Spectral domain analysis, modelling and transformation of sound

Axel Roebel

Spectral domain analysis, modelling and transformation of sound

research summary report submitted to the
Université Pierre et Marie Curie (UPMC) by

Axel Röbel

Analysis-Synthesis Team,
UMR STMS IRCAM-CNRS-UPMC
75004 Paris, France

for obtaining the qualification

Habilitation à Diriger des Recherches

defense: 05. december 2013

Jury:

Mr. Philippe Depalle	Professor, Schulich School of Music, McGill University, Canada	Reviewer
Mr. Thierry Dutoit	Professor, numediart, University of Mons, Belgium	Reviewer
Mr. Sylvain Marchand	Professor, Image & Son Brest, Université de Bretagne Occidentale, France	Reviewer
Mr. Christophe d'Alessandro	DR CNRS, LIMSI, France	Examiner
Mr. Patrick Flandrin	DR CNRS, ENS, France	Examiner
Mr. Bruno Gas	Professor, ISIR UPMC, France	Examiner
Mr. Anssi Klapuri	Professor, Tampere University of Technology, Finnland	Examiner
Mr. Xavier Serra	Professor, MTG, Pompeu Fabra University, Spain	Examiner

Version 1.06

CONTENTS

Summary

The following report describes the research that I have performed or directed over the last 20 years covering the period from my PhD thesis (A. Röbel 1993) until today. After an introduction discussing the evolution and invariants of my research activities in chapter 1, the research I have performed before joining IRCAM is presented in a short summary in chapter 2. The main part of the report will then describe my research on spectral domain analysis, modelling, and transformation of sound that has been conducted at IRCAM during the last 13 years.
First, the specificities of the research environment at IRCAM and the general objectives of my research activities will be briefly discussed in chapter 3. Then, signal analysis and signal transformation using the short time Fourier transform (STFT) and extensions of the phase vocoder algorithm are covered in chapter 4. Advanced methods for STFT based signal processing that are currently developed and not yet used in practical applications will be presented in chapter 5. The three methods covered include: signal adaptive resolution, texture transformation and source separation. My research on sinusoidal modelling taking into account non-stationary sinusoidal components is presented in chapter 6. Research results related to source-filter models, notably spectral envelope estimation and instrument timbre modelling, are described in chapter 7. The research related to the problem of fundamental frequency estimation for sounds containing single or multiple quasi harmonic sources are described in chapter 8. The following chapter 9 discusses my research on speech signal processing, covering the three research problems: glottal pulse parameter estimation, voice conversion, and singing synthesis.
Activities related to the development of signal processing libraries are described in chapter 10. In the final chapter of this document (chapter 11) I will discuss the four major directions that I expect to be important for my research during the next five years.

Résumé

Ce rapport décrit les recherches que j'ai réalisées ou dirigées pendant les 20 dernières années depuis ma thèse (A. RÖBEL 1993) jusqu'à aujourd'hui. Après une introduction, qui discute les évolutions et les invariantes de mes activités de recherche dans le chapitre 1, les recherches effectuées avant mon arrivée à l'IRCAM sont résumées brièvement dans le chapitre 2. Les chapitres suivants traitent ensuite du sujet principal de ce rapport, de mes recherches sur l'analyse, la modélisation et la transformation du signal sonore utilisant les représentations dans le domaine spectral, que j' ai effectuées pendant les dernières 13 années à l'IRCAM.
D'abord, la spécificité du milieu de la recherche à l'IRCAM et les objectifs généraux de mes activités de recherche sont décrits dans le chapitre 3. L'analyse et la transformation du signal utilisant la transformée de Fourier à court terme (TFCT) et des extensions de l'algorithme du vocodeur de phase sont couverts par le chapitre 4. Des méthodes avancées du traitement du signal basé sur la représentation TFCT qui sont actuellement en cours de développement et qui ne sont pas encore utilisées dans des application pratique sont présentées dans le chapitre chapitre 5. Les trois méthodes y discutées sont : la représentation et transformation avec résolution temps-fréquence adaptée au signal, la transformation des textures, et la séparation des sources. Mes recherches sur la modélisation sinusoïdale en tenant compte des composantes sinusoïdales non stationnaire sont présentées dans le chapitre 6. Les résultats de recherche liés aux modèles source-filtre : l'estimation de l'enveloppe spectrale et la modélisation du timbre de l'instrument, sont décrits dans le chapitre 7. Mes recherches concernant l'estimation de la fréquence fondamentalle des sons contenant une ou plusieurs sources quasi harmoniques sont décrites dans le chapitre 8. Le chapitre 9 discute de mes recherches sur le traitement des signaux de parole, en particulier les problèmes de la caractérisation de source glottique, de conversion du locuteur, et de synthèse de chant.
Les activités liées au développement de bibliothèques pour le traitement du signal sont décrites dans le chapitre 10. Le dernier chapitre de ce document (chapitre 11) présente les quatre grands axes que je considère comme importantes pour mes recherches pendant les cinq prochaines années.

Acknowledgments

Without a supporting and stimulating environment a researcher cannot be creative and will not produce any interesting results. Here below I will try to give credit to all the people that have established my environment in a manner that I feel was extremely supportive. This report is therefore dedicated to all those that have helped me to achieve the results that are the subject of the present report:

Xavier Rodet, who gave me the opportunity to join his team in 2000, and who has been, throughout the more than ten years of collaboration, a constant source of stimulation, both as a research collaborator and as a personal friend.

Alain Lithaud, *the analytic ear*, who has heard more of SuperVP test sounds than anybody else in the world, and who has validated as expert listener many, if not all, sound transformation algorithms I have developed.

The PhD students I have worked with that have entrusted me with the responsibility for an important part of their scientific development and that often have become good friends. I list them here approximately in the order of appearance: Chunghsin Yeh, Juanjo Burred, Fernando Villavicencio, Marco Liuni, Gilles Degottex, Wei-Chen Chang, Tien Ming Wang, Henrik Hahn, Stefan Huber, Henrik von Coler, Wei-Hsiang Liao, and Yu-Ren Chien.

The master students I have supervised and that all have contributed important facets to our understanding of sound representation and sound transformation.

Then there are all the friends and colleagues, temporary or permanent, in the analysis/synthesis team, always open to share their knowledge, to join a discussion about research or to help out in case of problems: Frédéric Cornu, Charles Picasso, Thomas Helie, Geoffroy Peeters, Sean O'Leary, Erdal Özbek, Yuki Mitsufuji, Pierre Lanchantin, Christophe Veaux, Snorre Farner, Miroslav Zivanovic.

Hugues Vinet, who is constantly trying to improve the position of the R&D department in the ever and rapidly changing societal and political environment in France.

The many people at IRCAM working behind the scenes that create the organizational foundation for our research activities and that are never mentioned in research reports: The system team that keeps our hardware in shape, the team of secretaries that help us organizing travels and scientific events, the human resource department that does all administrative tasks related to receiving visiting researchers and reminds us to go on holiday, the financial services that helps us spending money by moving it into the correct account and the building services team that keeps the physical work environment in the best possible shape.

Last but not least my family, Mei-Hua and Klarissa, the two women in my life that make the sun shine even in rainy Paris in November, my father who went early, my mother who is ready to go, and my brother who does everything in his own way.

... und jedem Anfang wohnt ein Zauber inne
der uns beschützt und der uns hilft zu leben...
- Hermann Hesse, Das Glasperlenspiel

Reference notation

To clearly distinguish the work of other people from the work of researchers under my supervision or work I performed myself references the related to these two groups of people will be formatted differently. Publications of myself and publications of co-workers that did work under my supervision will be cited using braces, e.g. (A. Röbel 2001c), and references not related to my own work will be cited using square brackets, e.g. [Parker and Chua 1987].
Moreover, references to different kind of projects will not use the (author name, year) citation format but will be numbered separately using distinguishing prefixes indicating the type of the reference. For a visiting postdoc this would for example give (PD4). The prefixes that are used are:

PhD : PhD thesis and projects with visiting PhD students,
MA : Master-thesis,
PD : Projects of visiting Post-docs that I supervised,
RP : Research projects that I supervised or in that I participated,
LI : Industrial software licenses,
SW : Software development projects that I worked on,
ES : External software.

CHAPTER 1

INTRODUCTION

The present report covers the time period starting with my PhD thesis (A. Röbel 1993) until today. It is natural that over a time span of 20 years the research topics evolve. Starting with a rather fundamental research problem in the PhD thesis (A. Röbel 1993) I changed to investigate more practically oriented signal models during the period as assistant professor at the Technical University of Berlin, and to an even stronger application oriented research direction at IRCAM. There are nevertheless a few constants that did not evolve over time.

Sound has been at the centre of my research since my master thesis. This is clearly a strong personal preference, which however will not be discussed further. Then, there is a relatively strong interest in the application of mathematics that can be observed in all my research activities. To establish the links between the purely theoretical world of mathematical theorems and conclusions and the real world is traditionally an approach that was used in physics. Today this method has been extended to many application domains (e.g. economics, computer science, statistics, game theory) including as well (digital) signal processing, and Fourier - or other signal transformations. The benefit here is the fact that given the preconditions of mathematical theorems are met, mathematical reasoning can be used to develop and test hypothesis, before experimental evaluation. When it comes to sound processing, then the important benefit of spectral domain processing is a rough structural similarity to the auditory system that decomposes sound waves into spectral bands as well.

One can speculate about the reason for this organization of the auditory system, may be the resulting spectral representation can be encoded most efficiently [Lewicki 2002]. For signal transformation, however, the efficiency of the sound representation is much less important than the coherence between the sound parameterization and the properties of the underlying physical sound source. It is relatively straightforward to see that coherent sound transformation is easiest if the different modes of vibration are resolved into individual sinusoidal components (A. Röbel 2010a). These considerations have led me to centre my research on spectral domain signal processing. This means signal processing after transformation of the signal into one of the many spectral domain representations or models, as the sinusoidal model, that derive from these representations. The constant frequency resolution of the STFT based representation is especially favourable for the quasi harmonic sounds that are encountered in music and speech, and therefore, the STFT representation plays a central role in the following report.

The main focus on spectral domain representation cannot and should not preclude work on time domain representations. This is especially the case for speech signal processing, where it is important to take into account the irregularities of the glottal pulse sequence, which can be achieved more easily if at least part of the sound model is handled in the time domain. Examples of this kind of representation will be briefly discussed in chapter 9.

CHAPTER 2

EARLY RESEARCH ACTIVITIES

The research topic I was working on during my PhD thesis and for the time directly after the PhD was related to a problem I had worked on during my master studies, and that had emerged a few years earlier: the reconstruction of the state space of nonlinear dynamical systems using delayed time series of the systems output signal [Takens 1981]. The basic theory that I was interested in at that time was the theory of nonlinear dynamical systems notably systems exhibiting chaotic behavior and the discovery that a reconstruction of the state space of any nonlinear stationary system can be obtained by means of constructing vectors from delayed samples of a sufficiently long time series of the output of the system [Parker and Chua 1987]. The problem I had investigated in my thesis was to learn the system's nonlinear dynamics in the reconstructed state space using adaptive nonlinear functions, notably artificial neural networks [Bishop 1995], and the method I had established during the thesis was able to faithfully reproduce chaotic dynamics from stationary time series (A. Röbel 1993). A main problem of the method was the extremely high computational costs for training the nonlinear models. To counter these costs I had developed a method to reduce the training data by means of dynamically selecting the most important and difficult data points from the available training data sets (A. Röbel 1994a,b). It is interesting to note that the algorithm I had proposed selects training data points that bear some resemblance with the support vectors in support vector regression [Smola and Schölkopf 2004] notably when using iterative procedures or chunked versions of the dataset to construct the support vector machine. The difference is, however, that in my algorithm I used the selected vectors to train the regression function while in support vector regression the regression function is directly based on the support vectors.
After the PhD Thesis I went through a number of different positions that will be described in the following sections.

2.1 GMD

The first position after my PhD was a PostDoc position at the GMD[1] in Berlin. There I worked under direction of Dr. G. Kock and Prof. S. Jaehnichen in a research group dedicated to artificial neural networks. Time series prediction with neural networks [Refenes 1995; Weigend and Gershenfeld 1993] was a very active area of research and the application of the nonlinear predictors for the synthesis of musical time series was seen as a very interesting extension of the group activities.
On of the problems I investigated at the GMD was the question of the estimation of characteristic measures of the system attractor (its dimension and its Lyapunov exponents) from the trained model (A. Röbel 1995a,c). Moreover I investigated into the application of the model for nonlinear prediction of real world time series (A. Röbel 1996), especially the prediction of ozone air pollution levels (van Praagh 1995). With respect to the application of the model to sounds generated by monophonic musical instruments I studied the use of a virtual control variable that did allow separating the time varying attractors of non-stationary dynamical systems (non-autonomous systems) into a sequence of autonomous attractors and I had made some successful experiments with individual notes of musical instruments (A. Röbel 1995b). I demonstrated that the control variable that I had proposed to unfold the non-stationary dynamics into a sequence of evolving attractors could be used for local time scale modification and time reversal (A. Röbel 1995b). I was able to successfully apply the model to speech signals (A. Röbel 1997). It is interesting to note

[1] today the GMD is part of the Fraunhofer Institute for Open Communication Systems (FOKUS)

that a very similar approach for the representation of time evolution has recently been used to represent time evolution of musical instrument timbre in our work on modelling the timbre space of musical instruments (see section 7.2).

