HAL
open science

On the use of voice descriptors for glottal source shape parameter estimation

Stefan Huber, Axel Röbel

To cite this version:

Stefan Huber, Axel Röbel. On the use of voice descriptors for glottal source shape parameter estimation. Computer Speech and Language, In press, 0885-2308, www.sciencedirect.com/science/article/pii/S0885230813000776. 10.1016/j.csl.2013.09.006 . hal-00865343

HAL Id: hal-00865343

https://hal.science/hal-00865343v1

Submitted on 21 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language xxx (2013) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

On the use of voice descriptors for glottal source shape parameter estimation☆

Stefan Huber*, Axel Roebel**

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

Received 24 April 2013; received in revised form 1 August 2013; accepted 18 September 2013

Abstract

This paper summarizes the results of our investigations into estimating the shape of the glottal excitation source from speech signals. We employ the Liljencrants–Fant (LF) model describing the glottal flow and its derivative. The one-dimensional glottal source shape parameter R_d describes the transition in voice quality from a tense to a breathy voice. The parameter R_d has been derived from a statistical regression of the R waveshape parameters which parameterize the LF model. First, we introduce a variant of our recently proposed adaptation and range extension of the R_d parameter regression. Secondly, we discuss in detail the aspects of estimating the glottal source shape parameter R_d using the phase minimization paradigm. Based on the analysis of a large number of speech signals we describe the major conditions that are likely to result in erroneous R_d estimates. Based on these findings we investigate into means to increase the robustness of the R_d parameter estimation. We use Viterbi smoothing to suppress unnatural jumps of the estimated R_d parameter contours within short time segments. Additionally, we propose to steer the Viterbi algorithm by exploiting the covariation of other voice descriptors to improve Viterbi smoothing. The novel Viterbi steering is based on a Gaussian Mixture Model (GMM) that represents the joint density of the voice descriptors and the Open Quotient (OQ) estimated from corresponding electroglottographic (EGG) signals. A conversion function derived from the mixture model predicts OQ from the voice descriptors. Converted to R_d it defines an additional prior probability to adapt the partial probabilities of the Viterbi algorithm accordingly. Finally, we evaluate the performances of the phase minimization based methods using both variants to adapt and extent the R_d regression on one synthetic test set as well as in combination with Viterbi smoothing and each variant of the novel Viterbi steering on one test set of natural speech. The experimental findings exhibit improvements for both Viterbi approaches.
© 2013 Elsevier Ltd. All rights reserved.

Keywords: Glottal source; Voice quality; R_d shape parameter; LF model; Viterbi smoothing

1. Introduction

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 one glottal flow cycle with an impulse train produces the glottal source waveform. The speech production model is completed by the convolution of the glottal excitation source with the impulse responses of the vocal-tract filter (VTF) as well as the radiation filters at lips and nostrils level. To

☆ This paper has been recommended for acceptance by RKM.

* Corresponding author. Tel.: +33 (0)1 4478 4966; fax: +33 (0)1 4478 1540.
** Principal corresponding author. Tel.: +33 (0)1 4478 4845.
E-mail addresses: stefan.huber@ircam.fr (S. Huber), axel.roebel@ircam.fr (A. Roebel).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.09.006

Please cite this article in press as: Huber, S., Roebel, A., On the use of voice descriptors for glottal source shape parameter estimation. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.09.006

solve the problem of estimating all components that are used in this model from a recorded speech signal a multitude of approaches have been proposed. However, for the moment none of these algorithms is sufficiently robust to allow for a reliable analysis of natural speech. In the following we will address a number of problems and investigate into means to obtain a physiologically consistent estimate of a glottal pulse shape parameter from synthetic and natural speech signals.

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 Normalized Amplitude Quotient (NAQ) (Alku et al., 2002) or the R_d parameter (Fant et al., 1994; Fant, 1995). We use the latter to parameterize the LF model (Fant et al., 1985) describing the glottal volume-velocity flow and its derivative.

In our experiments with natural speech signals we observed the importance to cover extreme adducted and abducted phonations. This requires to extend the normal R_d range [0.3, 2.7] (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 (Fant et al., 1994) for $R_d > 2.7$ is required to describe abducted phonations occuring 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.

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. To coherently cover an extended R_d range [0.1, 6.0] we use our equations recently proposed in Huber et al. (2012). In this study we discuss more in detail the range adaptation of the R_d parameter regression. Additionally, we introduce a second variant of the adapted and extended R_d regression.

Moreover, to avoid local instabilities of the R_d estimator we propose to apply Viterbi smoothing (Forney, 1973) since even with the extended R_d range the results are often perturbed by implausible jumps. However, each of our R_d estimation methods can, under some conditions, be systematically skewed. In addition, Viterbi smoothing cannot correct possibly skewed or biased R_d contours in longer time segments. Therefore we propose to steer the Viterbi algorithm by exploiting the covariation of other voice descriptors. The latter are used to train GMMs from which a second R_d estimate is predicted. This defines an additional prior R_d probability used to steer Viterbi smoothing.

To estimate R_d we use the phase minimization based paradigm established in Degottex et al. (2010), Degottex (2010) and Degottex et al. (2011). It is based on minimizing the mean squared phase errors present in the spectrum when matching synthesized glottal formants against a strictly harmonic representation of voiced speech. The R_d value that is used to synthesize the glottal formant resulting in the lowest remaining phase error is selected as estimated R_d value per frame. The baseline method called MSPD2I1 in Huber et al. (2012) and three recently in Huber et al. (2012) proposed phase minimization variants called MSPD2I0, MSPD2I2 and MSPD2IX are examined together with the two R_d regression variants on an objective evaluation test set of synthetic speech signals.

Additionally, we examine the performance of Viterbi smoothing and the novel Viterbi steering using the four phase minimization methods to estimate R_d on natural speech. The evaluation is similar to Fröhlich et al. (2001), Ó Cinnéide (2012) and Kane and Gobl (2013a). It is based on comparing the OQ estimates derived from the R_d curves (OQ_{R_d}) with the OQ estimates derived from an analysis of synchronously to the audio waveforms recorded EGG signals (OQ_{EGG}). The objective of this study is to determine the best performing parameterization for Viterbi smoothing and the novel Viterbi steering as well as the most robust phase minimization method.

The article is organized as follows. The glottal source shape parameter R_d and the LF model parameterization are detailed in Section 3. The basic model for the human speech production is introduced in Section 3. The phase minimization based methods to estimate the R_d parameter are explained in Section 4. The adaptation and the range extension of the R_d parameter regression are discussed in Section 5. The optimized R_d estimation using Viterbi smoothing and the novel Viterbi steering is explained in Section 6. The evaluation results are presented in Section 7 and 8. The summary and the conclusions about this work are given in Section 9.

2. LF model parameterization

The R waveshape parameters R_a, R_k and R_g parameterize the LF model (Fant et al., 1985, 1994; Fant, 1995). 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 pulse and describes its asymmetry. R_g is an inverse measure of the glottal pulse rise time (Fant et al., 2000).

The parameterization of R_d is determined by means of a statistical regression on values of the R waveshape parameter set (Fant and Liljencrants, 1994; Fant, 1995) that were observed for voiced natural speech (Gobl, 1988; Karlsson, 1990). It is based on exploiting systematic covariations of the R parameters present in the studied speech corpora. Fant introduced in Fant and Liljencrants (1994) and Fant (1995) equations to compute approximate R parameter values from an R_d value, denominated with the subscript p as 'predicted' R waveshape parameters R_{ap}, R_{kp}, and R_{gp}. We employ the term R_{*p} waveshape parameter set. The equations how to calculate 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) and 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 from the R waveshape parameter set. The reduced form OQ_i defines the open quotient

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

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

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

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 R_d parameter can be expressed as F_0-normalized glottal waveshape parameter (Fant, 1995)

$$R_d = \frac{U_0}{E_e} \cdot \frac{F_0}{110} = \left(\frac{1}{0.11}\right) \cdot \left(\frac{T_d}{T_0}\right). \tag{3}$$

The scaling factor 1/110 corresponds to $F_0 = 110$ Hz as a typical average in male speech (Gobl, 1988; Fant, 1997). The direct ratio in amplitude measure relates U_0 and E_e to the T_d-related Amplitude Quotient (AQ) in Alku and Vilkman (1996) and the R_d-related Normalized Amplitude Quotient (NAQ) in Alku et al. (2002).

3. Human 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

$$S(\omega) = G_{Rd}(\omega) \cdot C(\omega) \cdot L(\omega) \cdot H(\omega, F_0, D), \tag{4}$$

with ω being the angular frequency. Eq. (4) defines the deterministic part of the voice production model which is composed of a representation of the following components. The shape parameter R_d (Fant, 1995) parameterizes the Liljencrants–Fant (LF) (Fant et al., 1985) pulse model of the glottal flow ($G_{Rd}(\omega)$). 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 parameterized with the fundamental frequency F_0 and the delay between pulse sequence and frame center in terms of the phase delay D of the fundamental sinusoid is represented by $H(\omega, F_0, D)$.

