qualities. Roughly good voice quality associations can be concluded for both systems. Both follow roughly the ideal dashed grey line with the deviations increasing with higher changes.

The MOS synthesis quality evaluation shown in fig. 6.56 for *PSY* exhibits partially highest ratings up to an excellent synthesis quality of 5 for all but the "relaxed" and "very relaxed" voice quality characteristics with index +2 and +3. The evaluated mean synthesis quality MOS_μ=2.82 of SVLN is comparably lower than MOS_μ=3.38 for *PSY*. Stronger voice quality changes are assessed with less good MOS synthesis qualities for both systems.

Table 6.12: *VQ Test 2 Speaker Fernando - VQ voice and MOS sound quality summary*

Method	Δ VQ$_\mu$	Δ VQ$^2_\sigma$	MOS$_\mu$	MOS$^2_\sigma$
PSY	0.8312	0.6858	3.3766	0.9880
SVLN	0.9740	0.5967	2.8182	0.7462

In general, *PSY* received a lower deviation from the true voice quality rating and a higher MOS synthesis quality compared to SVLN, summarized in table 6.12.

Listening test setup:

Table 6.13: *VQ Test 2 Speaker BDL - Test indices per synthesis method and transformation index*

Method	-3	-2	-1	0	+1	+2	+3
PSY	10	11	04	06	02	08	07
SVLN	13	15	09	03	14	12	01

18 participants evaluated the re-synthesized sound examples of the first speech recording of the CMU Arctic

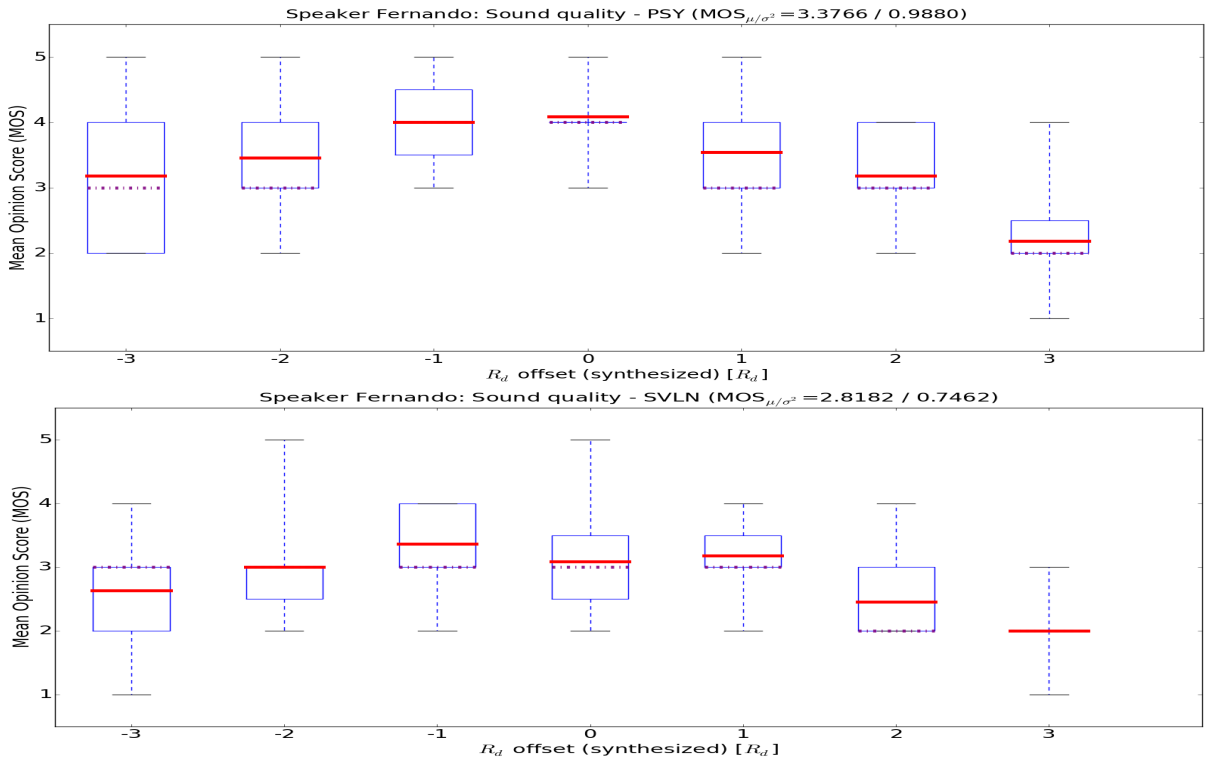


Figure 6.56: *VQ Test 2 Speaker Fernando - MOS synthesis quality rating results*

speaker BDL. The synthesized speech phrases can be found online via the following link for Speaker BDL [6]. Table 6.13 lists for each voice quality transformation for the listening test VQ test 2 the indices for speaker BDL. Each sound example can thus be heard to follow the presented work. The hidden original speech phrase is placed at test index 05.

6.6.5.2 VQ Test 2 Results - English male speaker

Transformation of the R_d^{gci} contour:

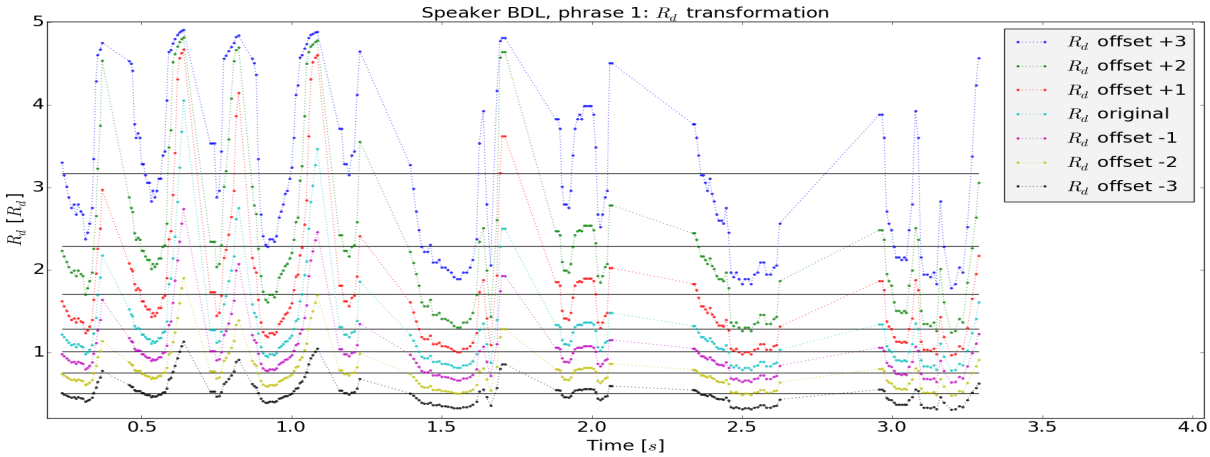


Figure 6.57: *VQ Test 2 Speaker BDL - Example of $R_d^{'gci}$ contour generation with soft saturation*

Fig. 6.57 shows the original and transformed $R_d^{'gci}$ contours utilized for the voice quality transformation listening test of speaker BDL.

Table 6.14 lists the OQ_μ and R_d^μ mean values of the original R_d^{gci} and the transformed $R_d^{'gci}$ contours. The ΔR_d^μ mean difference and the $R_d^{\sigma^2}$ variance columns to the right exhibit increasing values from the $R_d^{'gci'}$ contour with index -3

[6] Speaker BDL: http://stefan.huber.rocks/phd/tests/vqMisterBDL/

Table 6.14: *VQ Test 2 Speaker BDL - Example OQ and R_d values for voice quality transformation*

Voice quality (index)	OQ_μ	R_d^μ	$R_d^{\sigma^2}$	ΔR_d^μ
Very relaxed (+3)	0.8633	3.1688	0.8891	-0.8780
Relaxed (+2)	0.8047	2.2909	0.8049	-0.5823
Modal to relaxed (+1)	0.7469	1.7086	0.5924	-0.4203
Modal (original) (0)	0.6912	1.2883	0.2849	
Tense to modal (-1)	0.6372	1.0153	0.1356	-0.2730
Tense (-2)	0.5638	0.7546	0.0569	-0.2607
Very tense (-3)	0.4642	0.5086	0.0238	-0.2460

for a "very tense" to index +3 for a "very relaxed" voice quality. This reflects as well the JND threshold objective discussed in section 6.4.3.2.

Time and spectral domain examples:

Fig. 6.15 exemplifies the signal waveforms produced for the voice quality transformation test. The left column shows the time and the right column the spectral domain representation for each seven transformed and re-synthesized speech phrases of *PSY* for speaker BDL. Please note that the middle row does not imply a transformation. It shows the re-synthesized original speech phrase The time domain examples show the impact of the energy maintenance on the re-synthesized waveforms. The loudness has been altered and the sinusoidal content has been modified due to the transformed voice qualities towards tense or relaxed phonation types. Despite, the amplitude envelopes and peak time domain amplitude appear visually to not have changed to a huge extent. More interesting is the visual inspection of the spectral domain examples. The "very relaxed" voice quality of the first row contains less sinusoidal content in higher frequency regions than the "very tense" voice quality in the last row. Sinusoidal content appears by the harmonic interval defined by the pitch period contour over time and is shown in red colour. It can be distincted from the signal parts having less energy in yellow or green colour.

Listening test results:

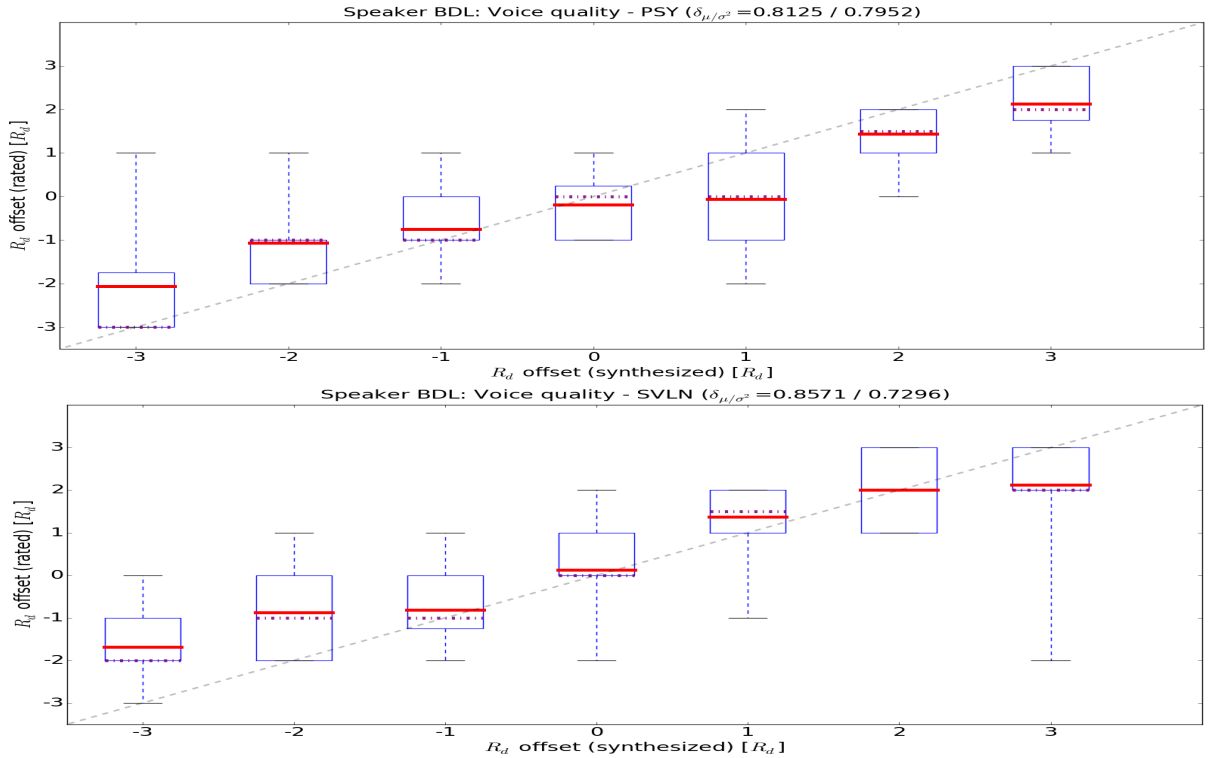
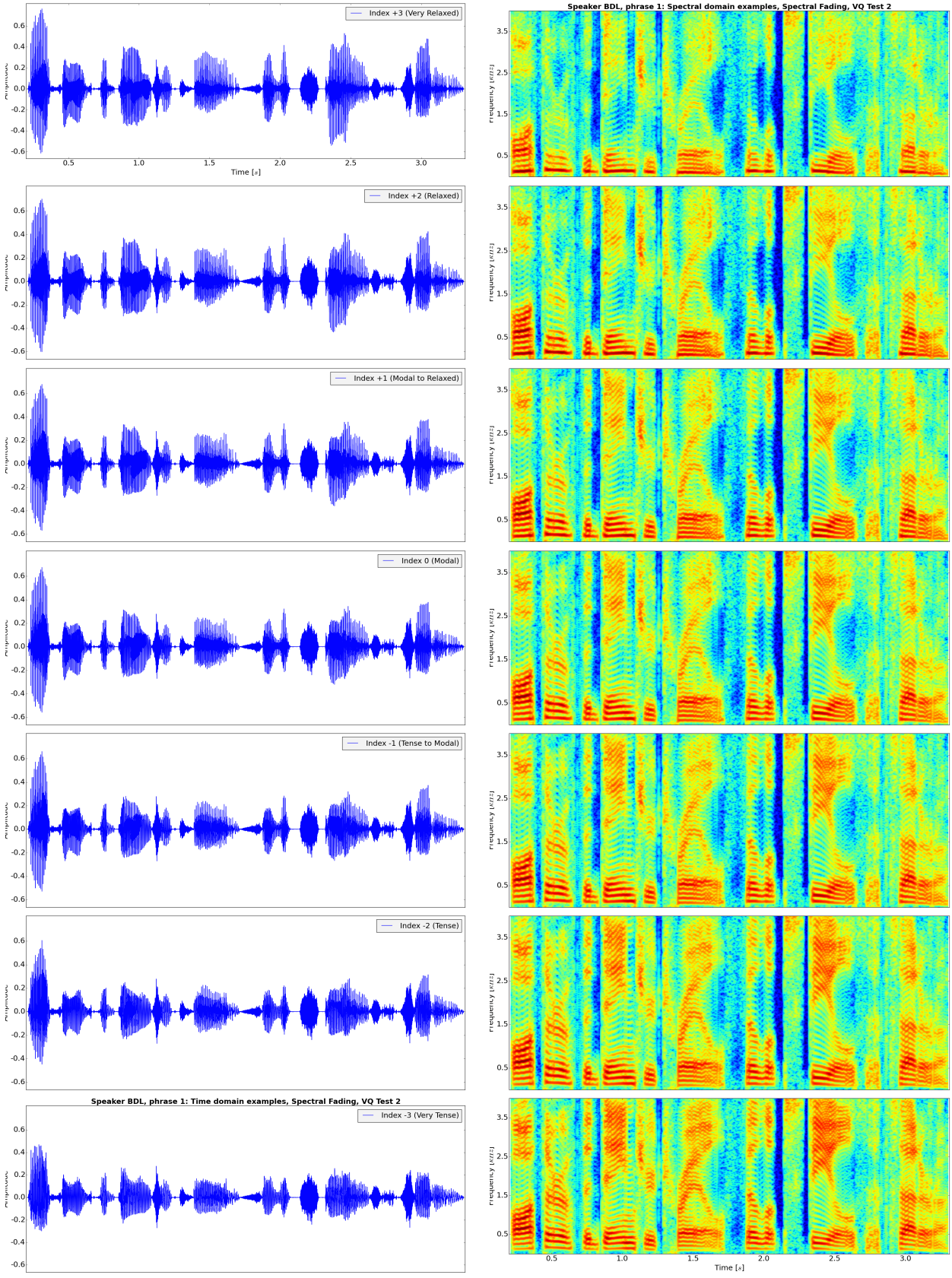


Figure 6.58: *VQ Test 2 Speaker BDL - Voice quality rating results*

Fig. 6.58 depicts the voice quality ratings for the proposed method *PSY* and the baseline method SVLN. The horizontal red (violet) lines reflect the mean (median) voice quality ratings of all participants per test phrase with the same indices as in table 6.14. The mean deviation value δ_μ=0.81 for *PSY* expresses a slightly lower disagreement of the listeners as with the slightly higher mean deviation value δ_μ=0.88 for SVLN. However, *PSY* has a higher variance for the voice quality rating than SVLN. The whisker in fig. 6.58 for SVLN at the voice quality index -2 (vertical) for the synthesized voice quality index +3 (horizontal) indicates that one test participant perceived

Table 6.15: *VQ test 2 Speaker BDL - Time and spectral domain examples of utilized speech waveforms*



a bigger disagreement with this transformed speech phrase synthesized by SVLN. The same is valid for *PSY* for voice quality index -3 having received a voice quality rating at index +1. Still, good voice quality associations can be again concluded for both systems on average, following roughly the ideal dashed line.

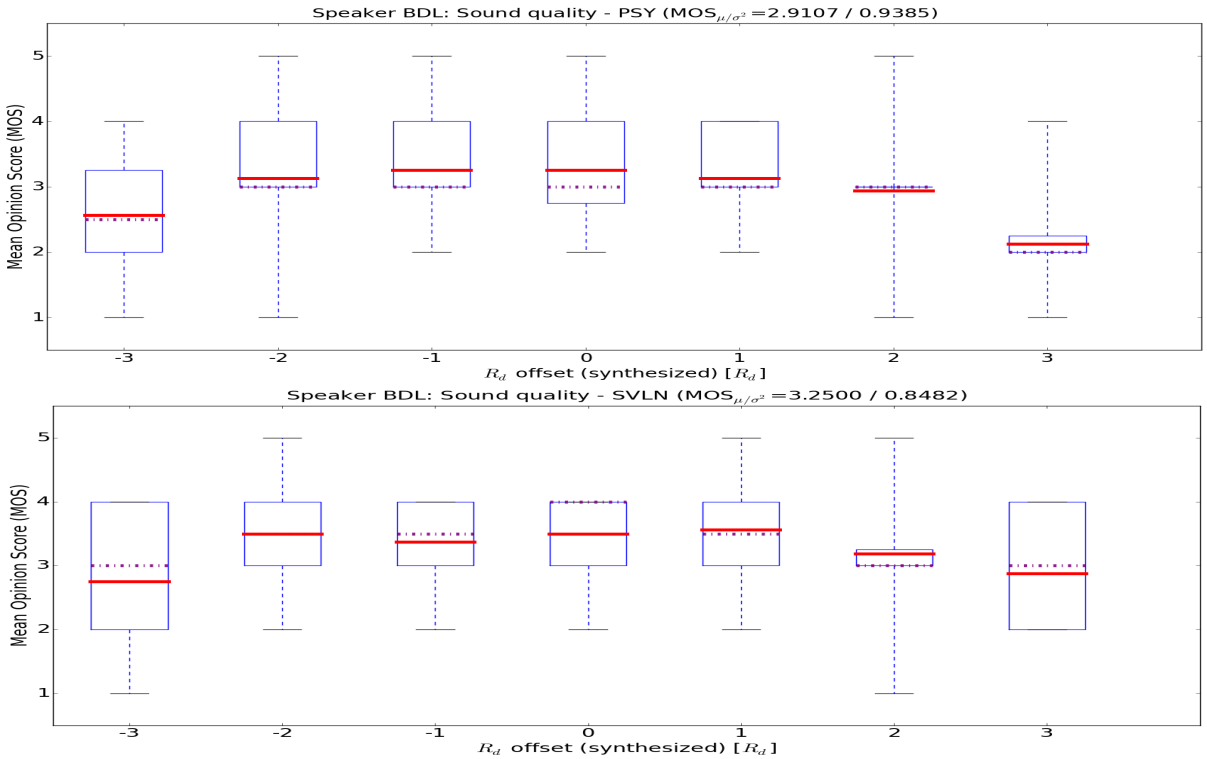


Figure 6.59: *VQ Test 2 Speaker BDL - MOS synthesis quality rating results*

The MOS synthesis quality evaluation for *PSY* depicted in fig. 6.59 indicates with MOS_{μ}=2.91 a lower synthesis quality as SVLN with MOS_{μ}=3.25. Also its variance is higher than the one of SVLN. Apparently, the participants judged the relaxed voice quality of index +2 with the lowest "poor" and the highest "excellent" MOS rating for both systems. The same observation is valid for voice quality index -2 for a tense voice quality synthesized by *PSY*.

Table 6.16: *VQ Test 2 Speaker BDL - VQ voice and MOS sound quality summary*

Method	Δ VQ$_{\mu}$	Δ VQ$^2_{\sigma}$	MOS$_{\mu}$	MOS$^2_{\sigma}$
PSY	0.8125	0.7952	2.9107	0.9385
SVLN	0.8571	0.7296	3.2500	0.8482

The test result summary for speaker BDL given in table 6.16 shows similar voice quality rating results for both systems. The MOS synthesis quality is lower for *PSY*. An explanation of this observation can be found in the following paragraph.

Creaky voice quality: GCI timing correction

Creaky voice segments are present in the utilized recording of speaker BDL, e.g. at the syllable endings of the speech segments around 1.60 to 1.72 and 3.08 to 3.30 seconds. This can be inspected via the links given at the beginning of this section. The study of [Drugman et al., 2013] estimates that speaker BDL produces creaky voice segments for about 60 % at the two phonemes preceding a pause. The presence of creak actually prohibits the utilization of the current implementation of the GCI time instant correction algorithm on voiced borders in *PSY*. The algorithm is presented as step 6 in section 6.2.1.3 and exemplified by fig. 6.3. The GCI timing correction cannot handle for the time being the aperiodic fundamental pulse sequence present in speech segments containing a creaky voice quality. The algorithm was therefore not activated to process the given speech phrase of speaker BDL.

The better synthesis quality of SVLN may be a result of neglecting the estimated GCI positions. SVLN first interpolates each $R_d^{'gci}$ contour to an R_d contour on its STFT time basis and then applies median smoothing. At each synthesis frame the glottal pulse is synthesized in the spectral domain according to the interpolated and

smoothed R_d contour. This removes the irregularly placed GCI time sequence of the creaky voice quality from voiced speech segments containing creak. In contrast, *PSY* maintains the GCI estimation and does not interpolate nor smooth from a given $R_d^{'gci}$ to an R_d contour. It relies on a glottal pulse sequence synthesized over the complete analyzed speech phrase. This design is one reason that *PSY* is able to achieve higher voice quality and MOS synthesis quality ratings for most of the evaluated examples. However, this is not the case for the given speech phrase of speaker BDL.

Creaky voice quality: GCI detection
The aperiodic pulse sequence given for the creaky voice segments may cause problems for *PSY*. Both evaluated speech systems do for the time being not contain an algorithmic solution to model creak. *PSY* tries on the one hand to maintain specific speech details in the synthesis by not interpolating and smoothing the estimated voice descriptor input set. It is on the other hand more prone to errors resulting into artefacts if some specific details cannot be estimated or handled precisely enough.

An example of a failing GCI detection can be inspected in fig. 6.9 representing the glottal pulse $g_{R_d}^{gci}$ and the spectral envelope sequence $\mathcal{T}_g$ for *PSY*. A missing or wrongly estimated GCI time instant produces a hole in the glottal pulse sequence around 3.185 seconds. It leads to a deformation or interruption of the harmonic sinusoidal sequence at this time instant. This failure is for example further transferred to the full-band VTF $C_{full}(\omega)$ illustrated in fig. 6.14 which exhibits a spike over the whole spectrum at the same time instant.

Signal spikes in original recording
Another interesting observation are two signal spikes found in the original speech recording of BDL phrase 1 at 1.568 and 1.658 seconds, illustrated as vertical dashed lines in the spectral plots of fig. 6.60. The spikes cannot be observed from the time domain plot at the top. It shows from top to down the corresponding signal segment of BDL in the time and the spectral domain. As well the spectra of the direct re-synthesis by *PSY* and SVLN without any transformations applied are depicted. The spikes may for example be caused by saliva present in the oral cavity of speaker BDL. *PSY* reproduces the two spikes. This leads to slightly audible artefacts for voice quality transformations towards a more relaxed voice quality. More sinusoidal content is removed from the original recording in this case. The signal spikes are more revealed. It leads to a higher perceptual sensation of an artefact. This may rise in the listener the opinion that these artefacts are a result of a possible algorithmic error. Their sound quality ratings for the given example may degrade accordingly. However, the true reason is a signal spike in the original speech recording.

SVLN in contrast does not reproduce such fine signal details. This is clearly visible in the lowest plot where the horizontal signal spike at 1.658 seconds is mostly removed. Only the higher energetic plop around 13 kHz remains present. However, the plop is actually smeared into neighbouring frames. *PSY* reproduces closely the plop in comparison to the original signal. On the one hand this is in general a good behaviour of the novel speech system *PSY*. But it leads in this case to the error-prone drawback discussed here.

The reasons for having chosen the speech phrase of BDL despite such known problems is the nevertheless reasonably good performance of both speech systems and the interesting finding with the signal spikes shown in fig. 6.60. Both systems could process speech segments containing the creaky voice quality without explicitly addressing its intrinsic signal behaviour. Also the synthesis quality did not deteriorate to an unacceptable amount. Another reason for choosing the first phrase of the CMU Arctic speaker BDL is the widespread degree of popularity of this speech phrase. It allows transparently following the presented research work. Please note that the behaviour concerning an amplification of such signal spikes in the context of voice quality transformation was also observed for other tested speech phrases not shown here.

6.6.5.3 VQ Test 2 Results - French female speaker

Transformation of the R_d^{gci} contour:
Fig. 6.61 depicts the original R_d^{gci} contour for speaker Margaux in cyan-coloured dots found in the middle. The transformed $R_d^{'gci}$ contours above and below are fit into the R_d range by applying again the soft saturation algorithm introduced in section 6.4.3.2. All illustrated $R_d^{'gci}$ contours are utilized for the listening test on voice quality transformation for speaker Margaux here presented.

Table 6.17 shows the OQ_μ and R_d^μ mean values for the voice quality transformation. The JND threshold objective discussed in section 6.4.3.2 is as well respected since the variance $R_d^{\sigma^2}$ listed in the 3^{rd} and the mean difference ΔR_d^μ listed in the 4^{th} column increase with the increasing test indices from -3 to +3.

Listening test setup:
15 participants evaluated the transformed and re-synthesized sound examples of the first speech recording of the French female speaker Margaux. The synthesized speech phrases can be found online via this link for Speaker

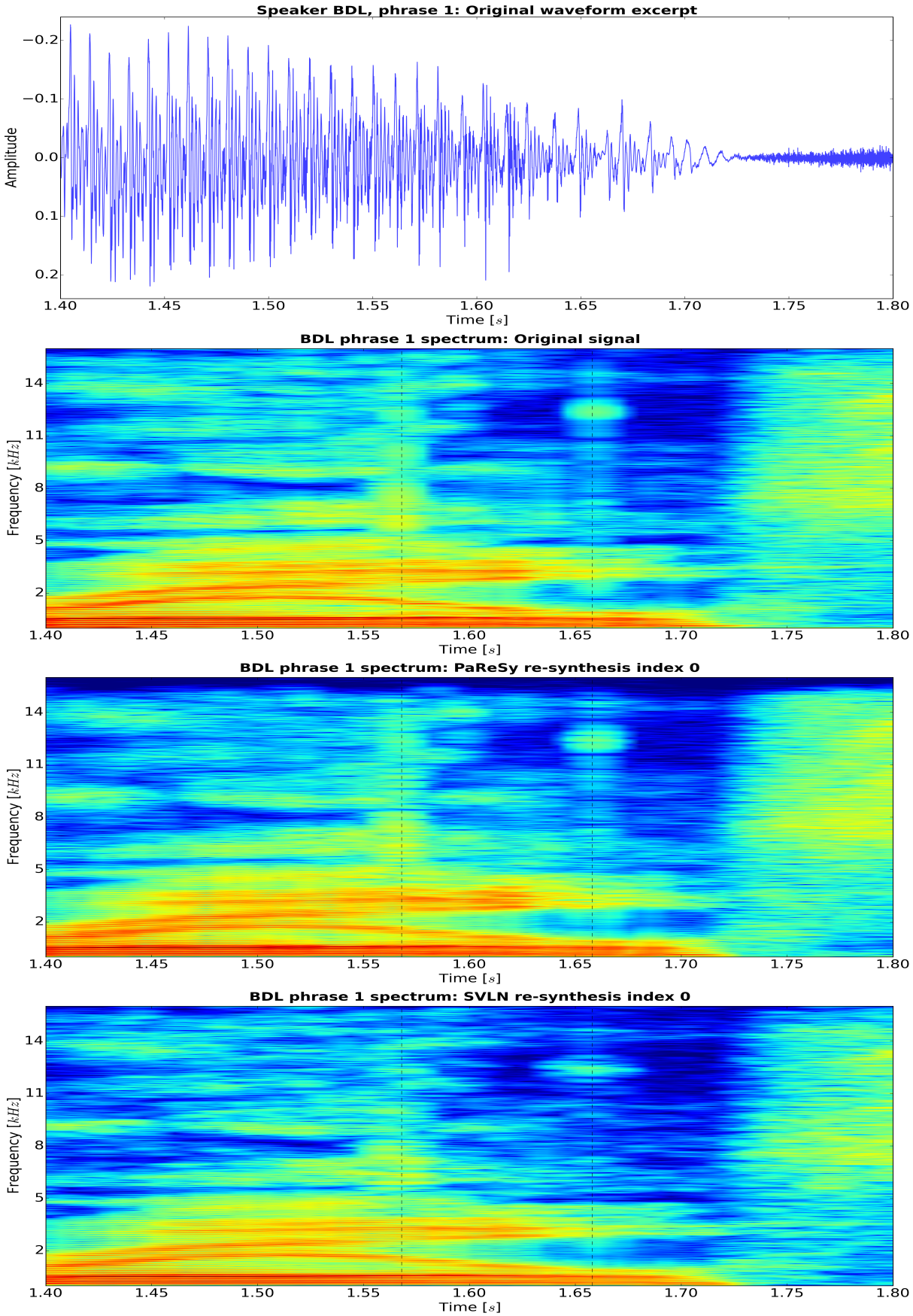


Figure 6.60: *VQ test 2 Speaker BDL - Signal spike transferred from original recording to re-synthesis*

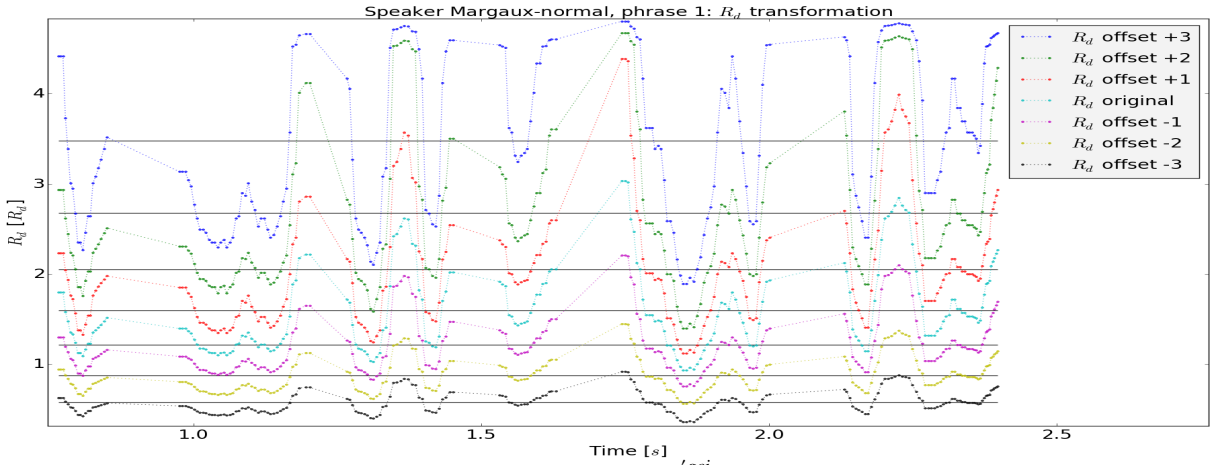


Figure 6.61: *VQ Test 2 Speaker Margaux - Example of $R_d^{'gci}$ contour generation with soft saturation*

Table 6.17: *VQ Test 2 Speaker Margaux - Example OQ and R_d values for voice quality transformation*

Voice quality (index)	OQ_μ	R_d^μ	$R_d^{\sigma^2}$	ΔR_d^μ
Very relaxed (+3)	0.8801	3.4793	0.8001	-0.8004
Relaxed (+2)	0.8360	2.6789	0.7341	-0.6276
Modal to relaxed (+1)	0.7889	2.0513	0.4410	-0.4558
Modal (original) (0)	0.7427	1.5954	0.2176	
Tense to modal (-1)	0.6858	1.2141	0.1045	-0.3813
Tense (-2)	0.6073	0.8724	0.0360	-0.3417
Very tense (-3)	0.4992	0.5767	0.0151	-0.2957

Table 6.18: *VQ Test 2 Speaker Margaux - Test indices per synthesis method and transformation index*

Method	-3	-2	-1	0	+1	+2	+3
PSY	14	04	07	10	08	12	03
SVLN	01	02	05	15	06	09	11

Margaux [7]. Table 6.18 lists per synthesis method and transformation index the listening test indices for each sound example being available online. The hidden original speech phrase is placed at test index 13.

Listening test results:

Fig. 6.62 depicts the voice quality ratings for the proposed method *PSY* and the baseline method SVLN. The horizontal red (violet) lines reflect the mean (median) voice quality ratings of all participants per test phrase with the same indices as in table 6.17. *PSY* achieves a comparably lower mean deviation δ_μ=0.76 and variance δ_σ^2=0.58 for the voice quality rating compared to SVLN with mean δ_μ=1.14 and variance δ_σ^2=0.94. The mean voice quality judgements for *PSY* follow for index -3 to index +1 within the voice quality range "very tense" to "modal"/"relaxed" very closely the ideal dashed grey line. It indicates that the listeners could very well identify the underlying perceptual sensation of the different synthesized voice qualities. Just the two relaxed voice qualities with index +2 and +3 exhibit a higher deviation from the ideal dashed line in grey colour. The baseline method SVLN could less well arise the perceptual sensation of different voice qualities to the listeners. More tense voice qualities and the direct modal re-synthesis without modification are more rated as modal and respectively relaxed voice quality.