2.2 Assistant professor for communication sciences

After about one and a half years at the GMD I left to join the department for communications science of the TU Berlin as an assistant professor. The position included the responsibility to teach the field of digital audio signal processing to the students of the communications science department and the sound engineering students from the high school of arts of Berlin. When I entered the department the research interests were very diverse but mostly kept a strong connection to sound engineering. The activities covered room acoustics, 3d sound field reproduction, multi channel recording, data compression, multi media, psychoacoustics and information theory. Accordingly, during my time at the electronic studio of the department of communication science, I started to work on more traditional sound modelling techniques, like sinusoidal models, and investigated into advanced signal processing techniques related to blind signal deconvolution, AD conversion and noise shaping, sound synthesis and sound transformation.
With respect to the learning of dynamic models I tried to improve my understanding of the behaviour of the model when applied to non-stationary dynamics. This was essential for example for the understanding of complete musical notes. I followed the procedure for the interpolation of 2 source vector fields that had been described in [Mettin and Mayer-Kress 1996] and which I could directly apply to the trained dynamical models to generate new interpolated attractor dynamics. It turned out that the interpolation procedure that I had implicitly used to model non-stationary dynamics was capable to generate attractors with intermediate sound characteristics (A. Röbel 1998a,b, 1999b). The main inconvenience was the fact that attractor changes that could be related to changes of the phase relations of the individual partials of an harmonic sound could during interpolation give rise to both: amplitude and frequency variations. Further studies revealed another problem that was related to the way the model would represent weakly inharmonic sounds. This problem was related to the fact that the modelling of time varying dynamics of musical sounds did suffer from the inherent ambiguity between high-dimensional attractors and time varying low-dimensional attractors (A. Röbel 2001c).
The problems related to the training of dynamical models for highly non-stationary systems, notably musical instruments, led me to search for representations with stronger prior constraints, that would require less training to track time varying dynamics. The most prominent candidate were non-stationary sinusoidal models. Accordingly, I started to investigate non-stationary signal models based on a sinusoidal representation [McAulay and Quatieri 1986; Xavier Rodet 1998; X. J. Serra and Smith 1990]. The model I was developing was integrating a set of continuous parameter contours and was designed to estimate the time varying parameter contours directly (A. Röbel 1999a). One of the main hypothesis was that the fact that the parameter contours can be estimated over long segments, the separability condition that is required for example for the application of reassignment methods [Auger and Flandrin 1995] could be relaxed. The investigation of non-stationary sinusoidal models was later continued at IRCAM and will be discussed in section 6.1.

2.3 Research scholarship at CCRMA

After about 4 years of work as assistant professor at the TU I had the possibility to take a sabbatical and I chose to do this in form of a research scholarship at CCRMA (RP1). During this scholarship I investigated into fundamental properties of the estimation of the sinusoidal model with continuous parameter contours, and developed the theoretical foundations that allowed creating an experimental system (Wright et al. 2000) that was finalized later at IRCAM[2].

[2] see section 6.1

CHAPTER 3

RESEARCH AT IRCAM

Before the subsequent chapters will describe the research that I have performed or directed during the time I have worked at IRCAM, the present chapter will describe the context of my research activities at IRCAM. First, this chapter will describe the research objectives that were central to nearly all the different research topics I have been working on during the last 13 years. And then it will provide a sort of technological context that is a summary of the technologies that had been developed at IRCAM by the time when I started to work there and that would at some point fall under my responsibility.

3.1 Research objectives

Scientific research at IRCAM has a clear mission: support the artistic and creative projects in music. This mission is achieved through different means. On one hand, composers and musical assistants (the RIM or *réalisateur en informatique musical*) may request new solutions for problems they encounter in their projects, on the other hand, scientific research may open new means (algorithms for conversion of speaker gender) that evoke new creative ideas and may act as a sort of inspiration for composers. This constant two way interaction with creative users, has important consequences for the scientific research at IRCAM. First, research results have to be implemented in a form that makes them usable by the creative users at IRCAM. The algorithms to be developed have to deliver high quality results not only in the controlled situation of a scientific experiment but also in situations where creative users try to explore the limits of the methods. The applications to be developed should preferably be self contained and independent of additional software such that Matlab implementations are generally not sufficient. As a consequence I invested a non negligible part of my work into development activities and these activities will be described briefly in chapter 10. Second, the algorithms that are developed are constantly evaluated and used in different contexts and often under extreme conditions by users that are rather critical with respect to sound quality. The exposure to concrete use cases generates valuable feedback such that algorithms can be continuously improved.
The analysis/synthesis team of IRCAM was directed by Xavier Rodet. This team is specialized in sound signal processing covering sound analysis and synthesis methods, signal models and signal representations, as well as specific problems related to speech or music signal analysis or transformation. When entering the analysis/synthesis team I became responsible for sound signal transformation and supporting technologies (phase vocoder and sinusoidal modelling, fundamental frequency estimation, STFT representation).

3.1.1 Sound transformation with intuitive controls

The term *sound transformation* covers many signal processing technics that are outside the scope of the research performed in the analysis/synthesis team: gain change, filtering, mixing, limiting, are good examples. The sound transformation problems that are treated in the following are characterized by the fact that the specification of the transformation is not given in terms of time domain or frequency domain operators, but in terms of a high-level descriptor, like sound duration. The transformation is then expected to change the specific descriptor respecting as much as possible the physical characteristics of the sound source without using an explicit physical model.
This kind of sound transformations leads to intuitive controls in the sense that all human beings develop their understanding of sounds by means of experience with real world sound sources. We all know how it sounds if a flute or a guitar is played slow or fast, more or less loud. An implementation of these kind

of transformations however is very complicated, because sound sources do not react linearly to changes of control parameters. As a simple example consider the flute and guitar note mentioned above. Playing longer notes on a flute will generally require non constant time stretch factors during attack and sustain. A similar problem exists for pitch changes. For physically coherent pitch modification the timbre of the instrument needs to be taken into account[1].
The perceptually relevant parameters of the sinusoidal components that have to be modified for physically coherent signal transformation are well known (instantaneous amplitude and frequency, and for certain signals the phase relations between sinusoids). These parameters can be exposed by means of a spectral domain signal representation. For noise components, the wind noise of a flute for example, a complete description of the perceptually relevant signal parameters was not known by the time I started to work on sound transformation. This constraint knowledge is an important reason for the fact most of the research on sound transformation with high-level controls has focussed on making things right for the sinusoidal components, and as well a reason for the fact that most of the research I have done in the last 13 years is related to sinusoidal components.
It is only a few years ago that perceptual experiments have revealed the set of statistical signal descriptors that are relevant for perception of noise and sound texture signals [J. McDermott et al. 2009; J. H. McDermott and E. P. Simoncelli 2011] and accordingly, research has now started to take into account these statistical descriptors for transformation of sound textures and noises [2].

3.2 Technological context

At IRCAM there did exist a number of sound transformation technologies that were either based on additive signal models or on the phase vocoder [Dolson 1986; Flanagan and Golden 1966]. The phase vocoder implementation at IRCAM was called *SuperVP* (Super Vocodeur de Phase) and it was extremely well received by the users (visiting composers, musical assistants (RIM) and the community of the IRCAM Forum). This success was partly due to the graphical user interface *AudioSculpt* that allowed visualization and editing of the spectral representation of the phase vocoder program as well as due to the numerous sound transformations that were available. The sound quality after transformation however was not considered sufficient. One of the main problems were severe timbre transformations whenever significant time stretching was applied [3].
The additive signal model was also used at IRCAM and improving this model was part of my responsibility as well, this however, was less important because compared to the state of the art the implementation existing at IRCAM was considered very efficient [Freed et al. 1992; G. Peeters and X. Rodet 1999; Xavier Rodet 1998; X. Rodet and P. Depalle 1992]. Unfortunately, the adaptive additive model I had developed so far presented a number of drawbacks for the users at IRCAM. First the optimization of the non-stationary adaptive sinusoidal model that I had developed was extremely time consuming and it was clear that the algorithm would always remain costly. Second it was considered to require a relatively high investment into research and development to obtain a robust and practically useful application. Third it was not evident that, compared to existing algorithms, the algorithm would provide significant benefits with respect to possible sound transformations that would justify the investment. Therefore, while I was asked to continue research and development of the existing algorithms for estimation and synthesis of sinusoidal models the research into the adaptive algorithm was considered to have lower priority,
Besides signal modelling and transformation there was research to be done in the area of signal analysis. These analysis were often motivated by the use of the results for sound transformation and manipulation. An essential problem was the estimation of the fundamental frequency, which is one of the perceptually most essential parameters of musical sounds. Accordingly, at IRCAM there did exist a long history of research related to fundamental frequency estimation [Doval and X. Rodet 1991, 1993] and I was expected to continue this research.

[1] see section 7.2
[2] see section 5.2 for related research efforts.
[3] see section 4.1

CHAPTER 4

SOUND PROCESSING USING STFT REPRESENTATION

Related projets

(SW2) A. Roebel et al. (2000). *SuperVP: command line tool and c++ library for audio treatment in real time or non real time based on an extended phase vocoder*. A. Roebel: Scientific direction and software development since 2000, F. Cornu: software development since 2007, P. Depalle: initial version before 2000.

This chapter will describe my research activities related to sound analysis and transformation using either an STFT representation or, derived from the STFT, a phase vocoder representation [Dolson 1986] of the sound signal. The work on phase vocoder based sound manipulation has been initiated at IRCAM by P. Depalle who deserves credit for many of the initial ideas and fundamental concepts that were leading to the first implementation of the phase vocoder based sound manipulation software *SuperVP* in 1993.
When I took responsibility of the project in 2000, the phase vocoder implementation did support time varying pitch and time stretch operations, as well as numerous frequency domain filters. Moreover, there was a nonlinear extended cross-synthesis module operational. As mentioned already, the sound quality obtained for time stretching operations was not satisfactory.
The improvement of the phase vocoder algorithm was my first research project at IRCAM. At the beginning I reviewed the mathematical foundations for STFT based signal processing and the phase vocoder[1]. Based on the understanding of the phase vocoder algorithm as an implicit sinusoidal model the fundamental weaknesses of the algorithm can be predicted easily. For the processing and transformation of quasi-stationary sinusoidal components one can expect to achieve very high quality results as soon as the phase synchronization between neighbouring bins is established. But for signal components that are not (or not well) covered by the sinusoidal model, notably onsets or noise, or sequences of pulses that require specific phase relations to keep their perceptual properties, sound transformation can be expected to result in artefacts. One of the central objectives of my research at IRCAM was to extend the phase vocoder based sound processing such that it could handle as many of these sound classes with very high quality. The results I have obtained so far will be discussed in the present chapter. Ongoing research activities will then be discussed in chapter 5. The presentation will cover the following topics:

- Sinusoids: According to the findings in [J. Laroche and Dolson 1999a; Puckette 1995], the spectral STFT bins contributing to a single sinusoid cannot be treated independently. This problem will be discussed in section 4.1,
- Transients: Parameter changes of the different sound sources that take place within a single analysis window require phase synchronization between all STFT bins that are affected by the parameter change. An extension of the phase vocoder coherently dealing with these kind of changes is described in section 4.2,
- Pulsed excitation: Speech signals are generated by a quasi periodic sequence of pulses that gives rise to a quasi harmonic set of sinusoids with a rather particular relation of phases between these sinusoids

[1] These mathematical foundations have later been published in form of lecture material (A. Röbel 2006c) following my stay as Edgar-Varèse guest professor at the Technical University of Berlin.

[Quatieri and McAulay 1992]. Neglecting these phase relations leads to more or less strong artefacts in transformed speech sounds. My work on the shape invariant signal transformation in the phase vocoder is described in section 4.4.

- Noise and sound textures: Because of the fact that perceptually these components are less important for most music instruments and for speech the problems related to perceptually transparent transformation (time stretching) of these sound components have been neglected for a rather long time. After the most important problems with sinusoidal components and transients have been solved, the manipulation of sound textures and noises has triggered dedicated research efforts that are described in section 5.2.

Besides the research directly related to extending the sound quality obtained with the phase vocoder another line of research is related to supporting methods. This line of research covers:

- Sinusoids/Noise classification: The distinction of noise and sinusoidal components is an important pre-processing step that is essential for the shape invariant signal transformation described in section 4.4. The algorithm for detection of sinusoidal components is described in subsection 4.3.1.
- Parameter selection: selecting the proper window size remains one of the major problems for the use of the STFT signal processing algorithms. As a solution to this problem I have initiated research on algorithms with automatic determination of the window size which will be described in section 5.1.
- Frequency domain transposition: The very high quality and rather efficient implementation of the phase vocoder in SuperVP has triggered considerable interest in using SuperVP in real time. To reduce the inherent latency related to the time stretching and resampling based approach to transposition an frequency domain implementation has been developed that is described in section 4.5.

Closely related to the present topic is the section about envelope estimation. However, due the strong relation between spectral envelop estimation and source filter modelling this topic will be discussed in the chapter 7.

4.1 Intra sinusoidal phase synchronization

As has been mentioned before, time stretching sinusoidal components with a classical phase vocoder does not lead to perceptually convincing results. This can be explained by the fact that the individual bins of the STFT that contribute to the same sinusoidal component are treated independently [Puckette 1995]. As a result of incoherencies due to inevitable analysis and processing errors, the phase coherence of the different bins will be lost over time. This desynchronization will result in cancelation of the contributions of the different bins such that the amplitude of the sinusoidal components will not be preserved. Different solutions to this problem had been described. The method proposed in [Puckette 1995] relies on averaging phases of neighbouring bins, but does not work for all cases. The method described in [J. Laroche and Dolson 1999a; Mark Dolson 2000] selects a master bin within each spectral peak, updates only the master bin with the phase vocoder algorithm and preserves the phase differences of all other bins in that peak. This can be seen as an implicite implementation of the FFT based synthesis of sinusoids [P. Depalle and X. Rodet 1995; Goodwin and X. Rodet 1994], where sinusoids are constructed from the spectral peaks. In an experimental investigation, I found that while the computational costs of the proposed solution are very small, it does not avoid synchronization problems completely. Therefore, I adopted an alternative approach that consisted of an refined selection of the bins to synchronize for a spectral peak and a reconstruction of spectral peaks from a sinusoidal model following the approach in [Goodwin and X. Rodet 1994]. Initialy only stationary sinusoidal eaks wer considered but the approach was later extended to take into account the frequency slope of the underlying sinusoids as described in section 4.5. Due to the fact that the number of bins to be treated with the phase vocoder algorithm is significantly increased, the new method is computationally a bit more costly than the method proposed in [J. Laroche and Dolson 1999a; Mark Dolson 2000]. The results obtained with this new bin synchronization strategy however, were perceptually very satisfying. Similar to the method presented in [J. Laroche and Dolson 1999a] time stretching of chirp signals with

the new method does only have a very minor effect on the chirp amplitude, such that the phase vocoder implementation behaves nearly equivalent as an explicit sinusoidal model.