Please cite this article in press as: Huber, S., Roebel, A., On the use of voice descriptors for glottal source shape parameter estimation. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.09.006

4. Glottal pulse parameter estimation

The R_d parameter estimation algorithms based on the objective functions for phase minimization that will be used in this study have been completely described in Degottex (2010), Degottex et al. (2011) and Huber et al. (2012). In this section we will only explain the basic scheme required for the following investigations presented in this paper.

The algorithm constructs a sinusoidal model of a speech signal frame. It is transformed into a harmonic model describing a single pitch period with a sampling rate that falls onto a harmonic grid at position $2K+2$. The harmonic model is assumed to be noise free for each harmonic k up to K. In the experiments presented in this study we determine the highest harmonic sinusoidal partial K by rounding the ratio 8000 Hz to F_0 to the nearest integer value. According to Eq. (4) the voice production model can then be simplified into

$$S(k) = G_{Rd}(k) \cdot C(k) \cdot L(k) \cdot e^{jkD} \quad k \in [0, 1, \ldots, K]. \tag{5}$$

We use the procedure described in Stylianou (2001) to construct $S(k)$ from a signal frame by means of finding the parameter set having minimum error. The fundamental frequency F_0 is estimated using the approach presented in Yeh and Roebel (2004). The algorithm proceeds by means of testing the minimum phase property of the VTF spectrum that is obtained for a sufficiently compact grid of R_d values:

$$\hat{C}_{Rd}(k) = \frac{S_k}{G_{Rd}(k) \cdot jk}. \tag{6}$$

The residual will represent for the correct R_d parameter the minimum phase transfer function of the vocal tract filter (Degottex et al., 2011). The accurate value of R_d is estimated on the convolutive residual which is described in detail in Degottex et al. (2011) and Huber et al. (2012). 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. 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.

An arbitrary delay D is introduced into the voice production model of Eqs. (4) and (5). 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). A 2nd order difference operator compensates the dependency on the pulse position but introduces a high-pass filter. Thus, subsequent integrations are required to suppress the influence of the high-pass filter.

The denomination Mean Squared Phase Differentiation Integration specifies the acronym MSPDI. The number of differentiation or integration steps are indicated by each subsequent number after D and respectively I. The application of first 2 differentiations and then 0, 1, and respectively 2 integration steps on the phase errors of the convolutive residual 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 method are combined by the method MSPD2IX in form of a weighted sum.

5. Adaptation variants of the extended range R_d regression

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 we could find in the literature illustrating these contours established by Fant for both the normal and the upper R_d range. 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. 1. 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).

In Huber et al. (2012) we proposed 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. Additionally, we extended the R_d range 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. We adapt the R_{gp} curves of the normal and the upper R_d range at the minimum of the convex function of R_{gp} for the normal R_d range at $R_d = 1.8476$. In Huber et al. (2012) we rounded

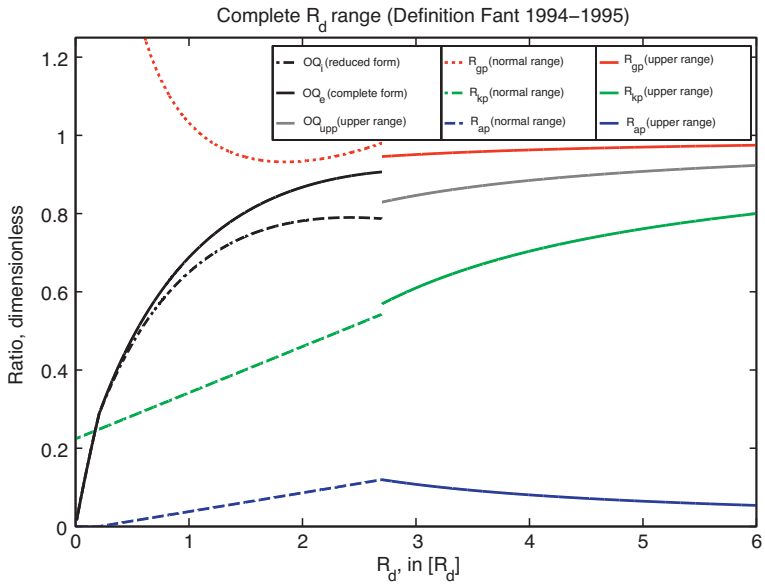

Fig. 1. Original waveshape $R*_p$ parameter contours.

this adaptation point to $R_d = 1.85$. An offset of 9.3552×10^{-3} has to be added to the R_{gp} contour of the upper R_d range to compensate for a remaining difference. Eqs. (7), (8), (9), and (10) follow our adaptation proposed in Huber et al. (2012), denoted here as variant 1. It requires to add an offset of -4.2753×10^{-2} to $R_{kp_{2.70}}$ of the upper R_d range.

In this paper we introduce adaptation variant 2 which takes into account that according to Fant et al. (1994) R_{kp} for the upper R_d range depends on R_{gp}. We adapt R_{kp} to depend on the upper range equation not at $R_d = 2.7$ as defined in Fant et al. (1994) but already at $R_d = 1.8476$ to be more conform with our introduced 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 $+4.2753 \times 10^{-2}$ adapts $R_{kp_{1.85}}$ at the upper R_d range accordingly. The contours of our proposed $R*_p$ parameter adaptation for both variants are shown in Fig. 2.

Please note that the OQ contour of the reduced form OQ_i derived from the original R_d regression of Fant (1995, 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. Our R_d regression adaptation variants suppress these ambiguities by establishing a strictly increasing OQ_i contour over the whole R_d range. We find for the reduced OQ_i form defined in Eq. (1) maximum values of $OQ_i = 0.90$ at $R_d = 6.0$ for variant 1 and $OQ_i = 0.95$ for variant 2.

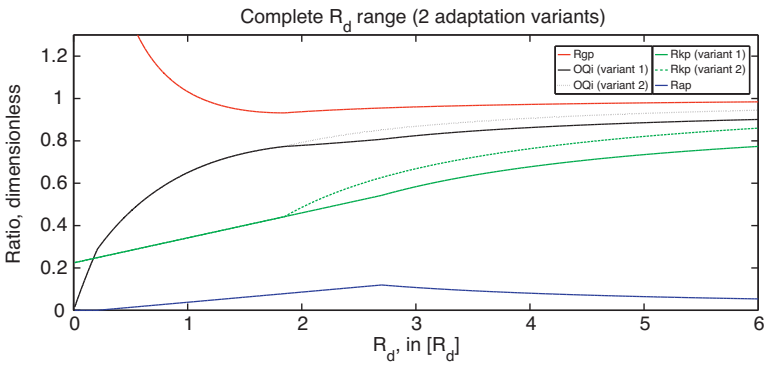

Fig. 2. Adapted waveshape $R*_p$ parameter contours.

Please cite this article in press as: Huber, S., Roebel, A., On the use of voice descriptors for glottal source shape parameter estimation. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.09.006

Our proposed set of equations with a more exact listing of the offset and border values as reported in Huber et al. (2012) to define the adaptation of the R_d waveshape parameter regression for the extended R_d range and both adaptation variants is:

$$R_{ap} = \begin{cases} 0 & \forall\ 0.01 \leq R_d < 0.21 \\ \dfrac{(-1 + 4.8 \cdot R_d)}{100} & \forall\ 0.21 \leq R_d \leq 2.70 \\ \dfrac{(32.3/R_d)}{100} & \forall\ 2.70 < R_d \leq 6.00 \end{cases} \tag{7}$$

$$OQ_{upp} = 1 - \frac{1}{(2.17 \cdot R_d)} \quad \forall\ 1.8476 \leq R_d \leq 6.00 \tag{8}$$

Variant 1:

$$R_{kp_{2.70}} = \begin{cases} \dfrac{(22.4 + 11.8 \cdot R_d)}{100} & \forall\ 0.01 \leq R_d \leq 2.70 \\ (2 \cdot R_{gp} \cdot OQ_{upp}) - 1.0428 & \forall\ 2.70 < R_d \leq 6.00 \end{cases} \tag{9}$$

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

Variant 2:

$$R_{kp_{1.85}} = \begin{cases} \dfrac{(22.4 + 11.8 \cdot R_d)}{100} & \forall\ 0.01 \leq R_d \leq 1.8476 \\ (2 \cdot R_{gp} \cdot OQ_{upp}) - 0.9572 & \forall\ 1.8476 \leq R_d \leq 6.00 \end{cases} \tag{11}$$

$$R_{gp} = \begin{cases} \dfrac{0.25 \cdot R_{kp_{1.85}}}{((0.11 \cdot Rd)/(0.5 + 1.2 \cdot R_{kp_{1.85}})) - R_{ap}} & \forall\ 0.01 \leq R_d \leq 1.8476 \\ 9.3552 \times 10^{-3} + \dfrac{596 \times 10^{-2}}{7.96 - 2 \cdot OQ_{upp}} & \forall\ 1.8476 < R_d \leq 6.00 \end{cases} \tag{12}$$

6. Viterbi smoothing and steering

6.1. Viterbi smoothing

In this paper we denominate the well accepted Viterbi algorithm (Forney, 1973) as standard Viterbi smoothing. We distinct it from our novel approach to steer the Viterbi algorithm which we denominate as Viterbi steering accordingly. Viterbi steering exploits the covariation of voice descriptors and voice quality features and will be introduced in Section 6.2.