However, despite the clearly better voice quality results for *PSY*, its synthesis quality is in contrast rated relatively low. The three voice qualities with index -3, -1 and 0 are rated simultaneously as "excellent" and "poor", depicted in fig. 6.63. This indicates that different participants received a different perception of the sound quality from the respective synthesized phrases. The same is valid for index +3 of SVLN. This observation is reflected in the low mean MOS_μ=2.71 and high variance MOS_σ^2=1.08 values. SVLN achieves with a higher mean MOS_μ=3.15 and lower variance MOS_σ^2=0.80 ratings better synthesis quality results. Both systems receives again lower rating results for higher voice quality changes.

In general, *PSY* received a lower deviation from the true voice quality rating but a lower MOS synthesis quality compared to SVLN, summarized in table 6.19. Two reasons explained in the following may indicate the cause for the lower sound quality rating of *PSY*.

a) *Unvoiced stochastic component $U(\omega)$*:

[7] Speaker Margaux: http://stefan.huber.rocks/phd/tests/vqMissM/

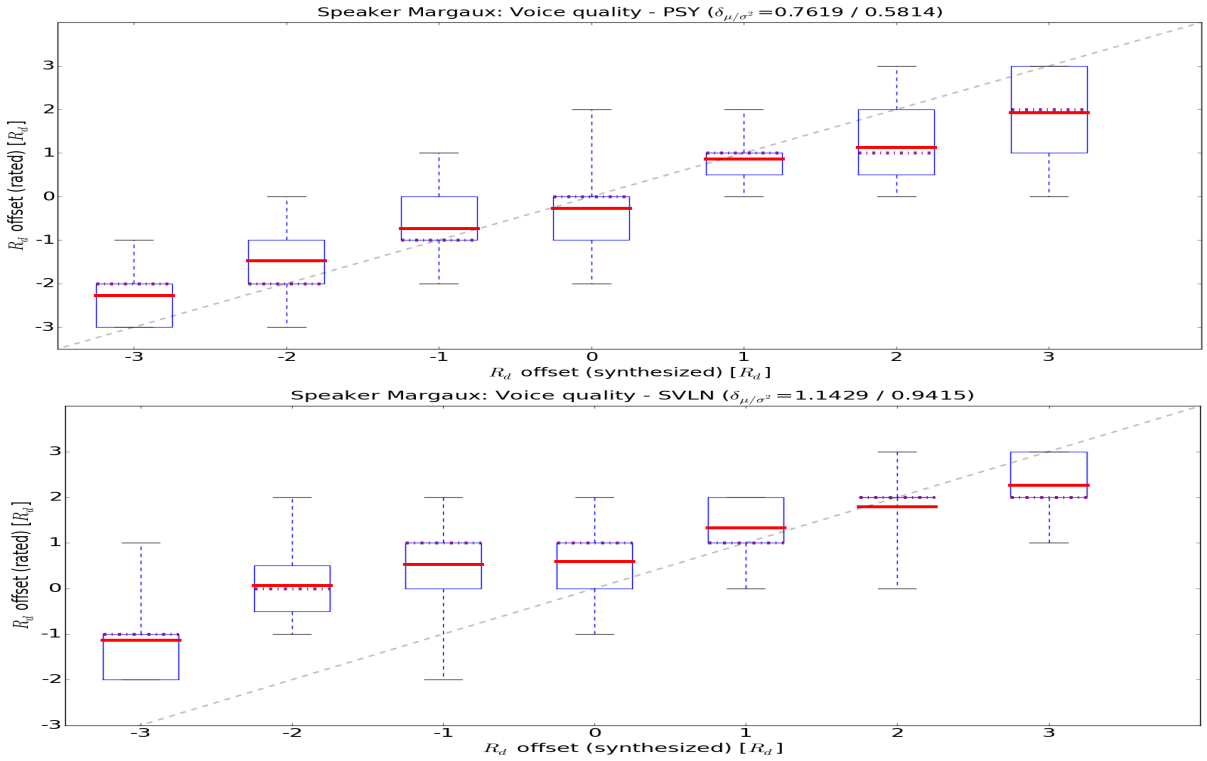


Figure 6.62: *VQ Test 2 Speaker Margaux - Voice quality rating results*

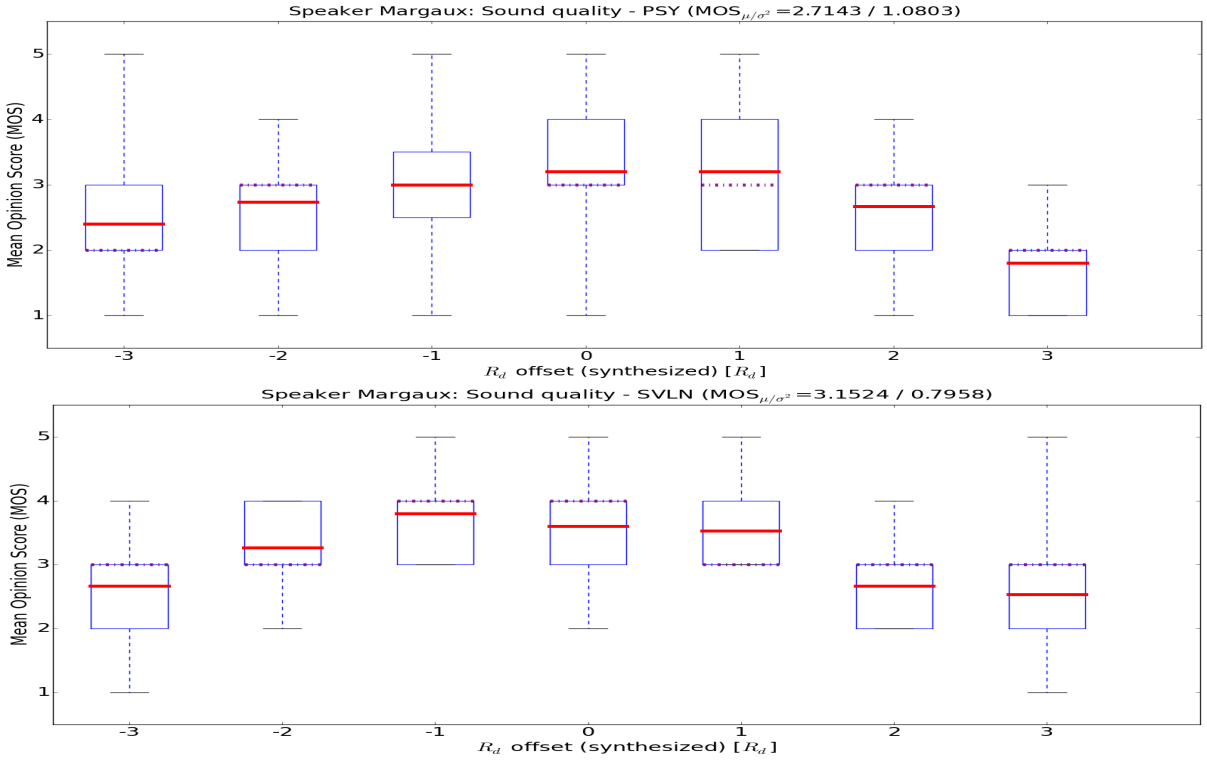


Figure 6.63: *VQ Test 2 Speaker Margaux - MOS synthesis quality rating results*

Table 6.19: *VQ Test 2 Speaker Margaux - VQ voice and MOS sound quality summary*

Method	Δ VQ$_\mu$	Δ VQ$^2_\sigma$	MOS$_\mu$	MOS$^2_\sigma$
PSY	0.7619	0.5814	2.7143	1.0803
SVLN	1.1429	0.9415	3.1524	0.7958

The voice quality transformations towards "relaxed" (index +2) and "very relaxed" (index +3) received the lowest sound quality ratings for *PSY*. The synthesis of less sinusoidal content in higher frequency regions due to the steep spectral slope of the voiced deterministic component $V(\omega)$ reveals more portions of the unvoiced stochastic component $U(\omega)$ for both speech systems. The different signal processing designs to estimate $U(\omega)$ and synthesize $u(n)$ lead consequently to different sound sensations of the unvoiced part for *PSY* and SVLN. *PSY* bases the construction of $U(\omega)$ on the DSM-based subtraction of $V(\omega)$ from $S(\omega)$ and the further processing steps explained in section 6.3. SVLN high pass filters a generated white noise signal and applies amplitude modulation according to equ. 3.10 to artificially create the modulations present in natural human speech. *PSY* in contrary retrieves natural modulations from the original speech recording. The two relaxed voice quality phrases can be reheard online via the link given above. The "relaxed" voice quality of *PSY* is listed at index 12, the "very relaxed" one at index 03. The SVLN example for "relaxed" is available at index 09 and "very relaxed" at index 11.

The examples of *PSY* suffer from a certain discontinuous, ticking or fragmented sounding sequence of different concatenated noise segments. The DSM-based subtraction being applied per frame may interfere over different phonemes. A subtraction of the estimated sinusoidal content with different energy levels over frames may lead to the described perceptual phenomena. A simple smoothing of the concatenated spectral envelopes $\mathcal{T}_{unv}$ describing $U(\omega)$ before the excitation with white noise may suppress the described deterioration from a natural and smooth sounding unvoiced waveform. Actually all and not just the relaxed sound examples of SVLN contain on the other hand a certain buzzy, metallic and reverberation like background sound. Concerning the test results for speaker Margaux, the drawbacks introduced with the unvoiced waveform $u(n)$ in *PSY* are obviously perceived as more annoying by the test participants than the general drawbacks of SVLN.

b) *Psycho-acoustic learning effect*:
The expectation underlying this notion is that the listeners completed this test successively, starting from test example 1 until the end. The test sequence for the first four example is:
Index 1 = SVLN "very tense"
Index 2 = SVLN "tense"
Index 3 = *PSY* "very relaxed"
Index 4 = *PSY* "tense"
etc.
The test sequence was randomized to hide from the participants any possibility to conclude the true test content being hidden behind the randomization. However, the indexation remains unchanged online. The test index 3 for a "very relaxed" voice quality of *PSY* is therefore unfortunately embedded between three "tense" voice qualities. Tense voice qualities are perceived as more pleasant by the listeners since they are rated with a higher MOS synthesis quality. The voiced component masks the possibly less pleasant sounding unvoiced component $U(\omega)$ explained for the preceding step a). The perception and cognition of the human auditory system is subject to the conditioning on perceptual repetitions [Henrich et al., 2003] and the formation of auditory memories [Agus et al., 2010]. This may even amplify in the perception of the listener the mentioned drawbacks introduced with the unvoiced component of *PSY* while the tense voice quality masks the drawback. Another unfortunate placing for *PSY* is given at index 14 for "very tense". It follows directly the hidden reference of the original recording. The latter has naturally a very high sound quality. Posing to the listeners directly afterwards a voice quality transformation containing expected drawbacks may lead to a worse judgement.

6.6.6 Hypothesis on modal voice quality

All three selected phrases are part of a speech corpus reflecting a normal (modal) voice quality. The two listening tests presented in sections 6.6.4 and 6.6.5 are based on the hypothesis that each selected original recording under evaluation constitutes consequently a normal voice quality. This hypothesis was evaluated by asking the test participants to additionally rate the voice quality of the original speech recording.

Fig. 6.64 shows that most participants agreed to rate the original speech recording and as well as its reference hidden within the listening test as normal/modal. However, it would be interesting to ask participants the same question but utilizing from an expressive speech corpus one phrase being definitely "very tense" and another one being definitely "very relaxed". The ratings shown in the three figures in 6.64 could be biased by the expectation of the listeners and not reflect the listening perception itself. Such possible bias is slightly indicated by the higher variance of the voice quality rating of the hidden original reference for each speaker. SVLN exhibits a slightly higher voice quality deviation from index 0 than *PSY*. The hidden original reference was well identified throughout the three listening tests. It received the highest MOS ratings of roughly MOS_{μ}=4.5 for each test.

Several participants asked for speech examples of different voice quality for the given speaker under evaluation in order to tune their perception for the listening test. However, problems arise consequently with a standard speech

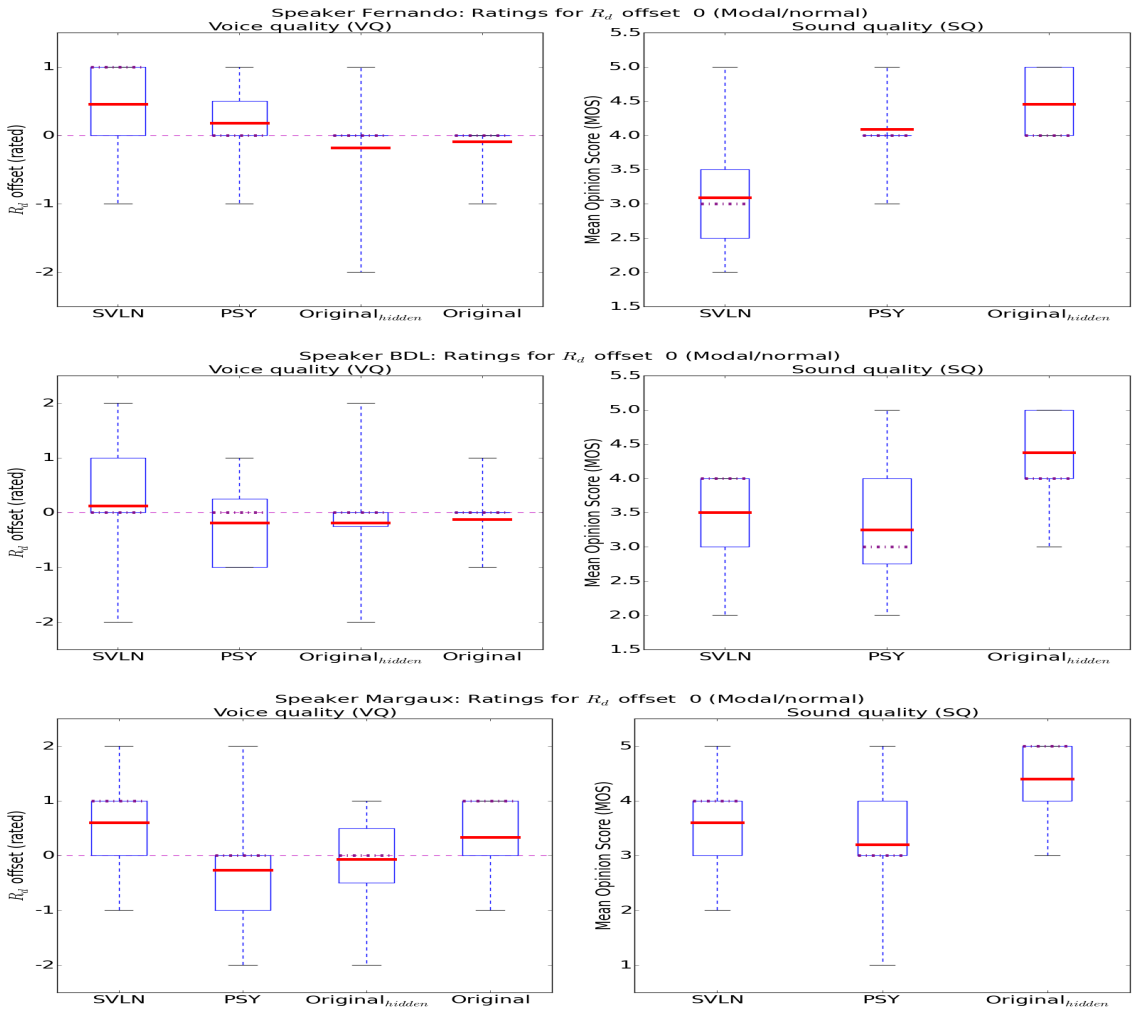


Figure 6.64: *VQ test 2 - Voice and synthesis quality rating for re-synthesis and original phrase*

corpus of exclusively modal voice quality for a listening test on voice quality transformation. No original speech recording can be given to demonstrate the desired effect of a more tense or a more relaxed voice quality. It would require an additional expressive speech corpus of the same speaker. The participants should therefore listen to all examples before making any ranking if an expressive speech corpus is not available. However, this introduction was not given to the participants for each listening test. The test results may therefore be biased since it is difficult to make a consistent decision on the phonation type of each presented voice quality.

6.7 Conclusions

PSY:
The presented speech analysis, transformation and synthesis framework *PSY* demonstrates for both voice quality transformation tests, presented in the sections 6.6.5 and 6.6.4, its possibility to arise the perceptual sensation of different voice qualities to the listener. It outperforms the baseline method SVLN especially in rating the different synthesized voice qualities. The reason may be the voice descriptor smoothing required by SVLN which does suppress certain fine speech details contained in the original signal. This in turn leads to a higher MOS rating of the synthesis quality of *PSY* for both speakers of VQ test 1 and for speaker Fernando of VQ test 2. The lower MOS-rated synthesis quality of *PSY* compared to SVLN for speaker BDL is explained by the creaky voice quality and the presence of signal spikes, both being discussed in section 6.6.5.2. The conclusion drawn for VQ test 2 for the female speaker Margaux suggests that supplementary work is required to further increase the synthesis quality of *PSY* by improving the synthesis of the stochastic signal part $u(n)$.

Spectral fading:
The findings presented with the subjective listening test of section 6.6.5 suggest that the proposed *PSY* synthesis variant "spectral fading" along with the R_d^{gci} transformation to generate different $R_d^{'gci}$ contours of section 6.4.3.2 is able to re-synthesize different versions of the analysed input speech phrase such that the perceptual sensation of different voice quality characteristics arose in the perception of a listener. Its MOS-assessed synthesis quality received partially very good judgements for minor changes in voice quality. Major voice quality changes are appraised of moderate quality for both the baseline method SVLN and the proposed method *PSY* spectral fading.

GMM-based RMS energy model:
The implemented model to predict RMS-based energies within the context of voice quality modifications is only applicable for transformations towards a more relaxed voice quality if no normalization is applied before synthesis to avoid clipping artefacts. This normalization has been omitted for the conducted listening test in section 6.6.4. The same normalization factor would have to be applied before synthesis to the speech phrase being transformed to a relaxed voice quality. This would have resulted in too low signal amplitudes being not anymore audible. The reason is that the GMM energy prediction is for the time being not generating sufficiently robust RMS values. However, a simple normalization to a certain level for all phrases before synthesis would maintain the same SNR ratio. The GMM energy model is well applicable if no listening test is required and the application permits the utilization of a heuristic normalization level. Each re-synthesized speech phrase would have the same amplitude level. The GMM energy contour prediction for the voiced and unvoiced signal components is in this case well reflected. Still, the presented GMM prediction has to be evaluated and possibly optimized per chosen speaker.

Pitch-adaptive STFT:
The energy model in *PSY* is based on a simple, robust and efficient RMS energy measure of the linear amplitude spectrum. Its current straight-forward implementation puts the constraint of a constant STFT window size in the analysis and synthesis. The possibility of a pitch-adaptive windowing in *PSY* should improve its quality.

Chapter 7

Contribution - *coVoC*: Concatenative Voice Conversion

he mind that opens up to a new idea never returns to its original size.

ALBERT EINSTEIN

7.1 Introduction

Please note that the novel *coVoC* system for VC which will be introduced in section 7.2 is not a contribution of the author to this thesis. It has been developed at IRCAM and is currently patent pending [Roebel, 2015]. The reasons that *coVoC* is not placed in the corresponding START chapter 4 for Voice Conversion are that

- the novel speech framework *PSY* proposed in chapter 6 has been established to further advance the means of *coVoC*,
- the discussions presented in the following are better structured such that the reader can more easily follow the ideas of advancing *coVoC* with *PSY*, instead of having to switch between chapters,
- the extension of *coVoC* by the additional transformation of voice descriptors by *PSY* combines both systems,

such that the representation within one chapter facilitates the given explanations.

Section 7.3 discusses the extensions implemented in the novel speech framework *PSY* to extend the VC means of *coVoC*. The transformation of an extended set of voice descriptors is conducted by *PSY* in addition to the spectral envelope conversion of *coVoC*. The aim is to augment the VC performance by addressing further features which proved in the literature to describe the characteristics of a speakers voice identity and which proved to contribute to the conversion of a speakers voice identity. However, the initial experiments conducted for the evaluation of the tests presented in section 7.4 could for the time being not validate the full potential of the proposed extension of *coVoC* by *PSY*. The subjective evaluation of section 7.4.2 exhibits only minor while the objective evaluation of section 7.4.3 exhibits at least reasonable advancements of the VC performance compared to the GMM baseline method presented in section 7.4.1.

7.2 System description - *coVoC*

7.2.1 Motivation

Statistical models destroy the natural signal contour of the modelled feature vectors, notable the spectral envelope encoding. The different trials to minimize the over-smoothing effect as discussed in section 4.6.3 cannot prevent

to loose the natural spectral envelope contour in the VC context.
Frequency Warping methods maintain to a huge extent the natural spectral envelope contour, but are not able to reasonably augment the conversion score towards the target speaker.
Vector Quantization, Codebook based and Frame Selection approaches employ to a huge extent the direct signal features from the target database. These means could maintain the natural signal contour. However, Vector Quantization and Codebook based methods suffer from the many concatenation points present. Each concatenation is prone to exhibit a signal difference being strong enough to provoke artefacts. The same reason applies to a lower extent to Frame Selection approaches. The collection and concatenation of frames from different speech segments having different phonetic and prosodic content is prone to brake the natural signal contour over frames [Uriz et al., 2011]. This results into lower synthesis qualities which affects additionally the target speakers voice identity in the converted phrase.

The ***coVoC* (concatenative Voice Conversion)** system is based on similar means to concatenate signal content being directly derived from the target database [Roebel, 2015]. The main difference being that segments comprising a complete phoneme are matched between the source phrase and the target database. Each selected target phoneme is exchanged with the source phonemes and concatenated. The most appropriate target phoneme sequence is selected according to the source phoneme sequence. The assembled phoneme sequence of the target speakers database is thus converted to the target speaker. Please note that not the phoneme itself but its corresponding spectral envelope sequence T_{sig} is concatenated and exchanged. The envelope sequence T_{sig} of each selected target phoneme is warped to the time length of the source phoneme. This constructs over phonemes the envelope sequence T''_{sig} that is converted to the target speaker. The original envelope sequence T_{sig} of the source phrase is replaced by the concatenated target envelopes T''_{sig}.

coVoC reduces the one-to-many feature correlation problem being present between speakers [Godoy et al., 2009]. The coverage of a whole phoneme unit removes redundancy compared to the one-to-many relations present between single evaluated frames, as discussed in section 4.3. The *coVoC* approach is still interfered by the one-to-many problem since certainly many different phonemes of the same phoneme type or articulation group are available from the target corpus for each phoneme of the source speakers phrase.

7.2.2 Concatenative Unit Selection for VC

The novel VC system denominated *coVoC* is based on the matching and concatenation of phoneme units. The **phoneme sequence matching** of *coVoC* consists conceptually of the three following main algorithmic steps:

1. Estimate a target phoneme sequence
2. Retrieve and concatenate the corresponding spectral envelopes
3. Apply it as spectral envelope filter on the source phrase signal

The *coVoC* algorithm for concatenative VC shares a high similarity with the exemplar-based VC method described in section 4.8.4. The main difference is the type of utilized units. The exemplar-based VC approach selects units which span over multiple frames. It does consequently not require to conduct an automatic phoneme detection with the additional requirement of textual information, as explained in section 4.2.1.

The algorithm *coVoC* explained in the following relies in contrast on the automatic phoneme border detection given by the algorithm *ircamAlign* described in section 4.2.1. This requires the spoken text in parallel to each speech phrase such that the linguistic structure can be encoded into the X-SAMPA alphabet. On the one hand, it is computationally more expensive and risks that the time instants of the detected phoneme borders are not perfect. On the other hand, the required optimization of the best performing length of considered consecutive frames as one exemplar unit as in [Wu et al., 2013] is per se avoided. The unit length employed by *coVoC* varies automatically according to the phoneme length. Assuming that the phoneme borders are estimated without errors, the locally adaptive unit length is always perfect. The exemplar-based VC method of [Wu et al., 2013] uses MFCC based filters from the signal for the content matching between the source and target speakers data. The *coVoC* method employs a spectral envelope estimator to construct the spectral features for the MFCC calculation.

The *coVoC* VC system is based on an unit selection algorithm of estimated phonemes. It works according to the following pseudo-algorithm 8:

The *coVoC* system employs a basic database-based unit selection paradigm. The application of concatenative synthesis techniques should augment the converted target stream to a high fidelity sounding output. Improvements of the VC quality in this supervised environment by the optimized selection of whole phoneme units instead of single spectral frames are expected. The complete phonetic content of the target corpus is clustered into articulation

Algorithm 8 - Concatenative VC (*coVoC*) based on phonetic content matching

for all phonemes in (source phrase, target corpus) **do**
Annotation 1: Automatically label and encode in X-SAMPA
Annotation 2: Detect and set time instants of phoneme begin and end
Organisation: Group similar phonemes into articulation groups
Stamp: Compute MFCC vectors on spectral envelopes and energies over frames
end for
for all phonemes in source phrase **do**
for all target phonemes of same articulation group **do**
Align phonemes: Apply Dynamic Time Warping using MFCCs Manhattan distance to walk through trellis
Matching error: Calculate matching error between aligned MFCC vectors
end for
Select: Sort by least matching error, keep N target phoneme candidates to generate Viterbi lattice
end for
for all selected N target phonemes **do**
Target cost: Set matching error as target cost factor
Concatenation cost: Compute mean difference between subsequent target phonemes
end for
Viterbi decoding: Estimate optimal phoneme sequence as best path with lowest cost through phoneme lattice
Sequence generation: Select spectral envelope sequence per target phoneme and concatenate as $T''_{sig}(\omega)$
Synthesis: Apply spectral envelope sequence $T''_{sig}(\omega)$ on source speakers speech phrase to be converted

groups. One articulation group may contain one or several phonemes as encoded by the X-SAMPA standard. Per phoneme of the source phrase, a list of N target phonemes of the corresponding articulation group is selected. The N target phonemes having the lowest matching error to the corresponding source phoneme are selected. A Dynamic Programming (DP) approach by means of Viterbi decoding optimally selects the best target phoneme sequence corresponding to the phoneme sequence of the source phrase. The overall lowest error over the complete phoneme sequence results from the assembly (*concatenation cost*) of phoneme pairs having the lowest error per phoneme switch (*target cost*). The optimized phoneme sequence resulting in the converted spectral envelope sequence $T''_{sig}(\omega)$ is applied to the speech phrase of the source speaker being selected for conversion. This should contribute to the naturalness and intelligibility of the final synthesis. The synthesis quality and the conversion score should reasonably augment compared to conventional VC methods. The evaluation on natural human speech presented in section 7.4 will examine these expectations.

7.2.3 System comparison

The major advantage of the *coVoC* system lies in its ability to handle smaller and non-parallel speech corpora within the VC context.

Non-parallel:
The *coVoC* system facilitates a non-parallel database approach. It does not require to learn the correlation between corresponding source and target speaker features as with conventional statistical VC approaches. The latter requires to train a model such that a conversion function can be derived from the learned data mapping. Instead, *coVoC* retrieves the relevant and best matching features directly from the target corpus.

Target-only:
Many text-independent or cross-lingual VC approaches, introduced in the sections 4.9.1 and 4.9.2, still require a corpus of the source speaker. The sentences do not necessarily have to be the same as in the parallel case. But these VC systems require a quasi-parallel alignment into artificial phonetic categories such that the correlation between source and target features can be reflected in the conversion function. The *coVoC* systems alleviates this aspect. It only requires the speech phrase of the source speaker intended to be converted, along with the corpus of the target speaker.

Phonetic encoding:
The novel VC system *coVoC* requires a phoneme annotation of the whole target speaker corpus and the single speech phrase of the source speaker used for conversion. The HMM-based annotation system *ircamAlign*, proposed in [Lanchantin, 2007, Lanchantin et al., 2008] and explained in section 4.2.1, is utilized to annotate phonemes and to set their begin and end time instants. It requires additionally the textual information of the spoken phonetic content per sentence. The estimation of the phoneme borders risks that the utilized annotation system does not cor-

rectly estimate the phoneme borders. The *coVoC* system minimizes that risk by introducing parameter β_{ext} which extents each estimated phoneme over its borders into the neighbouring phonemes. This results into an overlapping of phonemes such that *coVoC* is not designed as a hard clustering system. The parameterized phoneme overlapping allows with β_{ext} adjusting for a certain level of soft clustering. Moreover, the annotation into neighbouring phonemes optimally adapts the underlying unit length such that no fixed entity length leads to a hard clustering design.

7.3 System description - *coVoC* combined with *PSY*

7.3.1 Goals

The objective of the Concatenative VC system *coVoC* is to establish a Voice Conversion paradigm which is suited to provide a converted feature sequence describing the target speakers voice identity with high conversion score and synthesis quality. It should establish in combination with the speech system *PSY*, presented in the preceding chapter 6, a flexible VC framework with high quality. The combination of both systems should facilitate advanced voice modification possibilities known from HMM speech synthesis [Drugman et al., 2009b, King, 2010]. It should provide a higher synthesis quality compared to HMM synthesis [Lanchantin et al., 2010]. The aim is to achieve a high synthesis quality being close to a TTS Unit Selection speech system [Hunt and Black, 1996]. Additionally, the intended system should offer the usage of smaller database compared to TTS speech synthesis. Moreover, it should have the flexibility of *coVoC* like non-parallel and target-only explained in the preceding section 7.2.3. It should also constitute a very good capturing of the target speakers voice identity.

7.3.2 Intention

The *coVoC* system exhibits a very promising approach to conduct VC. However, the *coVoC* standalone system as described in the preceding section 7.2 applies only the converted spectral envelope sequence on the speech phrase of the source speaker. It neglects the additional conversion of other voice descriptors contributing to a speakers voice identity as discussed in section 4.7. Other VC system consider the conversion of remaining voice descriptors, such as the residual signal of section 4.7.1, the glottal excitation source of section 4.7.2, and several prosodic features discussed in section 4.7.3. Further means are consequently required to augment the VC performance especially in terms of the conversion score towards the target speakers voice identity. This is the reason why the novel speech framework *PSY* introduced in chapter 6 has been establish for this work. *PSY* allows conducting advanced Voice Conversion and Transformation tasks of selected voice descriptors.

7.3.3 Conversion and transformation of extended voice characteristics

Please note that the discussions presented in the following utilize two operators to distinct between the conversion or transformation of features from the source to the target speaker:
Operator $''$ = *coVoC*-based conversion
Operator $'$ = GMM-based transformation

The operator $''$ denotes the conversion of the spectral envelopes for the unvoiced and VTF spectra for the voiced part over the source speakers phoneme sequence using the *coVoC* algorithm. The converted spectral envelope sequences are utilized to generate $V''(\omega)$ and $U''(\omega)$.

The operator $'$ denotes the utilization of the corresponding GMM models of section 6.4.2 for the GMM-based prediction of the voice descriptors F'_{VU}, R'_d, E'_{voi} and E'_{unv}. Each model is trained on the target speakers database to predict the relevant voice descriptor contours from the converted $V''(\omega)$ and $U''(\omega)$ signals before synthesis. Respectively, F'_{VU} is predicted using the GMMs explained in section 6.4.2.4, R'_d is derived from the GMM of section 6.4.2.5, and the RMS-based energy modification reflected in E'_{voi} and E'_{unv} is conducted by means of the GMM modelling explained in section 6.4.2.3. Any signal component without the operators $'$ or $''$ denotes consequently the corresponding voice descriptor of the source speaker.

PSY re-utilizes the information of the phoneme matching and selection estimated by *coVoC* to generate $T''_{sig}(\omega)$. Instead of using $T''_{sig}(\omega)$ for synthesis it employs the corresponding VTF sequence $T''_{voi}(\omega)$ and the spectral envelope sequence $T''_{unv}(\omega)$. Both have been established in parallel at the analysis stage of the target speakers database beforehand. The $T''_{voi}(\omega)$ and $T''_{unv}(\omega)$ sequences are extracted with the same phoneme annotation and DTW alignment mechanism as *coVoC*. Additionally, *PSY* performs the transformation of several voice descriptors using the

GMM-based contour prediction of section 6.4.2 within the Voice Conversion context. The following set of voice descriptors to describe a target speakers voice identity are addressed:

- $\boldsymbol{V''(\omega)}$:
  coVoC-based conversion of the VTF sequence $T''_{voi}(\omega)$ used to synthesize the voiced component $V''(\omega)$
- $\boldsymbol{U''(\omega)}$:
  coVoC-based conversion of the unvoiced spectral envelope sequence $T''_{unv}(\omega)$ used to synthesize the unvoiced component $U''(\omega)$
- $\boldsymbol{F'_{VU}}$:
  Transformation of the source speakers Voiced / Unvoiced Frequency boundary F_{VU} contour to the target speakers F'_{VU} contour using the GMM F_{VU} model of section 6.4.2.4
- $\boldsymbol{R'_d}$:
  Transformation of the source speakers LF shape parameter R_d to the target speakers R'_d contour using the GMM R_d model of section 6.4.2.5
- $\boldsymbol{E'_{voi}}$ and $\boldsymbol{E'_{unv}}$:
  Transformation of the source speakers RMS-based voiced E_{voi} and unvoiced E_{unv} energies to the target speakers E'_{voi} and unvoiced E'_{unv} energy contours using the GMM energy models of section 6.4.2.3

However, the transformation of the fundamental frequency F_0 or the prosodic characteristics as indicated in section 4.7.3 have not yet been addressed within this thesis.

7.3.4 Risks

The *coVoC* standalone application of section 7.2 exchanges the spectral envelope of the source speaker by the envelope sequence $T''_{sig}(\omega)$ that is converted to the target speaker. The utilized speech signal of the source speaker constitutes a coherent signal waveform on which $T''_{sig}(\omega)$ is applied on. In contrast, the artificial feature combinations of $V''(\omega)$, $U''(\omega)$, F'_{VU}, R'_d, E'_{voi}, and E'_{unv} addressed by *PSY* introduces further risks to the synthesis system. On the one hand, the implied freedom is desired to further advance the conversion towards the target speaker. On the other hand, the artificial feature combination may result in a non-coherent description of a speech signal. The combined voice descriptors used for synthesis may
a) not reflect real speech data,
b) not sound like the voice identity of the target speaker,
c) lead to artefacts.
Further constraints are therefore necessary to assure a coherent description of the converted speech signal such that possible artefacts are minimized and the target speakers voice identity is correctly captured. The following sections explain the means implemented in *PSY* to coherently handle a converted and transformed set of voice descriptors. An examination of such coherent signal reconstruction when joining different voice descriptors is conducted with the evaluation of the different *PSY* variants discussed in section 7.4.