4.2 The phase vocoder algorithm for transient signal segments

Industrial licences

(LI2) A. Roebel and X. Rodet (2004). *MakeMusic*. Library for time stretching and pitch shifting with transient preservation for music signals,
(LI4) A. Roebel (2005). *Roni Music*. Library for time scaling and pitch shifting with transient preservation for music signals,
(LI5) Axel Roebel (2008). *NeoCraft*. Library for transposition and time-scaling with transient preservation for music signals
(LI8) Axel Roebel (2009a). *MXP4*. Library for time scaling and pitch shifting with transient preservation for music signals
(LI9) Axel Roebel (2009b). *UniversSons - MachFive 3*. Library for time scaling and pitch shifting with transient preservation for music signals
(LI12) Axel Roebel (2010b). *IRCAMTools-TRaX*. Development of a professional audio plugin for music and voice transformation with high level controls and high sound quality, in collaboration with the French software development company Flux in Orléans
(LI11) Axel Roebel (2010a). *OhmForce*. Use of supervp library for sample precise time scaling with transient preservation for music signals

The synchronization of the bins contributing to individual sinusoidal components was a significant step towards a general high-quality time-scaling algorithm. The improvements obtained for the transformation of quasi stationary sinusoids were very satisfying, but at the same time, they put the other problems of the phase vocoder algorithm into focus. This concerned notably the problem related to the processing of transient signal components. In the present context transient signal components are characterized by changes of signal characteristics within the time duration of the analysis window. The most important and problematic example are note onsets, that on one hand are perceptually very important and that on the other hand are completely destroyed in the classical phase vocoder algorithm if the onset takes place within a single analysis frame. Accordingly, the literature on transient signal preservation often treats the terms transient and onset as equivalent, while strictly speaking the transient signal class contains not only onsets but also many more cases (pitch transitions).
The detection, extraction and independent processing of transient signal components had become an important question and related research results began to appear around 2000. Some of the research was aiming to establish onset preservation strategies for the phase vocoder or similar approaches [Bonada 2000; Duxbury et al. 2001, 2002]. The basic idea was to use onset detection algorithms to determine transient segment of the signal, to reinitialize phases at the beginning of these time segments, and to suppress any transformation for a short segment to avoid artefacts due to phase modifications in the transient segments. Other researchers were aiming to improve transient representation in explicit sinusoidal models like for example the sinusoids and noise and transients model in [Levine and Smith 1998], or the Loris system proposed by [Fitz et al. 2000].
At IRCAM, the main interest was related to improve the phase vocoder algorithm such that artefacts related to the processing of note onsets would be reduced. The analysis of the algorithms proposed in [Bonada 2000; Duxbury et al. 2001] revealed two major problems. The first was related to the fact that the transient classification was using only temporal information. A time segment could therefore be only either transient or non transient and this prevents proper handling of time segments containing onsets and stationary notes at the same time. Another drawback of the proposed extensions was the fact that a local modification of user control parameters was required leading to complicated compensation strategies to preserve the overall transformation requested by the user.
The objective for this research was therefore, to find a method that would allow an appropriate treatment of note onsets and, at the same time, minimize adverse effects on stationary components happening syn-

chronously with the onsets events. Moreover, the desired algorithm would not require the modification of user supplied transformation parameters proposed in [Bonada 2000; Duxbury et al. 2001].
If no modification of transformation parameters is allowed then onset preservation comes down to reinitialization of the phases of all spectral bins that are part of a transient component at an appropriate position. Accordingly, two problems had to be solved:

1. detection: an appropriate transient detection algorithm had to be developed that would detect a transient event and that would establish a spectral mask that allows separating transient signal components.

2. preservation: the appropriate time for phase reinitialization had to be found that would allow optimal reconstitution of the original signal onset.

4.2.1 Onset position

To find an appropriate onset preservation strategy I used as test case a restricted class of transient signals consisting of sinusoids with onsets that are formed by a linear amplitude increase followed by saturation. Concerning the optimal position for phase initialization there exist two arguments that both, a priori, require the phase reinitialization to take place if the onset is close to the window centre.
The first reason concerns the exact reconstruction of the onset during synthesis. Because the transient will be reproduced without any error only in the frame in that the phases of the transient bins are reinitialized the reinitialization should take place when the impact of the reconstructed transient on the output signal is largest. Due to the existing analysis and synthesis windows the maximum impact will be achieved if the onset position is close to the centre of the analysis window.
The second argument is concerned with the position of the onset in the transformed signal. Reinitialization of the phases will produce the transient at the very same position where it was located in the analysis window. During time stretching the frame position will reflect the transformed time scale. The signal evolution within the analysis frame however, is not adapted and any existing offset of the onset from the frame centre of the original frame would therefore require additional modifications to adapt the onset position to the transformed signal. Unfortunately, the required adaptation may be very complicated or even impossible notably when during a time stretching operation of the signal the offset of the onset from the frame centre needs to be increased. In this case it can happen that part of the onset has to be moved outside of the frame, which would require very complicated operations. To avoid the need to reposition the onset the phase reinitialization should take place when the onset position is in the centre of the window.
A note has to be made regarding the onset position. The appropriate position to be used is the perceived onset position. Unfortunately, perceived onset positions can not be determined without ambiguity directly from the signal itself. This is true already for the restricted class of onsets that was studied here, but even more for real world note onsets that have a much more complicated time structure. In general it can be assumed that the perceived onset time depends on the form of the onset, the spectral content of the noite but as well on the musical context. Accordingly, in a recent study the perceived attack time was modelled by means of a probability distribution instead of a single position [M. J. Wright 2008]. Given that profound problems exist for the precise estimation of perceived attack times, I decided to simplify the problem by means of using objective measures to represent the time position of the transient events.
The experimental investigation of the impact of the time position of the phase reset for the restricted class of onsets revealed that the onset is reconstructed with a rather small error, when the phase reinitialization is done at the moment when the centre of the linear ramp is in the centre of the analysis frame (A. Röbel 2003a,b).

4.2.2 Detection of transient components

There exist many approaches to detect attack transients [Bello et al. 2005]. Most of the algorithms known by 2003 where based on the evolution of the signal energy (spectral flux) in frequency bands [Bonada 2000; A. Klapuri 1999; Levine 1998; Masri and Bateman 1996]. Some of these algorithms [Bonada 2000; A. Klapuri 1999] did use psychoacoustic arguments for the selection of bandwidths and thresholds. Other

detection functions were based on variations in the frequency trajectories of individual bins [Duxbury et al. 2001, 2002]. Some of the algorithms that require knowledge about the future signal evolution [Boeck et al. 2012; Lacoste and Eck 2007] impose inacceptable latency into the algorithm and can therefore not be accepted for on the fly detection of transient components in the phase vocoder. There have been very few algorithms that tried to determine precise frequency masks of onset events [Duxbury et al. 2001; X. Rodet and Jaillet 2001]. These algorithms did use detection functions working on individual bins. Detection thresholds are generally fixed, but in some cases the detection threshold takes into account the signal properties [Duxbury et al. 2001, 2002]. Since 2003 new algorithms with new detection functions based on variants of the spectral flux have been proposed [Bello et al. 2004; Boeck et al. 2012; Dixon 2006; S. Hainsworth and M. Macleod 2003], none of these is aiming to detect the time frequency positions of individual transient spectral components such that a separation of transient and non transient components can be performed in a signal frame.

While high spectral resolution was desirable for the transient masks the direct use of individual spectral bins [X. Rodet and Jaillet 2001] seems not necessary due to the fact that DFT bins are generally highly correlated. Moreover, an individual bin may give only limited information about the time evolution of the related signal component. A classical example is a chirp signal. The energy of a chirp signal is located at one end of the analysis window for one part of the spectral peak and on the other end of the analysis window for another part of the spectral peak. Given the algorithm to be developed aims preservation of note onsets a chirp signal, or a part of it, should not be confused with an onset. Spectral peaks represent the smallest spectral components that can be readily accessed, and they are not subject to the chirp/onset confusion mentioned before. Accordingly, I decided to use individual spectral peaks as the smallest spectral units to be used for the creation of onset masks.

A rather interesting approach for the solution of the detection problem was the technique proposed in the Loris system [Fitz et al. 2000]. There the problem of the representation of sinusoidal attacks was solved in a rather elegant manner by means of deriving the information about the sinusoidal onset locations directly from individual spectral peaks using the reassignment technique [Auger and Flandrin 1995]. The reassignment technique allows to calculate the mean-time [Cohen 1995] of the signal related to any sub-band of a DFT spectrum and in the Loris system this technique was applied to individual spectral peaks. Spectral peaks representing onsets can be detected by means of detecting peaks with mean-time above a threshold right of the window center. Unfortunately, many noise peaks exhibit mean-time above the threshold and therefore the transient detection cannot be based on the evaluation of individual peaks. For onsets of musical notes or drumbeats, however, there will generally exist multiple transient peaks with similar mean-time. Accordingly, I developed a statistical model that allows to detect the significant increase of te number spectral peaks with mean-time above the threshold and that allows to distinguish the appearance of transient peaks due to background noise from the sudden increase of the number of transient peaks due to a note onset event.

4.2.3 Onset preservation

Combining the two results: the algorithm for the detection of transient events and the algorithm for reseting the phase of transient events in polyphonic sounds requires the determination of the time position for that the phases of the onset events should be reset. Given that for the onset detection we already use the peak mean-time it is most straight-forward to use the mean-time as well for the determination of the time at that the pashes should be reset. Given that the spectral information related to the onset cannot be used before the phases of the spectral peaks contributing to the onset are reset, the signal present in front of an onset is artificially continued to avoid systematic silence in front of onsets.

4.2.4 Results

The algorithms that have been discussed in this section have been evaluated from two perspectives. The first one is the detection of onsets and the second one is the extraction and preservation of onsets and other transients from sound signals.

Onset detection

The algorithm was not specifically targeting the detection of note onsets as any abrupt change of the signal characteristics (fingering noise on guitar strings) should be detected to prevent denaturation of the signal during time stretching, it nevertheless turned out to provide state of the art performance when applied to onset detection problem. The onset detection algorithm has been evaluated repeatedly during MIREX evaluation campaigns [IMIRSEL 2005, 2006, 2007, 2009, 2010, 2011]. The algorithm turned out to perform very well and the best results I have obtained in 2011 belong to the best results reported over the MIREX history especially when confined to the category of algorithms that are causal.

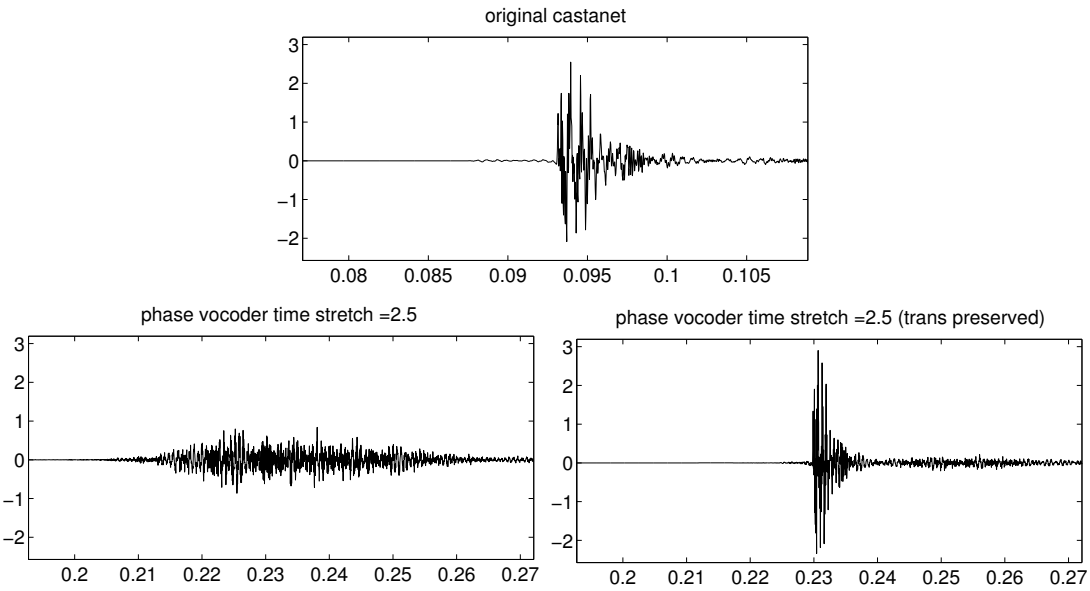

Figure 4.1: Comparison of an original castanet sound (top) with the same sound time stretched by a factor 2.5 obtained with standard phase vocoder (centre) and the new transient preservation algorithm (bottom).