The R_d parameter estimation operates frame-based by selecting at each analysis step the glottal pulse shape corresponding to the lowest remaining residual phase error. 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 shapes to an efficient 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., 2011) may systematically be mislead in such segments.

These random influences 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. We define the probabilistic model of standard Viterbi smoothing as follows:

Observation probability $P(O|X)$

We approximate the speech production model using a grid of R_d values. Each of the $N_{R_{di}}$ 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 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 $E_{Rd} = 0$ 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). We employ the SIGMA algorithm (Thomas and Naylor, 2009) to detect the GCIs. The evaluation of this work 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

$$L(p) = \sum_n \log(P(O|X_p(n)) \cdot P(X_p(n)|X_p(n-1))), \tag{13}$$

where n is the discrete time and p is a path through the state space of the process. We insert into Eq. (13) the log probability function E_{Rd} of the observation probability $P(O|X)$ and scale its distribution with parameter α_a. Respectively we include the probabilistic distribution $\Delta R_d/\Delta_n$ of the transition probability $P(X_n|X_{n-1})$ with its scale parameter γ_g to find

$$\overline{L}(p) = -\sum_n \alpha_a \cdot E_{R_{d_n}} + \gamma_g \cdot \frac{\Delta R_d}{\Delta_n} + C. \tag{14}$$

The term C 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. We can factor out the scaling factor γ_g of the log-likelihood of the transition which leaves the parameter $\alpha = \alpha_a/\gamma_g$ as control parameter. This defines the probability function that is used to perform Viterbi smoothing on all sequences p:

$$\overline{\overline{L}}(p) = -\sum_n \alpha \cdot E_{R_{d_n}} + \frac{\Delta R_d}{\Delta_n}. \tag{15}$$

The experimental setup of Section 8 and the test of Section 8.2.4 examines which value for α creates the best Viterbi smoothing results without the application of the novel Viterbi steering.

6.2. Viterbi steering

The novel steering of the Viterbi algorithm constitutes an extension of standard Viterbi smoothing presented in the preceding Section 6.1. Viterbi smoothing can augment the R_d estimation robustness but it cannot correct a systematic bias present in longer time segments. We observe a possible systematic bias for each R_d estimator 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. In this chapter we investigate into means to correct a possible systematic bias of the R_d estimation in the mentioned problematic regions.

The phase minimization paradigm (Degottex et al., 2011) 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. Therefore we examined the covariation of additional voice descriptors 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. We use the recordings of the speakers BDL, JMK and SLT of the CMU Arctic speech database (Kominek and Black, 2004) which provides simultaneously recorded speech waveforms and EGG signals. To estimate OQ from the corresponding EGG signals we use the DECOM method (Henrich et al., 2004). We derive the OQ_{EGG} contours from all available phrases of all three speakers for all voiced segments using the voicing decision described in Section 6.1.

As a proof-of-concept for the possibility to exploit specific speech signal features for the estimation of the R_d parameter we establish a machine learning approach to aid Viterbi smoothing. By using a statistical model we exploit 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). A function to predict a second OQ estimate from the voice descriptor set is derived from the trained statistical model.

6.2.1. Covariation voice descriptors

From an extensive analysis of different voice descriptors and combinations in-between them on the CMU databases we determined several voice descriptors. Each selected feature demonstrated to be highly positively correlated with the OQ_{EGG} reference, except the Voiced/Unvoiced Frequency (VUF) boundary (Roebel, 2010; Stylianou, 2001) which shares a negative correlation. All proposed voice descriptors are not influenced by a lower number of stable harmonic sinusoids and are therefore well suited to exploit their covariation with the shape of the glottal excitation source.

H1–H2: The amplitude difference in dB of the first two harmonic partials H1 and H2 (H1–H2). According to Henrich et al. (2001) it is a reliable spectral correlate of OQ. H1–H2 proved to contribute to the discrimination between breathy and tense voice qualities in Scherer et al. (2012). We measure the relation directly in the magnitude spectrum and do not apply inverse filtering to measure $H1^*$–$H2^*$ from the corresponding glottal source signal as in Fant (1995), Hanson (1995) and Henrich et al. (2001). This avoids possible problems with inverse filtering (Rothenberg, 1972; Alku, 1992; Drugman et al., 2008), while the direct measure of H1 in the magnitude spectrum is according to Keating and Esposito (2006) influenced by the first formant F1. To smooth the direct H1–H2 measure we apply a median filter of order 5.

3 MFCC bins: With the sum of the 3rd, 4th and 6th MFCC bin (Ellis, 2005) we seek to model the spectral slope (Scherer et al., 2012) or the spectral tilt (Murphy, 2001) to reflect the amplitude continuation of the spectral envelope which is correlated to a tense, modal or breathy phonation. However, neither a regression on the slope of the spectral peaks as in Scherer et al. (2012) nor the utilization of the spectral tilt measures R_{14} or R_{24} as in Murphy (2001) nor other related measurements were able to achieve an overall correlation to the OQ_{EGG} reference being as high and robust as the summed combination of the 3rd, 4th and 6th MFCC bin (Ellis, 2005). Apparently, the proposed summation of the three MFCC bins is less influenced by the variation of the vocal tract formants.

$\mathbf{F}_0$: 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 while 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 foundation of the voice not just originates the different voice qualities but serves as well as a determinant of F_0 and sound pressure level (SPL) contours. Speakers favour a particular pitch range for each phonation type (Laver, 1968; Childers and Lee, 1991). However, different studies (Laver, 1980; Maddieson and Hess, 1987; Hanson et al., 1990) have shown that the relation between pitch and different phonation types is speaker-dependent. We estimate F_0 using the monophonic F_0 implementation based on the principles described in Yeh and Roebel (2004).

VUF: The Voiced/Unvoiced Frequency (VUF) boundary (Roebel, 2010; Stylianou, 2001) correlates with the voiced/unvoiced energy ratio and the bandwidth of the glottal formant. The VUF is thus related to E_e which determines the amount of generated sinusoidal energy (Fant, 1995), and the noise energy level which originates from turbulences created at the glottis, for example due to an imperfect glottal closure and a high airflow rate (Childers and Lee, 1991).

Please cite this article in press as: Huber, S., Roebel, A., On the use of voice descriptors for glottal source shape parameter estimation. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.09.006

A tense voice parameterized by low R_d values originates a broad excitation spectrum with sinusoidal content present in higher frequency regions, while a relaxed voice parameterized by higher R_d values originates only few harmonic sinusoidal partials in lower frequency regions (Fant, 1995). The *VUF* is estimated by a) detecting sinusoidal peaks (Zivanovic et al., 2004), b) measuring the sinusoid vs. noise energy ratio (SNE) in sub-bands of fixed and constant bandwidth, c) selecting as *VUF* the highest band for that the SNE is above a given threshold. The measured *VUF* contour is smoothed using a median filter covering the time of half the length of the analysis window.

6.2.2. OQ_{GMM} 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, from an informal examination on the CMU Arctic databases we measure H1–H2 and OQ_{EGG} values below both limits. In Henrich et al. (2001) and Doval et al. (2006) it is shown that the relation between $H1^*$–$H2^*$ and OQ is additionally influenced by the asymmetry coefficient a_m representing the skewness of the glottal pulse. Moreover, recent studies (Kreiman et al., 2012, 2012b; Chen et al., 2013) suggest that the relationship is speaker-dependent, leading to positive and negative correlations, or situations where one parameter remains relatively constant while the other varies considerably.

We could not find an analytic formulation expressing the relation of OQ with any of the other voice descriptors in the literature, and as well not for the complete set of the proposed voice descriptors. Thus, we employ a GMM to model the relation of the covariation feature set with the OQ_{EGG} reference. A similar approach using Gaussian mixture modelling to predict glottal source signals has already been proposed in Thomas et al. (2009) and Gudnason et al. (2012). Per speaker we train one GMM on the covariation feature combination estimated on each voiced segment of the other two speaker databases and their corresponding OQ_{EGG} estimates.

We distinct two feature sets:

Model 1 refers to the utilization of the voice descriptors H1–H2, *VUF* and the sum of the 3rd, 4th and 6th MFCC bin.