7.3.5 VTF and spectral envelope conversion

STAte-of-the-ART research as discussed in section 4.7.1 has shown that the residual component carries relevant speaker-dependent information. It is therefore important to consider the modelling and transformation of the residual in a VC system. In contrast to most START works on VC as introduced in section 4.7.1, the residual is modelled differently in this work. It is split into two components. The voiced part of the residual contains the deterministic part of the glottal excitation source and the fundamental frequency F_0. It constitutes the excitation of the voiced component $V(\omega)$ discussed in section 6.2. Here, the unvoiced part of the residual reflects the unvoiced component $U(\omega)$ introduced in section 6.3.

The central idea of the *coVoC* approach is to completely avoid any averaging of different spectral envelopes, as it is the case for common statistical models. This ensures that the spectral envelopes to be used for the conversion are all valid envelopes of the target speaker. The converted VTF sequence $T''_{voi}(\omega)$ and the spectral envelope sequence $T''_{unv}(\omega)$ require contrary to the *coVoC* standalone version further means at the connection points of the concatenated phonemes in *PSY*. An additional smoothing has to be applied since the partially huge signal differences present at the concatenation borders proved to provoke artefacts at the synthesis stage of *PSY*. Please note that the *coVoC*

standalone version applies only the converted spectral envelope sequence $T''_{sig}(\omega)$ on the speech signal of the source speaker.

In contrast, each VTF sequence $T_{voi}(\omega)$ estimated on each speech phrase of the target speakers corpus is influenced by the corresponding R_d glottal source estimation used to extract the VTF sequence per target phrase. A converted VTF sequence $T''_{voi}(\omega)$ underlies consequently higher spectral variations than its corresponding converted spectral envelope sequence $T''_{sig}(\omega)$. This is especially the case at phoneme concatenation borders on which the R_d estimator may have more difficulties to estimate the true glottal source shape contained in the signal. In contrast, the R_d estimation performs more robust at the stable part of phonemes in their middle where usually more stable harmonic sinusoids are available. The spectral slope of the VTF is thus influenced by the R_d estimation which determines the shape and thus the spectral slope of the glottal pulse used to extract the VTF sequence.

Additionally, *PSY* has to employ an R_d contour to synthesize the voiced component $V''(\omega)$. The current implementation of *PSY* offers two possibilities within the VC context: The direct usage of the source speakers R_d contour, or the R'_d contour being transformed to the target speaker. This increases further the likelihood that higher changes of the spectral slope in short-time segments are present in the converted voiced component $V''(\omega)$. Most importantly, such possible higher fluctuations may lead to artefacts of each sinusoidal contour over the short-time segments around phoneme concatenation borders. Especially if two VTF sequences are joined from two phonemes having comparably different spectral slopes it is cumbersome to guarantee an artefact-free sinusoidal continuation.

The same is valid for the converted spectral envelope sequence $T''_{unv}(\omega)$ to synthesize the unvoiced component $U''(\omega)$. The sinusoidal detection and deletion may work less robust where less stable harmonic sinusoids are present, occurring predominantly at phoneme borders. The estimated spectral envelope $T_{unv}(\omega)$ of the unvoiced component of each analyzed speech phrase may not describe the true spectral contour of the stochastic signal part around phoneme borders. Joining then again at the phoneme concatenation borders spectral envelopes from different phonemes may lead to higher spectral changes such that artefacts are more likely to be introduced in the converted unvoiced component $U''(\omega)$ at synthesis.

The interpolation at phoneme concatenation borders in *PSY* is therefore contrary to *coVoC* necessary to minimize the possibility of too high spectral changes causing artefacts at the phoneme concatenation borders. The phoneme interpolation algorithm works as follows:
First, both sequences $T''_{voi}(\omega)$ and $T''_{unv}(\omega)$ are scaled to log zero mean such that no amplitude differences deteriorate the following algorithms.
Second, an interpolation of the VTF $T''_{voi}(\omega)$ and spectral envelope $T''_{unv}(\omega)$ sequences is applied at each concatenation point. Each interpolation segment covers a heuristically determined speech segment of 21 *ms* length around the phoneme switch. The interpolation is conducted for each spectral bin of the utilized DFT. This interpolation at concatenated phonemes to minimize amplitude differences is comparably computational expensive.
Third, an additional median smoothing filter of order 3 is applied on the interpolation results of each spectral bin. The median filter minimizes further possible amplitude differences around the interpolation borders.

Fig. 7.1 illustrates the explained interpolation and median smoothing algorithm applied to the converted VTF sequence $T''_{voi}(\omega)$. Fig. 7.2 illustrates the same means applied to the converted spectral envelope sequence $T''_{unv}(\omega)$. Both examples have been used in the evaluation on the French male speaker pair of section 7.4. The vertical lines at each second plot shown in the middle exemplify the concatenation borders.

7.3.6 Advanced energy handling

The conventional energy maintenance of *PSY* introduced in section 6.4.1.2 has been successfully applied for the voice quality transformation tests presented in section 6.6. However, the simple energy re-scaling to the RMS-based energy contour measured on the source speakers speech phrase within a VC application suffers from partially high energy bursts of the converted unvoiced component $U''(\omega)$. This introduces audible artefacts. Additional means are implemented in *PSY* to adjust the utilized RMS energy contour before the energy re-scaling described in equ. 6.19 is applied.

Two stages are used to modify the RMS energy contour E_{unv} estimated on the source speakers phrase. The first stage saturates the energy E_{unv} of the unvoiced component $U''(\omega)$ to heuristically determined thresholds. The second stage applies a median filter to smooth the energy contour accordingly.

Stage 1 - Saturation:
The threshold parameter θ_{voi} defines the cutoff level applied to E_{unv} to saturate the energy in mixed voiced and unvoiced segments. This suppresses artefacts introduced by noise bursts which are caused by an erroneous sinusoidal detection used to create the unvoiced components $U(\omega)$ of the analyzed corpus of the target speaker. An example of a speech segment with higher noise level can be inspected around ~1.0 seconds in fig. 7.2. However,

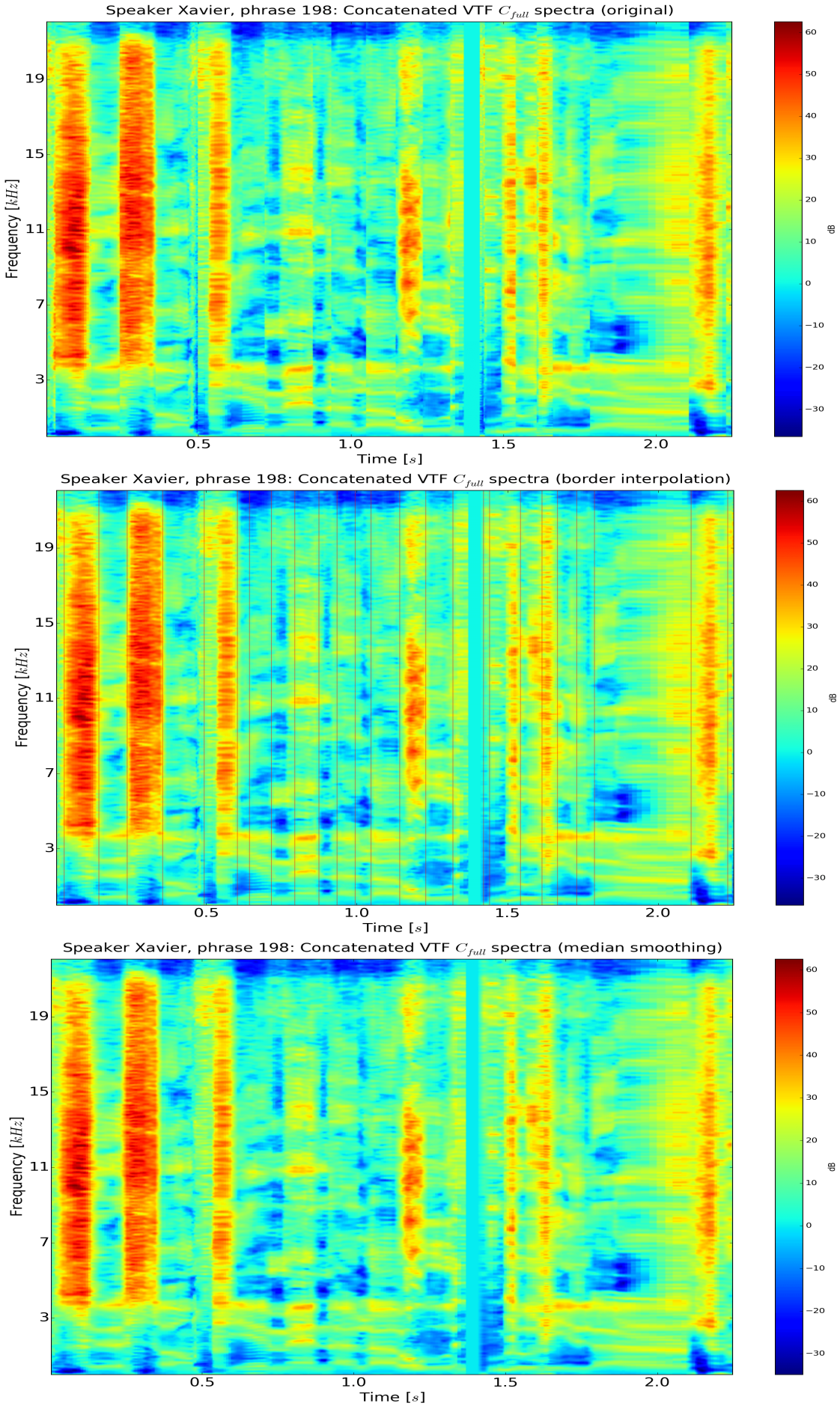


Figure 7.1: *Interpolation and median smoothing at concatenated $T''_{voi}(\omega)$ borders*

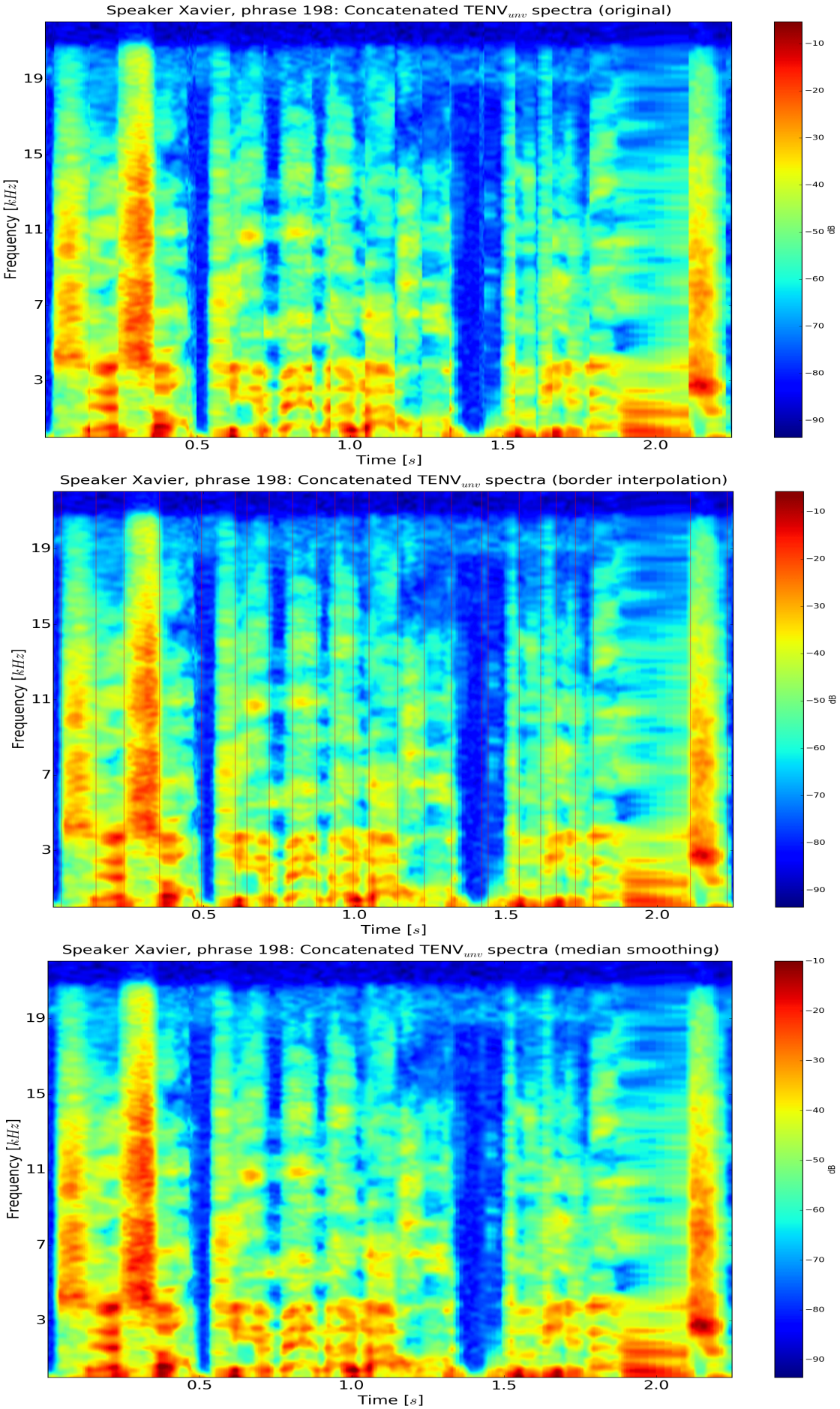


Figure 7.2: *Interpolation and median smoothing at concatenated $T''_{unv}(\omega)$ borders*

the application of the energy saturation isn't shown since it is applied after the interpolation and smoothing at phoneme concatenation borders.

Another heuristically determined energy cutoff threshold θ_{unv} defines the saturation level applied to E_{unv} in purely unvoiced segments. The additional threshold θ_{unv} is required to balance the unvoiced energy level E_{unv} between voiced and unvoiced segments.

Please not that the heuristic determination of the energy cutoff thresholds θ_{voi} and θ_{unv} can be robustly established after some synthesis trials with different levels. Once established it works well for one speaker pair.

However, the sole application of the energy threshold to the unvoiced component $U''(\omega)$ would lead to a deterioration of the overall signal energy E_{sig}. The energy difference between the original energy measure E_{unv} and either θ_{voi} or θ_{unv} is therefore attributed to the RMS energy of the voiced component E_{voi}. This maintains the original signal energy E_{sig}. The pseudo-algorithm 9 describes the required energy management implemented in *PSY* for VC:

Algorithm 9 - Advanced energy handling in *PSY* for VC

for all Synthesis frames k in voiced segments **do**
  if $E_{unv}(k) > \theta_{voi}$ **then**
    $E_{voi}(k) = E_{voi}(k) + E_{unv}(k) - \theta_{voi}$
    $E_{unv}(k) = \theta_{voi}$
  end if
end for
for all Synthesis frames k in unvoiced segments **do**
  if $E_{unv}(k) > \theta_{unv}$ **then**
    $E_{voi}(k) = E_{voi}(k) + E_{unv}(k) - \theta_{unv}$
    $E_{unv}(k) = \theta_{unv}$
  end if
end for

Please note that the voiced energy contour E_{voi} is only partially applicable in unvoiced segments. In theory, it should be zero. Practically, the window of the STFT used to synthesize $V''(\omega)$ generates voiced signal content at the unvoiced segments around voiced borders.

Stage 2 - Smoothing:
The requirement to apply empirically defined energy threshold levels implies not just manual user interaction. The hard saturation implied by the thresholds destroys to some extent the smooth energy contour of E_{unv} and E_{voi}. An additional median filter of order 10 is applied after the energy saturation to re-establish a smooth transition especially around voiced segment borders.

Please note that the advanced handling of the unvoiced energy E_{unv} becomes obsolete if the GMM-based energy prediction to transform E_{unv} towards the target speakers energy E'_{unv} is activated. In general it can be noted that an improvement of the algorithms of section 6.3.1 (Stochastic residual estimation) and 6.3.2 (Posterior filtering) used to estimate the unvoiced signal part shall minimize the mentioned energy problems with the converted unvoiced component $U''(\omega)$ at synthesis.

7.3.7 Advanced spectral fading synthesis

For the synthesis of $V''(\omega)$, the combination of $T''_{voi}(\omega)$ with a glottal source signal being parameterized by either the source speakers R_d contour or the R'_d contour that is transformed to the target speaker may lead to artefacts. The reason being that the original R_d contour of the target speaker used to extract $T_{voi}(\omega)$ leaves the latter with a certain spectral slope. The multiplication of $T''_{voi}(\omega)$ with the spectral representation of the glottal pulses at synthesis may result in a too steep and thus unnatural spectral slope contour. Moreover, the advanced energy handling for VC described in the preceding section 7.3.6 is not able to handle properly the constructed voiced component $V''(\omega)$.

A similar signal behaviour which lead to the implementation of the spectral fading synthesis variant of *PSY* described in section 6.5.5 is observed. The difference being that in the VC case artefacts are more prone to appear on voiced borders when transitioning from or to purely unvoiced segments. The solution to establish a coherent signal construction of the converted voiced component $V''(\omega)$ is to advance the means of the spectral fading synthesis. A simple linear increase of the linear ramp defining the slope m_{LP} of the low pass filter to fade out the voiced

component is not applicable. It would be effective all the time and nearly cut off the voiced signal content at the F_{VU}.

Consequently, the steepness m_{LP} of the low pass filter is adjusted dynamically. A cutoff frequency F_{boost} is set to 1 kHz. The modification of the m_{LP} slope becomes active if the frequency F_{VU} of the source speaker or the frequency $F_{VU}^{'}$ transformed to the target speaker is below F_{boost}. The parameter θ_{boost} permits the adjustment of the m_{LP} slope modification.

$$m_{LP}^{'} = \begin{cases} m_{LP} - (F_{boost} - F_{VU}) \,/\, \theta_{boost} & \forall \quad F_{VU}, F_{VU}^{'} < F_{boost} \\ m_{LP} & \forall \quad F_{VU}, F_{VU}^{'} \geq F_{boost} \end{cases} \tag{7.1}$$

Equ. 7.1 defines the slope modification for $m_{LP}^{'}$ of the low pass filter P_L for the synthesis of $V^{''}(\omega)$. The slope boosting leads to higher filter slopes and thus a higher attenuation of the voiced component. It becomes active if F_{VU} is below F_{boost}. If active, the modified filter slope $m_{LP}^{'}$ is more steep the lower the F_{VU} or $F_{VU}^{'}$ frequency is below F_{boost} and the lower it reaches 0 Hz. The voiced component diminishes with this more the lower the F_{VU} or $F_{VU}^{'}$ frequency boundary.

7.4 Evaluation on a French male speaker pair

7.4.1 GMM baseline method

A detailed description of the GMM-based baseline system for VC used for comparison with the *coVoC* system alone, and in combination with different *PSY* variants can be found in [Lanchantin and Rodet, 2010, Lanchantin and Rodet, 2011]. Please note that not the Dynamic Model Selection algorithm but the basic VC approach is utilized here. It uses the joint density model explained in section 4.6.1 and proposed in [Kain, 2001]. Full co-variance matrices are employed. The chosen LPC/LSF order to model the spectral envelope is set to 50 per spectral frame. 180 phrases or 90 % of the parallel corpus introduced in the following section 7.4.2 are employed to train a GMM being comprised of 16 Gaussian components.

The co-variance correction introduced in section 4.6.3.4 has been applied to produce the GMM-based baseline method example for this VC test. The paper of [Lanchantin et al., 2011b] includes two other VC approaches which were not utilized here. Please note that the Extended Conditional GMM separating the source model for recognition and the source-target model for conversion have not been applied as well. Also the local trajectory modelling by means of the Discrete Cosine Transform (DCT) based stylization has been avoided. The co-variance correction parameter λ_G is set to 0.9 to re-produce much but not the complete variance of the target speaker.

7.4.2 Subjective evaluation - Listening test

A subjective evaluation by means of a listening test evaluates the *coVoC* system alone and in combination with different *PSY* variants versus the GMM baseline method introduced in the preceding chapter 7.4.1. A male-to-male Voice Conversion is established in French language. Another phrase from the Hispanic male speaker "Fernando" (speaker F) already utilized in section 6.6 is chosen as target speaker. The source speaker "Xavier" (speaker X) is taken from the same parallel corpus established at IRCAM in [Lanchantin et al., 2008] for research on VC. This speaker pair was chosen as test setup since it had been utilized extensively at the laboratory to evaluate the VC performance on the former works establishing the GMM baseline method. The publications of [Lanchantin and Rodet, 2010] uses speaker Xavier as source speaker S and speaker Fernando as target speaker A. The follow-up study of [Lanchantin and Rodet, 2011] employs as well the same speaker pair. The corresponding listening test can be found via this link: "VC baseline test".

The listening test established for this work has been published on the following relevant mailing lists:
- Internally at IRCAM.
- To the Auditory list [1], McGill University, Montreal, Quebec, Canada.
- To the list "Parole" [2], established at the University of Avignon, France, for research communications within the francophone speech community.

41 participants have conducted the listening test. The test can be followed online via this link "VC from X to F" [3]. Two original recordings per source speaker X and target speaker F are presented on top of the webpage to the

[1] Auditory list: www.auditory.org
[2] Parole list: www.afcp-parole.org/
[3] Voice Conversion from Xavier to Fernando: http://stefan.huber.rocks/phd/tests/VoCoX2F/

listener to facilitate the conditioning of the human perception on both voice identities. The participants were asked to rate the synthesis quality of each presented phrase according to the Mean Opinion Score (MOS) scale of table 6.3. The conversion score to evaluate the VC performance in terms of Speaker Identity (SI) is based on another MOS scale according to the following table 7.1:

Table 7.1: *Voice identity rating indices and suggested characteristics*

Index	Voice identity characteristic
1	Source
2	Closer to source
3	Mix of source and target
4	Closer to target
5	Target

The novel *coVoC* system alone, presented in section 7.2, and *coVoC* in combination with different *PSY* variants, presented in section 7.3, are examined on their VC performance. Both are evaluated with respect to their performance versus the GMM baseline method introduced in section 7.4.1.

7.4.2.1 *coVoC* and *PSY* parameterization

The target phrase of the utilized parallel corpus has been excluded from the training data set for all evaluated systems. An overlapping of β_{ext} = 30 % into the neighbouring phonemes has been applied to establish the test examples with *coVoC*. The employed MFCC feature vectors used an order of 16 coefficients. 15 Gaussian mixture components were defined to train each corresponding model for the prediction of the transformed voice descriptors. The slopes of the low pass filter P_L and the high pass P_H filter of the utilized spectral fading synthesis are set to m_{LP}=-12 dB and respectively m_{HP}=-48 dB per octave. The m_{LP} slope modification introduced in section 7.3.7 is set to an empirically determined factor of θ_{boost}=10. All parameter settings have been established empirically by means of several consecutive conversion and synthesis trials towards the target speaker.

Table 7.2 lists the sequence of speech phrases presented to the listener online for evaluation. Please note that

Table 7.2: *VC Test from Xavier to Fernando - Test indices per Voice Conversion algorithm*

Index	Speaker	VC Method	Signal parts
1	Source X	Original recording	$S_{src}(\omega)$
2	Converted	*coVoC* and *PSY*	$V''(\omega)$, $U''(\omega)$, F_{VU}, R_d, E_{voi}, E_{unv}
3	Target F	*PSY* (Re-Synthesis)	$\hat{S}_{tar}(\omega)$
4	Converted	*coVoC* and *PSY*	$V''(\omega)$, $U''(\omega)$, F'_{VU}, R'_d, E_{voi}, E_{unv}
5	Converted	*coVoC* standard	$T''_{sig}(\omega)$
6	Source X	*PSY* (Re-Synthesis)	$\hat{S}_{src}(\omega)$
7	Converted	*coVoC* and *PSY*	$V''(\omega)$, $U''(\omega)$, F'_{VU}, R'_d, E'_{voi}, E'_{unv}
8	Converted	GMM baseline	$T''_{sig}(\omega)$
9	Target F	Original recording	$S_{tar}(\omega)$
10	Converted	*coVoC* and *PSY*	$V''(\omega)$, $U''(\omega)$, F'_{VU}, R_d, E'_{voi}, E'_{unv}
11	Converted	*coVoC* optimized	$T''_{sig}(\omega)$
12	Converted	*coVoC* and *PSY*	$V''(\omega)$, $U''(\omega)$, F'_{VU}, R'_d, E_{voi}, E_{unv}
13	Converted	*coVoC* and *PSY*	$V''(\omega)$, $U(\omega)$, F_{VU}, R_d, E_{voi}, E_{unv}

the column 'Signal parts' lists those signal parts being converted or transformed from the source to the target speaker. The rows with the original or re-synthesized speech phrase list the corresponding signal denomination for completeness (without having any signal part converted or transformed).

Only the test phrase with index 13 utilizes the original spectrum $U(\omega)$, the unvoiced stochastic component of source speaker X not being converted to target speaker F. All other converted examples utilize the spectral envelope sequence $\mathcal{T}''_{unv}$ converted to target speaker F to synthesize $U''(\omega)$, along with the converted VTF sequence $\mathcal{T}''_{voi}$ to synthesize the voiced component $V''(\omega)$. The utmost right column lists the relevant signal components utilized for synthesis by each *coVoC* and *PSY* approach.

The original recordings of source speaker X and target speaker F, denoted as $S_{src}(\omega)$ and $S_{tar}(\omega)$, are listed as test phrase 1 and 9. The examples of their direct re-synthesis of $\hat{S}_{src}(\omega)$ and $\hat{S}_{tar}(\omega)$ using *PSY* with its estimation of

the corresponding voiced $\hat{V}(\omega)$ and unvoiced $\hat{U}(\omega)$ components are listed as test phrase 6 and 3. Both examples are established without the application of any means to convert by $''$ or to transform by $'$.

Test phrase 8 presents the GMM baseline method explained in section 7.4.1. Test phrases 5 and 11 list the direct application of $\mathcal{T}''_{sig}$ being converted by *coVoC* on the spectrum $S(\omega)$ of source speaker X by means of the SuperVP phase vocoder of section 2.5.2. The test phrase 5 "*coVoC* standard" has been established in combination with the same parameter set that has been used to create the *coVoC* and *PSY* variants listed as test phrases 2, 4, 7, 10, 11, 12, and 13. The test phrase 11 "*coVoC* optimized" represents a further parameter optimization of the standard parameter set defined for the *coVoC* standalone version to better approximate the utilized algorithm on the speaker pair used for conversion. Please note that one parameter set can be optimal for the *coVoC* and *PSY* combination but may interfere the quality of the *coVoC* standalone version, and vice versa. Due to the different source excitation signals used, a converted spectral envelope sequence $\mathcal{T}''_{sig}$ for *coVoC* standalone or the corresponding $\mathcal{T}''_{voi}$ and $\mathcal{T}''_{unv}$ combination for *coVoC* and *PSY* results in a different synthesis quality and conversion score.

7.4.2.2 Hidden original and re-synthesized versions

Fig. 7.3 depicts the listening test results for the hidden original X_{orig} / F_{orig} and the hidden re-synthesized X_{resyn} / F_{resyn} speech phrases. The **Speaker Identity (SI)** rating according to table 7.1 is shown in the upper picture. The mean results are shown in continuous red lines. The median results are shown in dash-dotted violet lines. Both are for the original recordings as expected very close to their judgement, being 1 for the source speaker X_{orig} and 5 for the target speaker F_{orig}. One out of 41 participants has been apparently confused to rate test phrase 9 being the target speaker F_{orig} with 1 (source) instead of 5 (target). All other 40 participants could well identify the hidden original F_{orig} and rated it with 5.

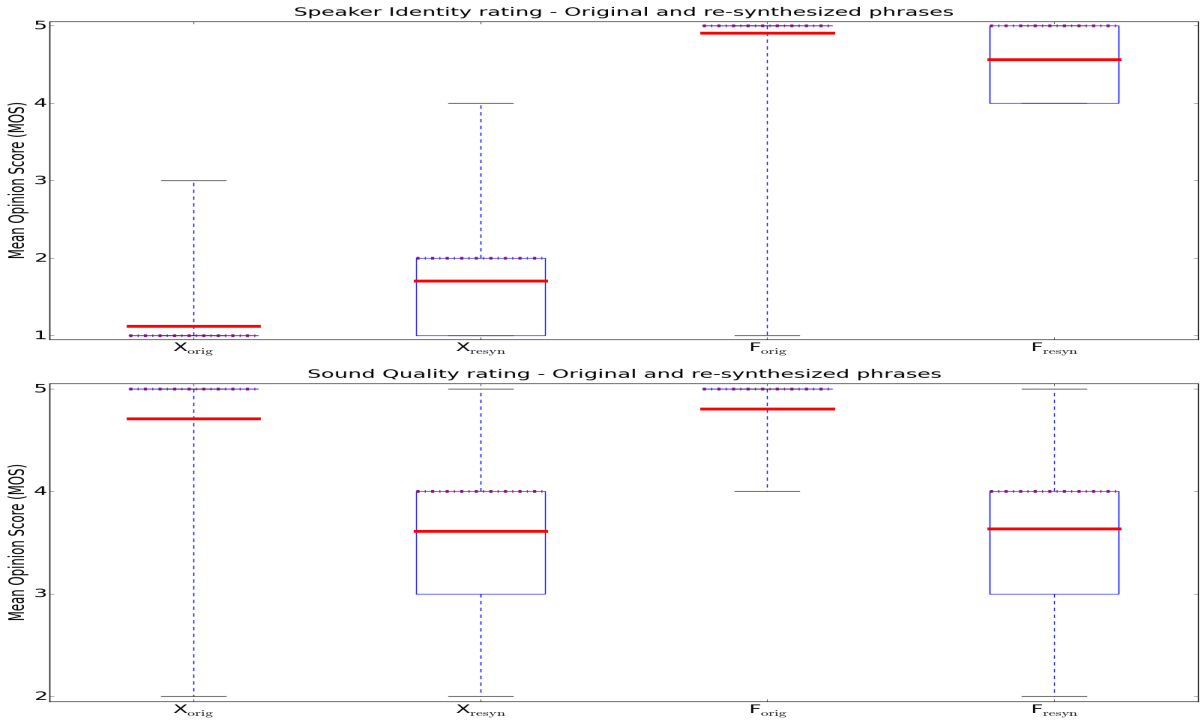


Figure 7.3: *VC listening test - Speaker Identity (SI) and Synthesis Quality (SQ) rating (Original phrases)*

The hidden re-synthesized versions X_{resyn} and F_{resyn} received less good but still acceptable ratings. The re-synthesis version F_{resyn} of the target speaker exhibits for the Speaker Identity a perfect median at 5 and a mean of SI-MOS=4.56. However, the SI median of the source speaker X_{resyn} is at 2 (Closer to Source), with SI-MOS=1.71. The listeners perceive due to the decline in synthesis quality introduced by the re-synthesis with *PSY* in general a shift in Speaker Identity away from the source speaker X and the target speaker F. The deviation in SI of the ideal 1 for X_{resyn} and the ideal 5 for F_{resyn} indicates the deterioration introduced by the analysis and re-synthesis of the speech phrase using *PSY* of chapter 6. The deviation indicates the best achievable performance of the VC variants using *PSY*. The results of the combination *coVoC* and *PSY* presented in the following have to be related to the results of X_{resyn} and F_{resyn}.

The **Sound Quality (SQ)** rating on the MOS scale exhibits a similar observation than its corresponding SI rating.

A deviation of the ideal synthesis quality 5 (Excellent) which is expected for an original speech recording of high quality [Lanchantin et al., 2008] is reported for the hidden original speech examples X_{orig} and F_{orig} by some listeners. The median of the MOS Sound Quality is for both original versions at an excellent 5. The re-synthesized versions of the source speaker X_{resyn} and the target speaker F_{resyn} received a mean of SQ-MOS=3.63 and SQ-MOS=3.61, with both having a median of SQ-MOS=4.

7.4.2.3 Results - GMM baseline and *coVoC* original versions

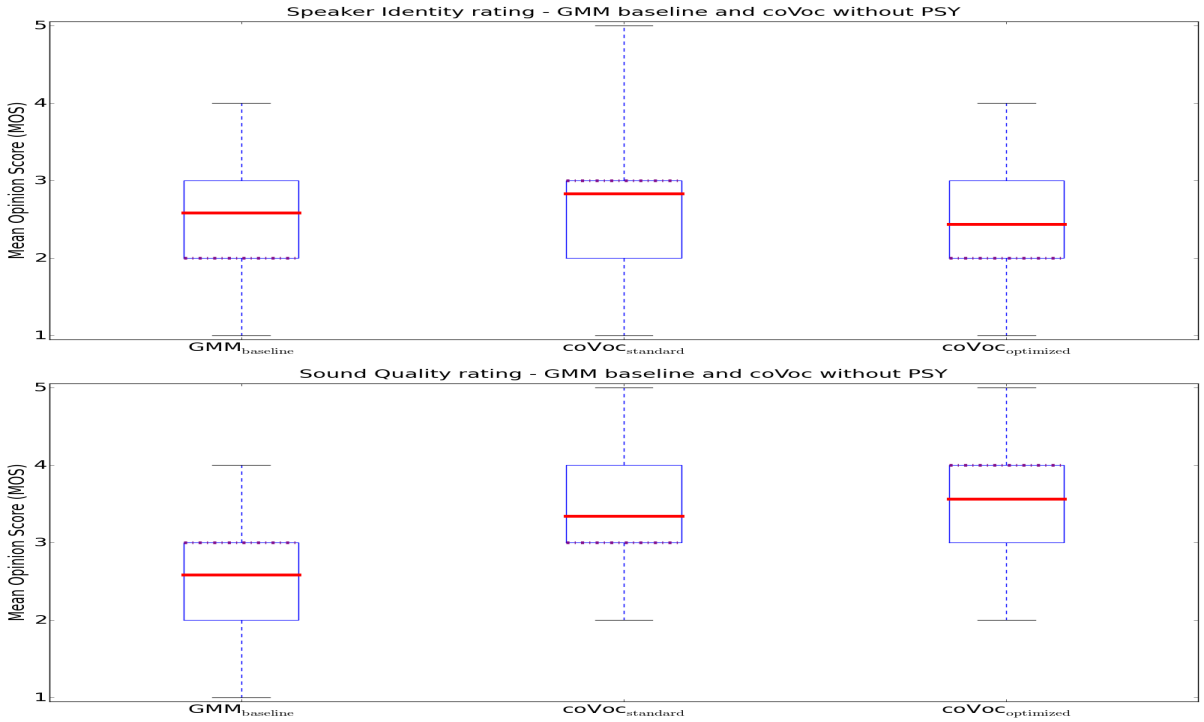


Figure 7.4: *VC test - Speaker Identity (SI) and Synthesis Quality (SQ) rating (GMM and coVoC)*

Fig. 7.4 illustrates the listening test results of the GMM baseline versus the *coVoC* standard and optimized versions. The desired improvements compared to the GMM approach are limited. Summarized in table 7.3, the *coVoC* standard version achieves a SI-MOS=2.83 being only marginally higher than SI-MOS=2.59 for the baseline. The *coVoC* optimized version receives an even lower SI-MOS=2.44! However, at least the synthesis quality of *coVoC* standard with SQ-MOS=3.34 and of *coVoC* optimized with SQ-MOS=3.56 are comparably higher than SQ-MOS=2.59 for the GMM baseline.