Transient preservation

Processing attack transients in the phase vocoder with the proposed algorithm results in significant improvements of attack quality. The algorithm has been integrated into SuperVP the phase vocoder application of IRCAM and is therefore available in all applications that use the SuperVP library or executable section 10.2. Due to the fact that the algorithm is selectively processing spectral peaks it is well suited for processing polyphonic sounds. To give a visual idea of the effect of the transient preservation I have chosen a monophonic castanet sound. The upper part of Figure 4.1 shows the time signal of a single beat within a sequence of castanet sounds. Beneath the result that has been obtained after time stretching the signal with a standard phase vocoder by a factor of 2.5 is shown. The destruction of the attack event is obvious. At the bottom of the figure the same signal has been time stretched by the same factor with transient preservation switched on. The attack is preserved and the sound characteristics of the attack are very close to the original attack. Some sound examples demonstrating the transient preservation and transient extraction capabilities of the algorithm are available online (A. Lithaud et al. 2008, see examples 5. and 6.).
The transient processing algorithm proposed above has been evaluated in a listening test in [Grofit and Lavner 2008]. Grofit and Lavner compared the results obtained with our algorithm (A. Röbel 2003a) with the results produced according to [Bonada 2000] and there own time domain based algorithm. The signals that have been used are a weakly polyphonic note sequence played with a pipa[2], an extract of polyphonic melodic rock music without vocals, an extract of a mildly polyphonic electric guitar and an extract of Bob

[2]chinese plucked string instrument with four strings

Dylan playing *Blowing in the wind* using a western guitar a harmonica and singing voice. As expected both frequency domain algorithms consistently and significantly outperformed the time domain algorithm in all cases. When comparing the two frequency domain algorithms one finds that for three out of four signals the algorithm described above received the highest ranking. Bonada's algorithm obtains better performance only for the Dylan extract. Because the Dylan song is the only sample with voice and because voice transformation with spectral domain models is problematic (see section 4.4) we conjecture that the phase vocoder implementation of Bonada may have less problems with voice. It is unclear, however, what may be the reason for this advantage.

4.3 Classification of sinusoidal and noise components

Related projets

> (PD3) Miroslav Zivanovic (2003). *Detection, estimation and extraction of non-stationary sinusoids in noise: application to musical signals*. Visting PostDoc Researcher at IRCAM, Jan. 2003 - Juin. 2003,
> (PhD2) C. Yeh (2008). "Multiple Fundamental Frequency Estimation of Polyphonic Recordings". Director X. Rodet, supervision A. Röbel. PhD thesis. Université Paris 6 (UPMC),
> (PD4) Miroslav Zivanovic (2006). *Improving state of the art strategies for automatic detection and classification of signal components into sinusoidal and noise components*. Visting PostDoc Researcher at IRCAM, Sep. 2006 - Fev. 2007.

The research on transient preservation had 2 important results. First, it demonstrated that it was possible to detect, and extract transient signal components in polyphonic signals. And based on this detection an algorithm had been devised, that allowed to significantly improve the perceived quality of time stretched sound signals with note onsets. Given that transient signal components had been made accessible directly in the STFT spectra the next question that emerged was related to the detection of the two other classes of signal components (sinusoids and noise) in the DFT spectra. At first this was a rather fundamental research problem not related directly to any application, but as will become clear later, this research had considerable impact in at least the two areas: speech transformation (section 4.4) and polyphonic fundamental frequency estimation (section 8.2). A considerable part of the research in this area has been done in collaboration with M. Zivanovic, who came as a Post doc researcher to IRCAM twice (PD3; PD4). Another part of this work has been done in collaboration with by C. Yeh in the context of his PhD thesis on the estimation of fundamental frequencies in polyphonic sounds (C. Yeh 2008; C. Yeh and A. Röbel 2006a,b).
The problem to be solved was to distinguish sinusoidal and noise components such that these components could be separated in a DFT spectrum and then individually treated. Unfortunately, as there are many different types of signal representations based on different classes of sinusoids the distinction between sinusoidal and noise components is somewhat ambiguous. This can be seen easily by means of considering the STFT representation of a signal. The STFT uses a superposition of a set of time windowed stationary sinusoidal components to represent arbitrary signals. Accordingly, a noise signal can be represented without error by means of a superposition of sinusoidal components. This is similar as the ambiguity related to the representation of an periodically AM modulated stationary sinusoid that can equivalently be represented by means of a set of stationary sinusoids with appropriate amplitudes, frequencies and phases. Given these ambiguities it is important as first step to properly define the two signal classes to be distinguished. The basic idea is similar to the approach that was used for the transient signals. There the transient signal class was defined in relation to the length of the analysis window. Similarly for the present case the ambiguity can be resolved by means of establishing additional constraints that are based on the resolution for the Fourier representation as follows: A sinusoidal component should be isolated and resolved as a individual spectral peak in the DFT spectrum. This constraint implies two important characteristics of the class of sinusoidal components:

1. A sinusoidal component is required to have sufficiently slow frequency and amplitude modulation to give rise to an individual spectral peak.

2. The amplitude of a sinusoidal component has to be sufficiently above the background noise and other sinusoidal components in the respective frequency band.

Given the additional constraint the objective then becomes to find spectral peaks that represent windowed sinusoids with limited amplitude and frequency modulation. The fact that again spectral peaks are used to segregate the signal spectrum has the additional benefit that the sinusoid and noise classification can be combined easily with the existing transient/non-transient classification.
A study of existing methods revealed that despite the fact that sinusoidal modelling and sinusoidal parameter estimation is a technique that is widely used for signal processing the distinction between sinusoidal and noise components in a DFT spectrum are considered in only very few publications [S. W. Hainsworth et al. 2001; Lagrange et al. 2002; Thomson 1982]. The majority of applications for sinusoidal modelling are based on the classical approach that estimates the complete sinusoidal trajectories and classifies sinusoids according to properties of the trajectories [X. J. Serra and Smith 1990]. The formation of the complete sinusoidal parameter trajectories, however, requires a high latency, which is a problem for real time applications. Moreover it adds the burden of trajectory forming to applications like the phase vocoder that otherwise don't require this processing step. Even if today there exist a few methods for the detection of non-stationary sinusoids from individual frames, all these algorithms require either the estimation of sinusoidal parameters including frequency and amplitude slopes, [Lagrange et al. 2002; Wells and D. Murphy 2007, 2010] or the use of multiple dedicated analysis windows [Thomson 1982; Wells and D. Murphy 2010], which makes the detection rather costly.

4.3.1 Sinusoidal signal class

The first step in this research was to precisely define the two signal component classes that would be considered for detection. It is common for sinusoidal models to consider sinusoids with slowly varying amplitude and frequency parameters [McAulay and Quatieri 1986; X. J. Serra and Smith 1990], (A. Röbel 2006a). For an investigation into the properties of the spectral peak classes, however, this requirement is not sufficient. To completely define the space of sinusoidal components we had to select concrete limits of the amplitude and frequency modulation rate and depths, and we had to specify a concrete form of the modulation laws. Following the discussion in (M. Zivanovic et al. 2007, 2008) we selected a sinusoid with sinusoidal amplitude and frequency modulation embedded in white noise as our reference sinusoid signal. Accordingly we used the following mathematical representation for our class of sinusoidal signals

$$x(n) = (1 + A_{AM}\cos(\Omega_{AM}n + \beta))\cos(\omega_0 n + A_{FM}\sin(\Omega_{FM}n + \alpha)) + r(n) \tag{4.1}$$

where $r(n)$ represents the additive white Gaussian noise. A_{AM} and Ω_{AM} are depth and rate of the amplitude modulation and A_{FM} and Ω_{FM} the respective values for the frequency modulation. The modulation parameters need to be limited such that the DFT spectrum contains its dominant peak around the sinusoidal frequency. The limits are discussed in (M. Zivanovic et al. 2007, 2008). α and β are arbitrary phase offsets that allow to control the phase relation between frequency and amplitude modulation and ω_0 is the centre frequency of the sinusoid which should not effect the descriptors of the sinusoidal class.

4.3.2 Descriptor definitions

To avoid any costly analysis in this study we have used only peak descriptors that don't make use of any parameters related to a sinusoidal model. The list of descriptors that were used can be found in (A. Röbel et al. 2004; M. Zivanovic et al. 2004), they all describe properties of the signal related to the spectral peak. Here I will discuss only the normalized peak bandwidth descriptor B_L. The mean frequency $\bar{\omega}$ and the bandwidth B give a rough idea of the concentration of the spectral density along the frequency grid. Considering L to be the number of samples in the spectral peak then the normalized bandwidth descriptor B_L can be defined as:

$$\bar{\omega} = \frac{\sum_k k|X(k)|^2}{\sum_k |X(k)|^2}, \tag{4.2}$$

$$B_L = \frac{B}{L} = \frac{\sum_k (k-\bar{\omega})^2|X(k)^2|}{L\sum_k |X(k)^2|}. \tag{4.3}$$

where the summation is done over all the bins in the spectral peak under investigation.

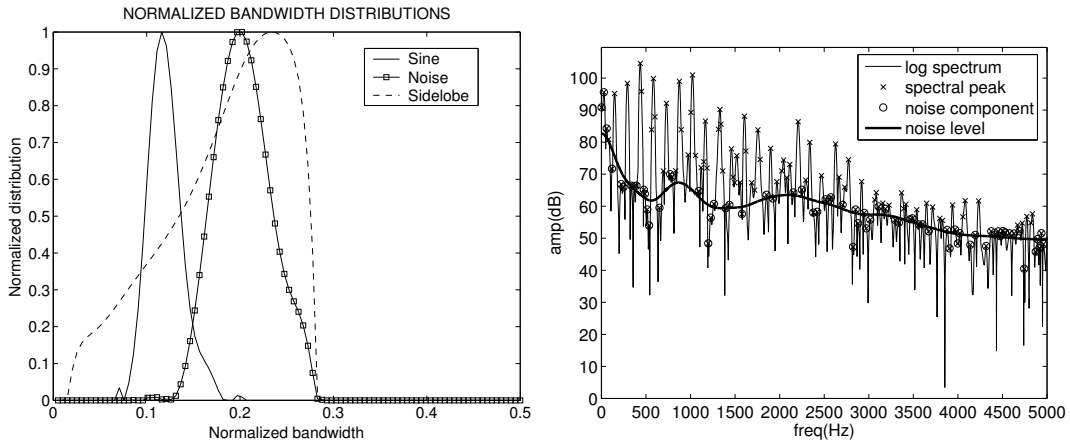

Figure 4.2: Results for sinusoids and noise classification. Distributions of normalized band width descriptor B_L (left) and noise floor estimated from a spectrum of polyphonic music (right).

For deriving the classification thresholds for the descriptors we did rely on a worst-case signal. The related test signal is a single AMFM-sinusoid in noise where both frequency and amplitude change in a sinusoidal manner. To resemble natural vibrato signals, the period of the frequency modulation is two times the period of the amplitude modulation. The characteristics of the test signal are:

- for amplitude modulation: modulation index 0.5,
- for frequency modulation: 200 Hz of frequency deviation.

The analysis window is a 50ms Hanning window and the frequency modulation period is 100ms. For calculating the DFT we use 4096-point FFT with the sample rate being 44100Hz. This scenario roughly reproduces the analysis conditions for the tenth harmonic of a 333Hz pitch tone under half tone vibrato extent. The performance depends on the level of the background noise. Here we will present the two cases of no noise, and peak SNR of 25db. Peak SNR represents average ratio between the maximum amplitude of the sinusoidal signal and the mean noise level.
The distribution of the B_L descriptor for the peak classes that have been obtained for the test signal is shown in Figure 4.2. For the sinusoidal distributions the descriptors were applied only to the largest peak in the spectrum for a total of 1100 time frames. The noise distributions were obtained by analysing all the peaks in the DFT of a white noise signal. To derive the sidelobe distributions we analysed all the sidelobe peaks of a stationary noise-free sinusoid. For ease of comparison all distributions are displayed normalized such that their maximum value is equal to one. As the threshold levels we are going to determine aim to preserve fractions of the distributions, this normalization does not affect the results.
From Figure 4.2 one sees that the B_L distributions for noise and the sinusoidal class that is considered have only very small overlap. An interesting question that comes up is what the descriptor B_L measures to achieve this nice performance. A first element to understanding this descriptor can be derived from the bandwidth formula given in [Cohen 1995, p. 16]. For any signal $s(t) = A(t)e^{(\phi(t))}$ with Fourier transform $S(w)$ being normalized such that $\int S(w)^2 dw = 1$ the bandwidth B is given by

$$B^2 = \int (w - \bar{w})S(w)^2 dw = \int (\frac{A'(t)}{A(t)})^2 A(t)^2 dt + \int (\phi'(t) - \bar{w})^2 A(t)^2 dt. \tag{4.4}$$

This shows that for continuous signals (and similar for discrete signals) the bandwidth is related to the amplitude variation and the variance of the phase slope with respect to its mean. If we apply this to the signal related to our peak then we understand that the bandwidth B is related to variation of amplitude and phase slope over time. To understand the normalization by L consider a signal containing two stationary sinusoids with frequency difference L. The bandwidth for this signal will be $B = L/4$ which in fact is the

maximum bandwidth a signal confined within a band of bandwidth L can have. Combining these results we can conclude that the normalized bandwidth descriptor characterises the amplitude and frequency modulation of the signal related to an observed peak with respect to the amplitude and frequency modulation of a signal with maximum bandwidth covering the same band.

4.3.3 Noise floor estimation

The work on the separation of sinusoidal and noise components has continued in the context of the estimation of multiple fundamental frequencies in polyphonic music (PhD2). One of the common errors we had observed during our work on multiple f0 estimation was the fact that F0 were inserted that explained mostly noise components. Therefore, the objective in that case was to detect signal components that should be explained by a fundamental frequency. The fundamental problem here is the fact that sinusoidal components will often overlap each other in polyphonic music. In fact none of the previously mentioned algorithms for sinusoidal component detection does allow to detect overlapping sinusoidal components. To be able to handle this case the constraint for the sinusoidal components would have to be relaxed, because by construction these sinusoids are no longer resolved. Because we did not expect to be able to find features that would allow separating overlapped sinusoids from noise, we changed the strategy and established a two-step approach instead. What distinguishes the overlapping sinusoids from noise in polyphonic music is the fact that these sinusoids are outliers with respect to the amplitude distribution of the noise. To be able to detect these outliers we assume that the noise spectrum follows a Rayleigh distribution. We can then test the amplitude distribution of the spectrum in different bands and compare the skewness of the measured spectrum with the skewness of the Rayleigh distribution, which is approximately $S_{ray} = 0.6311$ (C. Yeh and A. Röbel 2006a; C. Yeh et al. 2010). The basic idea of the algorithm was to first detect resolved sinusoidal components and to remove those from the spectrum using sinusoidal parameters estimated according to [M. Abe and Smith 2005] or (A. Röbel 2008). In the residual spectrum we evaluated all the bands with respect to the skewness and removed outliers in all those bands that had an exceedingly large skewness. For initially large positive skewness removing outliers will generally reduce the skewness. If this is not the case we did consider the band incompatible with a noise hypothesis in which case we consider it to be generated by overlapping sinusoids. The algorithm takes care of non-white residual spectra by means of normalization of the noise by its estimated frequency dependent noise level. An example of an estimated noise floor for a spectrum of polyphonic music is given in Figure 4.2.
While it is clear that the noise floor estimate achieved is very approximate, especially for very high polyphonies, it nevertheless establishes a means to distinguish important and less important parts of the spectrum and it was one of the elements that contributed significantly to our performance in polyphonic pitch estimation.