Model 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, 2011). We express the joint probability density $p(M, R|\lambda^{(Z)})$ of the feature set M (either model 1 or 2) and the OQ_{EGG} reference R, conditioned on the model parameters λ with the weights α_q, mean vector μ_q^Z and the covariance matrices Σ_q^Z over Q mixture sequences as

$$p(M, R|\lambda^{(Z)}) = \sum_{q=1}^{Q} \alpha_q \cdot N(M, R; \mu_q^{(Z)}, \Sigma_q^{(Z)}). \tag{16}$$

Q is set to 6 mixture components for model 1, and to respectively 8 mixture components for model 2. The function

$$F(m) = \sum_{q=1}^{Q} p_q^m m \cdot [\mu_q^r + \Sigma_q^{rm} \Sigma_q^{mm^{-1}} (m - \mu_q^m)] \tag{17}$$

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

6.2.3. Viterbi steering implementation

The prediction of OQ_{GMM} values from each GMM model per speaker defines the additional prior probability M_{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 set for model 1 or model 2.

Prior R_d probability $P(X|M)$

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 M 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 M_{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 M_{prior} probabilities.

Optimal Viterbi steering path

We insert the prior R_d probability into Eq. (13) to find

$$L(p) = \sum_n \log(P(O_{R_d}|X_p(n)) \cdot P(X_p(n)|M(n)) \cdot P(X_p(n)|X_p(n-1))). \tag{18}$$

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

$$\overline{L}(p) = -\sum_n \alpha_a \cdot E_{R_{d_n}} + \beta_b \cdot M_{prior_n} + \gamma_g \cdot \frac{\Delta R_d}{\Delta_n} + C \tag{19}$$

Again, the constant term C can be ignored by the Viterbi algorithm. We factor out the scaling factor γ_g 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.

$$\overline{\overline{L}}(p) = -\sum_n \alpha \cdot E_{R_{d_n}} + \beta \cdot M_{prior_n} + \frac{\Delta R_d}{\Delta_n} \tag{20}$$

The evaluation tests for the proposed novel Viterbi steering of Eq. (20) are presented and discussed in Section 8. We examine which values for α and β result in the highest performance to estimate R_d contours on natural speech using as evaluation data set the three CMU Arctic databases.

7. Objective evaluation on a synthetic test set

In this section we examine the performance of phase minimization algorithms to estimate R_d on a synthetic test set similar to Degottex et al. (2011) and 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 vocal tract filter, 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. To simulate the VTF influence we synthesize 16 synthetic vowels C_{VTF} using Maeda's digital simulator (Maeda, 1982). Each C_{VTF} is convolved with each glottal formant parameterized by an R_d value within the range [0.1, 6] on a grid of step size $R_d = 0.1$. We synthesize at 10 different F_0 values within the range [80, 293] Hz. 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 VUF boundary to F_0 determines how many stable harmonic sinusoids N are observable before being masked by noise. We evaluate the influence of N for the range [3, 8] by restricting the R_d estimation algorithm to observe N partials. Additionally we simulate the position of the window with respect to the period in time 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). The results are presented in a compact manner by adding up the bias and standard deviation of the R_d estimation errors as a function of the examined parameter.

7.1. Examination on dependency in F_0

Fig. 3 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 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

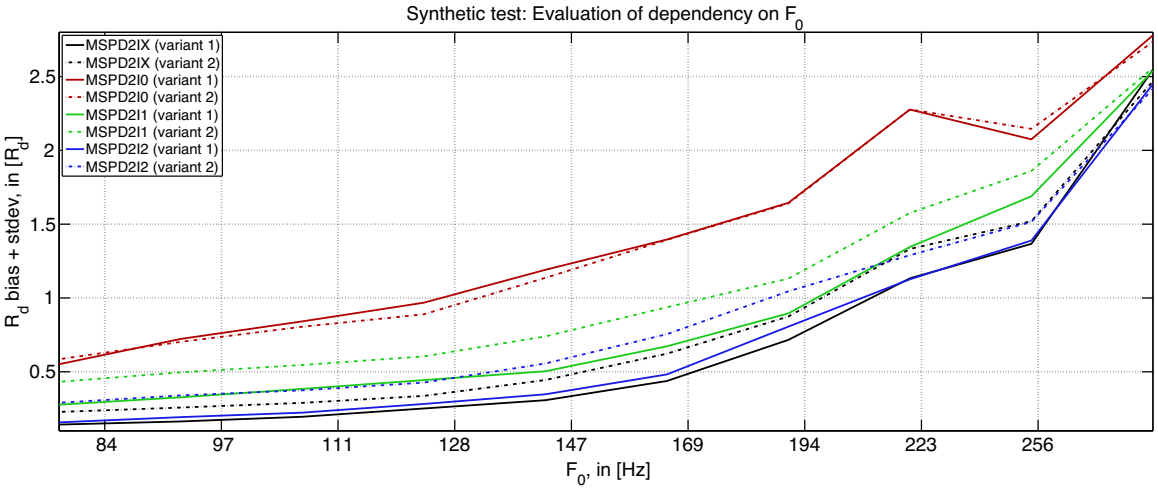

Fig. 3. Evaluation of R_d estimation, F_0-dependency.

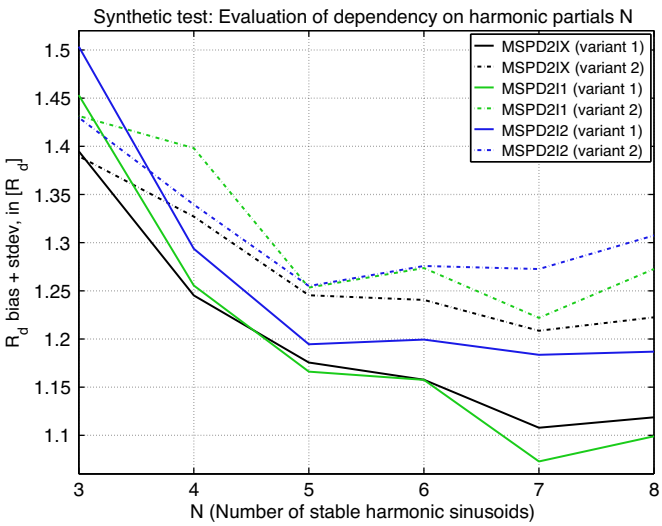

Fig. 4. Evaluation of R_d estimation, N-dependency.

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.

7.2. 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. We simulate these influences by varying N, shown in Fig. 4. For the variation of N depicted in Fig. 4 we set both the glottal source noise n^{σ_g} and the environmental noise n^{σ_e} to $n = 1$. This corresponds to the lowest noise level of −50 dB used in our synthetic test. With this low noise level the characteristics of natural speech are simulated but the misleading interference of noise is suppressed which is required to properly examine a different number of stable harmonic sinusoids N.

Since MSPD2I0 demonstrated to perform worst it is omitted in Fig. 4 to provide a more clear presentation. 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 cumulated amount of R_d estimation errors at $N=7$. This is justified by the fact that the harmonic sinusoids of $N \geq 8$ 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.

7.3. Examination on dependency in voice quality

Fig. 5 exemplifies how the objective function of each phase minimization based method is dependent on the phase differences of the LF model 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. Fig. 5 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 our proposed R_d range extension 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. 7. The R_d estimation error surfaces are spanned up frame-wise over time and reflect the behaviour of each objective function.

Finally we conclude that 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. When examining Fig. 5 by visual inspection it is apparent 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 underperform more for higher R_d values above $R_d > 2.7$.

8. Objective evaluation on natural speech

In this section we evaluate the four phase minimization methods (Huber et al., 2012) on natural 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.

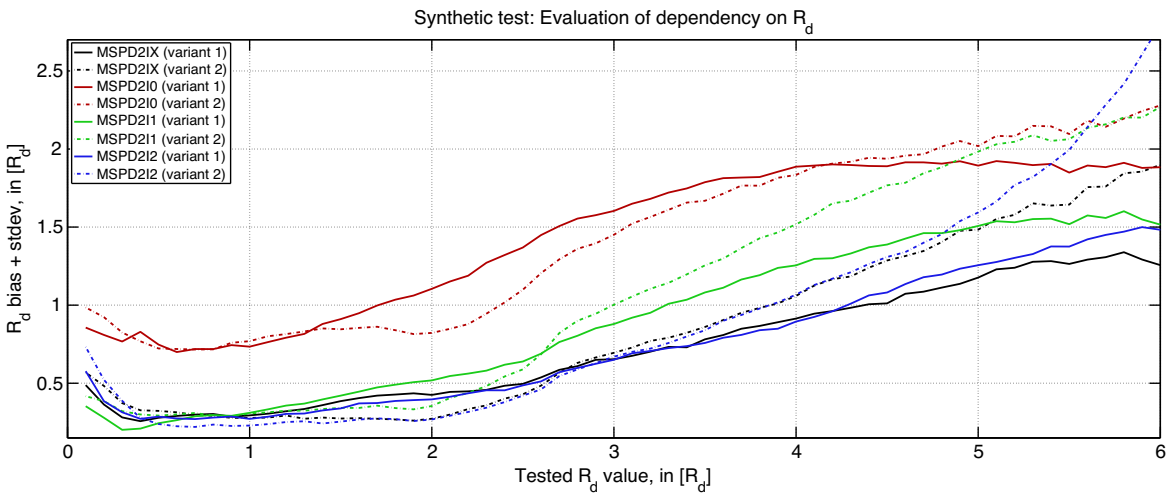

Fig. 5. Evaluation of R_d estimation, dependency on voice quality measured in R_d.