7.4.2.4 Results - *coVoC* and *PSY* versions

Fig. 7.5 depicts the SI and SQ results for the *PSY* variants introduced in section 7.3. The variant with index 13 to the most left utilizes the spectral envelope sequence $\mathcal{T}''_{voi}$ converted by *coVoC* to synthesize the converted voiced component $V''(\omega)$. It uses the non-converted spectral envelope sequence $\mathcal{T}_{unv}$ of the source speaker to synthesize the unvoiced component $U(\omega)$. Each further *coVoC* and *PSY* variant shown on the horizontal to the right utilizes the converted spectral envelope sequences $\mathcal{T}''_{voi}$ and $\mathcal{T}''_{unv}$ to synthesize the converted voiced $V''(\omega)$ and unvoiced $U''(\omega)$ components. The more to the right, the more voice descriptors are transformed towards the target speaker. The utmost left variant with index 13 utilizes a voice descriptor set being completely derived from the source speaker: F_{VU}, R_d, E_{voi} and E_{unv}. The *coVoC* and *PSY* variant to the most right with index 7 utilizes a voice descriptor set being completely transformed from the source towards the target speaker: F'_{VU}, R'_d, E'_{voi}, E'_{unv}. Table 7.2 summarizes the test phrases in ascending order as shown to the test participants online. Table 7.3 re-orders the phrases such that the test phrase with index 13 appears at the highest row as *coVoC* and *PSY* variant "c+P". It represents the least conversion towards target speaker F with only $V''(\omega)$ being converted. The lowest row with index 7 reflects the largest conversion towards target speaker F with all components being converted or transformed. The column "Method" lists in parentheses the features which have been converted or transformed towards the target speaker. The following three columns to the right list the mean (SI_μ), median ($\tilde{SI}$) and variance

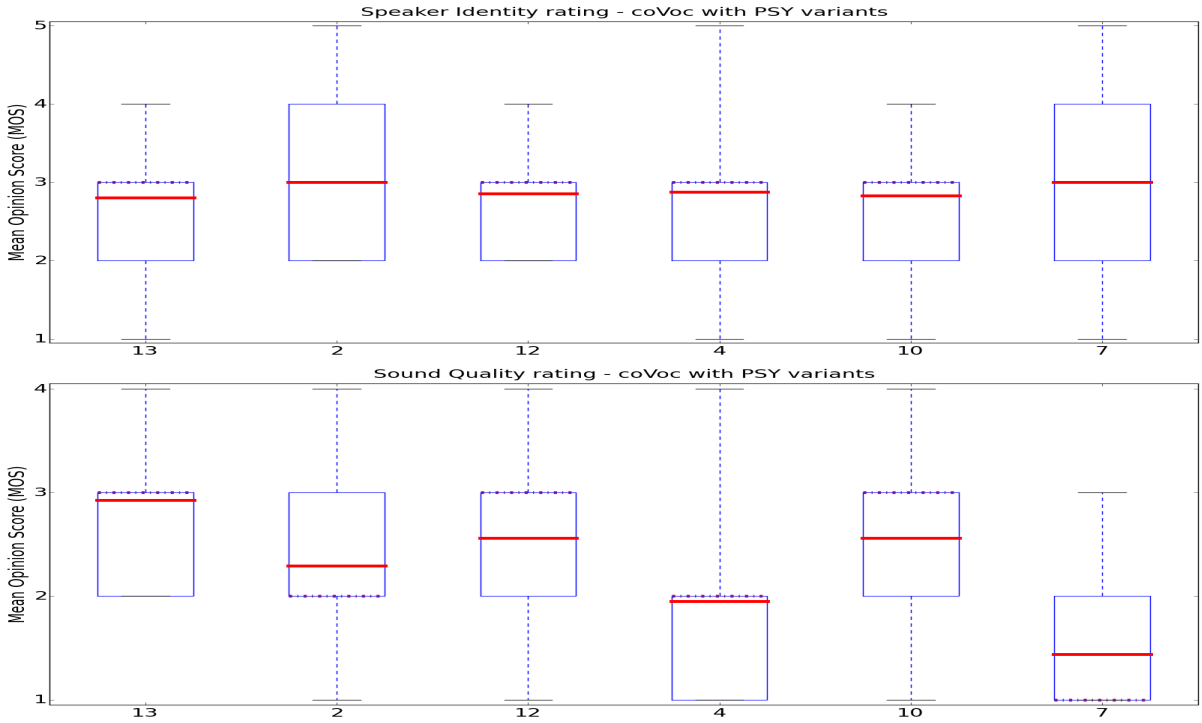


Figure 7.5: *VC test - Speaker Identity (SI) and Synthesis Quality (SQ) rating (coVoC and PSY variants)*

Table 7.3: *VC listening test - Speaker Identity (SI) and Synthesis Quality (SQ) rating*

Index	**Method (converted, transformed)**	$\mathbf{SI}_{\mu}$	$\tilde{\mathbf{SI}}$	$\mathbf{SI}_{\sigma^2}$	$\mathbf{SQ}_{\mu}$	$\tilde{\mathbf{SQ}}$	$\mathbf{SQ}_{\sigma^2}$
1	Source X original	1.12	1	0.39	4.71	5	0.63
6	Source X *PSY* re-synthesis	1.71	2	0.67	3.61	4	0.69
9	Target F original	4.90	5	0.62	4.71	5	0.63
3	Target F *PSY* re-synthesis	4.56	5	0.50	3.63	4	0.85
8	GMM baseline	2.59	2	0.88	2.59	3	0.76
5	*coVoC* standard	2.83	3	0.93	3.34	3	0.81
11	*coVoC* optimized	2.44	2	0.91	**3.56**	4	0.70
13	c+P (V'')	2.80	3	0.77	2.93	3	0.75
2	c+P (V'', U'')	**3.00**	3	0.86	2.29	2	0.74
12	c+P (V'', U'', F'_{VU})	2.85	3	0.75	2.56	3	0.70
4	c+P (V'', U'', F'_{VU}, R'_d)	2.88	3	0.77	1.95	2	0.73
10	c+P (V'', U'', F'_{VU}, E'_{voi}, E'_{unv})	2.83	3	0.79	2.56	3	0.73
7	c+P (V'', U'', F'_{VU}, R'_d, E'_{voi}, E'_{unv})	**3.00**	3	0.91	1.44	1	0.59

(SI_{σ^2}) of the Speaker Identity rating. The three columns to the most right list mean (SQ_{μ}), median ($\tilde{SQ}$) and variance (SQ_{σ^2}) of the Synthesis Quality rating.

The impact of adding more transformed voice descriptors to the synthesis towards the target speaker is only marginal given the evaluated speaker pair. The least conversion with index 13 exhibits a SI-MOS=2.80 while the largest conversion exhibits a SI-MOS=3.00, with all other *coVoC* and *PSY* variants achieving a Speaker Identity conversion score in-between. The contributions to the SI of the *PSY* variants do not achieve significant improvements. Moreover, the variant with the least conversion (index 13) applied achieves the highest SQ-MOS=2.93, while the variant with the largest conversion (index 7) applied exhibits the lowest SQ-MOS=1.44.

Please note that the very low SQ-MOS rating for index 7 has been expected. The combination of the predicted F'_{VU} and R'_d to predict in turn the energies E'_{voi} and E'_{unv} results for the time being in a feature combination which renders unfortunate energy values. This introduces higher and unnatural sounding energy contours which can be perceived as artefacts as a result of an algorithmic mistake. However, the listeners rated its Speaker Identity with SI-MOS=3.00 as the highest conversion score in the test. The other variant (index 2) which achieves as well the highest conversion score SI-MOS=3.00 does not apply any voice descriptor transformation. Each *coVoC* and *PSY* variant "c+P" exhibits a slightly higher SI-MOS conversion score than the GMM baseline method of section 7.4.1.

7.4.2.5 Results - General discussion

The following observations apply in general to all results of the listening test. They constitute further reasons why the results of the subjective evaluation are not completely reflecting the expectation of the presented algorithms.

Test index randomization:
Please note that the presented test results are sadly biased. The order of the utilized speech phrases was randomized once when setting up the listening test. However, the author did not set up a system online on the webpage which places the speech phrases in random order each time a participants starts the test, as it is for example done for the listening test of [Lanchantin et al., 2011b]. The psycho-acoustic learning effect as discussed in section 6.6.5.3 interferes therefore the results.

3^{rd} speaker effect:
The well-known 3^{rd}-speaker effect, discussed in section 4.7.1, is repeatedly reported in the VC research community, especially for the intra-gender Voice Conversion task [Erro, 2008]. Some listeners can not recognize neither the source nor the target speaker. The effect was as well reported by some participants of the VC listening test presented in this work.

Prosodic difference:
One major drawback of the conducted VC test is the huge difference in the prosodic characteristic between source and target speaker. As explained in the beginning of this section, the speaker pair has been chosen to reflect the former research work conducted at IRCAM. However, the huge prosodic differences mask to a certain extent the results perceptually since no conversion in prosody has been applied. The similar results of the *coVoC* and *PSY* variants can be partially attributed to the masking effect of the prosodic differences.

Perceptual masking effect:
The synthesis quality ratings for the combination of *coVoC* and *PSY* are comparably lower compared to the *coVoC* standalone version. As already discussed in the sections 7.3.4, 7.3.5, 7.3.6, and 7.3.7, the reasons are the higher risks to robustly establish a feature combination reflecting a signal waveform of natural human speech. The effect of transforming an extended feature set to the target speaker is thus masked by the introduced artefacts. A higher synthesis quality would lead to a better perception of the target speakers voice identity. Several listening test participants reported in correspondence that they could not perceive apparent differences between the presented speech phrases due to the reduced synthesis quality. The evaluation of the combined *coVoC* and *PSY* system could consequently not be rated reliably.

7.4.3 Objective evaluation - LSF distance measure

This section presents the test results of an objective evaluation corresponding to the subjective evaluation of the preceding section 7.4.2. The objective evaluation is based on the same speech phrases listed at and explained with the tables 7.2 and 7.3. In particular the GMM baseline of section 7.4.1, the *coVoC* standalone versions 'standard' and 'optimized', and all variants of *coVoC* combined with *PSY* are examined with an objective and not a subjective evaluation metric.

7.4.3.1 Test setup

Speech parameterization: LSF
A spectral envelope sequence is estimated on each speech phrase using the True Envelope estimator presented in section 2.6.4. A T''_{sig} sequence is estimated on each converted speech phrase. The TE sequences T_{src} and T_{tar} are estimated on the original source and target speech phrases. Each TE sequence is parameterized by LSF vectors, introduced in section 2.6.2. The LSF encoding efficiently reflects the vocal tract formants [Bäckström and Magi, 2006]. The LSF parameterization serves with this as perceptually meaningful objective distance measure [Zheng et al., 1998] and constitutes an important part of a speakers voice identity [Hasegawa-Johnson, 2000].

Distance measure: Mean Root-Mean-Square Error D_{LSF}
The objective evaluation employs a distance measure to evaluate the difference between parameters reflecting a speech signal [Vepa et al., 2002]. The distance D_{LSF} between two LSF vectors F and F' is defined as the RMS error over the LSF order M and the mean over the number of frames N [Kain and Macon, 2001]:

$$D_{LSF}(F, F') = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(F_m(n) - F'_m(n)\right)^2} \tag{7.2}$$

The distance measure of equ. 7.2 expresses the difference between each LSF coefficient pair $F_m(n)$ and $F'_m(n)$.

Evaluation metric: Performance Index P_{LSF}
The Performance Index P_{LSF} of equ. 7.3 was proposed in [Kain and Macon, 2001]:

$$P_{LSF} = 1 - \frac{D_{LSF}(F_{tar}, \hat{F}_{tar})}{D_{LSF}(F_{tar}, F_{src})}\left[\frac{(\text{Transformation error} D_{tc})}{(\text{Inter-speaker distance} D_{ts})}\right] \tag{7.3}$$

It evaluates the transformation error D_{tc} as the remaining difference between the LSF coefficients measured on the target F_{tar} versus the converted towards the target $\hat{F}_{tar}$ speech phrase. The transformation error is normalized by the inter-speaker distance D_{ts} being the difference between source F_{src} and target F_{tar} LSF coefficients estimated on the corresponding speech phrases. The final P_{LSF} evaluates the difference of 1 minus the ratio of transformation error D_{tc} and inter-speaker distance D_{ts}. Both D_{tc} and D_{ts} are computed using the D_{LSF} distance measure of equ. 7.2.

Inter-speaker distance D_{ts}:
The original source and target speaker speech recordings of the speech phrase chosen from the parallel data corpus constitute the evaluation basis. The LSF feature vectors F_{tar} and F_{src} are estimated on the original target and source speakers speech recordings. The Euclidean distance metric of equ. 7.2 determines the difference between F_{tar} and F_{src}. It expresses the LSF distance the conversion has to bridge in order to transform the source into the target speech phrase. It constitutes the lower bound of an achievable conversion towards the target speaker.

Transformation error D_{tc}:
The same distance metric evaluates the remaining LSF distance between the LSF target vector F_{tar} and the LSF vector $\hat{F}_{tar}$ estimated on the speech phrase being converted from the source to the target speakers voice identity.

Obviously, an optimal Performance Index P_{LSF} of 1.0 is achieved if the converted to the target speaker LSF vector $\hat{F}_{tar}$ equals F_{tar}, resulting in a remaining transformation error of D_{tc}=0. Likewise, the worst Performance Index P_{LSF} of 0.0 is given if the source phrase hasn't been converted at all such that $\hat{F}_{tar}$ equals F_{src} which results into a ratio of 1.0 between transformation error D_{tc} and inter-speaker distance D_{ts}.

Feature alignment: Per phoneme Dynamic Time Warping (DTW)
Many different prosodic behaviours exist to utter the same sentence. Even the same speaker underlies naturally differences in timing and F_0 contour when repeatedly uttering the same phrase. Moreover, higher prosodic difference exist between different speakers. The timing difference between source, target and converted to the target speech phrase have to be minimized by means of a time alignment such that the feature vectors describing each phrase are properly matched.

The algorithmic steps to align the source, target and converted to the target speaker speech phrases to a common time basis are given as follows:

1. Compute the spectral envelope sequences T^{src}_{sig} and T^{tar}_{sig} on the original speech recordings of the source and target speaker.
2. Compute MFCC, Delta-MFCC and Delta-Delta-MFCC feature vectors on T^{src}_{sig} and T^{tar}_{sig}.

3. Compute a DTW cost matrix by examining the pair-wise Euclidean distance between the three MFCC feature vectors calculated in the preceding step.

4. Apply phonetic constraints by setting the DTW matrix entry at each phoneme border to zero. Set the remaining entries to an error maximum.

5. Apply dynamic programming by means of DTW to determine the best path through the trellis having lowest cost and lowest distance.

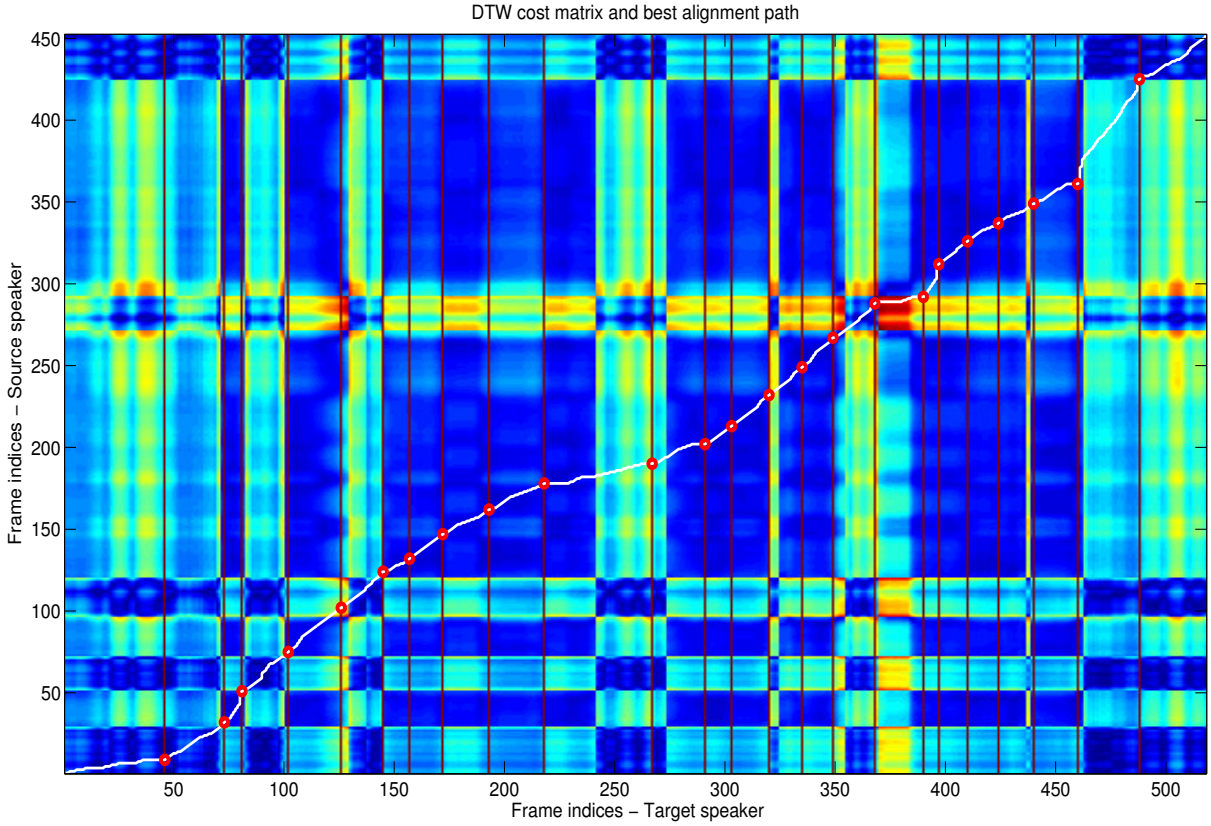


Figure 7.6: *Example of DTW cost matrix, phoneme borders (red) and best path (white) through DTW trellis*

Fig. 7.6 illustrates the DTW alignment estimated on the original speech recordings of the source and target speaker. Matrix entries in blue (red) colour indicate higher (lower) similarities between the extracted MFCC features and lower (higher) costs for the DTW algorithm. The y-axis reflects the source and the x-axis the target speakers frames. The vertical lines in dark red colour depict the phoneme borders being set to maximal cost. The phoneme borders are set according to the frame indices of the target speakers time basis. The circles in bright red colour reflect the phoneme borders being set to zero distance and highest similarity for both time basis. These phonetic constraints aid the dynamic programming approach in determining the best alignment path. The DTW algorithm has to path the phoneme border entries in the DTW cost matrix shown in white colour. On the one hand, the alignment is less prone to move too far away from the ideal line at the diagonal. On the other hand, the DTW is influenced by a possible erroneous phoneme border estimation. The latter is based on ircamAlign, presented in section 4.2.1. Fig. 7.7 shows the DTW alignment costs calculated from the source and the target speakers MFCC feature sequence. The DTW cost has to drop to zero each time it passes a phonetic constraint.

7.4.3.2 Test results

Table 7.4 summarizes the LSF distance measures calculated by using LPC orders of 50 and 20 for the Performance Indices P^{50}_{LSF} and respectively P^{20}_{LSF}. The objective evaluation results of the GMM baseline method, the standard and optimized *coVoC*, as well as the different *coVoC* and *PSY* variants are listed. The distances of the corresponding transformation error D_{tc} and inter-speaker distance D_{ts} are additionally listed. This better illustrates the impact of using a different LPC order for especially the inter-speaker distance D_{ts} which is constant for one speech phrase pair of source and target speaker.

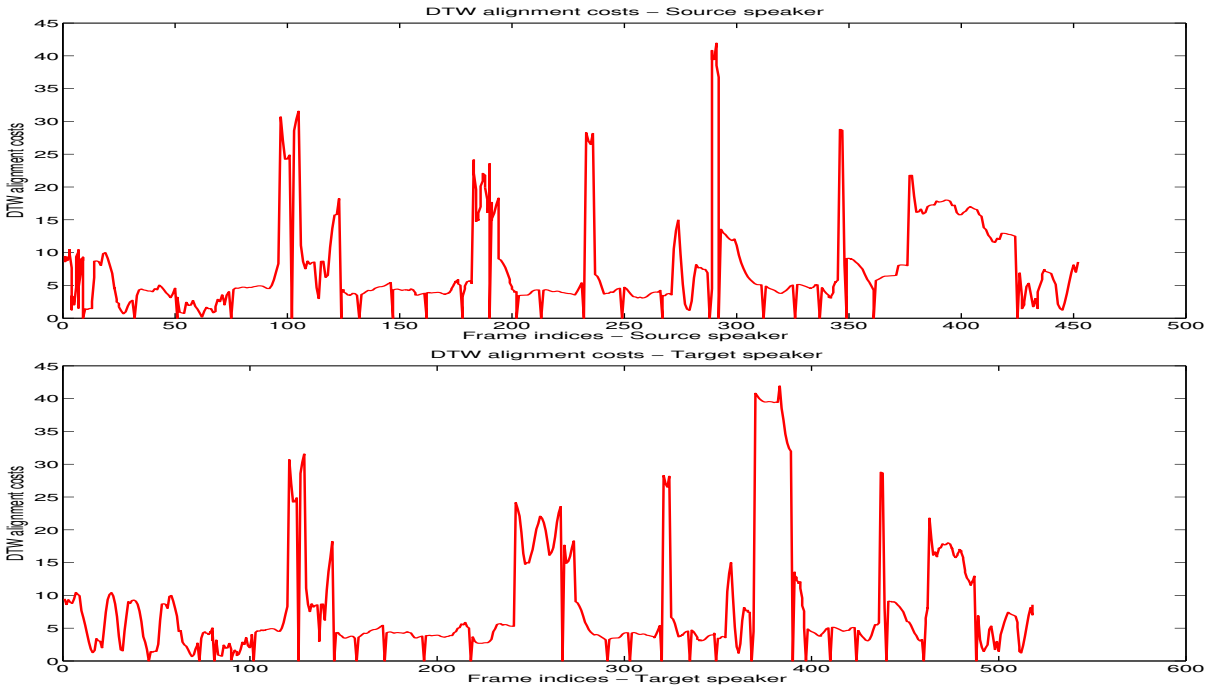


Figure 7.7: *Example of DTW alignment costs for speech phrases with time basis of source and target speaker*

Table 7.4: *VC objective evaluation - Performance Index P_{LSF}, as well as distances to target D_{tc} and D_{ts}*

Index	**Method (converted, transformed)**	P^{50}_{LSF}	D^{50}_{tc}	D^{50}_{ts}	P^{20}_{LSF}	D^{20}_{tc}	D^{20}_{ts}
3	*PSY* re-synthesis target F	*0.7696*	0.0086	0.0374	*0.8012*	0.0163	0.0819
6	*PSY* re-synthesis source X	*0.0107*	0.0370	0.0374	*0.0308*	0.0794	0.0819
8	GMM baseline	0.1160	0.0331	0.0374	0.1546	0.0693	0.0819
5	*coVoC* standard	**0.2047**	0.0298	0.0374	**0.2275**	0.0633	0.0819
11	*coVoC* optimized	0.1855	0.0305	0.0374	0.2187	0.0640	0.0819
13	c+P (V'')	0.1236	0.0328	0.0374	0.1289	0.0714	0.0819
2	c+P (V'', U'')	0.1378	0.0323	0.0374	0.1482	0.0698	0.0819
12	c+P (V'', U'', F'_{VU})	0.1657	0.0312	0.0374	0.1813	0.0671	0.0819
4	c+P (V'', U'', F'_{VU}, R'_d)	0.1675	0.0312	0.0374	0.1871	0.0666	0.0819
10	c+P (V'', U'', F'_{VU}, E'_{voi}, E'_{unv})	0.1760	0.0309	0.0374	0.1945	0.0660	0.0819
7	c+P (V'', U'', F'_{VU}, R'_d, E'_{voi}, E'_{unv})	0.1664	0.0312	0.0374	0.1849	0.0668	0.0819

Distances between original and re-synthesized speech phrases

The values P^{50}_{LSF}=0.7696 and P^{20}_{LSF}=0.8012 of table 7.4 constitute the upper bounds for the target speakers phrase. P^{50}_{LSF}=0.0107 and P^{20}_{LSF}=0.0308 constitute the lower bounds for the source speakers phrase. Both bounds reflect the chosen source and target speaker pair, the implied signal approximations of the utilized speech framework *PSY*, and the evaluation metric of LSF vectors using the P_{LSF} distance measure. The difference of 0.7696 minus 0.0107 for P^{50}_{LSF} and respectively 0.8012 minus 0.0308 for P^{20}_{LSF} refers to the amount of signal difference expressed by the LSF distance metric between both speech phrases. The VC algorithm has to apply this amount of signal difference to the source speakers signal such that it properly captures the target speakers voice identity.

An ideal analysis-resynthesis scheme would result in P^{50}_{LSF}=0 and P^{20}_{LSF}=0 for the source speakers phrase. Both measures reflect the diminution in quality introduced by the analysis and synthesis using the *PSY* speech system. The distances P^{50}_{LSF}=0.0107 and P^{20}_{LSF}=0.0308 indicate a high synthesis quality, with a low reduction in distance of only ∼1 % for P^{50}_{LSF} and ∼3 % for P^{20}_{LSF}.

Comparably huge differences of P^{50}_{LSF}=0.7696 and P^{20}_{LSF}=0.8012 are measured between the original versus the re-synthesized speech phrase for the target speaker. An ideal speech system and evaluation metric would achieve a P_{LSF} measure of 1.0. The huge reductions of ∼23 % for P^{50}_{LSF} and ∼20 % for P^{20}_{LSF} are primarily a result of the implied requirement to align the target to the source speakers phrase. The alignment is necessary to compensate the natural timing differences in phoneme length when uttering the same phrase such that the phonetic content of both speech phrases are matched in time. However, natural differences in prosody and articulation between speakers omit an ideal alignment. The matching of the LSF feature vectors between source and target speaker compares therefore consequently partially misaligned signal content. This leads to the observed higher reduction from an ideal P_{LSF}=1.0 for the target speaker phrase.

The original phrase of the target speaker and the phrases being converted to the target speaker have been aligned to the source speakers phrase. This is the reason why the measures P^{50}_{LSF} and P^{20}_{LSF} are far away from its ideal 1.0 for the target speaker. Likewise, both measure are very close to the ideal 0.0 for the source speaker.

Another reason for both measures to drift away from their ideal 0.0 and 1.0 P_{LSF} measure can be partially attributed to a related observation in [Erro, 2008]: Two similar spectra may result in different LSF vectors. The encoding of two spectra having a very similar perceptual content can result in some comparably different LSF coefficients (encoding a certain spectral region) of a LSF vector (encoding one complete spectral frame). Such differences may appear even if the perceptual difference is not perceivable to a listener. This constitutes an unavoidable drawback of the underlying evaluation metric based on LSF vectors and their distances.

Fig. 7.8 depicts the aligned LSF sequences being estimated on the original and re-synthesized speech phrases of the source and target speaker. The lower LPC order of 20 instead of 50 is more compact and thus better suited to illustrate the LSF feature vectors between source and target speech phrase, as well as between their original and their re-synthesized versions.

Only minor differences can be inspected visually between the LSF sequences estimated on the original and re-synthesized speech phrase of the target speakers. The first LSF coefficient shown in green colour exhibits bigger differences between the frames 420 to 470. However, the huge reduction in P^{50}_{LSF} and P^{20}_{LSF} from the ideal 1.0 resulting from a non-perfect DTW alignment which cannot be directly concluded from visual inspection.

Remaining distances between target and converted to the target speech phrases:

The highest proximity among the phrases converted to the target speaker is achieved by the standard *coVoC* method with P^{50}_{LSF}=0.2047 and P^{20}_{LSF}=0.2275. The optimized *coVoC* method achieves the overall second best performance with P^{50}_{LSF}=0.1855 and P^{20}_{LSF}=0.2187. Both comparably outperform the GMM baseline method whose results are P^{50}_{LSF}=0.1160 and P^{20}_{LSF}=0.1546.

The results of the objective evaluation corroborate the findings of the subjective evaluation between the GMM baseline method and the standard *coVoC* method using a default parameter set. The latter outperforms in both evaluations the baseline method. The optimized *coVoC* method achieves a lower (higher) performance in the objective (subjective) evaluations compared to the GMM baseline method.

The worst performing *coVoC* and *PSY* variant with index 13 which uses the unvoiced component $U(\omega)$ of the source speaker still outperforms with P^{50}_{LSF}=0.1236 the GMM baseline method which achieves P^{50}_{LSF}=0.1160. However, this is not the case for the Performance Index P^{20}_{LSF}. Only the *coVoC* and *PSY* variants using the F'_{VU} contour transformed to the target speaker outperform the baseline method for both Performance Index measures.

The sequence of *coVoC* and *PSY* variants as measured by the Performance Indices follows roughly the objective of this thesis: The objective evaluation exhibits a steady performance increase measured by both P_{LSF} metrics from the lowest conversion and transformation impact listed with index 13 to the highest impact listed with index 7. Starting from index 13 with only V'' being converted, the lowest performance of the *coVoC* and *PSY* variant exhibits P^{50}_{LSF}=0.1236 and P^{20}_{LSF}=0.1289. Adding different voice descriptor to convert or transform towards the

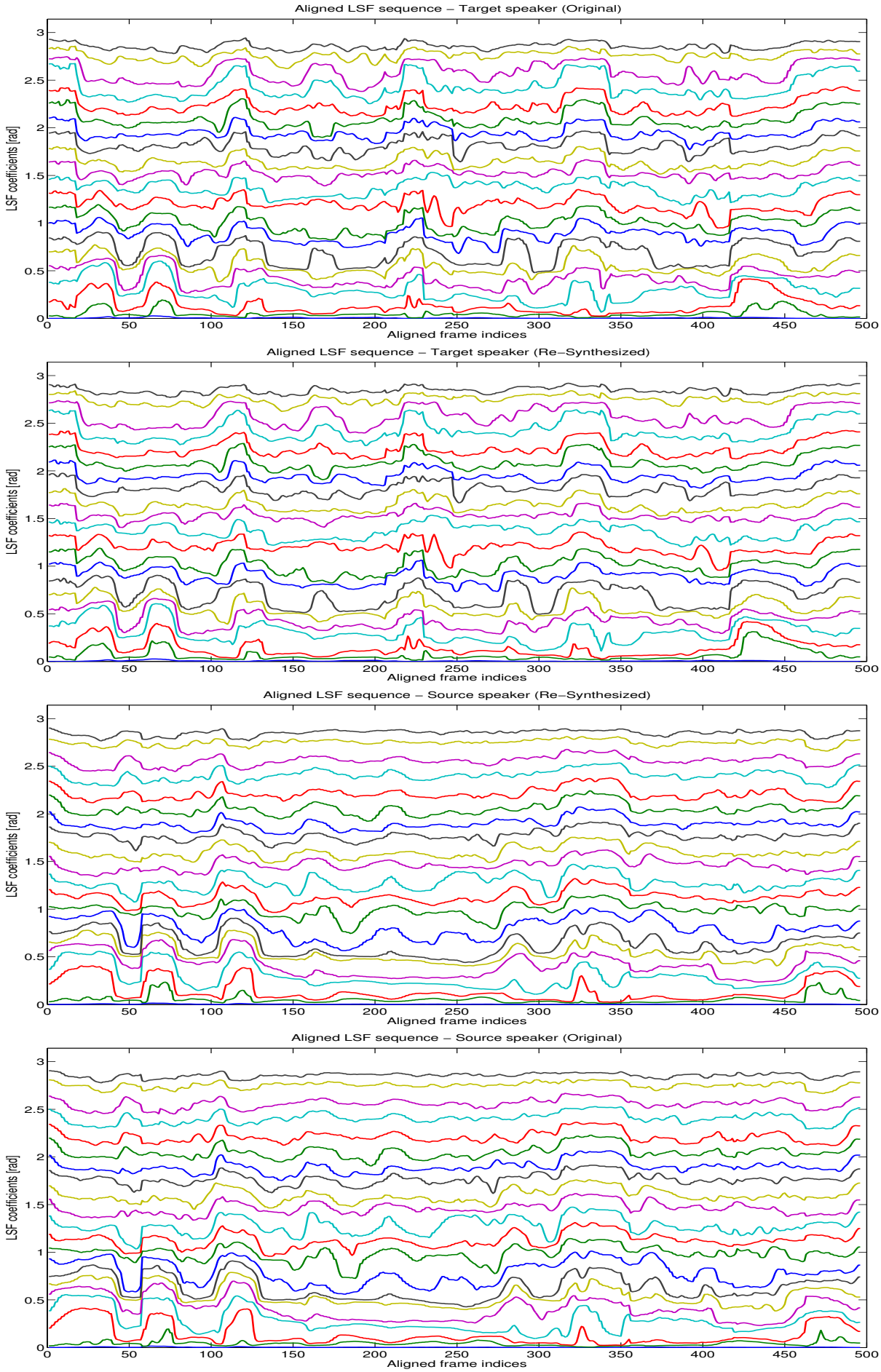


Figure 7.8: *LSF sequences, LPC order 20, original and re-synthesized speech phrases*

target speaker, both P_{LSF} measures increase. The *coVoC* and *PSY* variant with index 10 achieves the highest *coVoC* and *PSY* variant performance of P^{50}_{LSF}=0.1760 and P^{20}_{LSF}=0.1945. Only index 7 does contradict the thesis objective. Both $V^{''}$ and $U^{''}$ are converted. As well all voice descriptors $F^{'}_{VU}$, $R^{'}_{d}$, $E^{'}_{voi}$, and $E^{'}_{unv}$ are transformed. But the LSF distance measure is with P^{50}_{LSF}=0.1664 and P^{20}_{LSF}=0.1849 slightly lower.