4.4 Shape invariant processing

Related projects

(MA9) G. Champion (2004). “Application du modele additif shape invariant pour la transformation de la voix”. Rapport DEA Master ATIAM, supervision A. Roebel. MA thesis. Université Paris VI Pierre et Marie-Curie,
(MF2) Axel Roebel and Joshua Fineberg (2006-2007). *Creation of voices for the opera* Lolita *of J. Fineberg*. Transformation of the voice of the main actor into girls singing voices.

Industrial licenses

(LI12) Axel Roebel (2010b). *IRCAMTools-TRaX*. Development of a professional audio plugin for music and voice transformation with high level controls and high sound quality, in collaboration with the French software development company Flux in Orléans,
(LI6) Axel Roebel et al. (2008). *Xtranormal*. Library for voice transformation,
(LI10) Axel Roebel et al. (2010). *Xtranormal*. Library for voice transformation with high level control.

It is well known that the phase vocoder algorithm does not produce high quality results when applied to speech signals. The reason is the loss of the phase coherence of the harmonic sinusoids that will occur when the signal is processed (notably time stretched) without taking care of the phase relationships between the different sinusoids. This problem has been solved for the sinusoidal model in form of a shape invariant transformation proposed in [Quatieri and McAulay 1992]. A similar solution for the phase vocoder was presented in [J. Laroche 2003]. The increased interest in speech signal transformation at IRCAM led me to think about means to implement a shape invariant signal processing mode in our phase vocoder implementation. If possible, the algorithm should not require any additional information, like the fundamental frequency that was required for the algorithm proposed in [J. Laroche 2003].
The first steps of this research was an investigation into shape invariant processing with sinusoidal models (MA9). The idea was to establish a working system that could be used to compare results obtained later with a phase vocoder implementation. Moreover I wanted to study different means to establish vertical phase synchronization in the sinusoidal model. The results obtained were very convincing, and based on the experiences with shape invariant sinusoidal models I started to study means that would allow a modification of the phase vocoder algorithm. The shape invariant treatment tries to preserve the shape of the waveform and therefore it assumes at least a quasi-periodic signal. While this condition does not hold for all speech signals it nevertheless is a reasonable assumption for many speech signals, especially those that are not extremely expressive. For the shape invariant transformation in the phase vocoder we therefore suppose that the signal consists of voiced and unvoiced segments, and that the voiced signal segments are quasi periodic[3].
The phase coherence problem in the phase vocoder, as well as in the sinusoidal models, comes from the fact that frequency estimates of individual partials are used to predict the phase in the transformed signal. Small errors in the frequency estimates lead to slight phase deviation between different harmonics and those will accumulate over time leading to completely desynchronized phases. The method for shape invariant processing in the phase vocoder is very different from the solution proposed for sinusoidal models (A. Röbel 2010b). The standard phase vocoder is assuming non-harmonic sinusoids and in this case all sinusoids have to be updated independently. If the signal is quasi-periodic, then the procedure can be simplified significantly. In this case coherent phase update does not require integration of the frequencies over the complete sound. Instead phase modification can be done similar to basic synchronous overlap add schemes (SOLA) [Roucos and Wilgus 1985]. The principle ideas are then that for subsequent frames an optimal displacement position is determined such that cross correlation between the two frames is maximized. The correlation is done using only the sinusoidal components as determined by the algorithm described in section 4.3. If the correlation coefficient is high then the original phase of the input frame is used to time shift the frame to the desired position. The use of the original frame phase as a start position for the phase adaptation has the very nice side effect that the noise modulation in the high frequency region of voiced frames remains preserved which contributes to the high quality of the sounds produced during time stretching. If the correlation between the sinusoidal components of successive frames is too low, or if there aren't any sinusoidal components, then the standard phase vocoder phase update algorithm is used. While the standard phase vocoder is not designed to handle noise correctly, no better alternatives existed at that time. For noise signals the standard phase vocoder works better than doing overlap add without any phase adaptation. For a more appropriate scheme of noise processing in the phase vocoder I refer to ongoing research described in section 5.2 and W.-H. Liao et al. 2012. The results of that research may allow us in a near future to improve the handling of unvoiced segments in the phase vocoder.

4.4.1 Applications and Results

The shape invariant processing in SuperVP (SHIP) achieves a very significant improvement of the sound quality of transformed speech signals. The effect of the shape invariant transformation on a speech signal is demonstrated in Figure 4.3. In a few perceptual evaluations it has been compared to speech processing with TD-PSOLA [Moulines and Charpentier 1990], STRAIGHT [Kawahara 1997], and sinusoidal models [Y. Stylianou 2001]. Results strongly depend on the transformation and the speech examples but have to be

[3]For more refined speech transformation approaches we refer the reader to the section on glottal source parameter modification section 9.1

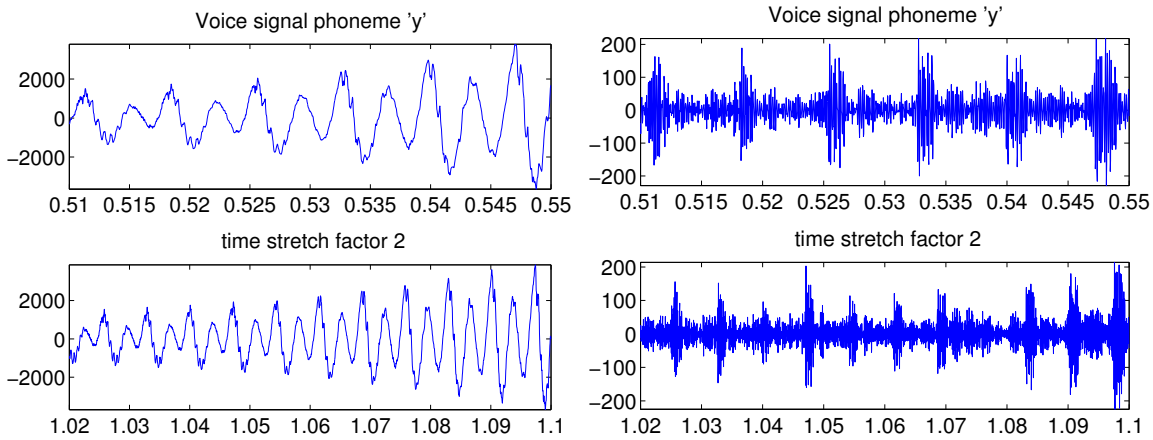

Figure 4.3: Example of a transformed waveform after time stretching by factor 2. The complete waveform (left) and the high frequency band (right) both exhibit clearly the preservation of the waveform.

seen on the background that the SHIP algorithm does no require any further analysis (like pitch markers). A rather weak point of the SHIP algorithm is the fact that the voiced/unvoiced frequency cannot be increased freely leading to too much unvoiced signal components when transposition down is requested. A similar problem has been discussed for FD PSOLA [Moulines and Charpentier 1990]. Transposition up yields generally much better results. The SHIP algorithm has been used in many artistic and commercial film projects that will be discussed in chapter 9.

4.5 Frequency domain transposition

Related projects

(RP2) *Projet ANR - Sample Orchestrator – 2006-2009* (2006). Task 3.1: Enhanced phase vocoder analysis and transformations in real time applications.

Industrial licenses

(LI9) Axel Roebel (2009b). *UniversSons - MachFive 3*. Library for time scaling and pitch shifting with transient preservation for music signals,
(LI10) Axel Roebel et al. (2010). *Xtranormal*. Library for voice transformation with high level control,
(LI12) Axel Roebel (2010b). *IRCAMTools-TRaX*. Development of a professional audio plugin for music and voice transformation with high level controls and high sound quality, in collaboration with the French software development company Flux in Orléans,

This research has been performed in the context of the sample orchestrator project (RP2) that had one its sub tasks dedicated to improve use of the phase vocoder based sound analysis and transformation for real time applications. One of the important problems of the phase vocoder based algorithm for transposition in real time applications is the fact that it induces computational costs that depend on the transposition parameter. The problem is clearly exposed in [J. Laroche and Dolson 1999b]. I summarize here the results: Costs are related to the number of frames that have to be processed to produce a certain number of samples. Costs grow inversely with time compression factor and proportionally with transposition factor for factors > 1. For man to children conversion (section 9.2) or speech to singing conversion (subsection 9.3.1) pitch shifting of 2 octaves or more can be required. Managing real-time applications with costs changing by a factor 2 or more is very difficult and therefore a solution was requested by many users at IRCAM.

The solution proposed in [J. Laroche and Dolson 1999b] is based on doing the transposition by means of shifting spectral peaks directly in the spectral domain representation. In that algorithm each spectral peak is supposed to be related to a sinusoidal component that generates the peak and the transposed spectrum is generated by means of shifting spectral peaks according to the estimated source frequency of the underlying sinusoidal component and the target frequency given by the time stretching factor and the source frequency. Shifting spectral peaks instead of scaling the frequency spectrum has the major advantage that the size of the signal segment does not change [Moulines and Jean Laroche 1995]. The source frequency is either determined by the centre frequency of the spectral bin containing the maximum amplitude of the peak or by means of finding the location of the maximum amplitude using the standard procedure for estimation of sinusoidal frequencies by means of quadratic interpolation of amplitudes [Amatriain et al. 2002].
The use of the algorithm proposed in [J. Laroche and Dolson 1999b] is problematic due to the fact that the algorithm is protected by patents [Jean Laroche and Mark Dolson 2003]. Therefore, I had to use alternative means and decided to use a method patented by IRCAM much earlier [P. Depalle and X. Rodet 1995; X. Rodet and P. Depalle 1992] that was dealing with the synthesis of sinusoidal components in the spectral domain. The basic idea is to go one step further in the analysis than Laroche and Dolson in [J. Laroche and Dolson 1999b] and to extract a complete set of sinusoidal parameters for all sinusoidal peaks (see the method discussed in section 4.3), and then to resynthesize these peaks in the spectral domain by means copying them from a pre-calculated set of sinusoidal peaks. The pre-calculated sinusoidal peaks are stored with sufficient frequency resolution such that they can be placed with minor artefacts on arbitrary target positions. The advantage of this system is the fact that the handling of non stationary sinusoids can take into account parameter estimates with reduced bias ([M. Abe and Smith 2004], (A. Röbel 2006b, 2007a,b, 2008)) and can modify the frequency slopes during transposition.
The quantitative evaluation of the algorithm on synthetic examples with known sinusoidal parameters has shown that the coherence of the synthesized spectra leads to significantly lower error of the transformed sinusoidal components when compared with the desired target components. The error measured against the known target sinusoidal component is up to 20dB smaller. On the other hand the perceptual evaluation has shown that the existing error in Laroche's method is perceptually not very important.
The research on the basic algorithm for spectral domain transposition is only a very small part of the research that is necessary to integrate the new algorithm into the phase vocoder framework. Many of the processing and analysis functions (e.g. section 4.2, section 4.4) required extensions and redesign to achieve nearly transparent results with both methods for transposition. This process converged to an acceptable result around 2010 such that the modified phase vocoder was available for two of our software projects (LI10; LI12). As a final remark it has to be said that the original approach to transposition has a number of advantages that lead to the fact that the quality obtained with transposition by means of resampling sounds slightly better. The two main differences are:

1. spectral domain transposition up will result in spectral holes for noise components that are perceptually very annoying,

2. spectral domain transposition of transient peaks will preserve all amplitude contours and all phase relations within the individual spectral peaks, but scale the frequency distance in between spectral peaks. In contrast to this with traditional time domain transposition the spectrum of transient components are scaled coherently for all bins. In the first case the waveform will be changed in a rather unpredictable manner, while in the latter case it will be stretched or compressed. Perceptually, the simple scaling of the signal waveform seems to be preferred.

For these two problems appropriate solutions that better approach the quality of the time domain transposition with the extended phase vocoder algorithm are still to be found.

CHAPTER 5

ADVANCED TOPICS IN STFT BASED SOUND REPRESENTATION AND TRANSFORMATION

All the problems related to STFT representations and phase vocoder based transformations that have been discussed so far stay rather close to the initial idea of the algorithm improving the results gradually by means of extending the classes of signals that can be manipulated (speech) or by means of improving the results obtained for one of the different classes of components that may be present in the signal.
In the following three sections I will discuss research directions that I have started during the last few years and that try to make a more radical change to the algorithm. While initial results have been obtained for all these topics most of them have not been developed so far that they would allow deployment for signal transformation in a user application. Over a few years I hope to be able to incorporate these new approaches into our algorithms such that they will become useful for music and sound production.
In the first of the following sections section 5.1 I will discuss research on time frequency representation with adaptive resolution that addresses the important problem of fixed time frequency resolution that is present in todays sound processing algorithms.
The section section 5.2 will address the problem of manipulating sounds that are outside of the scope of the models used for STFT based algorithms discussed so far. These are sounds that are not dominated by either tonal sources such that the sinusoidal model that is at the centre of the phase vocoder does no longer apply.
In the chapter on future research section 11.2 I will discuss research on manipulating individual sources in polyphonic music an approach that is now available in commercial software[1].