Please cite this article in press as: Huber, S., Roebel, A., On the use of voice descriptors for glottal source shape parameter estimation. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.09.006

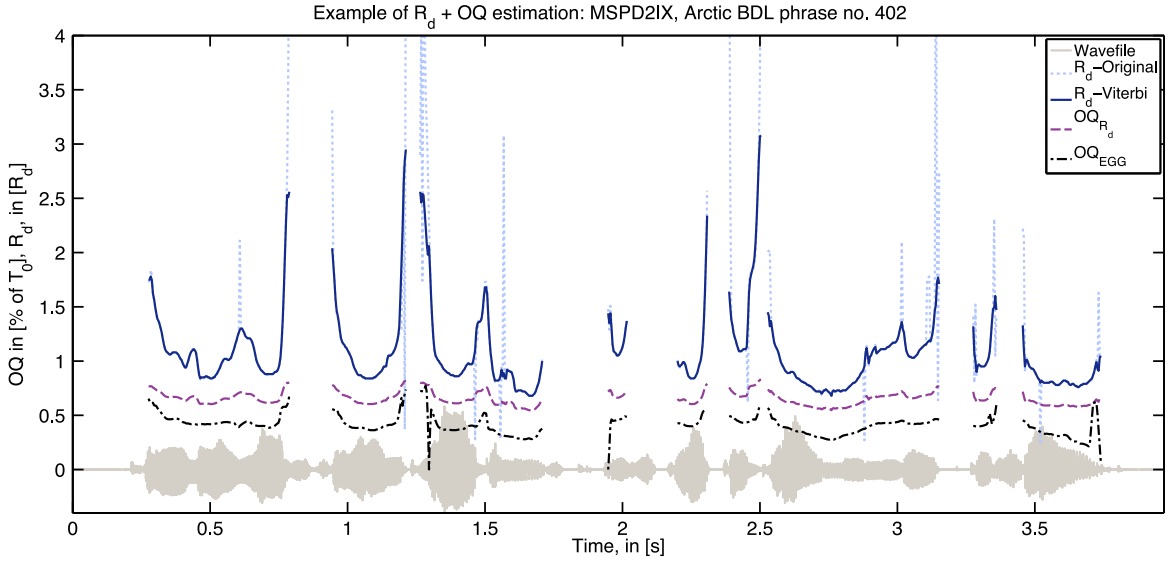

Fig. 6. Estimation example for method MSPD2IX, Viterbi smoothing, $\alpha = 0.47$, CMU Arctic database, speaker BDL, phrase number 402.

8.1. 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). For example, a reliable and exact determination of the time instant of glottal opening (GOI) can be more difficult and erroneous (Baken, 1992; Baer et al., 1983) than the estimation of GCIs. 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). 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 we choose as test scenario for natural human speech the OQ comparison using EGG and audio recordings because of its relatively easy setup and reasonable reliability in contrast to other methods. We compare the OQ contours derived from each R_d estimate (OQ_{R_d}) with the OQ contours estimated by the DECOM method (Henrich et al., 2004) on the corresponding EGG signals (OQ_{EGG}). The example of Fig. 6 shows the curves of the frame-based R_d estimator in light dotted, after Viterbi smoothing in dark solid, the from it derived OQ_{R_d} contour in dashed and the OQ_{EGG} reference in dash-dotted lines. We observe as in Childers and Lee (1991) and Herbst (2004) a general non-constant offset between the OQ_{R_d} and OQ_{EGG} contours due to the mentioned systematic bias of the OQ estimation by the EGG-based technique. We choose the phrase shown in Fig. 6 to exemplify that the EGG measure can be error-prone as the physiologically impossible jumps of OQ_{EGG} around 1.3 and 1.95 s illustrate.

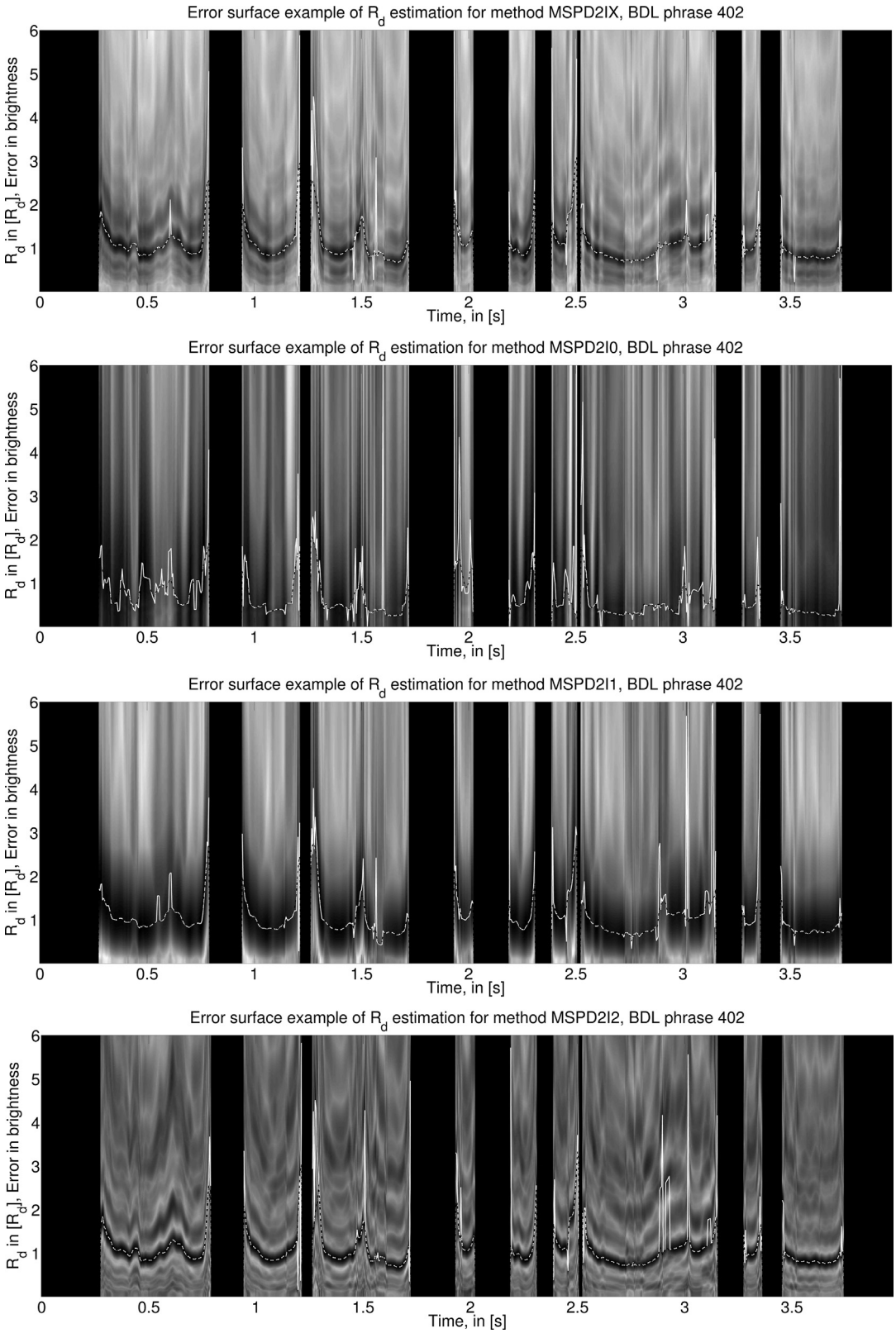

Fig. 7. R_d error surface examples, 4 phase minimization methods, standard Viterbi smoothing, $\alpha = 0.47$, BDL phrase 402.