Fig. 7.9 depicts on top the LSF sequence of the re-synthesized target speech phrase. It constitutes the upper bound of this test set given the *PSY* speech framework. The second LSF sequence reflects the conversion performance of the GMM baseline method. The influence of the over-smoothing effect introduced by the statistical model is visible, e.g. between the frames 150 and 300 or 350 to 450. The one but lowest LSF sequence illustrates the conversion performance of the standard *coVoC* method. The lowest LSF sequence shows the conversion results of the *coVoC* and *PSY* variant with index 10. It utilized the transformed $F^{'}_{VU}$, $E^{'}_{voi}$, and $E^{'}_{unv}$ contour for synthesis. This *coVoC* and *PSY* variant constitutes the best performing *coVoC* and *PSY* variant with P^{50}_{LSF}=0.1760 and P^{20}_{LSF}=0.1945.

Still, all presented algorithmic approaches do not properly capture the signal behaviour of the target speech phrase when converting the source phrase. This can be concluded by visual inspection of the selected LSF sequences. All do not exactly follow the corresponding LSF coefficient contours of the target phrase. This is reflected by all calculated P_{LSF} measures. The best performing conversion "*coVoC* standard" with P^{50}_{LSF}=0.2047 and P^{20}_{LSF}=0.2275 is still far from the performance measures P^{50}_{LSF}=0.7696 and P^{20}_{LSF}=0.8012 of the re-synthesized target speech phrases. This indicates that further work is required to properly convert a source speakers phrase to the target speaker to better capture its voice identity.

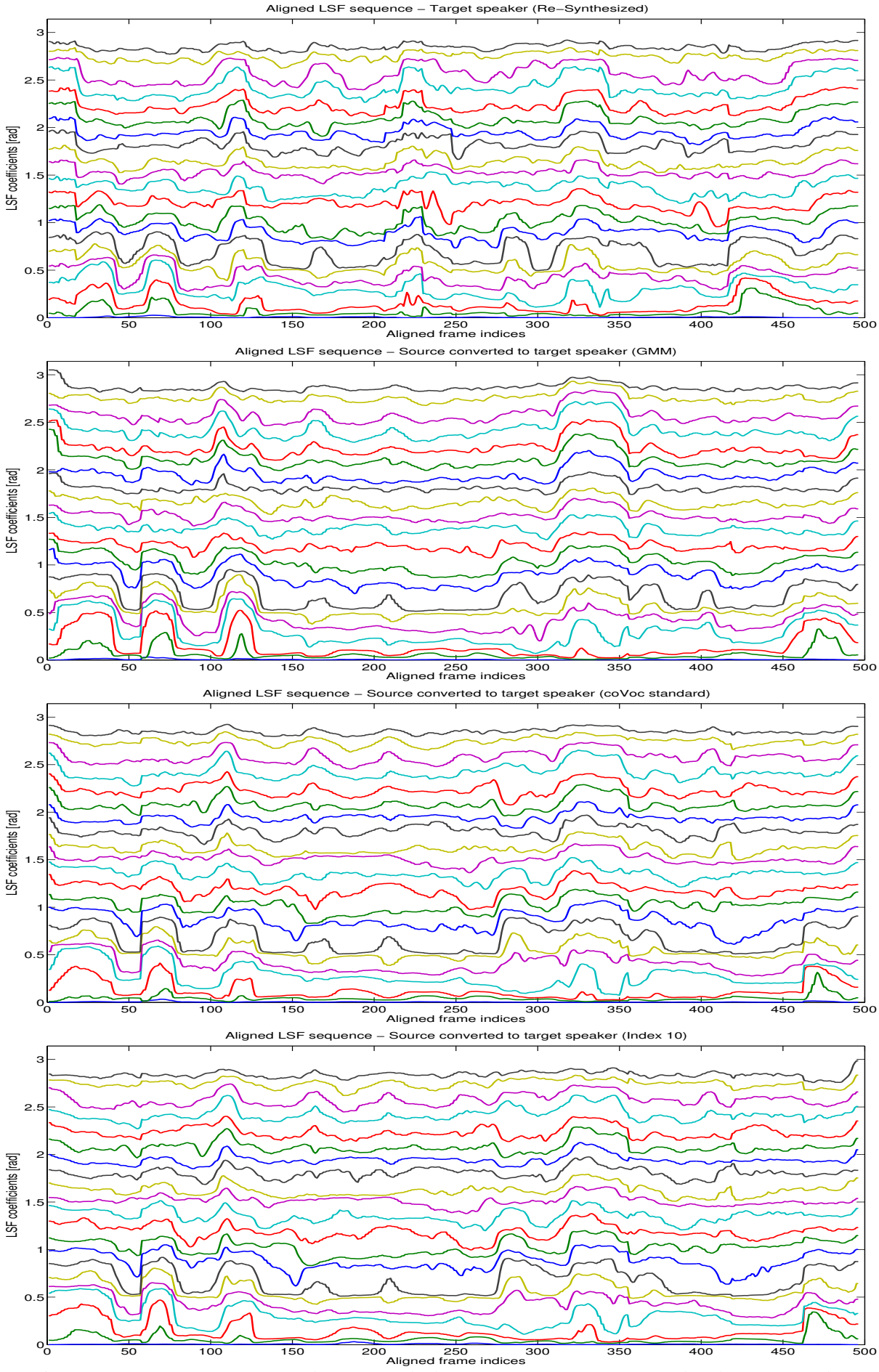


Figure 7.9: *LSF sequences, LPC order 20, speech phrases of target versus converted to the target speaker*

7.5 Conclusions

Major VC advancement:
The current status can be seen similarly to the Frame Selection approaches. The proof-of-concept analysis of [Helander et al., 2007] presented in section 4.8.3.2 concludes that Frame Selection methods are not eligible for VC. In contrast, several Frame Selection approaches discussed in section 4.8.3 proof the applicability of Frame Selection to augment the VC performance.

Despite the improvements of *coVoC* could only be validated to a minor extent by the test results discussed in section 7.4.2, its technological proof-of-concept has already been validated on different speaker pairs within several industrial projects at IRCAM. Please find further information via this link: IRCAM Voice Conversion Showcases [4]. The system *coVoC* constitutes thus a major advancement in Voice Conversion since reasonably good performing results have been achieved with other speaker pairs.

Further evaluation:
Due to the systematic problems of the presented listening test, described in section 7.4.2.5, further tests between different speaker pairs are required to properly evaluate both Voice Conversion and Voice Transformation systems. The novel *coVoC* standalone system of section 7.2 has to be further evaluated alone, and in combination with the transformation of the extended voice descriptor set using the *PSY* implementation, introduced in section 7.3.

Conversion and transformation of an extended voice descriptor set:
The construction of a speech phrase towards the target speaker by means of an artificial feature combination of converted voice descriptors proved to contribute to the VC task. Many research works, some of it presented in section 4.7, indicate that the voice identity of the target speaker is better captured the more features are converted or transformed. However, the risks presented in section 7.3.4 indicate that such means are difficult to handle. The following discussions of section 7.3.5 and 7.3.6 exemplify the requirement to implement further means assuring a synthesis of high quality. The extension of *coVoC* with the feature transformations conducted by *PSY* proved to be sensitive at synthesis. Further research work is required to assure a robust and more automatic handling of the combined features. This should augment the VC performance of *coVoC* combined with *PSY* since a lower synthesis quality masks perceptually the effect of transforming more voice descriptors towards the voice identity of the target speaker.

One-to-one speech phrase VC:
An interesting aspect of *coVoC* is the possibility to utilize in a parallel corpus setup as target corpus only the target speech phrase which corresponds to the source speech phrase. Only the single source and single target speech phrase are utilized for VC. Both phrases contain the same phonetic content and its evolution over time in the spoken sentence. This represents a non-realistic VC design. Such possibility does not exist in a real world application for VC. However, the parallel corpora established nowadays to design and evaluate a VC system provide this alternative. Conventional VC system would require a comparably higher amount of data to train their models on the underlying correlations between source and target feature vectors. The phoneme matching algorithm exempts *coVoC* from the requirement of having to evaluate a complete speech corpus. Conventional VC systems require as much data as available to properly model the acoustic space of the target speaker [Dutoit et al., 2007], or of both speakers [Kain, 2001, Stylianou et al., 1998].

The one-to-one phrase conversion can therefore be used to examine the performance of different parameterizations of the *coVoC* system. It performs in general better than the realistic design to exclude the existing target phrase from the target corpus to simulate a real world application. The target phoneme selection algorithm benefits from the same phonetic concatenation present between source and target speech phrase. The selection has only the freedom to choose between several phonemes if the chosen phrase contains several phonemes of the same X-SAMPA type. However, this freedom of phonetic choice presents an actual drawback, both in terms of the non-realistic one-to-one speech phrase design, and in the case of a one-to-many speech phrase setup given in real world applications. *coVoC* has to choose the next phoneme in the sequence if no second phoneme of the same type is available. This maintains the natural evolution of the concatenated spectral envelopes which is the reason for the better performance.

Sensitivity to phoneme annotation:
However, both the *coVoC* standalone version and its combination with *PSY* are for the time being rather sensitive to the phoneme annotation of the chosen source speakers phrase. Even marginal manual corrections of the phoneme annotation of the source speakers phrase may lead to a different selection of single phonemes from the target speaker corpus. This in turn may alter to a huge extent the perception of the re-synthesized target speech phrase at the corresponding speech segments of exchanged phonemes. The problem lies in the black box behaviour of the phoneme annotation and selection. If some selected target phonemes do not sound as desired and one corrects

[4] IRCAM Voice Conversion Showcases: http://anasynth.ircam.fr/home/media/covoc-voice-conversion-demo

the phoneme annotation of the source to change the target phoneme selection, it cannot be predicted if this manual correction increases or decreases the sound quality of the synthesized phrase and the speaker identity towards the target speaker.

This constitutes a drawback and an advantage of the *coVoC* system at the same time. On the one hand, it can result in a loop of manual phoneme corrections by means of trial-and-error to design the correct phoneme annotation of the source speaker to optimize the desired phoneme selection of the target speaker. On the other hand, the advantage is the possibility for corrections. The freedom of choice facilitates to achieve a VC of highest synthesis quality and conversion score which cannot be achieved with conventional VC systems.

Sensitivity to Viterbi parameterization:
The concatenation of matched phonemes is additionally sensitive to the parameterization of the Viterbi smoothing. The Viterbi weight to manually define the emphasis between single phoneme matches and the transition cost over the phoneme sequence constitutes certainly an implied attitude of the algorithm not posing a drawback of the VC system. The parameter to include a certain percentage of the neighbouring segments introduces another factor of sensitivity to the novel *coVoC* system. However, the mentioned sensitivity to the parameterization is an intrinsic behaviour of all concatenation techniques. Otherwise it can be seen again as an advantage to have the possibility to steer the system towards the desired results by means of several trials with different parameterizations.

Chapter 8

Summary and Outlook

God, who giveth a voice to all things, hath given us a voice:
He created you at first, and to Him are ye brought back.

THE HOLY QUR'AN (SURA XLI 523)

8.1 Conclusions

8.1.1 Estimation of the glottal excitation source

The adapted and extended R_d range of section 5.2 contributes to the robustness of the R_d estimator. The baseline method MSPD2I1 improved the R_d estimation results for the test set on natural human speech of section 5.6.4 from Pearsons correlation coefficient r=0.23 for the normal R_d range to r=0.29 and r=0.31 for the extended R_d range and both adaptation variants. The additive combination MSPD2IX of all phase minimization variants, introduced in section 5.3.3, improves along with the Viterbi smoothing, explained in section 5.4, the R_d estimation up to r=0.53, listed in table 5.5. This constitutes a reasonable increase of the R_d estimation performance compared to the baseline of ~130 %.

However, the promising Viterbi steering approach of section 5.5 could only advance the R_d estimation results to a marginal extent, only up to r=0.54. One reason can be that the R_d estimation given the employed test metric and the data set reached already a certain upper limit. This is indicated in section 5.6.4.5 by the Viterbi steering trial where the utilization of higher correlated voice descriptors shown in table 5.20 does not lead to any improvements.

The usage of R_d range adaptation variant 2 with the method MSPD2IX appears to provide sufficient means to robustly estimate the shape of the deterministic part of the glottal excitation source, parameterized by the LF shape parameter R_d. It remains to answer the following question for real world applications: Which parameterization of the proposed R_d estimator leads to an estimated R_d contour best reflecting the glottal source contained in the analyzed speech signal?

8.1.2 Speech model for voice transformation

The novel parametric speech framework *PSY* presented in chapter 6 is based on an extended source-filter model. It extends the means of the baseline speech model SVLN of section 3.8.4. *PSY* analyzes and synthesizes the unvoiced signal part with higher quality due to using a DSM approach, explained in section 6.3. It contains as well the more robust energy handling of section 6.4.1.2 and the full-band VTF extraction of section 6.2.3.4. Re-synthesized speech phrases are with this perceived more natural. No smoothing of the estimated voice descriptors is required, as explained in section 6.6.1.

The separated processing of the voiced and unvoiced signal parts and the usage of an extended set of voice descriptors allows for advanced speech transformation and conversion applications. The pulse-based speech excitation uses the estimated R_d contour. The latter is restricted to the GCI time instant. This allows for a computationally

low cost and thus fast synthesis. Additionally it achieves a synthesis of high quality. However, *PSY* is prone to suffer for the time being from R_d and GCI estimation errors at voiced segment borders. The different consecutive algorithmic steps explained in section 6.2.1.3 minimize these estimation drawbacks.

The GMM-based energy prediction of section 6.4.2.3 allows within the context of voice quality transformation or VC applications an energy alteration of the modified signal parts. However, it requires an energy re-scaling to an user-chosen level. Too huge transformations or conversions result for the time being in too huge energy modifications. A re-synthesized speech phrase may have a too low or a too high loudness level.

8.1.3 Voice Conversion

Thesis hypothesis:
The START VC system *coVoC*, introduced in section 7.2, has been extended by the means of the novel speech framework *PSY*, presented in chapter 6. The hypothesis of the presented thesis is that the target speakers voice identity should be better captured the more voice descriptors are converted or transformed towards the target speaker.

Test results on one pair of French mail speakers:
The objective and subjective evaluations of section 7.4 do not exhibit an improvement compared to the sole application of the *coVoC* VC approach. The combination of *coVoC* and *PSY* shows at least partially significant improvements compared to the GMM baseline method of section 7.4.1. Only marginal improvements for the *coVoC* standalone and with its combination with *PSY* compared to the GMM baseline method are reported for the listening test of the subjective evaluation in section 7.4.2. At least some higher improvements are observed for both evaluated systems with the LSF distance measure of the objective evaluation in section 7.4.3. However, the measured objective distance is still far from the theoretically upper bound listed in table 7.4. This indicates that both the *coVoC* standalone version and combined with *PSY* require further improvement to achieve a VC performance which captures well the target speakers voice identity. Additionally, more evaluations on different speaker pairs are necessary to better examine the performance of both systems. This shall evaluate more precisely the hypothesis of the thesis since several works are found in the literature supporting the idea [Kain, 2001, Rentzos et al., 2004, Machado and Queiroz, 2010, Pérez and Bonafonte, 2011].

Semi-automatic behaviour:
Please note that the proposed systems *coVoC* and *PSY* possibly require for the time being some manual corrections. The Voice Conversion system *coVoC* may require for some speech phrases an optimized handling of the phoneme annotation to minimize its sensitivity to the alignment. Different behaviours were observed for different speakers. Certain voices are difficult to convert due to their intrinsic particularities like very expressive speaking gestures. This leads to higher variations in F_0, the spectral envelope sequence, the concatenation of their phonetic content etc. Such high variations have to be reflected in the analyzed voice descriptors. Especially the processing of transients and the handling of borders between a purely unvoiced and a mixed voiced and unvoiced segment is prone to result into artefacts with the combination of *coVoC* and *PSY*. Please note that due to this problematic the F_0 and F_{VU} contours have been manually optimized on voiced borders for the listening tests presented in section 6.6 and 7.4.2. The *PSY* framework may produce unpleasant sounds if a glottal pulse is placed in a purely unvoiced segment.

Both systems *coVoC* and *PSY* are able to achieve a Voice Conversion of good synthesis quality, accompanied with a reasonably high conversion of the speaker identity towards the target speaker. However, due to the partially semi-automatic behaviour of *coVoC* and *PSY* it is for the time being not able to convert automatically a high amount of speech phrases over a whole corpus. It is suitable for the conversion of single speech phrases along with a time-intensive manual parameter adjustment.

Industrial usage:
The VC system *coVoC* has found usage in several projects at IRCAM to convert a voice talent into the speaker identities of famous celebrities. One animation film shown in public cinemas has used the converted voice of the deceased french actor Louis de Funes. Furthermore, voices for a documentary film about the trial of the French Maréchal Pétain and his contributors have been delivered by IRCAM. Further information on the industrial usage of the *coVoC* system can be found online: IRCAM Voice Conversion Showcases [1].

[1] IRCAM Voice Conversion Showcases: http://anasynth.ircam.fr/home/media/covoc-voice-conversion-demo

8.2 Future system improvements - *PSY*

8.2.1 The unvoiced stochastic component

The tests conducted in chapter 6 suggest that the current analysis and synthesis of the unvoiced stochastic component $U(\omega)$ in *PSY* is prone to deteriorate for certain cases the synthesis quality of the proposed speech framework. The combination of the currently separated approaches presented in section 6.3 could result into an improved synthesis quality. Notably the posterior filter below F_{VU} of section 6.3.2.3 could be extended with an algorithm to delete sinusoidal content around voiced borders. Either the threshold onsets approach of section 6.3.2.2 or the 2^{nd} step of the QHM algorithm explained in section 6.3.1.2 could be combined with the below F_{VU} filter.

8.2.2 Perceptual frequency and energy scaling

The human auditory perception does not analyse and abstract any sound on a linear but on a scale reflecting a logarithmic unit of measure [Härmä et al., 2000]. Loudness is not solely affected by sound pressure. The human auditory perception is able to distinct between even minor variations in frequency, amplitude, bandwidth and duration of single sinusoids. An approximation of the underlying mechanism of human auditory perception is proposed in [Union, 2001] by the highly recognized International Telecommunication Union. Their system for Perceptual Evaluation Of Speech Quality (PESQ) includes models for a perceptually-based frequency resolution, a middle-ear filter representing a frequency-dependent attenuation of spectral energy, a loudness-based spectral energy quantification, and a model of which perceptual bands are performed how by the inner ear.

Several proposed frequency and loudness scales approximate how the human perception processes the different rate and intensity levels of sounds continuously changing over time. The Mel scale is a perceptual scale of pitches having equal distance [Stevens et al., 1937]. The Bark scale is a subjective measurement of loudness [Zwicker, 1961]. The Equivalent Rectangular Bandwidth (ERB) models the filter bandwidths of human perception using rectangular band-pass filters [Moore and Glasberg, 1983]. The equal-loudness contour, standardized by International Organization for Standardization (ISO) as ISO 226:2003, measures the Sound Pressure Level (SPL) in dB such that a listener perceives a constant intensity level over the frequency spectrum [Robinson and Dadson, 1956].

The estimated spectral envelopes and most other processing executed in *PSY* is based on a linear frequency scale of the spectrum. Improvements of the analysis and synthesis quality may be achieved by employing one of the listed scales reflecting the human auditory perception. The GMM-based energy prediction which uses an RMS measure on the linear amplitude spectrum may benefit from a perceptual scaling as well. Especially since its current prediction leads to a too huge energy scaling which may produce clipping artefacts for transformations towards a tense voice quality. Still, clipping can be easily prevented by normalizing the waveform to an amplitude level within the range [-1, 1] before synthesis.

8.2.3 Improving the robustness of the glottal source shape parameter estimation

Perceptual scaling of the glottal source parameter space:
The error surface used to span the lattice for the Viterbi algorithm of section 5.4 is scaled on a linear basis. The 2^{nd} adaptation variant of the extended R_d range of section 5.2 exhibits slightly better results than adaptation variant 1 for the test on natural human speech. One reason could be the higher resolution of adaptation variant 2 in lower valued regions of the R_d range being perceptually more relevant [Henrich et al., 2003]. A warping of the glottal source parameter space according to perceptual means as in [van Dinther et al., 2004] to span the Viterbi lattice perceptually may improve the robustness of the R_d estimator. The warping could be applied to the R_d scale, the R waveshape parameters of the R_d regression, or the three parameters OQ, α_m and t_a interpreted along the R_d range.

Viterbi steering:
One drawback realized with the Viterbi steering attempt proposed in section 5.5 and the study of [Huber and Röbel, 2013] is that the overlapping of the phase minimization based error function and the GMM-based prediction results in two probability functions competing against each other. Additionally it causes further difficulty to determine two instead of only one best optimizing weighting factor for the Viterbi algorithm. The two probability surfaces should not be overlapped on the Viterbi lattice to overcome this drawback. The Viterbi steering prediction should be integrated as additional factor together with the phase minimization based error measure to span the Viterbi lattice.

Integration of additional time and spectral domain measures:
Other glottal source estimator algorithms could as well be integrated into the basic error measure spanning up the lattice for Viterbi decoding. The different glottal source estimators should first be evaluated separately on

their ability to add to the robustness of the currently available algorithm based on phase minimization and Viterbi smoothing. The SIM and the PowRd algorithms, explained in sections 3.7.1.2 and respectively 3.7.4, report to be robust against phase distortions. This could aid the basic phase minimization error measure in particular situations where it is more error-prone if e.g. less stable harmonic sinusoids are available, as summarized in section 5.3.4.

Specific LF parameter optimization:
The utilized LF parameter space is restricted to the subspace being defined by the R_d regression curve. Further optimization of the estimated glottal pulse shape can be achieved by implementing a 2^{nd} LF parameter estimation after an initial R_d estimate per frame. The estimated R_d value points to a subspace of the LF model parameter space. The corresponding R waveshape parameters can be derived from the R_d estimate. The R parameters can be varied within restricted borders around that subspace to better match the true glottal source shape contained in the analyzed signal frame. The SIM and the "DyProg-LF" methods discussed in section 3.7.1.2 and respectively 3.7.1.4 employ such means.

Extended VTF model reflecting zeros:
The additional attribution of zeros to the VTF model due to nasalization is already indicated in section 2.1 introducing the generic human voice production system. An Auto-Regressive Moving Average (ARMA) model [Makhoul, 1975] as a mixture of poles and zeros instead of an all-pole model representing the VTF could as well augment the robustness of the R_d estimator. It would capture the zeros from nasalized vowels and consonants which otherwise would have to be attributed to the error measure of the objective function for phase minimization. This introduces a bias into the R_d estimator which possibly leads to erroneous estimations. However, the correct selection of the model order per analysis frame remains to be addressed [Broersen, 2000, de Waele and Broersen, 2003, Broersen and de Waele, 2004].

8.2.4 Pitch-adaptive processing

The current approach in *PSY* with a fixed window size was chosen due to its facilitation of modelling the energy contours. The influence of an adapted window size at each analysis and synthesis step would have to be normalized in the energy model. However, drawbacks occur due to the fixed time-frequency grid set by the STFT analysis and synthesis step and window size [Griffin and Lim, 1983, Griffin and Lim, 1984]. Transient regions or sinusoidal content in otherwise stable segments of especially high-pitched voices may not be reconstructed with sufficient precision. An entropy-based adaptive STFT time-frequency grid as in [Liuni et al., 2013] proved to contribute to the analysis and synthesis quality in audio processing algorithms. An example of a pitch-adaptive processing approach for speech processing proposed in [Kawahara et al., 1999] is integrated into the STRAIGHT framework presented in section 2.5.1.

8.3 Future system improvements - *coVoC*

8.3.1 Abruptness in phoneme concatenation

The proposed *coVoC* system as a soft-clustering approach shares similar properties with the hard-clustering VC approaches using Vector Quantization and Codebook mapping, and the unit concatenation in TTS systems. Concatenation noises like clicks are reported in [Duxans et al., 2006, Duxans, 2006] which result from the assembly of speech units from different segments. In *coVoC*, similar artefacts may occur for certain phoneme combinations implying too strong changes within short-time segments at the concatenation points. The phoneme concatenation algorithm of the basic *coVoC* approach without *PSY* contains therefore means to interpolate the extracted spectral envelope sequence at their respective connection points. An extension of the default time length for interpolation could minimize such artefacts. However, this could corrupt on the other hand the natural spectral envelope continuation for other phoneme combinations where such artefacts are not present. Please note that these artefacts are more prone to occur if *PSY* utilizes the converted spectral envelope sequence from *coVoC* as basis to further advance the VC means by transforming other voice descriptors. The stand-alone version of *coVoC* is less sensitive to abrupt concatenations at phoneme borders. Further investigations are required to analyse which concatenated phoneme pairs are more prone to suffer from artefacts at the concatenation points. Informal observations suggest that fricatives should be treated by specialized means, e.g. a simple copy-and-paste procedure.

8.3.2 Segment preservation

On the one hand, the presented voiced and unvoiced processing approach of *PSY* suffers especially at voiced segment borders from an insufficient separation performance into $U(\omega)$ and $V(\omega)$. On the other hand, certain speech segments contain a special signal type which actually prohibits the employed modelling of $V(\omega)$ by an excitation with glottal pulses. Phonemes such as plosives and fricatives may not incorporate pulses whose shape is covered neither by the utilized R_d nor the complete LF parameter space. Transients are preserved in [Röbel, 2003] by directly copying and pasting the relevant speech segments. Similar means could preserve the natural signal continuation of the mentioned phoneme types. The transient preservation itself could possibly improve the synthesis quality of *PSY* if the sinusoidal detection fails as mentioned in chapter 6.

8.3.3 Transformation of prosodic features

The five dimensions of prosody proposed in [Pfitzinger, 2006] and discussed in section 4.7.3 are only partially reflected in the presented work. Further work has to be conducted to address the prosodic differences between speakers constituting a huge part of a speakers voice identity.

The modelling of intensity as measured by RMS is addressed by the energy model of *PSY* in 6.4.1. However, the GMM-based energy model suffers for the time being from prediction drawbacks and requires improvements. The transformation and conversion of voice quality by means of altering the glottal excitation source is successfully implemented in *PSY*. The transformation or conversion of the fundamental frequency F_0 contour, the timing structure such as syllable durations, and the reflection of further prosodic features remains to be addressed.

A method for unsupervised segmentation like [Obin et al., 2013] can be used to segment speech into syllabic units. Such units describe the prosodic strategies of a speaker: Accentuation and pause. The syllable is generally regarded as the minimum unit for the description of prosody. Syllabic segmentation does not need a-priori linguistic knowledge and is language independent. Methods to stylize local parameter trajectories, e.g. Discrete Cosine Transform (DCT), can be used to represent the prosodic contours on the syllable [Teutenberg et al., 2008, Mertens, 2004]. Tools for the automatic detection of prosodic prominence like accent can be used to help integrate a high-level representation of salient prosodic events [Obin et al., 2008, Obin et al., 2009, Obin, 2011]. Different levels of refinement of the prosodic model and different approaches for the integration of the prosodic features into the statistical model can be investigated to ensure that an efficient training of the model remains feasible.

The combination of syllabic segmentation [Obin et al., 2009, Obin et al., 2013] with the GMM-based acoustic feature classification should allow estimating, representing and mapping the intonation style from the source to the target speaker. This requires an a-priori segmentation into prosodic units (syllable) and the automatic detection of prosodic prominences [Obin et al., 2008, Obin et al., 2009]. Then, the joint prosodic contours are to be mapped on the prosodic units, modelled, and used to transform the speaking style of a speaker. The training of a joint speaking style model for the stylization of prominences can be based on syllable contours. During the transformation, the joint speaking style model can be used in parallel to or combined with the conventional timbre model to transform the speech characteristics to the target speaker. In particular, the long-term speech prosody variations have to be adequately combined with the short-term timbre variations. Different GMM versions representing prosodic, voice quality or VTF features have to be evaluated and compared such that an optimal compromise between dimensionality and quality can be obtained.

8.3.4 Modelling and conversion of more specific voice descriptors

8.3.4.1 Jitter and shimmer

Jitter is a frequency modulation resulting from slight aperiodicities of glottal pulses and is provoked by the acoustic properties of the vibration of the vocal folds [Mertens et al., 2012]. Shimmer results from cycle to cycle variations of the periods amplitude [Ghosh and Narayanan, 2011]. The vocal effect growl comes from simultaneous vibrations of the vocal folds and supra-glottal structures of the larynx [Bonada and Blaauw, 2013]. The vocal folds vibrate half periodically and generate sub-harmonic partials with varying magnitude and phase between different speakers or speaking styles. The sub-harmonics up to approximately 1.5 kHz have a higher magnitude when being closer to a harmonic partial. The sub-harmonics tend to follow the spectral shape of the harmonic sinusoids above this frequency. The applicability to model this behaviour within an analysis-transformation-synthesis scheme for VC has to be investigated.

8.3.4.2 Creaky voice quality

The established modelling of the glottal excitation source using the R_d parameterization of the LF model cannot properly handle speech segments containing creaky voice offsets. The well-established LF model, as most other glottal source shape models, is restricted to the vibration patterns present at the most common phonation types of human voice production. More special vocal fold vibration types like creaky voice require an extension of the glottal model. The creaky voice quality type requires its explicit detection and modelling [Drugman et al., 2013, Kane et al., 2013a].

8.4 Future research ideas

8.4.1 Non-linear system behaviour of the human voice production system

This section presents an open research question, based on an assumption of the author of this thesis. At least to his knowledge no evidence could be found in the literature which strongly supports or rejects the following discussion.

The LTI system theory of linear, zero, minimum, maximum and mixed phase signals is explained in chapter 2.2.1. The VTF behaves like a passive medium in quasi-stationary conditions. The impulse response of the Vocal Tract Filter (VTF) is assumed to be minimum phase. The vocal tract poles are due to this assumption of passivity considered as stable. In contrast, the glottal excitation source is assumed to be mixed phase [Degottex et al., 2011a, Drugman et al., 2011]. For certain cases with an abrupt determination without a return phase present (or with a return phase having a very high frequency), the glottal pulse can be (approximately) modelled exclusively by the maximum phase component. The glottal excitation source may only consist of a maximum phase part if a speaker utters a phrase with a more tense voice quality. The vocal cords behave in this condition as an active medium. This is described in discrete-time signal terms with a maximum phase.

A speaker who articulates speech with a more relaxed voice quality operates the vocal chords partially as an active (maximum phase) and partially as an passive (minimum phase) medium. The active maximum phase part occurs while the glottis is open. The terminology "active part" in correspondence to the open phase is mentioned in [Bozkurt, 2005, p.67]. The passive minimum phase part occurs while the return phase, when the glottis is about to close. If the glottal excitation source contains a return phase, then its signal is assumed to be mixed phase [Degottex, 2010].

Moreover, the VTF may contain a maximum phase part on boundary conditions like word, syllable or phoneme borders. The minimum phase hypothesis may be invalid if the vocal tract changes too fast its physical configuration. The quasi-stationarity condition may not anymore be given and the VTF may be unstable within short-time segments of fast changes. The study of [Zañartu et al., 2013] mentions a violated all-pole assumption if the glottal closure is incomplete. It may act like an active medium and contain thus a maximum phase part.

The open research question is thus if an active physical system can be considered in LTI signal terms as having a maximum phase part. The assumption is based on the view of electronic RLC circuit systems. The active components capacitor and inductor are able to absorb and insert energy like the anti-causal signal part in LTI systems terms. The passive component ohmic resistance constitutes the causal signal part in LTI systems terms. However, the causal / anti-causal LTI signal interpretation of a speech recording is based on a linear model [Doval et al., 2003]. In contrast, the interpretation of the human voice production on a physical basis as an active and / or passive medium, or its corresponding interpretation as an RLC circuit system, is based on a non-linear model.

The major problematic to validate the assumption are the difficulties to measure the airflow present at different levels of the human speech apparatus while a speaker is talking without the restriction of physical measurements in the vocal tract and the throat. The sub-, supra- and trans-glottal pressures have to be measured along with the speech recording and an EGG [Herbst, 2004] measuring the electric currents being simultaneously present in the corresponding physical parts such as muscles. The measurements are required to examine especially the behaviour of the acoustic coupling occurring during the glottal open phase [Titze et al., 2008].

8.4.2 Advanced glottal source estimation and a novel efficient parameterization

The perceptually motivated glottal source model, recently proposed in [Chen et al., 2013a], indicates that the speech community may start research into a novel glottal source model covering more glottal source shapes present in natural speech signals. Such an advanced model may incorporate recent research in view of the acoustic coupling in the human speech apparatus [Titze et al., 2008, Zañartu et al., 2013] to model the formant oscillations present

while the open phase, as already indicated in section 6.2.2.2. Further research could lead to consider the non-linear system behaviour discussed in the preceding section in an estimator and a model for the glottal excitation source. Another demand for a novel glottal source model may arise by modelling the creaky voice quality which may contain pulse shapes not being covered by current source models.

The parameterization of the LF model parameter space by the R_d parameter regression could be improved. The R_d regression curve was deduced from prominent parameter co-variations for Swedish male and female speakers in the 1980's [Gobl, 1988, Karlsson, 1990]. An interesting conjoint research approach could investigate into a similar but different regression curve reflecting additionally speakers of different ages, genders, and many different languages, along with different speaking styles of accent and expressivity. Especially the glottal source analysis on different languages could highlight that the currently language-dependent parameterization by R_d may lead to unfortunate deviations of the R_d-parameterized LF model.

8.4.3 Probabilistic objective evaluation

The objective evaluation of the different VC algorithms in section 7.4.3 constitutes a well-utilized evaluation metric of the VC research community [Kain, 2001, Erro, 2008, Lanchantin and Rodet, 2011]. One drawback of this evaluation metric is the required feature alignment , as discussed in section 7.4.3.2. An interesting alternative of an objective evaluation to measure the performance of different VC systems is to employ speaker verification frameworks. Such algorithms examine by probabilistic means the likelihood of single words or a whole speech phrase to belong to a speakers voice identity. A speaker verification system constitutes the counterpart of a VC system. Contrariwise, a VC system can be in turn utilized to evaluate a speaker verification system [Pellom and Hansen, 1999, Jin et al., 2008].

8.5 Future work applications

8.5.1 Human Voice Avatar Generation

The signal transformation algorithms that establish to transform voice identities into the existing target voice may not only be used for voice imitation. These parameters can also be used to control the transformation of existing voices into new synthetic voice identities or avatar voices. The goal of the voice avatar creation research is to provide means that will allow interpolating and extrapolating example voices that span an acoustic space of artificial voice characters. The acoustic space can be parameterized by means of the voice characterization features that are used for the Voice Conversion application.