5.1 Adaptive time-frequency resolution

Related projects

(RP2) *Projet ANR - Sample Orchestrator – 2006-2009* (2006). Task 3.1: Enhanced phase vocoder analysis and transformations in real time applications,
(PhD6) M. Liuni (2012). "Automatic Adaptation of Sound Analysis and Synthesis". Directors X. Rodet, et M. Romito, supervision A. Röbel. PhD thesis. Università di Firenze, Italie/Université Paris 6 (UPMC), France,

One of the key problems of the existing STFT or phase vocoder based signal transformation algorithms is the fact that the results of the transformation depend critically on the time and frequency resolution of the internal STFT representation. This resolution is controlled by the size of the analysis window. Many users, especially those without any background in signal processing, try to avoid adapting the window size, some users even work with a fixed window size for all sounds. Even if a user is willing to invest into finding the optimal window the problem is not solved because the same sound may require different time frequency

[1] see Melodyne DNA `http://www.celemony.com/cms/index.php?id=dna`.

resolutions at different times or in different bands. An example problem would be a piece of music containing spectrally dense guitar or piano chords in one segment and a rapid sequence of individual notes in another. The first segment requires long windows but the second segment short ones. A solution to this situation would be to allow the window size of the representation to vary over time. If these two segments happen at the same time but such that only weak spectral overlap exists between the different sources the problem is solvable with frequency dependent windows. Finally if they overlap spectrally then the only possible solution consists in separating the sources to achieve an appropriate parameterization of the representation for both sources. This last case can not be solved by means of changing the time frequency resolution and will be discussed separately in section 11.2. Before starting the discussion of our research on adaptive time-frequency resolution I have to note that adaptive resolution has to be distinguished from wavelet representations as proposed for example in [Bonada 2000; Evangelista et al. 2012]. Those have non-uniform time frequency resolution, which however, is nevertheless constant and not adapted to the signal[2]

Research related to signal representation and signal transformation with adaptive resolution has to deal with 2 problems. The first problem comes from the fact that the time varying resolution has to be specified. When today users do not want to select a scalar window size parameter for a given sound, one cannot assume that they would define a time varying or even frequency dependent resolution. Therefore, the first research problem is related to the automatic selection of an optimal time (and frequency) dependent window size.

For monophonic sound signals this problem can be addressed quite simply at least for harmonic sounds by means of linking the window size to the fundamental frequency of the sound. For polyphonic sounds the problem becomes much more involved and accordingly, this problem was the subject of the PhD thesis of Marco Liuni (M. Liuni 2012). In this thesis we investigated into algorithms for adapting time frequency resolution based besides other measures on the Rényi entropy of time frequency distributions and on the other hand into signal reconstruction from non uniform and non constant time frequency representations. Based on recent advances related to analysis/resynthesis with non-stationary Gabor frames [Dörfler 2011] and in collaboration with Monika Doerfler and Ewa Matusiak we have shown that perfect reconstruction from representations with non-stationary time frequency resolution cannot be achieved efficiently if resolution changes in both directions (time and frequency). For this case efficient approximations have been presented in (Marco Liuni et al. 2013) together with a rough understanding about the error bounds.

The second problem concerns only signal transformation algorithms. As the window size is no longer constant, the algorithms will need to take into account new degrees of freedom in the spectral representation. At least three levels of complexity can be distinguished

a. window size can change over time: Problems may arise when peaks need to be connected between frames with different resolution.

b. window size and DFT size change over time: The problem is basically very similar to the case a, besides the fact that the relation between bin position and frequency is no longer fixed. This problem is more a problem related to the development, because invariants that were present in the previous version of the algorithm are no longer valid.

c. time frequency resolution changes with time and frequency: Besides the problems mentioned under case (a) and (b) now additionally new strategies for manipulation of the peaks covering changes of frequency resolution have to be developed. This such that both representations produce coherent results (see for example [Bonada 2000]).

Research with respect to the representation and transformation with window size changing over time has been done in the (RP2) project (Vinet et al. 2011). In that project monophonic sounds where assumed (solo instruments and speech) and the window size was coupled to the f0 analysis removing all burden for window size selection from a potential user or algorithm. It turns out that the extension of the algorithms

[2] In [Bonada 2000] band sizes are slightly adaptive but this serves to improve connection of components between bands only, and not to change resolution.

remains relatively straightforward as long as the DFT size is kept fixed and only a few parts of the algorithms require changes. The consistency of the representation is in fact ensured by means of the phase locking within sinusoidal peaks.
Initial experiments related to transformation of polyphonic music using time dependent time frequency resolution derived automatically with the methods developed in the thesis of Marco Liuni (PhD6) have been performed in the same thesis (Marco Liuni et al. 2013). The experimental investigation has demonstrated that the adaptive resolution can achieve results that are equivalent or better than the results obtained with fixed resolution. Signal transformation under case c (time frequency resolution changes dynamically with time and frequency) has not yet been studied. As a next step the simplest case of adaptive resolution covering only time variation will be developed such that it can be used and evaluated in the context of everyday projects (AudioSculpt). If practical experience confirms the experimental results a next step would be to extend the phase vocoder algorithms to the cases b (time varying DFT size) and c (time varying frequency dependent resolution).

5.2 Sound textures

Related projects

> (RP8) *Projet ANR - PHYSIS, Physically informed and semantically controllable interactive sound synthesis – 2012-2015* (2012). Direction des travaux sur low level sound representation (WP3),
> (PhD12) W. H. Liao (ongoing). "Modelling and transformation of sound textures and environmental sounds". Directors X. Rodet et A. Su, supervision A. Roebel. PhD thesis. Université Paris 6 (UPMC) and National Cheng Kung University, Tainan, Taiwan,
> (MA19) Hugo Saulnier (2013). "Synthesis of Sound Textures". Stage Master 2 ATIAM, supervision A. Roebel and S. O'Leary. MA thesis. Université Paris VI Pierre et Marie-Curie.

Due to the overwhelming importance of music and speech signals for human communication nearly all of the existing sound transformation methods make at some point the assumption that sinusoidal components are present in the sound signal. This can be explicit as for example in the additive model [Amatriain et al. 2002] or the phase vocoder [Dolson 1986; J. Laroche and Dolson 1999a], or more subtle in the PSOLA method [Moulines and Charpentier 1990] where the window size is adapted to the local period and where time stretching or pitch shifting is achieved by means of creating regular pulses. There are however other classes of sounds not dominated by sinusoidal components and that contain a rather strong degree of randomness. These kind of sounds are often denoted as a *sound textures*, however, without there existing a formal and generally accepted definition of what is a sound texture [Strobl et al. 2006][3]. One of the early and probably most influential working definitions has been given in [Saint-Arnaud 1995; Saint-Arnaud and Popat 1998]. There sound textures are defined loosely to be constituted of atomic events, that are not necessarily time limited (turbulent wind noise) but that appear according to a high-level pattern that can be periodic (motor sounds) or random (rain), or both (waves). The high-level pattern and the fine structure of the sound should be time invariant when long time scales are observed. Randomness is generally considered an important factor. This definition is accompanied by examples that indicate that the number of atomic events is generally high, such that the events fuse into a common texture stream.
For these kind of sounds synthesis algorithms have been developed that allow synthesizing long sequences from short examples using either source-filter models in time and frequency domain [Athineos and D. Ellis 2003; Zhu and Wyse 2004], time domain superposition or granular synthesis [Lu et al. 2004; Diemo Schwarz and Norbert Schnell 2010], multi resolution representation [Dubnov et al. 2002; O'Regan and Kokaram 2007; Saint-Arnaud and Popat 1998], or components based synthesis [Verron 2010]. Most of these algorithms are based on some sort of replaying of the original sound using random variations to avoid the perception of repetition. The founding arguments for the existing methods are either the physical or the statistical properties of the sound generation process that in most cases was very close to the generative model discussed in [Saint-Arnaud 1995; Saint-Arnaud and Popat 1998]. Experimental evidence for the

[3] This is not very astonishing as the same holds true for example for music.

signal characteristics that are related to the perception of sound textures was not yet available. Recently, a detailed investigation into the relations between the signal characteristics and the perceived sound texture has been performed [J. McDermott et al. 2009; J. H. McDermott and E. P. Simoncelli 2011] providing insight into the statistics that have to be preserved to allow the recognition of sound textures. As a result of this study McDermott conjectures that moments (mean, variance, skewness and kurtosis) as well as cross correlation coefficients of signals energy in auditory filter bands as well as the same set of statistics for the modulation bands of the auditory filter signals are essential for texture recognition. This result is very important because it gives new insights into the perception of stationary aperiodic sounds, and these new insights will certainly lead to new approaches for synthesis and transformation of sound textures. While it seems clear that the proposed set of statistics will not be final the basis for developing sound textures synthesis and transformation algorithms has now become similar to the situation existing for periodic sounds for that the perceptually important characteristics (amplitude, frequency, and in some cases phase) are know for a very long time.

Based on the results of [J. McDermott et al. 2009; J. H. McDermott and E. P. Simoncelli 2011] I have started to investigate into new means for signal transformation. The main objective of this research is to find signal processing algorithms that allow manipulating or simply preserving perceptually relevant statistical sound characteristics. The first step in this direction is the PhD Thesis of Wei-Hsiang Liao (PhD12) where we investigate into using STFT based signal representation for noise and sound texture manipulation.

The use of the STFT is in contrast to many of the existing approaches that use wavelet based representations to better match the properties of the human auditory perception. The STFT representation however allows us to simplify the mathematical relations, and potentially to make use of existing technologies like transient and sinusoids detection. Moreover, in the long run, it simplifies the integration of the results back into existing algorithms such that the transformation of aperiodic components in music and speech signals may benefit eventually from the results obtained. The first problem that has been studied was directly related to the problem of time stretching or pitch shifting noise in the phase vocoder. We have investigated into the correlations of the STFT coefficients for STFT representation of noise and developed an algorithm for time stretching of Gaussian noises (W.-H. Liao et al. 2012) achieving significantly improved quality for time stretched noise. This initial algorithm requires only very few statistical descriptors of the STFT sub-bands to be preserved (variance, auto-correlation function, cross-correlation function). In the subsequent steps we investigate into time stretching of stationary environmental sound textures (Wei-Hsiang Liao et al. 2013). For the moment and at least for all the textures that were tried the algorithms allow generating long non repeating sequences of textures from examples of about 5-7s.

In a related project Hugo Saulnier (MA19) has investigated into using the algorithm proposed in [J. H. McDermott and E. P. Simoncelli 2011] for interpolation/morphing of sound textures. McDermott's algorithm uses a sound representation that is highly motivated by the auditory system. The comparison of the current version of our STFT based algorithm with the algorithm proposed by McDermott based on a representation using auditory bands and modulation spectra [4] has revealed that the STFT based algorithm is one order of magnitude faster than the version based on the auditory model [J. McDermott et al. 2009; J. H. McDermott and E. P. Simoncelli 2011]. Moreover, the convergence seems to be improved due to a more consistent implementation of the optimization procedure. Initial evaluation in informal listening tests has revealed that the quality of the sound textures generated with the STFT based algorithm is generally similar or better than the results obtained by McDermott's method. All results obtained so far are preliminary as the research is still in a rather early stage.

All these research activities are closely linked to the ANR project Physis (RP8) which is centred on the modelling, transformation and real-time synthesis of diegetic sounds for interactive virtual worlds (video games, simulations, serious games...) and augmented reality. IRCAMs contribution is oriented in two directions. The first objective is an automatic decomposition of textures into components (audio events with similar characteristics like little explosions in fire) together with statistical models that describe the appearance of these components. The interest of components based approaches is the fact that these would potentially allow synchronizing audio events with events in the visual scene. NMF based algorithms have been studied as potentially interesting approaches for the decomposition of the textures. For the moment, however, these algorithm have produced either bad signal representation or strongly dissected components.

[4] Thanks to J. McDermott for giving us access to his original implementation.

Second, we investigate into efficient manipulation of the perceptually relevant features [J. McDermott et al. 2009; J. H. McDermott and E. P. Simoncelli 2011] using time frequency representations of sounds. This research is done in the thesis of Wei-Hsiang Liao (PhD12).