Please cite this article in press as: Huber, S., Roebel, A., On the use of voice descriptors for glottal source shape parameter estimation. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.09.006

8.2. OQ test across speakers

8.2.1. Error surfaces of Viterbi smoothing

Fig. 7 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 functions generate an error curve per frame. 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 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 gray 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 speech example of speaker BDL. Side minima appear e.g. at ∼0.6 s at $R_d \approx 2.0$ and $R_d \approx 5.0$. Since these side minima are higher than the overall lowest error value, no unnatural jumps occur. The latter are present for MSPD2I2 at ∼2.9 s where the initial R_d estimate in white jumps three times from the apparently true R_d contour and its obvious error valley at $R_d \approx 1.0$ to misleading side minima at $R_d \approx 2.5$ and $R_d \approx 4.0$. 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 which results in a higher probability to produce physiological impossible jumps of the R_d estimate. MSPD2I1 suffers for example at ∼0.6 s 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. 7. 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 gray do not follow the true shape of the glottal source.

8.2.2. 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. We use r 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.

We use the root-mean-square error (*rmse*) as second evaluation metric. To avoid any impact of the bias that is present in the EGG-based OQ estimate we remove the mean between the OQ estimates for each voiced segment. Please note that the calculation of Pearsons product moment correlation coefficient r implies the removal of a possible bias between both evaluated sample distributions. In the following, the Rd adaptation variants 1 and 2 from Section 5 are evaluated for each test scenario.

Table 1
OQ test results, without Viterbi smoothing, Rd adaptation variant 1.

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

Table 2
OQ test results, without Viterbi smoothing, Rd adaptation variant 2.

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

Please cite this article in press as: Huber, S., Roebel, A., On the use of voice descriptors for glottal source shape parameter estimation. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.09.006

Table 3
OQ test results, median smoothing, order 5.

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

Table 4
Viterbi smoothing (optimal α-values in parentheses).

	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)

The results for each objective function estimating R_d without applying Viterbi smoothing are listed in Tables 1 and 2. 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 note that the baseline method MSPD2I1 of Degottex et al. (2011) 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. Since variant 2 outperforms variant 1 for each test set on natural speech, we will only show R_d adaptation variant 2 in the following.

8.2.3. *Smoothing with a moving average filter*

We examined different moving average filter types to evaluate 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 filter with order 5, shown in Table 3. It improves the estimated R_d contours of each phase minimization method only to a marginal extent. In the following we will investigate into means of establishing a more robust correction of the estimated R_d contours by utilizing different configurations of the Viterbi algorithm.

8.2.4. *Standard Viterbi smoothing*

The results of Viterbi smoothing without the utilization of the novel OQ_{GMM} prediction-based Viterbi steering are summarized in Table 4. The improvements of applying a dynamic programming algorithm to smooth the estimated glottal source shape curves are apparent when comparing its results with the ones of Tables 1 and 2 (without the application of Viterbi smoothing) and Table 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 *r*-correlation maxima and the *rmse*-error minima are shown per method. 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 $\alpha = 0.10$ to the *rmse*-minima.

In the following we will present the results of the novel steering of Viterbi smoothing 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 6. 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.

8.2.5. *Viterbi steering using GMM prediction*

We evaluate 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. The *rmse*-error and the *r*-correlation of each OQ_{GMM} prediction model using the feature set of model 1 are shown in Table 5 and respectively the feature set of model 2 in Table 6.

Table 5
Validation on training and test sets per speaker, model 1.

	$rmse_{training}$	$rmse_{test}$	$r_{training}$	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

Table 6
Validation on training and test sets per speaker, model 2.

	$rmse_{training}$	$rmse_{test}$	$r_{training}$	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

Table 7
Viterbi steering, model 1 (optimal α-values in parentheses).

	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)

Please note that by the straightforward training of a GMM using 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 correlations for r_{test} shown in Tables 5 and 6 being close to the performance of the signal processing based R_d estimation methods. The strong potential of the proposed OQ prediction using voice descriptors is indicated by the corresponding results on the training test set $r_{training}$ which by far outperform the results discussed in the following tests. However, the remaining principal problem is to overcome the speaker-dependency of the modelling to generalize better over speaker specific characteristics. This could be solved by employing more speaker databases 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. In Section 8.4 we will analyze more in detail the intrinsic characteristics of each speaker data set of the utilized CMU Arctic databases by examining the r-correlation and $rmse$-error results per speaker.

In the following sections we will examine 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 6 are shown in Table 7.

For this test we vary the Viterbi scaling parameter α while fixing the Viterbi scale β 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 the Table 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.

A one-way ANOVA analysis of the corresponding optimization of the scale parameter β for Viterbi steering demonstrated that β has no statistically significant influence on each of the evaluated phase minimization variants.

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

Table 8
Viterbi steering, model 2 (optimal α-values in parentheses).

	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)

Table 9
Viterbi steering, model 2 (optimal β-values in parentheses).

	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)

To further augment the robustness of the glottal source shape parameter estimation on natural speech we use as additional voice descriptor the fundamental frequency F_0 for the OQ_{GMM} model 2. A comparison of each *rmse*-error and each *r*-correlation value for each speaker between Table 5 (model 1, without F_0) and Table 6 (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, we fix the scale parameter β to 1.0 and vary the scale parameter α. Table 8 illustrates the results of the α-optimization for the Viterbi steering of model 2. We execute the one-way ANOVA analysis between the results of the corresponding test to optimize the scale parameter α of Viterbi steering using model 1 (illustrated in Table 7) with Viterbi steering using model 2 (listed in Table 8). 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 7) to $r = 0.5437$ (listed in Table 8).

Fixing the determined α-maxima in terms of the measured *r*-correlations, depicted in Table 8, to optimize the scale parameter β, whose results are illustrated in Table 9, does not exhibit statistically significant improvements. We show the β-optimization of Table 9 depicting 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. We compare the evaluation distributions of standard Viterbi smoothing (illustrated in Table 4) with the overall best Viterbi steering results, the β-optimization of model 2 (shown in Table 9). 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 contours of the estimated glottal source shape parameters demonstrates improvements to a statistically significant extent when comparing the results with standard Viterbi smoothing a) for both employed GMM models, b) for each of the four R_d estimation methods and c) for both R_d regression adaptation variants. Please note that we did only discuss the results for R_d regression adaptation variants 2 while 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 (Fant and Liljencrants, 1981) not just correlates with the contours of parameterized glottal excitation source shapes but can be exploited as covariation feature additionally to the other proposed voice descriptors to estimate glottal source parameters. However, this has to be evaluated on a bigger test set employing more speakers to evaluate the speaker-dependency.

Table 10
Comparison results of other methods.

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

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 covariation of other voice descriptors is a relatively robust manner to be used for glottal source parameter estimation. 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. As discussed in Section 8.2 and shown with Fig. 6, the EGG-based technique is not error-free. The employed evaluation metric is therefore limited and prevents the potential to achieve higher correlation and lower error measurements.

8.3. *OQ test of other methods across speakers*

We examine the results of the following other glottal source estimation algorithms on the same test set of the CMU Arctic databases to provide an objective comparison to our algorithm variants by employing the same evaluation metric. We denominate the method introduced in Kane et al. (2012), Kane and Gobl (2013a) as **DyProg-LF**. It is based on first estimating GCI locations using a modified version of the SE-DREAMS approach of Drugman et al. (2012b) described in Kane and Gobl (2013a). Then the iterative adaptive inverse filtering method (Alku, 1992) is used to compute the glottal source signals on which error values are calculated. The error criteria are based on measuring the correlation of synthesized LF model pulses parameterized over a grid of R_d values and matched to the source signals in the spectral and the time domain. The temporal and spectral errors form a target cost. The continuity of the parameter trajectories over frames forms a transition cost. 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.

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.

As third method we evaluate the power spectrum based method of Ó Cinnéide (2012) called **PowRd**. 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 determines the filter order and the coefficients of the vocal tract filter as well as the scale parameter E_e of the included LF voice source model. The PowRd method is based on the SIM approach of Fröhlich et al. (2001) and is thus as well robust to phase disturbances and the window position of the analyzed frame with respect to the period in time.

For all three methods we measure as well a non-constant offset between the OQ curves estimated from the audio recordings versus the OQ curves derived from the corresponding EGG recordings. It confirms the observation of a possible systematic OQ bias by the EGG-based method discussed in Section 8.1.

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 our DECOM basis, leading 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 10 confirms to a certain extent the conclusions we drew from our results. Dynamic programming immensely improves the results of glottal source shape parameter estimation by suppressing unnatural jumps in short-time segments. On

Table 11
Mean μ and standard deviation σ of speaker characteristics.

	BDL	JMK	SLT
F_{0_μ}	121.83 Hz	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

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 smoothing algorithm.

8.4. *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.

Table 11 shows the mean μ and standard deviation σ for the fundamental frequency F_0 and the *OQ* derived from the EGG signals, measured on all voiced segments and all phrases for each speaker. 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. This observation for BDL corroborates the findings from Drugman et al. (2012). JMK demonstrates the highest mean open quotient OQ_{EGG_μ} with the lowest variance σ^2.

From informal listening tests we perceive 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., 2012) 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 7 to a less good glottal source shape parameter estimation performance.