The interactive creation of novel voice identities, instead of converting into the perceived timbre of the target voice, requires the parametric modelling of a modular VC approach in order to morph between individual speakers voices. Two possibilities to create new voice characters or artificial speaker avatars can be investigated. Both can additionally be combined. Both should allow the generation of new voices from characteristic voice features providing the user control over the acoustic voice character space.

First, the relations of different voice characters and their individual quality within a voice character space have to be modelled, defined by parameters describing speaker individualities. Voice descriptors reflecting the timbre and the behavioural speaking style of different target speakers can be interpolated and extrapolated in a well defined manner between the given voice identities.

Second, new voice identities can be assembled from different speakers from the parameters or signal parts required to construct the speech signal of one human voice. Applying only partial Voice Conversions leads to voice characteristics that are assemblies of different parts of all voices that are present in the voice character space.

Both approaches require to establish a speech corpus comprising different voice identities. Parameters like the LF glottal source shape parameter R_d and the F_{VU} frequency boundary can be used directly to transform the voice quality. The consideration of prosodic stylization models may require more complex control strategies.

Line Spectral Frequencies (LSF) parameterize the spectral envelope. While the interpolation of spectral envelope parameters has been reported to work well using the LSF envelope representation [Paliwal, 1995], a straightforward interpolation of the LSF parameters in cases of significant differences of the formant structure may result in perceptually less convincing results. The proper separation of glottal source and VTF characteristics is expected to have a beneficial effect on the results obtained by means of interpolation. The spectral contour of the VTF is not interfered by the glottal source characteristics as it is the case for the spectral envelope.

The perceptual space to utilize such a voice character interpolation can be constructed using Multi-Dimensional

Scaling (MDS). It defines a timbre space based on similarity measures of sound characters, proposed by Grey in [Grey, 1977]. MDS reduces the feature dimensionality to retrieve an efficient information projection [Prandoni, 1994]. Further works to establish a sound character space may use Kohonen Self-Organizing Maps [G. DePoli, 1997] or an Euclidean-based distance metric [M. Slaney, 2005].

8.5.2 VC applied to the Singing Voice

A voice synthesizer provides mostly only subtle pronunciation controls to steer individual formants. It does not enable the control of a more opened or closed vowel pronunciation [de Poli et al., 2007]. Recently, the glottal source estimation algorithm, established by the works for this thesis [Huber et al., 2012, Huber and Röbel, 2013] and discussed in chapter 5, has been successfully utilized in [Röbel et al., 2012] to transform the voice quality of a singing voice example. The analysis, transformation and synthesis of the singing voice for different purposes achieved recently good results in [Janer, 2008]. The study of [Bonada, 2008] accomplished astonishing good results. Listeners could not distinct between an artificial singer from the software using a sample database or a real world recording. The latter builds the research basis for the well-known singing voice synthesizer Vocaloid [2] from the Japanese company Yamaha. It operates by means requiring the input of the lyrics in text form, and the melody contour in form of MIDI notes. Its basic system requires a higher learning curve to let the system synthesize a singing voice with an expressivity level being close to a human singer.

The VocaListener technology proposed in [Nakano and Goto, 2009, Nakano and Goto, 2011] mimics pitch, dynamics and the corresponding voice timbre changes of a singer. It separates voice timbre information and phonetic content to construct a timbre space for the modelling of expressive singing styles.

The HMM-based parameter generation system for singing voice synthesis proposed in [Saino et al., 2006] is designed to mimic voice quality and singing style of one dedicated singer. The approach is inspired by TTS synthesis methods. Instead of including speech relevant contextual factors like syllable and accentuation as for TTS it considers the corresponding symbolic representation as musical notation for singing, such as phonemes, tones, duration and position.

The systems *coVoC* for VC and *PSY* for speech analysis, transformation and synthesis could be extended by such means to establish a similar system for the modelling of a singers timbre space and the creation of novel singing characteristics. A basic version of the speech system *PSY* has already been integrated into the singing voice synthesizer of the ANR project ChaNTeR [3] for a general synthesis scheme. It is prone to be easily extended and combined with *coVoC* for the singing voice by similar means , as discussed in chapter 7.

8.5.3 Cross-Lingual VC

The novel VC approach *coVoC* presented in chapter 7 constitutes a promising starting point to conduct cross-lingual VC. The *coVoC* system relieves from having to construct a quasi-parallel corpus from an actual non-parallel corpus, as it is the case for common cross-lingual VC systems. The big advantage is that only the one single phrase of the source speaker selected for conversion is required for the conversion. It is not necessary to provide a complete corpus of the source speaker. Any voice imitator acting as source speaker could utter the desired phrase in any native language such that the *coVoC* system matches from the corpus of the selected target speaker in any foreign language the relevant phonetic content. The voices of famous celebrities acting in big blockbuster movies from Hollywood could for example be converted. The celebrities utter in their original English voice. A translation of the spoken text into any other language is required. The translated text is uttered in the other language by a chosen voice imitator having a similar prosodic speaking style than the celebrity. The phoneme matching of *coVoC* selects the best matching phonetic content from the celebrity corpus.

However, this still requires to construct a speech corpus of the celebrity person which may be cumbersome to compile. Some celebrity voices must be assembled from many different movies. This risks that different recording conditions are mixed into the same corpus. This could deteriorate the sound quality. Additionally, further means have to be implemented to address the problem of the missing subset of phonetic coverage between two languages, as discussed in section 4.9.2. An initial straight-forward solution is to simply select the phoneme being closest in a perceptual distance metric. The closest phoneme type in the utilized X-SAMPA phonetic alphabet could be employed.

[2] Vocaloid singing voice synthesizer: http://www.vocaloid.com/en/
[3] ChaNTeR: anasynth.ircam.fr/home/projects/anr-project-chanter/

8.5.4 Voice Conversion comparison

A currently pending problem in the Voice Conversion research community is the lack of an objective comparison between the different approaches and implementations presented by different researchers. For the time being, most research publications on VC are based on internal test and evaluation schemes established by the corresponding research group. Providing publicly available test sets including parallel and non-parallel corpora, different languages and speakers of different gender and age to let researchers conduct and present VC on pre-defined evaluation measures would provide more detailed insight. It would give a better overview about the currently available VC systems.

The well-known MIREX challenge [4] represents a well established comparison procedure to evaluate different algorithmic approaches. However, VC does not fit thematically to Information Retrieval. Better suited to host a comparative VC competition would be a new dedicated challenge for Voice Conversion, for example at the IEEE Audio and Acoustic Signal Processing Technical Committee [5].

Best prone to compare VC systems is the Blizzard Speech Synthesis Challenge [6], [7]. Established in 2005 [Black and Tokuda, 2005], it invites participants to build a corpus-based synthetic voice on the same data provided. Listening tests are conducted to evaluate the speech quality among the different synthesis systems. Most systems are common TTS synthesizers based on either concatenative speech synthesis using unit selection, or HMM-based speech synthesizers. Since 2007 the organizers explicitly mention as challenge rule that the usage of external data, required for a VC type system, is allowed.

[4] Music Information Retrieval Evaluation eXchange: www.music-ir.org
[5] AASP Challenges: IEEE AASP TC Challenges
[6] Blizzard Challenge: www.synsig.org/index.php/Blizzard_Challenge
[7] Blizzard Challenge: www.festvox.org/blizzard/

Chapter 9

Annex

There is music in Light as in Sound if the Heart be awake and the Ears keen.

Murshid Samuel L. Lewis - Nada Brahma

9.1 List of publications

The following research works have been presented at internationally recognized conferences or published at an internationally recognized journal during the course of the work for this thesis.

9.1.1 International peer-reviewed conference papers

The works presented in section 5.2 concerning the adapted and extended R_d range, as well the improvements of the phase minimization method in section 5.3.3 have been published at the international conference Interspeech (ISCA) in 2012:

```
@inproceedings{Huber12,
      author = {S. Huber and A. Röbel and G. Degottex},
       title = {Glottal source shape parameter estimation using phase minimization variants},
   booktitle = {13th Annual Conference of the International Speech Communication Association (Interspeech ISCA)},
     address = {Portland, Oregon, USA},
        year = {2012},
       pages = {1644-1647},
        issn = {1990-9772},
         url = {http://hal.archives-ouvertes.fr/hal-00773352},
         url = {http://articles.ircam.fr/textes/Huber12a/index.pdf},}
```

The works presented in section 6.4 on voice quality transformation and its evaluation using the semi-automatic R_d transformation presented in section 6.6.5 have been published at the international conference Interspeech (ISCA) in 2015:

@inproceedings{Huber15IS,
author = {S. Huber and A. Röbel},
title = {On glottal source shape parameter transformation
using a novel deterministic and stochastic speech analysis and synthesis system},
booktitle = {16th Annual Conference of the International Speech Communication Association (Interspeech ISCA)},
address = {Dresden, Germany},
year = {2015},
pages = {289-293},
url = {http://hal.archives-ouvertes.fr/hal-01185326},
url = {http://architexte.ircam.fr/textes/Huber15b/index.pdf},}

An extended version of the works presented above with the Interspeech paper includes additionally the GMM-based energy prediction of section 6.4.2.3, and one part of the listening test of section 6.6.4 using the time domain mixing variant for synthesis. This study on voice quality transformation has been published at the international Sound and Music Computing Conference (SMC) in 2015:

@inproceedings{Huber15SMC,
author = {S. Huber and A. Röbel},
title = {Voice quality transformation using an extended source-filter speech model},
booktitle = {12th Sound and Music Computing Conference (SMC)},
address = {Maynooth, Dublin, Ireland},
year = {2015},
pages = {69-76},
url = {http://hal.archives-ouvertes.fr/hal-01185324},
url = {http://architexte.ircam.fr/textes/Huber15a/index.pdf},}

9.1.2 International peer-reviewed journal article

The works presented in the sections 5.4 and 5.5 about Viterbi smoothing and steering have been published at the international journal Computer Speech & Language in 2013:

@article{Huber13,
author = {S. Huber and A. Röbel},
title = {On the use of voice descriptors for glottal source shape parameter estimation},
journal = {Computer Speech & Language},
year = {2013},
volume = {28},
number = {5},
pages = {1170-1194},
issn = {0885-2308},
url = {http://hal.upmc.fr/hal-00865343},
doi = {http://dx.doi.org/10.1016/j.csl.2013.09.006},
url = {http://www.sciencedirect.com/science/article/pii/S0885230813000776},
publisher = {Elsevier Ltd.},
address = {Oxford, UK},}

9.2 Funding acknowledgements

Please note that the speech framework *PSY* has been integrated into the singing voice synthesizer of the ANR project ChaNTeR [1] while a follow-up grant after the Ph.D. studies. The author was financed before by a CIFRE contract as a collaboration between the research institute IRCAM and the company Acapela Group.

9.3 Future information

Please have a look at the personal webpage http://stefan.huber.rocks/phd/. The author may put there further information or sound examples concerning the topic of this doctoral thesis in the future.

[1] ChaNTeR: anasynth.ircam.fr/home/projects/anr-project-chanter/

Acknowledgements

They listen to the preachers of death, and themselves preach afterworlds. Listen rather, my brothers, to the voice of the healthy body, it is a more honest and pure voice. More honestly and purely speaks the healthy body, perfect and squared-off: it speaks of the meaning of the earth.
Thus spoke Zarathustra.

Friedrich Nietzsche - Thus Spoke Zarathustra

Finally my thanks go ...

- to my supervisor Axel Roebel. His wise sense for any signal processing and statistical modelling related task and its difficulties, along with his constant attendance on and availability for any technical question, constitutes a sparkling spiritual sun to receive ones dedicated doctoral tan.

- to my former supervisor Xavier Rodet whose smart suggestions and kind backup to any technical and bureaucratic issues helped to finalize this thesis with sufficient gratification.

- to all the people of the Acapela Group for having given me the support I required to survive the time of these studies: Vincent Pagel, Fabrice Malfrere, Olivier Deroo, Geoffrey Wilfart, Olga Gordeeva, and all other colleagues. But not to forget Marius Cotescu for fruitful political discussions, Paul "WannaBeBelle" for one delicious croissant in the early morning, and François "Rock Star" for giving me the basis to practise daily my weak language skills in all possible cultural facets.

- to my two thesis reviewers Christophe d'Alessandro and Yannis Stylianou for having had the patience to read through the whole of my thesis and thus for spending a lot of their precious lifetime on this.
- to all other jury members for having kindly spend some of their precious lifetime to examine my minor contributions: Antonio Bonafonte, Marius Cotescu (again), Thomas Drugman, and Olivier Adam.

- to Pierre "LaChanteuse" for having given me tons of hints about VC, and for having provided a decent GMM-based Voice Conversion system for comparison purposes.

- to Gilles "Glottotex" for having provided a pretty good glottal source estimation algorithm, while having managed to leave room for further improvements such that I could easily advance on it.

- to my all the time almost always barely starving yet ex-colleagues Sean O'Leary, Sam Perry, and Alex Baker from the barn for having managed to leave me some of the most delicious foods on all those free buffets we liggers went for.

- to my (ex-) colleagues Nicolas "OberRobObin", 'enrik 'ahn (Orianne), Marco the adaptive PingPong-Papa Liuni, Luc "D'Ardaillon", Maria "GodSendABeauty", Lise Regnier, and all other remaining geeks down in the cave of IRCAM's AnaSynth team without sunlight underneath the beautifully colourful Stravinsky Fountain.

- to the god of IRCAM's welcome desk: Bruno "La bella flora" had not just had to welcome me at IRCAM quite some time, but he also had to overcome tons of jokes I placed on him or on life in general. It was always a pleasure to enter IRCAM with a laughter. Mille grazie, bella donna.

- to John Kane and Alain O'Cinneide for their kind help on the journal paper about glottal source estimation.

- to Python: I started to love you already while my first studies in the early years of this millennium and played around with you over the years. But ever since we switched from Matlab to Python, I just loved to rediscover how much love you are eligible for to spent on you.

Thanks go as well to all my friends in general. For all that good, fun and especially Rock 'N' Roll times we have had in the past, and more amazing awesome happenings to experience in the future.

Further special thanks have to go to my parents for all their kind support throughout my whole lifetime. Especially to my daddy who is the only person on earth being able to repeatedly fool and joke on me.

Very special thanks go to Martina Natalie Strohn for her comprehensive support and assistance while all that time, and for all that time, energy and nerves we both spent and lost on it.

Last but not least special thanks go to music in general for getting me sane out of this one. And to life, that thing people call 'god', the cosmos and the universe for producing those lovely patterns we observe with all their amazing facets in nature, vegetation, sounds, and our signals.

Stefan Huber
Paris, France, Europe, the Earth, the Milky Way, the Universe, and beyond

Abbreviation list

$\mathcal{T}$	True Envelope
$\mathcal{T}_{LPC}$	True Envelope Linear Predictive Coding
E_e	Excitation energy at GCI
F_0	Fundamental Frequency
F_m	Maximum Voiced Frequency
F_{VU}	Voiced/Unvoiced Frequency boundary
NAQ	Normalized Amplitude Quotient
coVoC	concatenative Voice Conversion
ACF	Auto-Correlation Function
AIC	Akaike Information Criterion
AMDF	Average Magnitude Difference Function
ANN	Artificial Neural Network
ARMAX	Auto-Regressive Moving Average with eXogenous input
ARX	Auto-Regressive with eXogenous input
BIC	Bayesian Information Criterion
CART	Classification and Regression Tree
CVQ	Constrained Vector Quantization
DAP	Discrete All-Pole
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DFW	Dynamic Frequency Warping
DMS	Dynamic Model Selection
DSM	Deterministic plus Stochastic Model
DSP	Digital Signal Processing
DTFT	Discrete-Time Fourier Transform
DTW	Dynamic Time Warping
GCI	Glottal Closure Instant
GMM	Gaussian Mixture Model
GOI	Glottal Opening Instant
GV	Global Variance
HMM	Hidden Markov Model
HNM	Harmonic plus Noise Model
JND	Just Noticable Difference
LF	Liljencrants and Fant model
LP	Linear Prediction
LPC	Linear Predictive Coding
LSF	Line Spectral Frequencies
LTI	Linear Time-Invariant
MAP	Maxmimum A-Posteriori
MDS	Multi-Dimensional Scaling
MFCC	Mel-Frequency Cepstral Coefficients
MSE	Mean Square Error
PSOLA	Pitch-Synchronous Overlap-and-Add
PSY	Parametric Speech SYnthesis
QHM	Quasi-Harmonic Model
RBF	Radial Basis Function
RMS	Root Mean Square

RMSE	Root-Mean-Square Error
RPS	Relative Phase Shift
SNE	Sinusoidal versus Noise Energy ratio
SPL	Sound Pressure Level
SPT	Spectral Parameter Trajectory
STFT	Short-Time Fourier Transform
STRAIGHT	Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum
SVLN	Separation, Vocal tract, Liljencrants-Fant model, Noise
TTS	Text-To-Speech
VC	Voice Conversion
VQ	Vector Quantization
VTF	Vocal Tract Filter
VTLN	Vocal Tract Length Normalization
ZZT	Zeros of the Z-transform

List of Tables

List of Figures

Bibliography

[Abe, 1991] Abe, M. (1991). A segment-based approach to voice conversion. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 765–768. 36, 38

[Abe et al., 1988] Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H. (1988). Voice conversion through vector quantization. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 655–658. 38, 49

[Abe and Smith, 2005] Abe, M. and Smith, J. (2005). Am/fm rate estimation for time-varying sinusoidal modeling. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 201–204, Philadelphia, PA, USA. 108

[Abe and Smith, 2004] Abe, M. and Smith, J. O. (2004). Am/fm rate estimation and bias correction for time-varying sinusoidal modeling. Technical Report STAN-M-118, Stanford University, Department of Music. 108

[Agus et al., 2010] Agus, T. R., Thorpe, S. J., and Pressnitzer, D. (2010). Rapid formation of robust auditory memories: Insights from noise. *Neuron*, 66(4):610–618. 159

[Alku et al., 1999] Alku, P., , Vintturi, J., and Vilkman, E. (1999). On the linearity of the relationship between the sound pressure level and the negative peak amplitude of the differentiated glottal flow in vowel production. *Speech Communication*, 28(4):269–281. 20

[Alku, 1992] Alku, P. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2-3):109–118. 26, 62, 114

[Alku et al., 2002a] Alku, P., Bäckström, T., and Vilkman, E. (2002a). Normalized amplitude quotient for parametrization of the glottal flow. *Journal of the Acoustical Society of America*, 112(2):701–710. 21, 22

[Alku and Vilkman, 1996] Alku, P. and Vilkman, E. (1996). Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering. *Speech Communication*, 18(2):131–138. 22

[Alku et al., 2002b] Alku, P., Vintturi, J., and Vilkman, E. (2002b). Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation. *Speech Communication*, 38(3-4):321–334. 116

[Allen, 1977] Allen, J. (1977). Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(3):235–238. 123

[Almeida and Silva, 1984] Almeida, L. B. and Silva, F. M. (1984). Variable-frequency synthesis: An improved harmonic coding scheme. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, San Diego, California, USA. 12

[Arslan, 1999] Arslan, L. M. (1999). Speaker transformation algorithm using segmental codebooks (stasc). *Speech Communication*, 28(3):211–226. 36, 38, 44

[Bäckström, 2004] Bäckström, T. (2004). *Linear predictive modelling of speech: Constraints and line spectrum pair decomposition*. PhD thesis, Helsinki University of Technology Laboratory of Acoustics and Audio Signal Processing. 16

[Bäckström and Magi, 2006] Bäckström, T. and Magi, C. (2006). Properties of line spectrum pair polynomials: A review. *Signal processing*, 86(11):3286–3298. 16, 40, 177

[Backstrom et al., 2007] Backstrom, T., Magi, C., and Alku, P. (2007). Minimum Separation of Line Spectral Frequencies. *IEEE Signal Processing Letters*, 14(2):145–147. 37

[Baer et al., 1983] Baer, T., Lofqvist, A., and McGarr, N. (1983). Laryngeal vibrations: A comparison between high-speed filming and glottographic techniques. *Journal of the Acoustical Society of America*, 73(4):1304–1308. 7, 70

[Baken, 1992] Baken, R. J. (1992). Electroglottography. *Journal of Voice*, 6(2):98–110. 7, 70

[Bechet, 2001] Bechet, F. (2001). Lia phon: un systeme complet de phonétisation de textes. *Traitement automatique des langues*, 42(1):47–67. 36

[Bernoulli, 1738] Bernoulli, D. (1738). *Hydrodynamica*. J. H. Decker, Strasbourg. 7

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. 38

[Black and Tokuda, 2005] Black, A. W. and Tokuda, K. (2005). The blizzard challenge 2005: Evaluating corpus-based speech synthesis on common datasets. In *9th European Conference on Speech Communication and Technology (Interspeech ICSLP)*, pages 77–80, Lisbon, Portugal. 193

[Boersma, 1993] Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proc. of the Institute of Phonetic Sciences, University of Amsterdam*, volume 17, pages 97–110. 90

[Bonada, 2000] Bonada, J. (2000). Automatic technique in frequency domain for near-lossless time-scale modification of audio. In *Proc. of International Computer Music Conference (ICMC)*, pages 396–399. 13

[Bonada, 2004] Bonada, J. (2004). High quality voice transformations based on modeling radiated voice pulses in frequency domain. In *Proc. of the 7th International Conference on Digital Audio Effects (DAFx)*, pages 291–295. 119

[Bonada, 2008] Bonada, J. (2008). *Voice Processing and Synthesis by Performance Sampling and Spectral Models*. PhD thesis, Music Technology Group (MTG), Universitat Pompeu Fabra (UPF), Barcelona, Spain. 85, 192

[Bonada and Blaauw, 2013] Bonada, J. and Blaauw, M. (2013). Generation of growl-type voice qualities by spectral morphing. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 6910–6914, Vancouver, Canada. 189

[Bozkurt, 2005] Bozkurt, B. (2005). *New spectral methods for analysis of source/filter characteristics of speech signals*. PhD thesis, Faculté Polytechnique De Mons, Mons/Bergen, Belgium. 9, 27, 96, 190

[Bozkurt et al., 2004] Bozkurt, B., Doval, B., d'Alessandro, C., and Dutoit, T. (2004). Zeros of z-transform (zzt) decomposition of speech for source-tract separation. In *8th International Conference on Spoken Language Processing (Interspeech ICSLP)*, Jeju Island, Korea. 25, 27

[Båvegård and Fant, 1994] Båvegård, M. and Fant, G. (1994). Notes on glottal source interaction ripple. *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, 35(4):063–078. 96

[Broersen, 2000] Broersen, P. (2000). Finite sample criteria for autoregressive order selection. *IEEE Transactions on Signal Processing*, 48(12):3550–3558. 188

[Broersen and de Waele, 2004] Broersen, P. and de Waele, S. (2004). Finite sample properties of arma order selection. *IEEE Transactions on Instrumentation and Measurement*, 53(3):645–651. 188

[Cabral et al., 2008] Cabral, J., Renals, S., Richmond, K., and Yamagishi, J. (2008). Glottal spectral separation for parametric speech synthesis. In *9th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 1829–1832, Brisbane, Australia. 29, 84, 86

[Cabral and Carson-Berndsen, 2013] Cabral, J. P. and Carson-Berndsen, J. (2013). Towards a better representation of the envelope modulation of aspiration noise. In Drugman, T. and Dutoit, T., editors, *Advances in Nonlinear Speech Processing*, volume 7911 of *Lecture Notes in Computer Science*, pages 67–74. Springer Berlin Heidelberg. 14, 84

[Camacho, 2007] Camacho, A. (2007). *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*. PhD thesis, University of Florida, USA. 11

[Camacho and Harris, 2008] Camacho, A. and Harris, J. G. (2008). A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, 124(3):1638–1652. 11

[Campbell and Mokhtari, 2003] Campbell, N. and Mokhtari, P. (2003). Voice quality: the 4th prosodic dimension. In *International Congress of Phonetic Sciences*, pages 2417–2420, Barcelona, Spain. 22, 23

[Catford, 1977] Catford, J. C. (1977). *Fundamental Problems in Phonetics*. Indiana University Press. 77

[Chan and Hui, 1996] Chan, C. and Hui, W. (1996). Wideband re-synthesis of narrowband celp-coded speech using multiband excitation model. In *4th International Conference on Spoken Language Processing (Interspeech ICSLP)*, Philadelphia, PA, USA. 13, 102

[Charpentier and Stella, 1986] Charpentier, F. J. and Stella, M. G. (1986). Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 11, pages 2015–2018, Tokyo, Japan. 37

[Chen et al., 2013a] Chen, G., Garellek, M., Kreiman, J., Gerratt, B. R., and Alwan, A. (2013a). A perceptually and physiologically motivated voice source model. In *15th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 2001–2005, Singapore. 20, 190

[Chen et al., 2013b] Chen, G., Samlan, R. A., Kreiman, J., and Alwan, A. (2013b). Investigating the relationship between glottal area waveform shape and harmonic magnitudes through computational modeling and laryngeal high-speed videoendoscopy. In *14th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 3216–3220, Lyon, France. 63

[Chen et al., 2003] Chen, Y., Chu, M., Chang, E., Liu, J., and Liu, R. (2003). Voice conversion with smoothed gmm and map adaptation. In *8th European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland. 40, 41

[Chetouani et al., 2009] Chetouani, M., Faundez-Zanuy, M., Gas, B., and Zarader, J. (2009). Investigation on lp-residual representations for speaker identification. *Pattern Recognition*, 42(3):487 – 494. 29

[Childers, 1995] Childers, D. (1995). Glottal source modeling for voice conversion. *Speech Communication*, 16(2):127–138. 5, 25, 29, 45

[Childers et al., 1985] Childers, D., Yegnanarayana, B., and Wu, K. (1985). Voice conversion: Factors responsible for quality. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 10, pages 748–751. 38

[Childers et al., 1986] Childers, D. G., Hicks, D. M., Moore, G. P., and Alsaka, Y. A. (1986). A model for vocal fold vibratory motion, contact area, and the electroglottogram. *Journal of the Acoustical Society of America*, 80(5):1309–1320. 70

[Childers and Lee, 1991] Childers, D. G. and Lee, C. K. (1991). Vocal quality factors: analysis, synthesis, and perception. *Journal of the Acoustical Society of America*, 90(5):2394–410. 23, 62, 70, 115

[Childers et al., 1977] Childers, D. G., Skinner, D., and Kemerait, R. (1977). The cepstrum: A guide to processing. *Proceedings of the IEEE*, 65(10):1428–1443. 25, 27

[Childers and Wong, 1994] Childers, D. G. and Wong, C.-F. (1994). Measuring and modeling vocal source-tract interaction. *IEEE Transactions on Biomedical Engineering*, 41(7):663–671. 96

[Ciobanu et al., 2012] Ciobanu, A., Zorila, T., Negrescu, C., and Stanomir, D. (2012). Maximum voiced frequency estimation for voice conversion used in text-to-speech systems. In *19th Int. Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 476–479. 12

[Colton and Conture, 1990] Colton, R. H. and Conture, E. G. (1990). Problems and pitfalls of electroglottography. *Journal of the Voice Foundation*, 4(1):10–24. 70

[Cotescu and Gavat, 2010] Cotescu, M. and Gavat, I. (2010). A study on the influence of prosody and excitation source model on synthetic speech. In *8th Int. Conference on Communications (COMM)*, pages 127–130. 86

[d'Alessandro, 2006] d'Alessandro, C. (2006). Voice source parameters and prosodic analysis. In *Methods in Empirical Prosody Research*, pages 63–87. Edited by Stefan Sudhoff, Denisa Leternovà, Roland Meyer, Sandra Pappert, Petra Augurzky, Ina Mleinek, Nicoale Richter, Johannes Schliesser, Walter de Gruyter, Berlin, New York. 19, 23, 91, 95, 116

[d'Alessandro et al., 2007] d'Alessandro, C., Bozkurt, B., Doval, B., Dutoit, T., Henrich, N., Tuan, V., and Sturmel, N. (2007). Phase-based methods for voice source analysis. In *Advances in Nonlinear Speech Processing*, volume 4885 of *Lecture Notes in Computer Science*, pages 1–27. Springer Berlin Heidelberg. 20, 21, 27, 114

[d'Alessandro et al., 1998] d'Alessandro, C., Darsinos, V., and Yegnanarayana, B. (1998). Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. *IEEE Transactions on Speech and Audio Processing*, 6(1):12–23. 14, 86, 107, 109

[d'Alessandro and Sturmel, 2011] d'Alessandro, C. and Sturmel, N. (2011). Glottal closure instant and voice source analysis using time−scale lines of maximum amplitude. *Sadhana − Academy Proceedings in Engineering Sciences, Indian academy of science, Springer*, 36(5):601–622. 22

[Darch et al., 2007] Darch, J., Milner, B., Almajai, I., and Vaseghi, S. (2007). An investigation into the correlation and prediction of acoustic speech features from mfcc vectors. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages IV–465–IV–468, Honolulu, Hawaii, USA. 63

[de Cheveigne and Kawahara, 2002] de Cheveigne, A. and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111. 11

[de Goetzen et al., 2000] de Goetzen, A., Bernardind, N., and Arfib, D. (2000). Traditional implementations of a phase-vocoder: The tricks of the trade. In *Proc. of the 7th International Conference on Digital Audio Effects (DAFx)*, pages 1–7, Verona, Italy. 27

[de Poli et al., 2007] de Poli, G., Avanzini, F., Klapuri, A., and Serra, X. (2007). *SMC Roadmap v1.0*. The S2S2 Consortium. 192

[de Waele and Broersen, 2003] de Waele, S. and Broersen, P. (2003). Order selection for vector autoregressive models. *IEEE Transactions on Signal Processing*, 51(2):427–433. 16, 188

[Degottex, 2010] Degottex, G. (2010). *Glottal source and vocal tract separation*. PhD thesis, Institut de Recherche et Coordination Acoustique/Musique (IRCAM), Université de Pierre et Marie Curie (UPMC), Université de Paris 6, Paris, France. 9, 10, 20, 21, 22, 25, 28, 30, 31, 57, 58, 59, 60, 64, 82, 85, 87, 114, 129, 190

[Degottex et al., 2013] Degottex, G., Lanchantin, P., Röbel, A., and Rodet, X. (2013). Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis. *Speech Communication*, 55(2):278–294. 11, 31, 32, 84, 85, 94, 101, 114

[Degottex et al., 2009a] Degottex, G., Röbel, A., and Rodet, X. (2009a). Glottal closure instant detection from a glottal shape estimate. In *Proc. 13th International Conference on Speech and Computer (SPECOM)*, pages 226–231, St. Petersburg, Russia. 22, 28

[Degottex et al., 2009b] Degottex, G., Röbel, A., and Rodet, X. (2009b). Shape parameter estimate for a glottal model without time position. In *13th International Conference on Speech and Computer (SPECOM)*, pages 345–349, St. Petersburg, Russia. 28

[Degottex et al., 2010] Degottex, G., Röbel, A., and Rodet, X. (2010). Joint estimate of shape and time-synchronization of a glottal source model by phase flatness. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5058–5061, Dallas, USA. 22, 28, 37, 86, 90

[Degottex et al., 2011a] Degottex, G., Röbel, A., and Rodet, X. (2011a). Phase minimization for glottal model estimation. *IEEE Transactions on Acoustics, Speech, and Language Processing*, 19(5):1080–1090. 28, 31, 45, 57, 58, 59, 60, 61, 67, 73, 82, 87, 91, 190

[Degottex et al., 2011b] Degottex, G., Röbel, A., and Rodet, X. (2011b). Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5128–5131, Prague, Czech Republic. 25, 31, 114

[Degottex and Stylianou, 2013] Degottex, G. and Stylianou, Y. (2013). Analysis and synthesis of speech using an adaptive full-band harmonic model. *IEEE Transactions on Acoustics, Speech, and Language Processing*, 21(10):2085–2095. 14, 85

[del Pozo and Young, 2008] del Pozo, A. and Young, S. (2008). The linear transformation of lf glottal waveforms for voice conversion. In *9th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 1457–1460, Brisbane, Australia. 5, 29, 33, 45

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38. 38

[Desai et al., 2010] Desai, S., Black, A. W., Yegnanarayana, B., and Prahallad, K. (2010). Spectral mapping using artificial neural networks for voice conversion. *IEEE Transactions on Audio, Speech & Language Processing*, 18(5):954–964. 4

[Desai et al., 2009] Desai, S., Raghavendra, E. V., Yegnanarayana, B., Black, A. W., and Prahallad, K. (2009). Voice conversion using artificial neural networks. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3893–3896, Taipei, Taiwan. 4

[Ding et al., 1995] Ding, W., Kasuya, H., and Adachi, S. (1995). Simultaneous estimation of vocal tract and voice source parameters based on an arx model. *IEICE Transactions*, pages 738–743. 30

[Dirac, 1930] Dirac, P. A. M. (1930). *The Principles of Quantum Mechanics*. Oxford University Press, New York, USA. 9

[Dognin et al., 2009] Dognin, P., Goel, V., Hershey, J., and Olsen, P. (2009). A fast, accurate approximation to log likelihood of gaussian mixture models. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3817–3820, Taipei, Taiwan. 117

[Doval and d'Alessandro, 1999] Doval, B. and d'Alessandro, C. (1999). The spectrum of glottal flow models. Technical Report LIMSI 99-07, Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (Orsay), Orsay. 20, 21

[Doval et al., 1997] Doval, B., d'Alessandro, C., and Diard, B. (1997). Spectral methods for voice source parameters estimation. In *5th European Conference on Speech Communication and Technology (Eurospeech)*. 27

[Doval et al., 2003] Doval, B., d'Alessandro, C., and Henrich, N. (2003). The voice source as a causal/anticausal linear filter. In *Proceedings of Voice Quality: Functions, analysis and synthesis (Voqual'03)*, Geneva, Switzerland. 9, 20, 27, 96, 190

[Doval et al., 2006] Doval, B., d'Alessandro, C., and Henrich, N. (2006). The spectrum of glottal flow models. *Acta Acustica united with Acustica*, 92(6):1026–1046. 19, 20, 21, 55, 62, 63

[Dromey et al., 1992] Dromey, C., Stathopoulos, E. T., and Sapienza, C. M. (1992). Glottal airflow and electroglottographic measures of vocal function at multiple intensities. *Journal of Voice*, 6(1):44 – 54. 70

[Drugman, 2011] Drugman, T. (2011). *Advances in Glottal Analysis and its Applications*. PhD thesis, Faculté Polytechnique De Mons, Mons/Bergen, Belgium. 8

[Drugman et al., 2011] Drugman, T., B.Bozkurt, and T.Dutoit (2011). Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation. *Speech Communication*, 53(6):855–866. 9, 19, 23, 25, 27, 190