5.3 Source separation

Related projets

(MA14) F. Rigaud (2010). "Séparation de la partie percussive d'un morceau de musique". Stage Master 2 ATIAM, supervision M. Lagrange, A. Roebel and G. Peeters. MA thesis. University Paris VI – Pierre et Marie-Curie,
(PhD9) T. M. Wang (ongoing). Visting PhD student at IRCAM in 2011, Research on separation of drum signals from polyphonic music, supervision A. Roebel, C. Yeh, M. Lagrange. PhD thesis. National Cheng Kung University, Tainan, Taiwan,
(MA16) A. Bonnefoy (2012). "Transcription de la partie percussive d'un morceau de musique". Stage Master 2 ATIAM, supervision M. Lagrange, A. Roebel and G. Peeters. MA thesis. Université Paris VI Pierre et Marie-Curie,
(RP7) Yuki Mitsufuji and Axel Roebel (2011-2012). *Collaborative research project with Sony Japan*. Source separation in multichannel audio recordings.
(RP5) *Projet FP7-ICT-2011, 3DTVS, 3DTV Content Search – 2011-2014* (2011). Direction des travaux sur 3D Audio & Multi Modal Content Analysis and Description (WP4)
(RP4) Chunghsin Yeh et al. (2010-2012). *Automatic midi annotation of polyphonic music*. Ableton. 2010-2012
(MA20) Jordi Pons Puig (2013). "Source separation for music signals". MA thesis. Universitat Politècnica de Catalunya · BarcelonaTech (UPC),

Audio source separation is a problem that has been investigated quite extensively in the literature. Early approaches were based on ICA [Cardoso 1997] but the related techniques were not easily applicable, because they do not allow searching for specific sources and because they pose severe constraints on the number of independent sources that can be separated. More recently nonnegative tensor factorization (NTF) and nonnegative tensor deconvolution (NTD) techniques have been proposed [Lee and Seung 2000]. These techniques can be used with dictionaries that are adapted to separate specific sources as for example musical instruments and seem therefore to have great potential for extracting musical instruments from polyphonic music [Dessein et al. 2012; Virtanen 2007]. Moreover they allow to combine the information available in multiple channels such that the redundancy present in multichannel recordings can easily be exploited [Févotte and Ozerov 2010; FitzGerald et al. 2005, 2008; Ozerov and Févotte 2010].
My research interests related to source separation have their origin in the Audio2Note project. There was on one hand the question whether the audio transcription provided by the A2N algorithm could be used for extraction of the individual notes. And on the other hand the objective to suppress the drum track from polyphonic music to eventually improve the note transcription. Unfortunately, due to time constraints, the signal separation part had been dropped from the A2N project so that the research was interrupted in a rather early stage. Consequently the research on music signal separation has been and is performed without dedicated funding. As these research efforts have started recently, the results are considered as preliminary studies.
A first investigation into drum separation has been performed in the master thesis of François Rigaud (F. Rigaud 2010). In this thesis we developed a new detection function for drum beat events in the spectral domain by means of onset detection coupled with a rather simple criterion concerning the amplitude evolution after the instrument onset. Signal separation was then performed by means of binary masking (). The method has been refined later by means of including information about harmonic structure that should not be present in a drum onset (PhD9). In the master thesis of Antoine Bonnefoy (MA16) we investigated into the performance of NMF based algorithms, notably convolutive NMF or nonnegative matrix deconvolution (NMD), for drum transcription on the sounds separated before by the drum extraction algorithm.

5.3.1 Multichannel audio

Due to the redundancy in the sound signals one can expect that sound separation using multi channel audio signals would allow achieving improved quality of the separated signals when compared to single channel cases. Many of the musical signals are stereo and therefore it would be very interesting to be able to benefit from the additional information.
A first attempt in this direction was undertaken in the collaborative research project proposed by Sony Inc. Japan (RP7). In this project a visiting researcher, Y. Mitsufuji, investigated into techniques for source separation using multichannel audio recordings. As a result of this collaboration with Y. Mitsufuji we have developed a sound source separation method that can be tuned to extract sources in a selected direction by means of providing spatial cues similar to beam forming with a microphone array using however only a stereo signal (Y. Mitsufuji and A. Röbel 2013)[5].
Signal separation in multichannel audio is a central question in the 3DTVS project (RP5) in that the analysis synthesis team is responsible for detecting and indexing different audio events in 3D TV multichannel audio scenes. In this project I work with Marco Liuni on approaches based on non-negative matrix factorization for detection of selected real world sound events (moving cars, flying helicopters, gunshots) in the 3D TV multichannel audio. Methods based on matrix factorization have been shown to provide good results for detection of overlapping audio events in music [Dessein et al. 2012] and everyday sound events [Cotton and Daniel PW Ellis 2011]. Building on these results the idea of the 3DTVS project was to investigate into the use of extensions of the NMF framework for multichannel audio, nonnegative tensor factorization (NTF) and nonnegative tensor deconvolution (NTD) [FitzGerald et al. 2005], to the problem of audio event detection. One of the main problems with the existing approaches is the fact that the mixing position of the sources is assumed to be fixed. One of our contributions in this area is the development of segmental or online versions of NTF and NTD audio event detection algorithms similar to [Duan et al. 2012]. The term online here refers to the fact that the audio stream is cut into segments (with a duration of 2s). Currently we develop more refined methods allowing the continuous tracking of objects across the channel matrix. Another contribution is a new adaptive detection method that avoids having to select fixed detection thresholds and significantly improves detection performance across different movies.
The initial experimental results with the simple class of gunshot audio events have demonstrated the very significant detection improvement that can be obtained when using the nonnegative tensor deconvolution with multichannel audio. Compared to NMF on a single channel audio downmix of the multichannel audio the NTF algorithm working on multichannel audio improved the detection accuracy by 50% in F-measure from 0.52 to 0.79. The introduction of target basis covering multiple analysis frames (NTD) improved further to 0.86 F-measure. Experiments with the detection of running cars are currently under study. Initial results seem to indicate similar improvements as with gunshots.
While the detection of general real world sound events like cars and gunshots might seem a bit far from IRCAM's central interests I would like to note that the research performed in the 3DTVS project has given us the opportunity to gather important experience with recent matrix factorization technics and to develop software for signal separation that in turn will be applied to problems like music instrument separation. An initial study in this area will be performed in the master thesis of Jordi Pons Puig (MA20) starting in September 2013 that will be described in the chapter on future work in section 11.2.

[5] A patent application has been submitted.

CHAPTER 6

SINUSOIDAL MODELLING

Sinusoidal modelling is one of the key technics for music and speech sound analysis and transformation. Besides the work on implicit sinusoidal models (signal representations based on sinusoidal models without representing sinusoids explicitly like the phase vocoder) and the use of sinusoidal models for instrument models and fundamental frequency analysis I have invested some effort into more fundamental work related to the parameter analysis of non-stationary sinusoids. Due to space constraints I will give a more detailed discussion for only one of the research topics from this area that I consider to be the most interesting and innovative contribution and will shortly summarize my research on other questions related to sinusoidal parameter estimation at the end of the section.

6.1 Adaptive trajectory model

Related projets

(RP1) A. Roebel (2000). *Adaptive additive synthesis of non-stationary sounds*. Research scholarship at CCRMA, DFG project, Ref RO2277/1-1,
(RP2) *Projet ANR - Sample Orchestrator – 2006-2009* (2006). Task 3.1: Enhanced phase vocoder analysis and transformations in real time applications

A mathematical formulation of a non-stationary sinusoidal model is given by

$$x(n) = \sum_{i=0}^{M} A_i(n) \cos(\varphi_i(n)). \tag{6.1}$$

All the $M + 1$ partials have time varying amplitude $A_i(n)$ and a phase function $\varphi_i(n)$ that generally has varying slope (time varying frequency). The creation of sinusoidal model for a given signal requires the determination of the sinusoids that are present in the signal and the determination of parameter trajectories for those sinusoids. Nearly all the existing algorithms for sinusoidal modelling approach the problem in two steps. In a first step the peaks of the discrete Fourier spectrum of the windowed time signal are detected and then, in the second step, the sinusoidal parameter trajectories are formed by means of connecting these peaks [Amatriain et al. 2002; X. J. Serra and Smith 1990]. In most cases the parameter evolution is assumed to be sufficiently slow such that for the estimation of amplitude and frequency parameters the time variation can be ignored [Marques and Almeida 1986; McAulay and Quatieri 1986; X. J. Serra and Smith 1990].
The two-step approach has clear disadvantages, as it requires independent detection and tracking of sinusoidal candidates limiting the performance of each of the two stages by the fact that the results of the other stage are not taken into account. The motivation for my research on sinusoidal modelling with continuous parameter trajectories was to establish a method that would avoid this two-step approach.
The basic idea was to integrate tracking and parameter estimation into an adaptive algorithm that would adapt continuous parameter contours of sinusoidal components by means of minimization of the error when comparing the model to the observed signal. Very few similar approaches exist even today. An example is [Ding and Qian 1997], where a very similar problem is addressed using however a model that was significantly more contained as it requires the sinusoids to stay close to a centre frequency. Another example is [Day and Godsill 2002] that uses a Bayesian framework to estimate parameters for harmonically

organized note models. Note, however, that these note models are limited to have frequency trajectories with limited modulation extend, which can be considered too constraint to allow any practical applications. The sinusoidal phase and amplitude parameter trajectories of the adaptive algorithm were represented using B-splines [Day and Godsill 2002; de Boor 1987; Ding and Qian 1997]. This allows the parameter trajectories to be piecewise polynomial functions $p(n)$ of arbitrary order o that are expressed by linear superposition of B-splines of the same order following

$$p_o(n) = \sum_i B_i b_i(n). \tag{6.2}$$

Here B_i is the weighting parameter of the i-th B-spline of order o, $b_i(n)$. Note that B-splines are functions with local support; hence they are non-zero only in a connected and bounded region. The description of the polynomial trajectories by means of B-splines renders the mathematical treatment simple and straightforward, but no results were available that could direct the proper choice of the B-spline parameters (segment size, and spline order). As important result of my investigation I could show that the frequency resolution of the representation is determined by the Fourier spectrum of the basic splines $b_i(n)$ in a very similar manner as it is determined by the spectrum of the analysis window for the STFT based sinusoidal parameter estimation (A. Röbel 2001a,b, 2006a). This is a very important result for the proposed model, because it allows controlling the selection of the parameter trajectory model based on theoretically sound principles. Moreover I could demonstrate experimentally that by means of regularization of the parameter trajectories the bounds of the estimation error can be lowered, which is due to the fact that the signal segment that is used is effectively increased. The computational complexity of the method is unfortunately rather high, such that it could not be used for practical applications. It is interesting to note however, that due to recent results related to polynomial sinusoidal parameter contour estimation with the distribution derivative method [Betser 2009] there seems now to exist a non-iterative method that may allow to estimate the continuous parameter contours for example in form of B-splines directly.

Aside this investigation into adaptive approaches to sinusoidal modelling I did research on the improvement of the estimation of parameters of non-stationary sinusoids. The estimation of the basic sinusoidal parameters amplitude, phase and frequency has received a lot of research efforts. As mentioned above the classic estimators simply neglected the time variation of the parameters. More recent analysis methods take the time variation of these parameters into account and allow reducing the analysis error [M. Abe and Smith 2005; Betser 2009; Marchand and Philippe Depalle 2008; Marques and Almeida 1989; G. Peeters and X. Rodet 1999; Peleg and Porat 1991; Wen and Mark Sandler 2009]. For onsets of sinusoids other models have been proposed [Fitz et al. 2000; Levine 1998; Levine and Smith 1998]. In my own contributions I was aiming to estimate the frequency slope of a non-stationary sinusoid (A. Röbel 2006b, 2007a,b, 2008).

Other problems related to sinusoidal modelling that I have been working on, notably in collaboration with Chunghsin Yeh, was the question how the superposition of multiple sinusoids should be integrated into the polyphonic fundamental frequency estimator (Chunghsin Yeh and Axel Roebel 2009). While the phase vocoder in its original form does not make use of an explicit sinusoidal model recent forms of frequency domain transposition [J. Laroche and Dolson 1999b] that significantly improve the efficiency of the phase vocoder transposition benefit from explicit sinusoidal models. In the context of the ANR project (RP2) I have notably shown that taking into account the frequency slope of the sinusoidal components allows to improve the objective coherence of transformed sinusoidal components (Vinet et al. 2011).

CHAPTER 7

SOURCE FILTER MODEL

The source filter model is of major importance for the transformation of sound. This importance is related to the fact that the model allows segregating the sound signal into two parts, the excitation signal and the resonator filter, that both have direct links to physical properties of real world sound sources. Accordingly, by means of exposing these two parts in a signal transformation method we provide means for rather intuitive transformation of sound signals. The *source* represents the excitation signal that relates to perceptual sound characteristics like pitch, noise level, inharmonicity. All of these are determined by physical properties of the exciting oscillator. The *filter* represents the resonator that describes the sound colour, which means here the way its energy is distributed in the frequency domain.
In the following two sections I will discuss my research related to the estimation of the filter component generally denoted as *spectral envelope* section 7.1, and research on modelling the timbre space of musical instruments sound with an extended source filter model section 7.2. The use of the source filter model for speech transformation will be discussed in the chapter on chapter 9.

7.1 Spectral envelop estimation

Related projects

> (PhD5) Fernando Villavicencio (2010a). "Conversion de la voix de haute qualité". Director X. Rodet, supervision A. Röbel. PhD thesis. Paris : Université Paris 6 (UPMC).
> moreover see sections on voice conversion (section 9.2), and speech to singing transformation (subsection 9.3.1.

The availability for an efficient and robust estimation of the spectral envelope became important first for my work on speech signal transformation. While the use of the AR model for spectral envelope estimation of speech has a physical foundation [Markel and Gray 1976], unfortunately for the important case of voiced sound segments the estimates that are obtained using the AR model are strongly biased. The bias can be reduced significantly as long as the model order is known and the spectral peaks that define the envelope are properly selected [El-Jaroudi and Makhoul 1991]. The need for the selection of the peaks as well as the determination of an appropriate model order however is a critical problem. This problem motivated the research into the True Envelope (TE) envelope estimator that had been initially proposed in [Imai and Y. Abe 1979]. The TE estimator is based on iterative cepstral smoothing. It automatically selects the set of peaks that are coherent with the model order and investigation into the TE estimator did show that despite the iterative procedure this estimator could be implemented extremely efficient. Moreover I was able to show that an appropriate order can be derived from the fundamental frequency of the sound signal taking into account the fact that the spectral envelop is sampled by the spectral peaks such that the sampling theorem can be used to select an adequate number of cepstral coefficients that perform near optimal band limited interpolation of the sampled envelope (A. Röbel and X. Rodet 2005a; A. Röbel et al. 2007). The experimental investigation into the comparison of the TE and standard LPC as well as bias corrected DAP estimators has been done in collaboration Fernando Villavicencio, who later used the TE estimator for advanced voice conversion[1] in his PhD thesis (PhD5). The TE estimator is one of the key technologies for the gender transformation of speech signals (see chapter 9).