In the following sections we will examine the *OQ* comparison test results per speaker. We will only show the results for each speaker using R_d adaptation variant 2 which demonstrates the better performance throughout the whole *OQ* comparison test. To restrict the huge amount of test results conducted for this study we only discuss important findings based on Pearsons correlation metric r.

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.

8.4.1. *OQ test results for speaker BDL*

The results of the synthetic test discussed in Section 7 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 the three evaluated speakers the highest r-correlations, listed in Table 12. He has the lowest mean open quotient OQ_{EGG_μ} 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

Table 12
BDL r-correlation results (α- or β-values in parentheses).

	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)

Table 13
BDL comparison results of other methods.

	DyProg-LF	Strik-LF	PowRd
r	0.6606	0.3268	0.3315

Table 14
JMK r-correlation results (α- or β-values in parentheses).

	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)

Table 15
JMK comparison results of other methods.

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

R_d estimation robustness for speaker BDL. The other three methods DyProg-LF, Strik-LF, and PowRd achieve as well the best R_d estimation results for speaker BDL, illustrated in Table 13.

8.4.2. 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_{EGG_μ} measured and his weak vocal effort perceived lead to less good R_d estimation results. By informal visual inspection of the OQ_{EGG} contours for speaker JMK we realize that the EGG-based OQ reference exhibits as well more physiological impossible movements compared to the other speakers OQ_{EGG}.

The Viterbi steering of model 2 using F_0 demonstrates only slight improvements for all but the worst performing method MSPD2I0, depicted in Table 14. 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 to establish performant R_d estimation results for speaker JMK, shown in Table 15. The PowRd method without the utilization of dynamic programming achieves a similar performance to the DyProg-LF approach using dynamic programming.

8.4.3. OQ test results for speaker SLT

The R_d estimation results lie for speaker SLT with the proportionally highest F_{0_μ} in the range measured for speaker JMK.

No method could benefit from F_0 as voice descriptor for the OQ_{GMM} prediction model 2, shown in Table 16. It indicates that the rather large difference in F_{0_μ} between the employed training data set of the male speakers BDL

Table 16
SLT r-correlation results (α- or β-values in parentheses).

	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)

Table 17
SLT comparison results of other methods.

	DyProg-LF	Strik-LF	PowRd
r	0.3638	0.1206	0.1204

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 17 benefits from the application of its used dynamic programming approach and achieves nearly a similar performance compared to the phase minimization variants using Viterbi smoothing and steering.

9. Summary and conclusions

9.1. Summary

In this paper we proposed a novel technique we denominate Viterbi steering to aid the estimation of glottal source shape parameters. It is based on exploiting the covariation of other voice descriptors to steer and optimize the glottal source estimation using Viterbi smoothing. This work presents additionally an in-depth analysis of glottal source estimation using phase minimization variants by discussing the results of two extensive objective evaluation tests on synthetic and natural human speech signals. Moreover, two variants to adapt and extent the range of the glottal source shape parameter R_d to parameterize the LF glottal source model are discussed.

9.2. Conclusions

From the works presented in this study we draw the following conclusions: It has been shown that the exploitation of covariation voice features is able to increase the robustness of the estimation of glottal excitation source parameters describing the voice quality of human speech for the employed data set. The utilization of a machine learning approach to classify voice qualities is able to aid and to possibly outperform the signal processing paradigms known to date to estimate glottal source signals in the future, if implemented in a more sophisticated manner. For example, the voice descriptor F_0 for the OQ_{GMM} prediction model 2 was utilized without a possible normalization by its mean F_{0_μ} per speaker. It could especially render for speaker SLT with the highest F_{0_μ} a better estimation performance for Viterbi steering of model 2. Also, the performance of the novel Viterbi steering approach using the GMM-based R_d-predictor should improve by the utilization of a bigger database covering more speaker characteristics in the trained models. Moreover, additional voice quality features as in Kane and Gobl (2013b), the consideration of an intensity measure like a relative SPL originated form the subglottal pressure (Laver, 1968; Vilkman et al., 1999; Fant and Kruckenberg, 2005) or the Harmonic Richness Factor (HRF) (Childers and Lee, 1991) to reflect a higher sinusoidal signal content related to tense voices and a higher noise signal content related to breathy voices should contribute to the OQ_{GMM} prediction robustness. The importance to smooth estimated glottal source parameters over time was validated as in Vincent et al. (2007), Kane et al. (2012) and Kane and Gobl (2013b) by the experimental findings. The proposed steering and the utilization of standard Viterbi smoothing demonstrates improvements to a statistically significant extent. The methods MSPD2I2 and MSPD2IX proposed in our last study (Huber et al., 2012) perform better than their corresponding baseline approach MSPD2I1 established in the preceding works (Degottex et al., 2010, 2011; Degottex, 2010). The

estimation of voice qualities with a comparatively higher relaxed phonation and less vocal effort poses more difficulties. The R_d adaptation variant 2 proposed in this study showed to provide 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 where values of R_d occur predominantly in the normal R_d range.

Acknowledgements

The main author is supported by CIFRE contract no. 2011/2011, a collaboration between the research institute IRCAM and the company Acapela Group. He would like to thank Dr. John Kane for the preparation of the test results for the DyProg-LF and Strik-LF method as well as for the kind feedback and discussion concerning the utilized speaker databases. Furthermore, he would like to thank Dr. Alan Ó Cinnéide for the disposal of and the kind help on his PowRd method. Also, thanks go to Dr. Olga Gordeeva from the Acapela Group for the analysis of the speaker databases. Finally, he is very grateful for the kind attendance by his supervisor Dr. Axel Röbel.

References

Alku, P., 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Communication 11 (2-3), 109–118.

Alku, P., Bäckström, T., Vilkman, E., 2002. Normalized amplitude quotient for parametrization of the glottal flow. Journal of the Acoustical Society of America 112 (2), 701–710.

Alku, P., Vilkman, E., 1996. Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering. Speech Communication 18 (2), 131–138.

Baer, T., Lofqvist, A., 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.

Baken, R.J., 1992. Electroglottography. Journal of Voice 6 (2), 98–110.

Catford, J.C., 1977. Fundamental Problems in Phonetics. Indiana University Press.

Chen, G., Samlan, R.A., Kreiman, J., Alwan, A., 2013. 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 ICSLP), Lyon, France.

Childers, D.G., Hicks, D.M., Moore, G.P., 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.

Childers, D.G., Lee, C.K., 1991. Vocal quality factors: analysis, synthesis, and perception. Journal of the Acoustical Society of America 90 (5), 2394–2410.

Colton, R.H., Conture, E.G., 1990. Problems and pitfalls of electroglottography. Journal of the Voice Foundation 4 (1), 10–24.

Degottex, G., 2010. Glottal source and vocal tract separation. IRCAM, Paris (PhD thesis).

Degottex, G., Roebel, A., Rodet, X., 2010. Joint estimate of shape and time-synchronization of a glottal source model by phase flatness. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, USA, pp. 5058–5061.

Degottex, G., Roebel, A., Rodet, X., 2011. Phase minimization for glottal model estimation. IEEE Transactions on Acoustics, Speech and Language Processing 19 (5), 1080–1090.

Doval, B., d'Alessandro, C., Henrich, N., 2006. The spectrum of glottal flow models. Acta Acustica united with Acustica 92 (6), 1026–1046.

Dromey, C., Stathopoulos, E.T., Sapienza, C.M., 1992. Glottal airflow and electroglottographic measures of vocal function at multiple intensities. Journal of Voice 6 (1), 44–54.

Drugman, T., Dubuisson, T., Moinet, A., d'Alessandro, N., 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), Porto, Portugal, pp. 202–207.

Drugman, T., Kane, J., Gobl, C., 2012. Modeling the creaky excitation for parametric speech synthesis. In: 13th Annual Conference of the International Speech Communication Association (Interspeech ICSLP), Portland, Oregon, USA.

Drugman, T., Thomas, M., Gudnason, J., Naylor, P., 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.

Ellis, D.P.W., 2005. PLP and RASTA (and MFCC, and inversion) in Matlab, online web resource: http://www.ee.columbia.edu/dpwe/resources/matlab/rastamat/.

Fant, G., 1995. The LF-model revisited. Transformation and frequency domain analysis. In: Quarterly Progress and Status Report, vol. 36, no. 2–2. Department of Speech, Music and Hearing, KTH, pp. 119–156.

Fant, G., 1997. The voice source in connected speech. Speech Communication 22 (2-3), 125–139.

Fant, G., Kruckenberg, A., 2005. Covariation of subglottal pressure f0 and intensity. In: Interspeech 2005 – Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, pp. 1061–1064.

Fant, G., Kruckenberg, A., Liljencrants, J., Båvegård, M., 1994. Voice source parameters in continuous speech. Transformation of LF-parameters. In: 3rd International Conference on Spoken Language Processing (ICSLP), Yokohama, Japan, pp. 1451–1454.