[Drugman et al., 2009a] Drugman, T., Bozkurt, B., and Dutoit, T. (2009a). Complex cepstrum-based decomposition of speech for glottal source estimation. In *10th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 116–119. 9, 27, 45

[Drugman et al., 2008] Drugman, T., Dubuisson, T., Moinet, A., d'Alessandro, N., and Dutoit, T. (2008). Glottal source estimation robustness - a comparison of sensitivity of voice source estimation techniques. In *IEEE International Conference on Signal Processing and Multimedia Applications (SIGMAP)*, pages 202–207, Porto, Portugal. 62, 67, 68, 91, 137

[Drugman and Dutoit, 2010] Drugman, T. and Dutoit, T. (2010). Chirp complex cepstrum-based decomposition for asynchronous glottal analysis. In *11th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 657–660. 27

[Drugman and Dutoit, 2012] Drugman, T. and Dutoit, T. (2012). The deterministic plus stochastic model of the residual signal and its applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):968–981. 14, 15, 29, 84, 85, 108, 109

[Drugman et al., 2012a] Drugman, T., Kane, J., and Gobl, C. (2012a). Modeling the creaky excitation for parametric speech synthesis. In *13th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, Portland, Oregon, USA. 77

[Drugman et al., 2014] Drugman, T., Kane, J., and Gobl, C. (2014). Data-driven detection and analysis of the patterns of creaky voice. *Computer, Speech, & Language*, 28(5):1233–1253. 25, 139

[Drugman et al., 2013] Drugman, T., Kane, J., Raitio, T., and Gobl, C. (2013). Prediction of creaky voice from contextual factors. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7967–7971, Vancouver, Canada. 87, 131, 154, 190

[Drugman et al., 2009b] Drugman, T., Moinet, A., Dutoit, T., and Wilfart, G. (2009b). Using a pitch-synchronous residual codebook for hybrid hmm/frame selection speech synthesis. In *Proc.IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3793–3796, Taipei, Taiwan. 165

[Drugman and Stylianou, 2014] Drugman, T. and Stylianou, Y. (2014). Maximum voiced frequency estimation: Exploiting amplitude and phase spectra. *Signal Processing Letters, IEEE*, 21(10):1230–1234. 12

[Drugman et al., 2012b] Drugman, T., Thomas, M., Gudnason, J., Naylor, P., and Dutoit, T. (2012b). Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):994–1006. 22, 26, 77

[Dutoit et al., 2007] Dutoit, T., Holzapfel, A., Jottrand, M., Moinet, A., Perez, J., and Stylianou, Y. (2007). Towards a voice conversion system based on frame selection. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 513–516, Honolulu, Hawaii, USA. 48, 183

[Dutoit and Leich, 1993] Dutoit, T. and Leich, H. (1993). An analysis of the performances of the mbe model when used in the context of a text-to-speech system. In *3rd European Conference on Speech Communication and Technology (Eurospeech)*, pages 531–534, Berlin, Germany. 13

[Duxans, 2006] Duxans, H. (2006). *Voice Conversion applied to Text-to-Speech Systems*. PhD thesis, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. 4, 50, 188

[Duxans and Bonafonte, 2006] Duxans, H. and Bonafonte, A. (2006). Residual conversion versus prediction on voice morphing systems. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 85–88, Toulouse, France. 44

[Duxans et al., 2004] Duxans, H., Bonafonte, A., Kain, A., and van Santen, J. P. H. (2004). Including dynamic and phonetic information in voice conversion systems. In *8th International Conference on Spoken Language Processing (Interspeech ICSLP)*, Jeju Island, Korea. 41, 42

[Duxans et al., 2006] Duxans, H., Erro, D., Pérez, J., Diego, F., Bonafonte, A., and Moreno, A. (2006). Voice conversion of non-aligned data using unit selection. In *TC-Star Workshop on Speech to Speech Translation*, Barcelona, Spain. 50, 188

[El-Jaroudi and Makhoul, 1991] El-Jaroudi, A. and Makhoul, J. (1991). Discrete all-pole modeling. *IEEE Transaction on Signal Processing*, 39:411–423. 17

[Elenius and Blomberg, 2010] Elenius, D. and Blomberg, M. (2010). Dynamic vocal tract length normalization in speech recognition. In *Proceedings from Fonetik*, pages 29–34. 46

[Ellis, 2005] Ellis, D. P. W. (2005). Plp and rasta (and mfcc, and inversion) in matlab. Online web resource. 62

[En-Najjary et al., 2003] En-Najjary, T., Rosec, O., and Chonavel, T. (2003). The impact of spectral modeling on the performance of a voice conversion system. *Proc. of the GRETSI Symposium on Signal and Image Processing*. 16

[Erro, 2008] Erro, D. (September 2008). *Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models*. PhD thesis, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, Barcelona, Spain. 14, 16, 17, 39, 40, 41, 44, 46, 47, 50, 85, 176, 180, 191

[Erro and Moreno, 2007] Erro, D. and Moreno, A. (2007). Weighted frequency warping for voice conversion. In *8th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 1965–1968, Antwerp, Belgium. 4, 14, 46

[Erro et al., 2010a] Erro, D., Moreno, A., and Bonafonte, A. (2010a). Inca algorithm for training voice conversion systems from nonparallel corpora. *IEEE Transactions on Audio, Speech & Language Processing*, 18(5):944–953. 50

[Erro et al., 2010b] Erro, D., Moreno, A., and Bonafonte, A. (2010b). Voice conversion based on weighted frequency warping. *Trans. Audio, Speech and Lang. Proc.*, 18(5):922–931. 14, 47

[Erro et al., 2008] Erro, D., Moreno, A., and Polyakova, T. (2008). On combining statistical methods and frequency warping for high-quality voice conversion. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4665–4668, Caesars Palace, Las Vegas, Nevada, USA. 38, 46

[Every and Szymanski, 2004] Every, M. and Szymanski, J. (2004). A spectral-filtering approach to music signal separation. In *Proc. of the 7th International Conference on Digital Audio Effects (DAFx)*, pages 197–200, Naples, Italy. 12

[Every and Szymanski, 2005] Every, M. and Szymanski, J. (2005). Separation of overlapping impulsive sounds by bandwise noise interpolation. In *Proc. of the 8th International Conference on Digital Audio Effects (DAFx)*, pages 20–22, Madrid, Spain. 12

[Every and Szymanski, 2006] Every, M. and Szymanski, J. (2006). Separation of synchronous pitched notes by spectral filtering of harmonics. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1845–1856. 12

[Every, 2006] Every, M. R. (2006). *Separation of Musical Sources and Structure from Single-Channel Polyphonic Recordings*. PhD thesis, University of York, UK. 12

[Fant, 1960] Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton, The Hague, Netherlands. 7, 8, 18

[Fant, 1961] Fant, G. (1961). A new anti-resonance circuit for inverse filtering. *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, 2(4):001–006. 26

[Fant, 1979] Fant, G. (1979). Glottal source and excitation analysis. *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, 20(1):085–107. 20

[Fant, 1981] Fant, G. (1981). The source filter concept in voice production. *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, 22(1):021–037. 7, 8, 16, 18, 29, 86

[Fant, 1995] Fant, G. (1995). The lf-model revisited. transformation and frequency domain analysis. *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, 36(2-3):119–156. 19, 20, 21, 22, 31, 53, 54, 55, 57, 62, 96, 101, 114, 115

[Fant, 1997] Fant, G. (1997). The voice source in connected speech. *Speech Communication*, 22(2-3):125–139. 19, 21, 22, 53, 54, 55, 63, 90, 101

[Fant and Ananthapadmanabha, 1982] Fant, G. and Ananthapadmanabha, T. (1982). Truncation and superposition. *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, 23(2-3):001–017. 96

[Fant and Kruckenberg, 1996] Fant, G. and Kruckenberg, A. (1996). Voice source properties of the speech code. *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, 37(4):045–056. 115, 116

[Fant and Kruckenberg, 2005] Fant, G. and Kruckenberg, A. (2005). Covariation of subglottal pressure, f0 and intensity. In *9th European Conference on Speech Communication and Technology (Eurospeech ICSLP)*, pages 1061–1064, Lisbon, Portugal. 81, 115, 116

[Fant and Kruckenberg, 2007] Fant, G. and Kruckenberg, A. (2007). Co-variation of acoustic parameters in prosody. *Proceedings of Fonetik, Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, 50(1):1–4. 118

[Fant et al., 1994] Fant, G., Kruckenberg, A., Liljencrants, J., and Båvegård, M. (1994). Voice source parameters in continuous speech. transformation of lf-parameters. In *3rd International Conference on Spoken Language Processing (ICSLP)*, pages 1451–1454, Yokohama, Japan. 20, 21, 22, 53, 54, 55, 62

[Fant et al., 2000] Fant, G., Kruckenberg, A., Liljencrants, J., and Båvegård, M. (2000). Acoustic-phonetic studies of prominence in swedish. *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, 41(2-3):001–052. 20, 21, 116, 118

[Fant and Liljencrants, 1994] Fant, G. and Liljencrants, J. (1994). Data reduction of lf voice source parameters. 21, 53

[Fant et al., 1985] Fant, G., Liljencrants, J., and Lin, Q.-G. (1985). A four-parameter model of glottal flow. *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, 26(4):001–013. 20, 21, 24, 53, 57, 87, 96, 114

[Fant and Lin, 1987] Fant, G. and Lin, Q.-G. (1987). Glottal source - vocal tract acoustic interaction. *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, 28(1):013–027. 96

[Felps et al., 2009] Felps, D., Bortfeld, H., and Gutierrez-Osuna, R. (2009). Foreign accent conversion in computer assisted pronunciation training. *Speech Communication*, 51(10):920 – 932. Spoken Language Technology for Education Spoken Language. 44

[Flanagan and Golden, 1966] Flanagan, J. L. and Golden, R. M. (1966). Phase vocoder. *The Bell System Technical Journal*, 45(9):1493–1509. 15

[Forney, 1973] Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278. 60

[Fröhlich et al., 2001] Fröhlich, M., Michaelis, D., and Strube, H. (2001). Sim - simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. *Journal of The Acoustical Society of America*, 110(1):479–88. 26, 64, 70

[Fujisaki et al., 1981] Fujisaki, H., Fant, G., and Liljencrants, J. (1981). Dynamic characteristics of voice fundamental frequency in speech and singing. acoustical analysis and physiological interpretations. *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, 22(1):001–020. 76

[Fujisaki and Ljungqvist, 1986] Fujisaki, H. and Ljungqvist, M. (1986). Proposal and evaluation of models for the glottal source waveform. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 31, pages 2.1–2.4, Atlanta, Georgia, USA. 20

[Fulop, 2003] Fulop, S. A. (2003). An accurate means for measuring formants. Technical report, Dept. of Linguistics, Dept. of Computer Science, The University of Chicago, USA. 47

[G. DePoli, 1997] G. DePoli, P. P. (1997). Sonological models for timbre characterization. *Journal of New Music Research*, 26(2). 192

[Geman et al., 1992] Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Comput.*, 4(1):1–58. 117

[Gerratt and Kreiman, 2001] Gerratt, B. and Kreiman, J. (2001). Measuring voice quality with speech synthesis. *Journal of The Acoustical Society of America*, 5(110):2560 – 2566. 24

[Ghosh and Narayanan, 2011] Ghosh, P. K. and Narayanan, S. S. (2011). Joint source-filter optimization for robust glottal source estimation in the presence of shimmer and jitter. *Speech Communication*, 53(1):98 – 109. 23, 189

[Gillett, 2003] Gillett, B. (2003). Transforming voice quality and intonation. Master thesis, University of Edinburgh. 114

[Gobl, 1988] Gobl, C. (1988). Voice source dynamics in connected speech. *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, 29(1):123–159. 21, 22, 191

[Gobl, 2003] Gobl, C. (2003). *The voice source in speech communication*. PhD thesis, Department of Speech, Music and Hearing, KTH, Stockholm. 20, 21, 22

[Gobl and Chasaide, 1992] Gobl, C. and Chasaide, A. N. (1992). Acoustic characteristics of voice quality. *Speech Communication*, 11(4-5):481–490. 23, 62

[Godoy, 2011] Godoy, E. (2011). *Spectral Envelope Transformation for High-Quality Voice Conversion*. PhD thesis, Orange Labs, Télécom Bretagne, France; Université de Rennes, Université européenne de Bretagne, France. 37, 39, 41, 43, 46, 51

[Godoy et al., 2009] Godoy, E., Rosec, O., and Chonavel, T. (2009). Alleviating the one-to-many mapping problem in voice conversion with context-dependent modeling. In *10th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 1627–1630, Brighton, United Kingdom. 40, 41, 49, 163

[Godoy et al., 2010a] Godoy, E., Rosec, O., and Chonavel, T. (2010a). On the use of spectral peak parameters in voice conversion. In *11th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, Makuhari, Chiba, Japan. 46

[Godoy et al., 2010b] Godoy, E., Rosec, O., and Chonavel, T. (2010b). On transforming spectral peaks in voice conversion. In *The Seventh ISCA Tutorial and Research Workshop on Speech Synthesis*, pages 68–73, Kyoto, Japan. 46

[Godoy et al., 2011] Godoy, E., Rosec, O., and Chonavel, T. (2011). Spectral envelope transformation using dfw and amplitude scaling for voice conversion with parallel or nonparallel corpora. In *12th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 673–676, Florence, Italy. 47

[Godoy et al., 2012] Godoy, E., Rosec, O., and Chonavel, T. (2012). Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora. *IEEE Transactions on Audio, Speech & Language Processing*, 20(4):1313–1323. 4, 37, 46, 47

[Gordon and Ladefoged, 2001] Gordon, M. and Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29(4):383 – 406. 24

[Gramming et al., 1988] Gramming, P., Sundberg, J., Ternström, S., Leanderson, R., and Perkins, W. H. (1988). Relationship between changes in voice pitch and loudness. *Journal of Voice*, 2(2):118 – 126. 118

[Grey, 1977] Grey, J. (1977). Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, 61(5). 192

[Griffin and Lim, 1983] Griffin, D. and Lim, J. (1983). Signal estimation from modified short-time fourier transform. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 8, pages 804–807, Boston, Massachusetts, USA. 13, 123, 188

[Griffin et al., 1988] Griffin, D. W., , and Lim, J. S. (1988). Multiband excitation vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(8):1223–1235. 13

[Griffin, 1987] Griffin, D. W. (1987). *Multi-Band Excitation Vocoder*. Rle technical report no. 524, Massachusetts Institute of Technology, Cambridge, MA, USA. 13

[Griffin and Lim, 1984] Griffin, D. W. and Lim, J. S. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, & Signal Processing*, 32(2):236–243. 13, 32, 123, 131, 188

[Gu and Tsai, 2014] Gu, H.-Y. and Tsai, S.-F. (2014). Improving segmental gmm based voice conversion method with target frame selection. In *The 9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 483–487, Singapore. 48

[Gu and Tsai, 2015] Gu, H.-Y. and Tsai, S.-F. (2015). A voice conversion method combining segmental GMM mapping with target frame selection. *Journal of Information Science and Engineering*, 31(2):609–626. 48

[Gudnason et al., 2012] Gudnason, J., Thomas, M. R. P., Ellis, D. P. W., and Naylor, P. A. (2012). Data-driven voice source waveform analysis and synthesis. *Speech Communication*, 54(2):199–211. 63

[Hanson et al., 1990] Hanson, D., Gerratt, B., and Berke, G. (1990). Frequency, intensity, and target matching effects on photoglottographic measures of open quotient and speed quotient. *Journal of Speech and Hearing Research*, 33:45–50. 62, 70

[Hanson, 1995] Hanson, H. (1995). Individual variations in glottal characteristics of female speakers. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 772–775, Detroit, Michigan, USA. 62

[Hanzlíček and Matoušek, 2008] Hanzlíček, Z. and Matoušek, J. (2008). On using warping function for lsfs transformation in a voice conversion system. In *Proc. 9th International Conference on Signal Processing (ICSP)*, volume 3, Beijing, China. 16

[Härmä et al., 2000] Härmä, A., Karjalainen, M., Savioja, L., Välimäki, V., Laine, U. K., and Huopaniemi, J. (2000). Frequency-warped signal processing for audio applications. *Journal of the Audio Engineering Society*, 48(11):1011–1031. 187

[Hasegawa-Johnson, 2000] Hasegawa-Johnson, M. (2000). Line spectral frequencies are poles and zeros of the glottal driving-point impedance of a discrete matched-impedance vocal tract model. *Journal of the Acoustic Society of America*, 108(1):457–460. 177

[Helander and Nurminen, 2007] Helander, E. and Nurminen, J. (2007). A novel method for prosody prediction in voice conversion. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages IV–509–IV–512, Honolulu, Hawaii, USA. 46

[Helander et al., 2007] Helander, E., Nurminen, J., and Gabbouj, M. (2007). Analysis of lsf frame selection in voice conversion. In *Proc. 12th International Conference on Speech and Computer (SPECOM)*, pages 651–656, Moscow, Russia. 16, 48, 183

[Helander et al., 2008a] Helander, E., Nurminen, J., and Gabbouj, M. (2008a). Lsf mapping for voice conversion with very small training sets. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4669–4672, Caesars Palace, Las Vegas, Nevada, USA. 16

[Helander et al., 2008b] Helander, E., Schwarz, J., Nurminen, J., Silen, H., and Gabbouj, M. (2008b). On the impact of alignment on voice conversion performance. In *9th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 1453–1456, Brisbane, Australia. 37, 39

[Helander et al., 2010] Helander, E., Virtanen, T., Nurminen, J., and Gabbouj, M. (2010). Voice conversion using partial least squares regression. *IEEE Transactions on Audio, Speech & Language Processing*, 18(5):912–921. 38, 39, 40

[Henrich et al., 1999] Henrich, N., d'Alessandro, C., and Doval, B. (1999). Glottal open quotient estimation using linear prediction. In *Proc. International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, pages 12–17. 55

[Henrich et al., 2001] Henrich, N., d'Alessandro, C., and Doval, B. (2001). Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data. In *7th European Conference on Speech Communication and Technology (Eurospeech ICSLP), 2nd Interspeech Event*, pages 47–50, Aalborg, Denmark. 62, 63

[Henrich et al., 2004] Henrich, N., d'Alessandro, C., Doval, B., and Castellengo, M. (2004). On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *Journal of the Acoustical Society of America*, 115(3):1321–1332. 61, 70

[Henrich et al., 2003] Henrich, N., d'Alessandro, C., Doval, B., Castellengo, M., Sundin, G., and D.Ambroise (2003). Just noticeable differences of open quotient and asymmetry coefficient in singing voice. *Journal of Voice*, 17(4):481–494. 24, 120, 149, 159, 187

[Herbst, 2004] Herbst, C. (2004). Evaluation of various methods to calculate the egg contact quotient. Diploma thesis in music acoustics, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden. 70, 190

[Hezard et al., 2012] Hezard, T., Hélie, T., Caussé, R., and Doval, B. (2012). Analysis-synthesis of vocal sounds based on a voice production model driven by the glottal area. In *Acoustics 2012*, Nantes, France. 20

[Hezard et al., 2013] Hezard, T., Hélie, T., and Doval, B. (2013). A source-filter separation algorithm for voiced sounds based on an exact anticausal/causal pole decomposition for the class of periodic signals. In *14th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 54–58, Lyon, France. 28

[Hill and Lewicki, 2007] Hill, T. and Lewicki, P. (2007). Statistics: Methods and applications. StatSoft, Tulsa, Oklahoma, USA. 75

[Hsia et al., 2007] Hsia, C.-C., Wu, C.-H., and Wu, J.-Q. (2007). Conversion function clustering and selection using linguistic and spectral information for emotional voice conversion. *IEEE Transactions on Computers*, 56(9):1245–1254. 40

[Huber and Röbel, 2013] Huber, S. and Röbel, A. (2013). On the use of voice descriptors for glottal source shape parameter estimation. *Computer, Speech, & Language*, 28(5):1170 – 1194. 11, 30, 45, 53, 54, 55, 62, 65, 74, 81, 83, 85, 90, 91, 96, 101, 136, 137, 187, 192

[Huber and Röbel, 2015a] Huber, S. and Röbel, A. (2015a). On glottal source shape parameter transformation using a novel deterministic and stochastic speech analysis and synthesis system. In *16th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 289–293, Dresden, Germany. 86

[Huber and Röbel, 2015b] Huber, S. and Röbel, A. (2015b). Voice quality transformation using an extended source-filter speech model. In *12th Sound and Music Computing Conference (SMC)*, pages 69–76, Maynooth, Dublin, Ireland. 86

[Huber et al., 2012] Huber, S., Röbel, A., and Degottex, G. (2012). Glottal source shape parameter estimation using phase minimization variants. In *13th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, 1990-9772, pages 1644–1647, Portland, Oregon, USA. 53, 54, 55, 57, 58, 59, 60, 67, 68, 69, 70, 82, 90, 91, 101, 192

[Hunt and Black, 1996] Hunt, A. J. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 373–376, Atlanta, GA. 47, 50, 165

[Ishi et al., 2008] Ishi, C., Sakakibara, K.-I., Ishiguro, H., and Hagita, N. (2008). A method for automatic detection of vocal fry. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):47–56. 25

[Itakura and Saito, 1968] Itakura, F. and Saito, S. (1968). Analysis synthesis telephony based upon the maximum likelihood method. In *Proc. 6th of the International Congress on Acoustics*, pages 17–20. 29

[Janer, 2008] Janer, J. (2008). *Singing-driven interfaces for sound synthesizers*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain. 192

[Jin et al., 2008] Jin, Q., Toth, A. R., Black, A. W., and Schultz, T. (2008). Is voice transformation a threat to speaker identification? In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4845–4848. 191

[Kafentzis et al., 2012] Kafentzis, G. P., Pantazis, Y., Rosec, O., and Stylianou, Y. (2012). An extension of the adaptive quasi-harmonic model. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4605–4608, Vancouver, Canada. 14

[Kain, 2001] Kain, A. (2001). *High resolution Voice Transformation*. PhD thesis, OGI School of Science and Engineering at Oregon Health and Science University, Portland, Oregon. 4, 39, 41, 44, 49, 171, 183, 186, 191

[Kain and Macon, 2001] Kain, A. and Macon, M. W. (2001). Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 813–816, Salt Palace Convention Center, Lake City, Utah, USA. 44, 177

[Kane, 2013] Kane, J. (2013). *Tools for analysing the voice - Developments in glottal source and voice quality analysis*. PhD thesis, Phonetics and Speech Laboratory, Trinity College, Dublin, Ireland. 8, 27

[Kane et al., 2013a] Kane, J., Drugman, T., and Gobl, C. (2013a). Improved automatic detection of creak. *Computer, Speech, & Language*, 27(4):1028–1047. 25, 87, 190

[Kane and Gobl, 2013a] Kane, J. and Gobl, C. (2013a). Automating manual user strategies for precise voice source analysis. *Speech Communication*, 55(3):397–414. 26, 64, 77

[Kane and Gobl, 2013b] Kane, J. and Gobl, C. (2013b). Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1170–1179. 83, 90

[Kane et al., 2013b] Kane, J., Scherer, S., Aylett, M., Morency, L.-P., and Gobl, C. (2013b). Speaker and language independent voice quality classification applied to unlabeled corpora of expressive speech. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada. 23

[Kane et al., 2012] Kane, J., Yanushevskaya, I., Chasaide, A. N., and Gobl, C. (2012). Exploiting time and frequency domain measures for precise voice source parameterisation. In *Proc. of the 6th International Conference on Speech Prosody*, Shanghai, China. 26, 83

[Karlsson, 1990] Karlsson, I. (1990). Voice source dynamics for female speakers. In *First International Conference on Spoken Language Processing (ICSLP)*, pages 69–72, Kobe, Japan. 21, 191

[Kawahara, 1997] Kawahara, H. (1997). Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1303–1306, Munich, Bavaria, Germany. 15

[Kawahara et al., 2001] Kawahara, H., Estill, J., and Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In *MAVEBA*, pages 59–64, Firentze, Italy. 13, 15

[Kawahara et al., 1999] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3–4):187–207. 15, 188

[Kawahara et al., 2008] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. (2008). Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3933–3936, Caesars Palace, Las Vegas, Nevada, USA. 15, 30

[Keating and Esposito, 2006] Keating, P. A. and Esposito, C. M. (2006). Linguistic voice quality. In *Proc. of the 11th Australasian International Conference on Speech Science and Technology*, pages 85–91, University of Auckland, Auckland, New Zealand. 23, 24, 62

[Keogh and Ratanamahatana, 2005] Keogh, E. and Ratanamahatana, C. A. (2005). Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3):358–386. 37

[Keogh and Pazzani, 2001] Keogh, E. J. and Pazzani, M. J. (2001). Derivative dynamic time warping. In *First SIAM International Conference on Data Mining (SDM)*, volume 1, pages 5–7. 37

[King, 2010] King, S. (2010). A beginners' guide to statistical parametric speech synthesis. Technical report, The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK. 165

[Klatt and Klatt, 1990] Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of Americax*, 87(2):820–857. 20, 23, 24, 26, 86

[Kominek and Black, 2004] Kominek, J. and Black, A. W. (2004). The cmu arctic speech databases. In *Proc. of the 5th ISCA Speech Synthesis Workshop*, pages 223–224. 61, 64, 70, 85, 91, 96, 128

[Kortekaas and Kohlrausch, 1997] Kortekaas, R. and Kohlrausch, A. (1997). Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform manipulation technique using single-formant stimuli. *The Journal of the Acoustical Society of America (JASA)*, 101(4). 30

[Kreiman et al., 1992a] Kreiman, J., Gerratt, B., Precoda, K., and Berke, G. (1992a). Individual differences in voice quality perception. *Journal of Speech and Hearing Research*, 35(3):512–520. 23

[Kreiman et al., 1992b] Kreiman, J., Gerratt, B., Precoda, K., and Berke, G. (1992b). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research*, 33:103–115. 23

[Kreiman et al., 2012a] Kreiman, J., Iseli, M., Neubauer, J., Shue, Y.-L., Gerratt, B. R., and Alwan, A. (2012a). The relationship between open quotient and h1(*)-h2(*). *The Journal of the Acoustical Society of America (JASA)*, 124(4):2495. 63

[Kreiman et al., 2012b] Kreiman, J., Shue, Y.-L., Chen, G., Iseli, M., Gerratt, B. R., Neubauer, J., and Alwan, A. (2012b). Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *The Journal of the Acoustical Society of America*, 132(4):2625–32. 63

[Ladefoged, 1996] Ladefoged, P. (1996). *Elements of Acoustic Phonetics, 2nd edition*. The University of Chicago Press, Ltd. London. 7

[Lanchantin, 2007] Lanchantin, P. (2007). ircamalign: Système d'étiquetage et d'alignement de signaux de parole. Technical report, Sound Analysis/Synthesis Team, Institut de Recherche et Coordination Acoustique/Musique (IRCAM). 36, 164

[Lanchantin et al., 2010] Lanchantin, P., Degottex, G., and Rodet, X. (2010). A hmm-based speech synthesis system using a new glottal source and vocal-tract separation method. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4630–4633, Dallas, USA. 31, 165

[Lanchantin et al., 2011a] Lanchantin, P., Farner, S., Veaux, C., Degottex, G., Obin, N., Beller, G., Villavicencio, F., Huber, S., Peeters, G., Röbel, A., and Rodet, X. (2011a). Vivos voco: A survey of recent research on voice transformations at ircam. In *Proc. of the 14th Int. Conference on Digital Audio Effects (DAFx)*, pages 277–285, Paris, France. 45, 119

[Lanchantin et al., 2008] Lanchantin, P., Morris, A. C., Rodet, X., and Veaux, C. (2008). Automatic phoneme segmentation with relaxed textual constraints. In *Proc. of the 6th Int. Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). 36, 128, 164, 171, 174

[Lanchantin et al., 2011b] Lanchantin, P., Obin, N., and Rodet, X. (2011b). Extended conditional gmm and covariance matrix correction for real-time spectral voice conversion. Technical report, Sound Analysis/Synthesis Team, Institut de Recherche et Coordination Acoustique/Musique (IRCAM). 42, 44, 171, 176

[Lanchantin and Rodet, 2010] Lanchantin, P. and Rodet, X. (2010). Dynamic model selection for spectral voice conversion. In *11th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 1720–1723, Makuhari, Chiba, Japan. 41, 63, 117, 171

[Lanchantin and Rodet, 2011] Lanchantin, P. and Rodet, X. (2011). Objective evaluation of the dynamic model selection method for spectral voice conversion. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5132–5135, Prague, Czech Republic. 41, 63, 117, 171, 191

[Laroche, 2003] Laroche, J. (2003). Frequency-domain techniques for high-quality voice modification. In *Proc. of the 6th Int. Conference on Digital Audio Effects (DAFx)*, pages 322–328. 15

[Laroche and Dolson, 1999] Laroche, J. and Dolson, M. (1999). New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 15

[Laroche et al., 1993a] Laroche, J., Stylianou, Y., and Moulines, E. (1993a). Hnm: A simple efficient harmonic plus noise model for speech. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, page 169–172. 13

[Laroche et al., 1993b] Laroche, J., Stylianou, Y., and Moulines, E. (1993b). Hns: Speech modification based on a harmonic + noise model. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 550–553, Minneapolis, Minnesota, USA. 13

[Laver, 1994] Laver, J. (1994). *Principles of Phonetics*. Cambridge Textbooks in Linguistics. Cambridge University Press. 23

[Laver, 1968] Laver, J. D. M. (1968). Voice quality and indexical information. *International Journal of Language and Communication Disorders*, 3(1):43–54. 23, 62, 81

[Laver, 1980] Laver, J. D. M. (1980). *The Phonetic Description of Voice Quality*, volume 31. Cambridge University Press. 23, 62

[Li et al., 2010] Li, Y., Zhang, L., and Ding, H. (2010). An algorithm for chinese voice conversion based on phonetic gaussian mixture model. In *3rd Int. Congress on Image and Signal Processing (CISP)*, volume 7, pages 3490–3494. 41

[Liuni and Röbel, 2013] Liuni, M. and Röbel, A. (2013). Phase vocoder and beyond. *Musica/Tecnologia*, 7(0):73–89. 15

[Liuni et al., 2013] Liuni, M., Röbel, A., Matusiak, E., Romito, M., and Rodet, X. (2013). Automatic adaptation of the time-frequency resolution for sound analysis and re-synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):959–970. 188

[Liénard and Barras, 2013] Liénard, J.-S. and Barras, C. (2013). Fine-grain voice strength estimation from vowel spectral cues. In *14th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 128–132, Lyon, France. 23, 45, 62, 115, 118

[Lu and Smith, 1999] Lu, H.-L. and Smith, J. O. (1999). Joint estimation of vocal tract filter and glottal source waveform via convex optimization. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 26

[Lu and Smith, 2001] Lu, H.-L. and Smith, J. O. (2001). Estimating glottal aspiration noise via wavelet thresholding and best-basis thresholding. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 11–14. 86

[M. Slaney, 2005] M. Slaney, H. Terawasa, J. B. (2005). Perceptual distance in timbre space. In *Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display*. 192

[Machado and Queiroz, 2010] Machado, A. F. and Queiroz, M. (2010). Voice conversion - a critical survey. In *Proc. of the 7th Sound and Music Computing (SMC) Conference*, pages 291–298, Barcelona, Spain. 4, 5, 39, 50, 51, 186

[Maddieson and Hess, 1987] Maddieson, I. and Hess, S. (1987). The effects of f0 of the linguistic use of phonation type. In *11th International Congress of Phonetic Sciences (ICPhS)*, volume 67, page 112–118. UCLA Working Papers in Phonetics. 62

[Maeda, 1982] Maeda, S. (1982). A digital simulation method of the vocal-tract system. *Speech Communication*, 1(3-4):199–229. 11, 67

[Maia and Stylianou, 2014] Maia, R. and Stylianou, Y. (2014). Complex cepstrum factorization for statistical parametric synthesis. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3839–3843, Florence, Italy. 86

[Makhoul, 1975] Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580. 16, 26, 29, 87, 188

[Marasek, 1997] Marasek, K. (1997). Egg & voice quality. Online web resource. 7, 25, 70

[Markel and Gray, 1976] Markel, J. and Gray, A. (1976). *Linear Prediction of Speech*, chapter 12, pages 278–284. Communication and cybernetics. Springer Verlag, New York. 31, 57, 87

[Markov and S.Nakagawa, 1999] Markov, K. P. and S.Nakagawa (1999). Integrating pitch and lpc-residual information with lpc-cepstrum for text-independent speaker recognition. *Journal of the Acoustical Society of Japan*, 20(4):281–291. 29

[Mathews et al., 1961] Mathews, M. V., Miller, J. E., and David, J. E. E. (1961). Pitch synchronous analysis of voiced sounds. *Journal of the Acoustical Society of America*, 33(2):179–186. 87

[McAulay and Quatieri, 1986] McAulay, R. J. and Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, volume 34, pages 744–754. 12

[Mertens et al., 2012] Mertens, C., Grenez, F., and Schoentgen, J. (2012). Analysis of vocal tremor and jitter by empirical mode decomposition of glottal cycle length time series. In *13th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 1634–1637, Portland, Oregon, USA. 189

[Mertens, 2004] Mertens, P. (2004). The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In *Proc. of Speech Prosody*, pages 549–552, Nara, Japan. B. Bel & I. Marlien. 189

[Moore and Glasberg, 1983] Moore, B. C. J. and Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, 74(3):750–753. 187

[Mouchtaris et al., 2007] Mouchtaris, A., Agiomyrgiannakis, Y., and Stylianou, Y. (2007). Conditional vector quantization for voice conversion. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 505–508. 37

[Murphy, 2001] Murphy, P. J. (2001). Spectral tilt as a perturbation-free measurement of noise levels in voice signals. In *7th European Conference on Speech Communication and Technology (Eurospeech ICSLP), 2nd Interspeech Event*, pages 1495–1498, Aalborg, Denmark. 62

[Murthy and Yegnanarayana, 1991] Murthy, H. A. and Yegnanarayana, B. (1991). Speech processing using group delay functions. *Elsevier Signal Processing*, 22:259–267. Department of Computer Science and Engineering, Indian Institute of Technology, Madras-600 036, India. 8

[Murthy and Yegnanarayana, 1999] Murthy, P. S. and Yegnanarayana, B. (1999). Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals. *IEEE Transactions on Speech and Audio Processing*, 7(6):609–619. 19

[Nakano and Goto, 2009] Nakano, T. and Goto, M. (2009). Vocalistener: A singing-to-singing synthesis system based on iterative parameter estimation. In *6th Sound and Music Computing Conference (SMC)*, pages 343–348, Porto, Portugal. 192