Fant, G., Kruckenberg, A., Liljencrants, J., Båvegård, M., 2000. Acoustic-phonetic studies of prominence in Swedish. In: Quaterly Progress and Status Report, vol. 41, no. 41. Department of Speech, Music and Hearing, KTH, Stockholm, Sweden, pp. 001–052.

Fant, G., Liljencrants, J., 1994. Data reduction of LF voice source parameters. In: Working Papers 43. Department of Linguistics and Phonetics, Lund University, Sweden, pp. 062–065.

Fant, G., Liljencrants, J., Lin, Q.-G., 1985. A four-parameter model of glottal flow. In: Quaterly Progress and Status Report, vol. 26, no. 4. Department of Speech, Music and Hearing, KTH, Stockholm, Sweden, pp. 001–013.

Fant, H.F., Liljencrants, J., 1981. Dynamic characteristics of voice fundamental frequency in speech and singing. acoustical analysis and physiological interpretations. In: Quaterly Progress and Status Report, vol. 22, no. 1. Department of Speech, Music and Hearing, KTH, Stockholm, Sweden, pp. 001–020.

Forney, G.D., 1973. The Viterbi algorithm. Proceedings of the IEEE 61 (3), 268–278.

Fröhlich, M., Michaelis, D., 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), 479–488.

Gobl, C., 1988. Voice source dynamics in connected speech. In: Quarterly Progress and Status Report, vol. 29, no. 1. Department of Speech, Music and Hearing, KTH, Stockholm, Sweden, pp. 123–159.

Gobl, C., 2003. The voice source in speech communication. Department of Speech, Music and Hearing, KTH, Stockholm (PhD thesis).

Gobl, C., Chasaide, A.N., 1992. Acoustic characteristics of voice quality. Speech Communication 11 (4-5), 481–490.

Gudnason, J., Thomas, M.R.P., Ellis, D.P.W., Naylor, P.A., 2012. Data-driven voice source waveform analysis and synthesis. Speech Communication 54 (2), 199–211.

Hanson, D., Gerratt, B., 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.

Hanson, H., 1995. Individual variations in glottal characteristics of female speakers. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 772–775.

Henrich, N., d'Alessandro, C., Doval, B., 1999. Glottal open quotient estimation using linear prediction. In: Proceedings of the International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, pp. 12–17.

Henrich, N., d'Alessandro, C., 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), 2nd Interspeech Event, Aalborg, Denmark, pp. 47–50.

Henrich, N., d'Alessandro, C., Doval, B., 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.

Herbst, C., 2004. Evaluation of various methods to calculate the egg contact quotient. Department of Speech, Music and Hearing, KTH, Stockholm, Sweden (diploma thesis in music acoustics).

Hill, T., Lewicki, P., 2007. Statistics: Methods and Applications. StatSoft, Tulsa, Oklahoma, USA.

Huber, S., Roebel, A., Degottex, G., 2012. Glottal source shape parameter estimation using phase minimization variants. In: 13th Annual Conference of the International Speech Communication Association (Interspeech ICSLP), 1990-9772. Portland, Oregon, USA.

Kane, J., Gobl, C., 2013a. Automating manual user strategies for precise voice source analysis. Speech Communication 55 (3), 397–414.

Kane, J., Gobl, C., 2013b. Wavelet maxima dispersion for breathy to tense voice discrimination. IEEE Transactions on Audio, Speech, and Language Processing 21 (6), 1170–1179.

Kane, J., Yanushevskaya, I., Chasaide, A.N., Gobl, C., 2012. Exploiting time and frequency domain measures for precise voice source parameterisation. In: Proceedings of the 6th International Conference on Speech Prosody, Shanghai, China.

Karlsson, I., 1990. Voice source dynamics for female speakers. In: First International Conference on Spoken Language Processing (ICSLP), Kobe, Japan, pp. 69–72.

Keating, P.A., Esposito, C.M., 2006. Linguistic voice quality. In: Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology, University of Auckland, Auckland, New Zealand, pp. 85–91.

Kominek, J., Black, A.W., 2004. The cmu arctic speech databases. In: Proceedings of the 5th ISCA Speech Synthesis Workshop, pp. 223–224.

Kreiman, J., Iseli, M., Neubauer, J., Shue, Y.-L., Gerratt, B.R., Alwan, A., 2012. The relationship between open quotient and h1(*)–h2(*). Journal of the Acoustical Society of America 124 (4), 2495.

Kreiman, J., Shue, Y.-L., Chen, G., Iseli, M., Gerratt, B.R., Neubauer, J., Alwan, A., 2012b. Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. Journal of the Acoustical Society of America 132 (4), 2625–2632.

Lanchantin, P., Rodet, X., 2010. Dynamic model selection for spectral voice conversion. In: 11th Annual Conference of the International Speech Communication Association (Interspeech ICSLP), Makuhari, Chiba, Japan.

Lanchantin, P., Rodet, X., 2011. Objective evaluation of the dynamic model selection method for spectral voice conversion. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic, pp. 5132–5135.

Laver, J.D.M., 1968. Voice quality and indexical information. International Journal of Language and Communication Disorders 3 (1), 43–54.

Laver, J.D.M., 1980. The Phonetic Description of Voice Quality, vol. 31. Cambridge University Press.

Maddieson, I., Hess, S., 1987. The effects of f0 of the linguistic use of phonation type. In: 11th International Congress of Phonetic Sciences (ICPhS), No. 67. UCLA Working Papers in Phonetics, pp. 112–118.

Maeda, S., 1982. A digital simulation method of the vocal-tract system. Speech Communication 1 (3-4), 199–229.

Marasek, K., 1997. Egg & Voice Quality, online web resource: http://www.ims.uni-stuttgart.de/institut/arbeitsgruppen/phonetik/EGG/frmst1.htm.

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), 2nd Interspeech Event, Aalborg, Denmark, pp. 1495–1498.

Please cite this article in press as: Huber, S., Roebel, A., On the use of voice descriptors for glottal source shape parameter estimation. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.09.006

Obin, N., 2012. Cries and whispers – classification of vocal effort in expressive speech. In: 13th Annual Conference of the International Speech Communication Association (Interspeech ICSLP), Portland, Oregon, USA.

Ó Cinnéide, A., 2012. Phase-distortion-robust-voice-source analysis. Dublin Institute of Technology, Dublin, Ireland (PhD thesis).

Ó Cinnéide, A., Dorran, D., Gainza, M., 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 ICSLP), Florence, Italy, pp. 57–60.

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.

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.

Roebel, A., 2010. Shape-invariant speech transformation with the phase vocoder. In: 11th Annual Conference of the International Speech Communication Association (Interspeech ICSLP), Makuhari, Chiba, Japan, pp. 2146–2149.

Rothenberg, M., 1972. A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. Journal of the Acoustical Society of America 53 (6), 1632–1645.

Sapienza, C.M., Stathopoulos, E.T., 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.

Scherer, S., Kane, J., Gobl, C., Schwenker, F., 2012. Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. Computer Speech and Language 27 (1), 263–287.

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.

Strik, H., Cranen, B., Boves, L., 1993. Fitting a LF-model to inverse filter signals. In: Third European Conference on Speech Communication and Technology (Eurospeech), Berlin, Germany.

Stylianou, Y., 2001. Applying the harmonic plus noise model in concatenative speech synthesis. IEEE Transactions on Speech and Audio Processing 9 (1), 21–29.

Thomas, M.R.P., Gudnason, J., Naylor, P.A., 2009. Data-driven voice source waveform modelling. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Taipei, Taiwan, pp. 3965–3968.

Thomas, M.R.P., 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.

Titze, I., 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.

Titze, I.R., 1994. Principles of Voice Production. Prentice Hall, Englewood Cliffs, NJ, pp. 354.

Vilkman, E., Lauri, E.-R., Alku, P., Sala, E., 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.

Vincent, D., Rosec, O., Chonavel, T., 2007. A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and hnm modeling. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Honolulu, Hawaii, USA, pp. 525–528.

Walker, J., Murphy, P., 2007. A review of glottal waveform analysis. Progress in Nonlinear Speech Processing, 1–21.

Yeh, C., Roebel, A., 2004. A new score function for joint evaluation of multiple f0 hypothesis. In: 7th International Conference on Digital Audio Effects (DAFx), Naples, Italy, pp. 234–239.

Zivanovic, M., Roebel, A., Rodet, X., 2004 September. A new approach to spectral peak classification. In: Proceedings of the 12th European Signal Processing Conference (EUSIPCO), Vienna, Austria, pp. 1277–1280.

Please cite this article in press as: Huber, S., Roebel, A., On the use of voice descriptors for glottal source shape parameter estimation. Comput. Speech Lang. (2013), http://dx.doi.org/10.1016/j.csl.2013.09.006