[Nakano and Goto, 2011] Nakano, T. and Goto, M. (2011). Vocalistener2: A singing synthesis system able to mimic a user's singing in terms of voice timbre changes as well as pitch and dynamics. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 453–456, Prague, Czech Republic. 192

[Naylor et al., 2007] Naylor, P. A., Kounoudes, A., Gudnason, J., and Brookes, D. M. (2007). Estimation of glottal closure instants in voiced speech using the dypsa algorithm. *IEEE Trans. Audio, Speech, and Language Processing*, 15(1):34–43. 22

[Nordstrom et al., 2008] Nordstrom, K. I., Tzanetakis, G., and Driessen, P. F. (2008). Transforming perceived vocal effort and breathiness using adaptive pre-emphasis linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1087–1096. 25, 29

[Nose and Kobayashi, 2011] Nose, T. and Kobayashi, T. (2011). Speaker-independent hmm-based voice conversion using quantized fundamental frequency. *Speech Communication*, pages 973–985. 5, 45

[Obin, 2011] Obin, N. (2011). *MeLos: Analysis and modelling of speech prosody and speaking style*. PhD thesis, Institut de Recherche et Coordination Acoustique/Musique (IRCAM), Université de Pierre et Marie Curie (UPMC), Université de Paris 6, Paris, France. 189

[Obin, 2012] Obin, N. (2012). Cries and whispers - classification of vocal effort in expressive speech. In *13th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, Portland, Oregon, USA. 77

[Obin et al., 2013] Obin, N., Lamare, F., and Roebel, A. (2013). Syll-o-matic: An adaptive time-frequency representation for the automatic segmentation of speech into syllables. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, British Columbia, Canada. 189

[Obin et al., 2008] Obin, N., Rodet, X., and Lacheret-Dujour, A. (2008). French prominence: A probabilistic framework. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3993–3996, Las Vegas, NV, USA. 189

[Obin et al., 2009] Obin, N., Rodet, X., and Lacheret-Dujour, A. (2009). A syllable-based prominence model based on discriminant analysis and context-dependency. In *13th Proc. of International Conference on Speech and Computer (SPECOM)*, pages 97–100, St.-Petersburg, Russia. 189

[Obin et al., 2014] Obin, N., Roebel, A., and Bachman, G. (2014). On automatic voice casting for expressive speech: Speaker recognition vs. speech classification. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 950–954, Florence, Italy. 51

[Ó Cinnéide, 2012] Ó Cinnéide, A. (2012). *Phase-Distortion-Robust Voice-Source Analysis*. PhD thesis, Dublin Institute of Technology, Dublin, Ireland. 29, 64, 76

[Ó Cinnéide et al., 2011] Ó Cinnéide, A., Dorran, D., Gainza, M., and Coyle, E. (2011). A frequency domain approach to arx-lf voiced speech parameterization and synthesis. In *12th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 57–60, Florence, Italy. 60

[Oppenheim et al., 1976] Oppenheim, A., Kopec, G., and Tribolet, J. (1976). Signal analysis by homomorphic prediction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):327–332. 8, 9, 10

[Oppenheim and Schafer, 1975] Oppenheim, A. and Schafer, R. (1975). *Digital Signal Processing*. Prentice-Hall. 8, 9, 130

[Oppenheim and Schafer, 1989] Oppenheim, A. and Schafer, R. (1989). *Discrete-Time Signal Processing*. Prentice-Hall. 130

[Oppenheim et al., 1999] Oppenheim, A., Schafer, R., and Buck, J. (1999). *Discrete-time Signal Processing (2nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. 8, 10

[Oppenheim et al., 1968] Oppenheim, A., Schafer, R., and Jr., T. S. (1968). Nonlinear filtering of multiplied and convolved signals. *Proceedings of the IEEE*, 56(8):1264–1291. 9, 27

[Oppenheim, 1978] Oppenheim, A. V., editor (1978). *Applications of Digital Signal Processing*, chapter Digital Processing of Speech, pages 117–168. Prentice-Hall. 8, 9, 10

[Orlikoff, 1991] Orlikoff, R. F. (1991). Assessment of the dynamics of vocal fold contact from the electroglottogram: Data from normal male subjects. *Journal of Speech, Language, and Hearing Research*, 34(5):1066–1072. 70

[Paalanen et al., 2005] Paalanen, P., Kamarainen, J.-K., Ilonen, J., and Kälviäinen, H. (2005). Feature representation and discrimination based on Gaussian mixture model probability densities - practices and algorithms. Technical Report 95, Department of Information Technology, Lappeenranta University of Technology. 117

[Paliwal, 1995] Paliwal, K. (1995). Interpolation properties of linear prediction parametric representations. In *4th European Conference on Speech Communication and Technology (Eurospeech)*, pages 1029–1032, Madrid, Spain. 16, 191

[Paliwal and Atal, 1993] Paliwal, K. K. and Atal, B. S. (1993). Efficient vector quantization of lpc parameters at 24 bits/frame. *IEEE Transactions on Speech and Audio Processing*, 1(1):3–14. 48

[Panchapagesan and Alwan, 2009] Panchapagesan, S. and Alwan, A. (2009). Frequency warping for vtln and speaker adaptation by linear transformation of standard mfcc. *Computer Speech & Language*, 23(1):42 – 64. 46

[Pantazis et al., 2008] Pantazis, Y., Rosec, O., and Stylianou, Y. (2008). On the properties of a time-varying quasi-harmonic model of speech. In *9th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 1044–1047, Brisbane, Australia. 14, 106, 107

[Pantazis et al., 2010a] Pantazis, Y., Rosec, O., and Stylianou, Y. (2010a). Analysis/synthesis of speech based on an adaptive quasi-harmonic plus noise model. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4246–4249, Dallas, Texas, USA. 14

[Pantazis et al., 2010b] Pantazis, Y., Rosec, O., and Stylianou, Y. (2010b). On the robustness of the quasi-harmonic model of speech. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4210–4213, Dallas, Texas, USA. 14

[Pantazis et al., 2011] Pantazis, Y., Rosec, O., and Stylianou, Y. (2011). Adaptive am-fm signal decomposition with application to speech analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2):290–300. 14, 108

[Pantazis and Stylianou, 2008] Pantazis, Y. and Stylianou, Y. (2008). Improving the modeling of the noise part in the harmonic plus noise model of speech. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4609–4612, Caesars Palace, Las Vegas, Nevada, USA. 14, 108, 109

[Pearson, 1900] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175. 71, 118, 142

[Peeters, 2004] Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, Sound Analysis / Synthesis Team, IRCAM. 81, 83

[Pellom and Hansen, 1999] Pellom, B. L. and Hansen, J. H. L. (1999). An experimental study of speaker verification sensitivity to computer voice-altered imposters. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 837–840. 191

[Percybrooks and Moore, 2007] Percybrooks, W. S. and Moore, E. (2007). New algorithm for lpc residual estimation from lsf vectors for a voice conversion system. In *8th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, volume 1138C, pages 1977–1980. 16, 44

[Percybrooks and Moore, 2012] Percybrooks, W. S. and Moore, E. (2012). A hmm approach to residual estimation for high resolution voice conversion. In *13th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 90–93, Portland, Oregon, USA. 44

[Pérez and Bonafonte, 2005] Pérez, J. and Bonafonte, A. (2005). Automatic voice-source parameterization of natural speech. In *9th European Conference on Speech Communication and Technology (Eurospeech ICSLP)*, pages 1065–1068, Lisbon, Portugal. 26

[Pérez and Bonafonte, 2009] Pérez, J. and Bonafonte, A. (2009). Towards robust glottal source modeling. In *10th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 68–71. 26

[Pérez and Bonafonte, 2011] Pérez, J. and Bonafonte, A. (2011). Adding glottal source information to intra-lingual voice conversion. In *12th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 2773–2776. 29, 45, 186

[Pfitzinger, 2006] Pfitzinger, H. R. (2006). Five dimensions of prosody: Intensity, intonation, timing, voice quality, and degree of reduction. In *Speech Prosody Abstract Book, Studientexte zur Sprachkommunikation, Band 40*, pages 6–9, Dresden, Germany. TUDpress. 23, 45, 113, 189

[Pilkington et al., 2011] Pilkington, N., Zen, H., and Gales, M. J. F. (2011). Gaussian process experts for voice conversion. In *12th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 2761–2764, Florence, Italy. 43

[Prandoni, 1994] Prandoni, P. (1994). An analysis-based timbre space. 192

[Prasanna et al., 2006] Prasanna, S. M., Gupta, C. S., and Yegnanarayana, B. (2006). Extraction of speaker-specific excitation information from linear prediction residual of speech. *Speech Communication*, 48(10):1243 – 1261. 29

[Puckette, 1995] Puckette, M. (1995). Phase-locked vocoder. In *Proc. of the IEEE Conference on Applications of Signal Processing to Audio and Acoustics*. 15

[Qiao et al., 2011] Qiao, Y., Tong, T., and Minematsu, N. (2011). A study on bag of gaussian model with application to voice conversion. In *12th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 657–660, Florence, Italy. 40

[Quatieri, 2002] Quatieri, T. F. (2002). *Discrete-time speech signal processing: Principles and Practice*. Prentice-Hall, Uper Saddle River, NJ, USA. 7, 9

[Rabiner and Schafer, 1978] Rabiner, L. and Schafer, R. (1978). *Digital Processing of Speech Signals*. Prentice Hall Signal Processing Series. 7

[Rajput et al., 2012] Rajput, P., Khanna, R., Lehana, P., and Singh, J. B. (2012). Effect of dynamic time warping on alignment of phrases and phonemes. *International Journal on Natural Language Computing (IJNLC)*, 1(3). 37, 39

[Rao and Yegnanarayana, 2006] Rao, K. and Yegnanarayana, B. (2006). Voice conversion by prosody and vocal tract modification. In *Proc. of the 9th International Conference on Information Technology (ICIT'06)*, pages 111–116. 5, 39, 45

[Rao and Yegnanarayana, 2007] Rao, K. S. and Yegnanarayana, B. (2007). Modeling durations of syllables using neural networks. *Computer Speech & Language*, 21(2):282–295. 46

[Rentzos et al., 2003] Rentzos, D., Vaseghi, S., Turajlic, E., Yan, Q., and Ho, C.-H. (2003). Transformation of speaker characteristics for voice conversion. *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'03)*, pages 706–711. 5, 45

[Rentzos et al., 2004] Rentzos, D., Vaseghi, S., Yan, Q., and Ho, C.-H. (2004). Voice conversion through transformation of spectral and intonation features. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 21–24, Montreal, Quebec, Canada. 5, 39, 45, 186

[Röbel, 2003] Röbel, A. (2003). A new approach to transient processing in the phase vocoder. In *6th International Conference on Digital Audio Effects (DAFx)*, pages 344–349, London, United Kingdom. 189

[Röbel, 2008] Röbel, A. (2008). Frequency-slope estimation and its application to parameter estimation for non-stationary sinusoids. *Computer Music Journal*, 32-2:68–79. 108

[Röbel, 2010a] Röbel, A. (2010a). Between physics and perception: Signal models for high level audio processing. In *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx)*, Graz, Austria. 13

[Röbel, 2010b] Röbel, A. (2010b). A shape-invariant phase vocoder for speech transformation. In *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx)*, pages 208–305. 12, 15, 37

[Röbel, 2010c] Röbel, A. (2010c). Shape-invariant speech transformation with the phase vocoder. In *11th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 2146–2149. 11, 37, 90

[Röbel et al., 2012] Röbel, A., Huber, S., Rodet, X., and Degottex, G. (2012). Analysis and modification of excitation source characteristics for singing voice synthesis. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5381–5384, Kyoto, Japan. 192

[Röbel et al., 2007] Röbel, A., Villavicencio, F., and Rodet, X. (2007). On cepstral and all-pole based spectral envelope modelling with unknown model order. *Elsevier, Pattern Recognition Letters*, 28(11):1343 – 1350. 17, 45, 94

[Röbel et al., 2004] Röbel, A., Zivanovic, M., and Rodet, X. (2004). Signal decomposition by means of classification of spectral peaks. In *Proc. of the International Computer Music Conference (ICMS)*, pages 446–449, Miami, Florida, USA. 106, 108, 110, 111

[Robinson and Dadson, 1956] Robinson, D. W. and Dadson, R. S. (1956). A re-determination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, 7(5):166. 187

[Roebel, 2015] Roebel, A. (2015). covoc: Voice conversion by means of concatenation of spectral envelopes. patent application. In preparation. 162, 163

[Rosenberg, 1971] Rosenberg, A. (1971). Effect of glottal pulse shape on the quality of natural vowels. *The Journal of the Acoustical Society of America*, 49(1A):583–590. 20

[Rothenberg, 1972] Rothenberg, M. (1972). A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *The Journal of the Acoustical Society of America*, 53(6):1632–1645. 26, 62

[Saino et al., 2006] Saino, K., Zen, H., Nankaku, Y., Lee, A., and Tokuda, K. (2006). An hmm-based singing voice synthesis system. In *9th International Conference on Spoken Language Processing (Interspeech ICSLP)*. 192

[Sapienza et al., 1998] Sapienza, C. M., Stathopoulos, E. T., and Dromey, C. (1998). Approximations of open quotient and speed quotient from glottal airflow and egg waveforms: Effects of measurement criteria and sound pressure level. *Journal of Voice*, 12:31–43. 70

[Saratxaga et al., 2009] Saratxaga, I., Hernáez, I., Erro, D., Navas, E., and Sanchez, J. (2009). Simple representation of signal phase for harmonic speech models. *Electronics Letters*, 45(7):381–383. 13

[Saratxaga et al., 2010] Saratxaga, I., Hernáez, I., Navas, E., Sainz, I., Luengo, I., Sanchez, J., Odriozola, I., and Erro, D. (2010). Ahotransf: A tool for multiband excitation based speech analysis and modification. In *Proc. of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA). 13

[Scherer et al., 2012] Scherer, S., Kane, J., Gobl, C., and Schwenker, F. (2012). Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. *Computer, Speech, & Language*, 27(1):263–287. 62

[Schwarz, 1998] Schwarz, D. (1998). Spectral envelopes in sound analysis and synthesis. Fakultät informatik, universität stuttgart, germany, IRCAM, Paris. 15

[Schwarz and Rodet, 1999] Schwarz, D. and Rodet, X. (1999). Spectral envelope estimation, representation, and morphing for sound analysis, transformation, and synthesis. In *ICMC: International Computer Music Conference*, Peking, China. 15

[Serrà, 2011] Serrà, J. (2011). *Identification of versions of the same musical composition by processing audio descriptions*. PhD thesis, Universitat Pompeu Fabra, Barcelona. 6

[Serra, 1989] Serra, X. (1989). *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University. 13, 85

[Serra, 1997] Serra, X. (1997). *Musical Sound Modeling with Sinusoids plus Noise*, pages 91–122. Swets and Zeitlinger. 13

[Serra and Smith, 1990] Serra, X. and Smith, J. (1990). Spectral modeling synthesis - a sound analysis/synthesis based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14:12–24. SMS. 13

[Silén et al., 2013] Silén, H., Nurminen, J., Helander, E., and Gabbouj, M. (2013). Voice conversion for non-parallel datasets using dynamic kernel partial least squares regression. In *14th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 373–377, Lyon, France. 50

[Smith, 2007] Smith, J. O. (2007). *Introduction to Digital Filters with Audio Applications*. http://ccrma.stanford.edu/~jos/filters/. online book. 8, 9, 10

[Smith, ssed] Smith, J. O. (accessed (date accessed)). *Mathematics of the Discrete Fourier Transform (DFT)*. http://ccrma.stanford.edu/~jos/mdft/. online book, 2007 edition. 8

[Snell and Milinazzo, 1993] Snell, R. C. and Milinazzo, F. (1993). Formant location from lpc analysis data. *IEEE Transactions on Speech and Audio Processing*, 1(2):129–134. 16

[Sorin et al., 2015] Sorin, A., Shechtman, S., and Pollet, V. (2015). Coherent modification of pitch and energy for expressive prosody implantation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 118

[Stevens et al., 1937] Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190. 187

[Strik, 1998] Strik, H. (1998). Automatic parameterization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. *Journal of the Acoustical Society of America*, 103:2659–2669. 28, 76, 114

[Strik et al., 1993] Strik, H., Cranen, B., and Boves, L. (1993). Fitting a lf-model to inverse filter signals. In *3rd European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, pages 103–106, Berlin, Germany. 26, 28, 76, 114

[Strik and Boves, 1992] Strik, S. and Boves, L. (1992). On the relation between voice source parameters and prosodic features in connected speech. *Speech Communication*, 11(2-3):167–174. 114

[Sturmel et al., 2009] Sturmel, N., d'Alessandro, C., and Rigaud, F. (2009). Glottal closure instant detection using lines of maximum amplitudes (loma) of the wavelet transform. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4517–4520, Taipei, Taiwan. 22

[Stylianou, 1996] Stylianou, Y. (1996). *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, Ecole Nationale Supèrieure des Télécommunications, Paris, France. 4, 13, 14, 30, 37, 38

[Stylianou, 2001] Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 9(1):21–29. 12, 13, 37, 39, 57

[Stylianou, 2009] Stylianou, Y. (2009). Voice transformation: A survey. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3585–3588, Washington, DC, USA. IEEE Computer Society. 4, 5, 25, 37

[Stylianou and Cappe, 1998] Stylianou, Y. and Cappe, O. (1998). A system for voice conversion based on probabilistic classification and a harmonic plus noise model. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 281–284, Seattle, Washington, USA. 38

[Stylianou et al., 1998] Stylianou, Y., Cappe, O., and Moulines, E. (1998). Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech & Audio Processing*, 6(2):131–142. 39, 40, 183

[Stylianou et al., 1995] Stylianou, Y., Laroche, J., and Moulines, E. (1995). High quality speech modification based on a harmonic+noise model. In *4th European Conference on Speech Communication and Technology*, Madrid, Spain. 13

[Sündermann et al., 2005a] Sündermann, D., Bonafonte, A., and Ney, H. (2005a). A study on residual prediction techniques for voice conversion. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 13–16, Philadelphia, PA, USA. 44

[Sündermann et al., 2005b] Sündermann, D., Hoege, H., Bonafonte, A., Ney, H., and Black, A. (2005b). Residual prediction based on unit selection. In *Proc. of the ASRU 2005, 9th IEEE Automatic Speech Recognition and Understanding Workshop*, San Juan, Puerto Rico. 44

[Sündermann et al., 2006a] Sündermann, D., Hoege, H., Bonafonte, A., Ney, H., Black, A., and Narayanan, S. (2006a). Text-independent voice conversion based on unit selection. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France. 48

[Sündermann et al., 2005c] Sündermann, D., Hoege, H., Bonafonte, A., Ney, H., and Duxans, H. (2005c). Residual prediction. In *Proc. of the ISSPIT 2005, 5th IEEE International Symposium on Signal Processing and Information Technology*, Athens, Greece. 44

[Sündermann et al., 2006b] Sündermann, D., Hoege, H., and Fingscheidt, T. (2006b). Breaking a paradox: Applying vtln to residuals. In *Proc. of the ITG 2006, 7th Symposium on Speech Communication of the Information Technology Society*, Kiel, Germany. 46, 85

[Sündermann and Ney, 2003] Sündermann, D. and Ney, H. (2003). Vtln-based voice conversion. In *Proc. of the ISSPIT 2003, 3rd IEEE International Symposium on Signal Processing and Information Technology*, Darmstadt, Germany. 46

[Sündermann et al., 2003] Sündermann, D., Ney, H., and Hoege, H. (2003). Vtln-based cross-language voice conversion. In *In Proc. of the ASRU 2003, 8th IEEE Automatic Speech Recognition and Understanding Workshop*, Virgin Islands, USA. 46

[Taylor, 2009] Taylor, P. A. (2009). *Text-to-Speech Synthesis*. Cambridge University Press, Cambridge, UK. 50

[Teutenberg et al., 2008] Teutenberg, J., Watson, C., and Riddle, P. (2008). Modelling and synthesising f0 contours with the discrete cosine transform. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 3973–3976, Las Vegas, U.S.A. 189

[Thati et al., 2012] Thati, S. A., Bollepalli, B., Bhaskararao, P., and Yegnanarayana, B. (2012). Analysis of breathy voice based on excitation characteristics of speech production. In *International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5. 23

[Thomas et al., 2009a] Thomas, M. R. P., Gudnason, J., and Naylor, P. A. (2009a). Data-driven voice source waveform modelling. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3965–3968, Taipei, Taiwan. 63

[Thomas et al., 2009b] Thomas, M. R. P., Gudnason, J., and Naylor, P. A. (2009b). Detection of glottal closing and opening instants using an improved dypsa framework. In *Proc. European Signal Processing Conference*. 22

[Thomas and Naylor, 2009] Thomas, M. R. P. and Naylor, P. A. (2009). The sigma algorithm: A glottal activity detector for electroglottographic signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8):1557–1566. 61

[Tian et al., 2014] Tian, X., Wu, Z., Lee, S. W., and Chng, E. (2014). Correlation-based frequency warping for voice conversion. In *The 9th International Symposium on Chinese Spoken Language Processing*, pages 211–215, Singapore. 47

[Titze, 1984] Titze, I. (1984). Parameterization of the glottal area, glottal flow, and vocal fold contact area. *The Journal of the Acoustical Society of America*, 75(2):570–580. 20

[Titze, 1994] Titze, I. R. (1994). *Principles of Voice Production*. Prentice Hall, Englewood Cliffs, N.J. 77

[Titze, 2004] Titze, I. R. (2004). Theory of glottal airflow and source-filter interaction in speaking and singing. *Acta Acustica united with Acustica 90*, 90:641–648. 96

[Titze et al., 2008] Titze, I. R., Riede, T., and Popolo, P. (2008). Nonlinear source–filter coupling in phonation: Vocal exercises. *Journal of the Acoustical Society of America*, 123:1902–1915. 18, 96, 190

[Titze and Talkin, 1979] Titze, I. R. and Talkin, D. (1979). A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation. *Journal of the Acoustical Society of America*, 66:60–74. 70

[Toda et al., 2005] Toda, T., Black, A., and Tokuda, K. (2005). Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 9–12, Philadelphia, PA, USA. 39, 43

[Toda et al., 2007] Toda, T., Black, A., and Tokuda, K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech & Language Processing*, 15(8):2222–2235. 43

[Toda et al., 2001] Toda, T., Saruwatari, H., and Shikano, K. (2001). Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 841–844. 5

[Toda and Young, 2009] Toda, T. and Young, S. (2009). Trajectory training considering global variance for hmm-based speech synthesis. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4025–4028, Taipei, Taiwan. 43

[Tooher and McKenna, 2003] Tooher, M. and McKenna, J. G. (2003). Variation of the glottal lf parameters across f0, vowels, and phonetic environment. In *Proc. of ISCA Tutorial and Research Workshop on Voice Quaility Functions, Analysis and Synthesis (VOQUAL '03)*, pages 41–46. 24

[Tooher et al., 2008] Tooher, M., Yanushevskaya, I., and Gobl, C. (2008). Transformation of lf parameters for speech synthesis of emotion: regression trees. In *Proc. of the 4th Int. Conf. on Speech Prosody*, pages 705–708, Campinas, Brazil. 24, 25, 114

[Tribolet, 1977] Tribolet, J. (1977). A new phase unwrapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(2):170–177. 27

[Turk and Arslan, 2005] Turk, O. and Arslan, L. (2005). Donor selection for voice conversion. In *13th European Signal Processing Conference*, pages 1–4. 51

[Turk et al., 2007] Turk, O., Arslan, L., and Deutsch, F. (2007). Automatic donor ranking and selection system and method for voice conversion. US Patent App. 11/376,377; Patent Publication No.: US 2007/0027687. 51

[Turk and Arslan, 2009] Turk, O. and Arslan, L. M. (2009). Automatic source speaker selection for voice conversion. *The Journal of the Acoustical Society of America*, 125(1):480–491. 51

[Union, 2001] Union, I. T. (2001). Perceptual evaluation of speech quality (pesq). *Itu-t recommendation p.862, ITU-T.* 187

[Uriz et al., 2009a] Uriz, A., Agüero, P., Bonafonte, A., and Tulli, J. C. (2009a). Voice conversion using k-histograms and frame selection. In *10th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 1639–1642. 49

[Uriz et al., 2008] Uriz, A., Agüero, P., Erro, D., and Bonafonte, A. (2008). Voice conversion using frame selection. Technical report, Reporte Interno Laboratorio de Comunicaciones, Facultad de Ingeniería, Universidad Nacional de Mar del Plata (UNMdP), Argentina. 49

[Uriz et al., 2011] Uriz, A., Agüero, P., Tulli, J., Castiñeria, J., González, E., and Bonafonte, A. (2011). Voice conversion using k-histograms and residual averaging. In *XIV Reunión de Trabajo en Procesamiento de la Información y Control (RPIC)*, pages 102–107. 44, 49, 163

[Uriz et al., 2009b] Uriz, A., Agüero, P., Tulli, J., González, E., and Cávez, A. B. (2009b). A comparison between gmm and non-gmm models applied in voice conversion. In *38th JAIIO - Argentine Symposium on Computing Technology (AST)*, pages 31–44. 49

[Uriz et al., 2009c] Uriz, A., Agüero, P., Tulli, J., González, E., and Cávez, A. B. (2009c). Voice conversion using frame selection and warping functions. In *XIII Reunión de Trabajo en Procesamiento de la Información y Control (RPIC)*, pages 417–422, Rosario, Argentina. 49

[Vaidyanathan, 2008] Vaidyanathan, P. P. (2008). *The theory of linear prediction*. Synthesis Lectures on Signal Processing 3. San Rafael, CA: Morgan & Claypool Publishers. xiv, 183 p. 16

[Valbret et al., 1992] Valbret, H., Moulines, E., and Tubach, J.-P. (1992). Voice transformation using psola technique. *Speech Communication*, 11(2-3):175–187. 30, 46

[van Dinther, 2003] van Dinther, R. (2003). *Perceptual aspects of voice-source parameters*. PhD thesis, CIP-Data library Technische Universiteit Eindhoven, The Netherlands. 21

[van Dinther et al., 2004] van Dinther, R., Kohlrausch, A., and Veldhuis, R. (2004). A method for analysing the perceptual relevance of glottal-pulse parameter variations. *Speech Communication*, 42(2):175–189. 21, 24, 120, 187

[Veldhuis, 1998] Veldhuis, R. (1998). A computationally efficient alternative for the liljencrants–fant model and its perceptual evaluation. *The Journal of the Acoustical Society of America*, 103(1):566–571. 20

[Vepa et al., 2002] Vepa, J., King, S., and Taylor, P. (2002). Objective distance measures for spectral discontinuities in concatenative speech synthesis. In *7th International Conference on Spoken Language Processing (Interspeech ICSLP)*, Denver, Colorado, USA. 177

[Ververidis and Kotropoulos, 2008] Ververidis, D. and Kotropoulos, C. (2008). Gaussian mixture modeling by exploiting the mahalanobis distance. *IEEE Transactions on Signal Processing*, 56(7-1):2797–2811. 39, 117

[Vilkman et al., 1999] Vilkman, E., Lauri, E.-R., Alku, P., Sala, E., and Sihvo, M. (1999). Effects of prolonged oral reading on f0, spl, subglottal pressure and amplitude characteristics of glottal flow waveforms. *Journal of Voice*, 13(2):303 – 312. 81

[Villavicencio et al., 2006] Villavicencio, F., Röbel, A., and Rodet, X. (2006). Improving lpc spectral envelope extraction of voiced speech by true-envelope estimation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 869–872, Toulouse, France. 4, 17, 94

[Villavicencio et al., 2007] Villavicencio, F., Röbel, A., and Rodet, X. (2007). All-pole spectral envelope modelling with order selection for harmonic signals. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 49–51. 4, 16, 17

[Villavicencio et al., 2008] Villavicencio, F., Röbel, A., and Rodet, X. (2008). Extending efficient spectral envelope modelling to mel-frequency based representation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1625–1628, Caesars Palace, Las Vegas, Nevada, USA. 17

[Villavicencio et al., 2009] Villavicencio, F., Röbel, A., and Rodet, X. (2009). Applying improved spectral modelling for high quality voice conversion. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4285–4288. 17, 41, 45

[Vincent et al., 2005] Vincent, D., Rosec, O., and Chonavel, T. (2005). Estimation of lf glottal source parameters based on an arx model. In *9th European Conference on Speech Communication and Technology*, pages 333–336. 30

[Vincent et al., 2007] Vincent, D., Rosec, O., and Chonavel, T. (2007). A new method for speech synthesis and transformation based on an arx-lf source-filter decomposition and hnm modeling. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, page 525–528, Honolulu, Hawaii, USA. 25, 30, 45, 83, 84, 90

[Vondra and Vích, 2010a] Vondra, M. and Vích, R. (2010a). Modification of the glottal voice characteristics based on changing the maximum-phase speech component. In *COST 2102 Int. Training School*, pages 240–251, Budapest, Hungary. 28

[Vondra and Vích, 2010b] Vondra, M. and Vích, R. (2010b). Speech modeling using the complex cepstrum. In *Third COST 2102 Int. Training School*, pages 324–330, Caserta, Italy. 28

[Walker and Murphy, 2007] Walker, J. and Murphy, P. (2007). A review of glottal waveform analysis. *Progress in nonlinear speech processing*, pages 1–21. 25, 60

[Wells, 1997] Wells, J. C. (1997). Sampa computer readable phonetic alphabet. In Gibbon, D., Moore, R., and Winski, R., editors, *Handbook of Standards and Resources for Spoken Language Systems*, pages Part IV, section B. Berlin and New York: Mouton de Gruyter. 36

[Wu, 1983] Wu, C. F. J. (1983). On the convergence properties of the em algorithm. *The Annals of Statistics (Institute of Mathematical Statistics)*, 11(1):95–103. 39

[Wu et al., 2006] Wu, C.-H., Hsia, C.-C., Liu, T.-H., and Wang, J.-F. (2006). Voice conversion using duration-embedded bi-hmms for expressive speech synthesis. *IEEE Transactions on Audio, Speech, & Language Processing*, 14(4):1109–1116. 4, 46

[Wu et al., 2010] Wu, Z., Kinnunen, T., Chung, E., and Li, H. (2010). Text-independent f0 transformations with non-parallel data for voice conversion. In *11th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 1732–1735, Makuhari, Chiba, Japan. 5, 45, 46

[Wu et al., 2013] Wu, Z., Virtanen, T., Kinnunen, T., Chng, E., and Li, H. (2013). Exemplar-based unit selection for voice conversion utilizing temporal information. In *14th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 3057–3061, Lyon, France. 49, 163

[Ye and Young, 2004a] Ye, H. and Young, S. (2004a). High quality voice morphing. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 44

[Ye and Young, 2004b] Ye, H. and Young, S. J. (2004b). Voice conversion for unknown speakers. In *8th International Conference on Spoken Language Processing (Interspeech ICSLP)*, Jeju Island, Korea. 36

[Yeh, 2008] Yeh, C. (2008). *Multiple Fundamental Frequency Estimation of Polyphonic Recordings*. PhD thesis, Institut de Recherche et Coordination Acoustique/Musique (IRCAM), Université de Pierre et Marie Curie (UPMC), Université de Paris 6, Paris, France. 12, 32

[Yeh and Röbel, 2004] Yeh, C. and Röbel, A. (2004). A new score function for joint evaluation of multiple f0 hypothesis. In *7th International Conference on Digital Audio Effects (DAFx)*, pages 234–239, Naples, Italy. 57, 62

[Yeh et al., 2010] Yeh, C., Röbel, A., and Rodet, X. (2010). Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals. *IEEE Transactions on Audio, Speech and Language Processing*, 18(6):1116–1126. 90

[Young et al., 2006] Young, S. J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (2006). *The HTK Book*. Cambridge University Press, Cambridge, England, UK. 36

[Yue et al., 2008] Yue, Z., Zou, X., Jia, Y., and Wang, H. (2008). Voice conversion using hmm combined with gmm. In *Congress on Image and Signal Processing (CISP)*, volume 5, pages 366–370. 40

[Yutani et al., 2008] Yutani, K., Uto, Y., Nankaku, Y., Toda, T., and Tokuda, K. (2008). Simultaneous conversion of duration and spectrum based on statistical models including time-sequence matching. In *9th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 1072–1075, Brisbane, Australia. 46

[Zañartu et al., 2013] Zañartu, M., Ho, J. C., Mehta, D. D., Hillman, R. E., and Wodicka, G. R. (2013). Acoustic coupling during incomplete glottal closure and its effect on the inverse filtering of oral airflow. *Proceedings of Meetings on Acoustics, Acoustical Society of America*, 19(1). 96, 190

[Zen et al., 2008] Zen, H., Nankaku, Y., and Tokuda, K. (2008). Probabilistic feature mapping based on trajectory hmms. In *9th Annual Conference of the International Speech Communication Association (Interspeech ISCA)*, pages 1068–1071, Brisbane, Australia. 43

[Zen et al., 2011] Zen, H., Nankaku, Y., and Tokuda, K. (2011). Continuous stochastic feature mapping based on trajectory hmms. *IEEE Transactions on Audio, Speech & Language Processing*, 19(2):417–430. 4, 43

[Zen et al., 2004] Zen, H., Tokuda, K., and Kitamura, T. (2004). A viterbi algorithm for a trajectory model derived from hmm with explicit relationship between static and dynamic features. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 837–840, Montreal, Quebec, Canada. 43

[Zen et al., 2007] Zen, H., Tokuda, K., and Kitamura, T. (2007). Reformulating the hmm as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech & Language*, 21(1):153–173. 43

[Zheng et al., 1998] Zheng, F., Song, Z., Li, L., Yu, W., Zheng, F., and Wu, W. (1998). The distance measure for line spectrum pairs applied to speech recognition. In *5th International Conference on Spoken Language Processing (Interspeech ICSLP), Incorporating The 7th Australian International Speech Science and Technology Conference*, volume 3, pages 1123–1126, Sydney Convention Centre, Sydney, Australia. 177

[Zivanovic and Röbel, 2008] Zivanovic, M. and Röbel, A. (2008). Adaptive threshold determination for spectral peak classification. *Computer Music Journal*, 32(2):57–67. 11, 12, 109

[Zivanovic et al., 2004] Zivanovic, M., Röbel, A., and Rodet, X. (2004). A new approach to spectral peak classification. In *Proc. of the 12th European Signal Processing Conference (EUSIPCO)*, pages 1277–1280, Vienna, Austria. 11, 12, 107, 109, 110, 111

[Zwicker, 1961] Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2):248–248. 187