[1] see section 9.2

7.2 Instrument models

Related projects

> (PhD1) J. J. Burred (2008). "From Sparse Models to Timbre Learning: New Methods for Musical Source Separation". Visiting PhD student at IRCAM in 2006, Research on modeling of spectral envelopes of musical instruments for musical source separation, supervision A. Roebel. PhD thesis. Communication Systems Group, Technical University of Berlin,
> (RP3) *ANR project - Sample Orchestrator II – 2010-2013* (2010). Supervision of research on modeling timbre spaces of musical instruments by means of extended parameterized source filter models (WP2),
> (MA15) H. Hahn (2010). "Generalisierte, grundtonabhängige Modelle für quasi-harmonische Instrumente". Magisterarbeit (Master), supervision A. Roebel. MA thesis. Technische Universität Berlin, Allemagne,
> (PhD8) H. Hahn (ongoing). "Synthèse et transformation des sons basés sur des modèles de type source-filtre étendu pour les instruments de musique". Director X. Rodet, supervision A. Roebel. PhD thesis. Université Paris 6 (UPMC).

The question of the representation of the timbre space of a musical instrument arises in different contexts:

1. Sound source identification: For polyphonic pitch (see section 8.2) and instrument sound separation (see section 5.3) the knowledge of the spectral envelope of the playing musical instruments would allow establishing constraints on the sound sources and therefore could potentially lead to improved detection and or separation performance.

2. Sound transposition: Today's sound transformation algorithms (see chapter 4) allow transpositions of more than a factor 2 without introducing artefacts. The problem here is related to the fact that the transformed signals do not merge acoustically with the untransformed sounds of the same sound source because for real world sound sources the sound colour will change with the pitch. The problem exists for musical instruments as well as for speech. For the former this situation creates severe constraints for the use of signal transformation algorithms for example in software samplers.

To address these situations I investigated into the possibilities to establish instrument models that describe the timbre space of either a single instrument or a class of instruments over the full pitch and intensity range. The development of this kind of model is a very ambitious task, and especially for sound source identification the impact of room acoustics and constitutes an important problem for the use of an instrument model. Considering sound transposition there exist only few results that allow describing the instrument dependent transformations that have to be applied during transposition of individual notes. Dudas has investigated into the optimal relation between transposition factors for the pitch and spectral envelope [Dudas 2002]. Recent advances in physical modelling of the wave propagation in the trumpet have shown that the nonlinear effects accompanying level changes can be modelled with a high degree of precision by means of a simple structure of filters and instantaneous nonlinear operations [Hélie and Roze 2008; Hélie and Smet 2008]. These models, however, are not available for all instruments and the question of adapting these models to particular instruments has not yet been addressed.
My contributions to the first problem have been developed in the context of the PhD thesis of Juan-José Burred (PhD1). The problem of this thesis was to use a priori information of the sources to improve sound source separation for music. I supervised J.J. Burred during his stay at IRCAM during which we developed a representation of the timbre space of musical instruments that was used in the thesis for source separation and was evaluated as well for instrument recognition (J. J. Burred et al. 2006; J. J. Burred et al. 2010; Juan José Burred and Axel Roebel 2010; Juan José Burred et al. 2009). The prototype curves in the instrument model space representing the average timbre evolution over individual notes for the different instruments are displayed in Figure 7.1. The diameter of the tubes represents the standard deviation σ over the set of notes of each instrument. For better visualization the diameter is scaled by 0.1 to avoid the overlap that otherwise is present between the different instrument prototypes. Due to space constraints I will not discuss the details of the instrument prototypes. I note however, that the investigation revealed the importance to find a representation that allows to efficiently take into account spectral features that are attached to partial

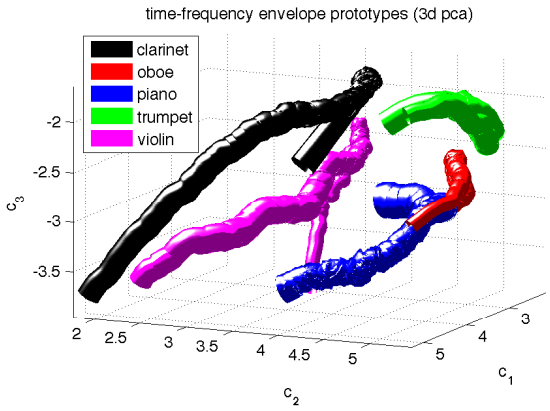

Figure 7.1: Average instrument note prototype curves in instrument timbre space. The width of the tubes represents 10% of the instrument specific variance.

position and therefore move with the fundamental frequency (even partials missing in clarinet) and features that are related to frequency and do not move with fundamental frequency (formant like features).

The research related to the second problem was performed in the master thesis of Henrik Hahn (MA15), and later in his PhD thesis (PhD8) in the ANR Project Sample Orchestrator II (SOR2) (RP3). The idea was to derive an extended source filter model of the timbre space of an individual musical instrument parameterized by intensity and pitch. The target application in the SOR2 project was the use of these models to extend the sound possibilities of sampling based synthesis applications by means of improved signal transformations. To achieve this the model was expected to correctly predict the instrument timbre as a function of pitch and intensity.

The instrument models are built by means of analysing a corpus of individual notes recorded from a single instrument, similar to what is used in sample based midi synthesizers. From this set two parameterized models of the spectral envelops of the noise and the sinusoidal components have to be constructed. In both cases the envelope is controlled by the parameters: pitch f_0, global intensity L that encodes the final note intensity (e.g. pp, mf, ff), and local intensity l that is used to establish the energy time envelope of a complete note. The local and global intensity are redundant, however they are necessary because the sound samples used in a midi sampler are generally normalized in energy so that the energy cannot be used to distinguish between note intensities. The note intensity is therefore derived from the instrument note annotations that are used to organize the instrument samples in the sampler database.
The spectral envelope models for both noise and tonal sound components need to be derived. Following a rather classical additive model we start in a first step to separate sinusoidal and noise components and each of the notes is split into an attack and a release segment. Based on the previous study on musical instrument modeling for sound separation mentioned above we used two components to represent the sinusoidal amplitudes A for a partial k. The first one, denoted excitation colour E, is a function of the note segment s (attack or release), the partial index k and the two further user parameters L, l An example for the importance of this model component is the clarinet for that partial amplitude depends strongly on the number k of a harmonic partial. The second, denoted resonator filter F, depends only on the frequency of the partial $\omega = kf_0$ [2]. The parts of the envelope model are represented in log amplitude using tensor

[2]The terminology source and filter component is taken as a convention. In reality the relations are rather complex such that a simple interpretation of source and filter parts as excitation signal and resonator filter is not possible. The plug position of a plugged string for example creates an envelope that depends only on frequency, but nevertheless is part of the excitation signal, while the physical position of the piano strings depend on the fundamental frequency but nevertheless this position will impact the resonator filter.

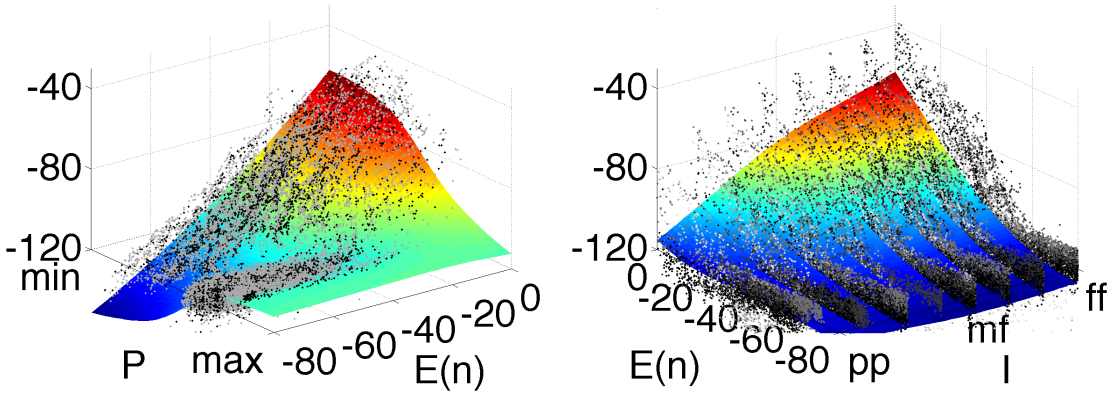

Figure 7.2: Model surface for the amplitude of the 9th partial of the piano model in the release phase. Partial amplitue as a function of pitch and local intensity (left), or local intensity and global intensity for medium pitches (right).

B-splines as follows

$$E(s,k,L,l) = \sum_i \sum_j \gamma_{i,j}^{k,s} B_i(L) B_j(l). \tag{7.1}$$

$$F(kf_0) = \sum_r \lambda_r W(kf_0) \tag{7.2}$$

$$A(s,k,f_0,L,l) = E(s,k,f_0,L,l) + F(k,f_0). \tag{7.3}$$

Here $\gamma_{i,j}^{k,s}$ is the weight for the tensor B-spline basis formed by the product of 2 univariate B-spline basis $B_i(L)$ and $B_j(l)$ and λ_r are the coefficients of the univariat B-spine basis $W(kf_0)$.
The B-spline parameters are trained on all available instrument samples minimizing the mean squared error of the log amplitude representation. The representation of the amplitude by means of a product creates an ambiguity because both components can be multiplied by a pair of constant factors without changing the result. These kinds of ambiguities are handled by means of regularization constraints that are added to the cost function during model training. The specific problem can be solved by means of a regularization term that evaluates the deviation of the log mean amplitude of the resonator filter from 0dB. The equations shown here describe the most basic setup. One may want the excitation colour E and the resonator filter F to depend on f_0 (e.g. for a piano where string positions and characteristics change). These kinds of extensions introduce additional ambiguities that we were able to handle with additional regularisation constraints. Further details of the models can be found in (Henrik Hahn and Axel Roebel 2013; H. Hahn and A. Roebel 2012).
The trained models for the harmonic component represent the amplitude of a partial as a function of pitch, global intensity, and local intensity as for example displayed for the 9th partial of a piano odel in Figure 7.2. The models are used to create white residual noise and tonal excitation signals. These excitation signals contain all note features that are not contained in the envelope models. This means they contain all the variations that the instrumentalist used when playing the individual notes. Pitch shifting those excitation signals is uncritical because they are white. Accordingly all the excitation residuals can be used for synthesis of all notes, which allows introducing variation into sequences of the same note without adding any additional recordings. The model structure being independent of the musical instrument hybrid instruments can be created by means of using excitation residuals from one instrument to excite the envelopes from another.
The results obtained so far are very satisfactory. We have been able to establish instrument models for wind, brass and string instruments (including plucked and struck strings) that allow using nearly arbitrary excitation residuals for each pitch and global intensity. The hybridizations obtained by means of combining excitation residuals and envelope model from different instruments create very convincing sounds, where

the components of both source instruments appear to fuse into a new perceived instrument. The stability of the perceived hybrid instrument over the complete range of pitches and intensities remains to be studied.

CHAPTER 8

FUNDAMENTAL FREQUENCY ESTIMATION AND MUSIC TRANSCRIPTION

In many musical cultures pitch, besides time position and duration, is one of the important features that is used to structure and organize sounds into music. This importance of pitch for music is one of the reasons that explain the importance of the fundamental frequency estimation for music signal processing. Another reason is the fact that the spectral resolution that is required to access individual sinusoidal components of the harmonic structure of the sounds generated by the pitched musical instrument depends on the fundamental frequency of the sound. Research on fundamental frequency or F_0 estimation has therefore been a central part of my research since my arrival at IRCAM. The following sections give a short overview of the results obtained in this domain.

8.1 Fundamental frequency estimation for monophonic signals

Related projects

(MA5) S. Starke (2002). "Bestimmung der Grundfrequenz mit Hilfe des Algorithmus zur Maximierung der Likelihood der harmonischen Auswahl von Maxima". Diplomarbeit (Master), supervision A. Roebel. MA thesis. Technische Universität Berlin, Allemagne,
(MA6) S. Schulz (2002). "Bestimmung der Grundfrequenz mit Hilfe wahrscheinlichkeitstheoretischer Bewertung von Signalspektren unter Verwendung spektraler Energiedichtemodelle". Diplomarbeit (Master), supervision A. Roebel. MA thesis. Technische Universität Berlin, Allemagne,
(MA7) M. Krauledat (2003). "Fundamental frequency estimation". Leonardo Internship, director X. Rodet, supervision A. Roebel. MA thesis. Westfälische Wilhelmsuniversität Münster,
(LI7) A. Roebel and X. Rodet (2008). *MakeMusic*. Library for fundamental frequency estimation for monophonic instrumental sounds,
(RP4) Chunghsin Yeh et al. (2010-2012). *Automatic midi annotation of polyphonic music*. Ableton. 2010-2012,
(MA17) L. Dale (2012). "Automatic Note Detection in Monopohonic Sound Files". Sciences de l'Ingénieur, Master Mécanique (M1), orientation Acoustique, supervision A. Roebel. MA thesis. University Paris VI – Pierre et Marie-Curie.

The fundamental frequency, or F_0, of a periodic signal can be defined as the inverse of the smallest of the infinite set of time shifts that leave the signal invariant [de Cheveigné and Kawahara 2002]. Strictly speaking, periodic signals cannot exist in the real world but more important for practical applications is the fact that real world sound sources are generally not stationary. For signals that contain small variations of the period within a time scale of a few periods the term quasi-periodic signal is used. The problem for estimating the F_0 of a quasi-periodic sound is the fact that the notion of small variations is ambiguous, because there are often multiple possibilities to explain a deviation of the period. Moreover, the possible

evolution of the F_0 requires to choose an appropriate observation time duration, which itself restricts the F_0s that can be estimated.
Building upon the previous research at IRCAM [Doval and X. Rodet 1991, 1993] I started my work with the objective to develop a robust and configurable F0 estimation algorithm. Due to the fundamental ambiguity mentioned above I was aiming to find a method that could easily be adapted to different characteristics of the input sound (different instruments, different noise level). The basic idea was to improve the existing criteria that were used in the existing F_0 estimation algorithms and that were based on properties of real world musical instruments (dominance of the harmonic components, smoothness of the spectral envelope). In a series of master thesis and internships (Jens Starke (MA5), Stefan Schulz (MA6), and Matthias Krauledat (Krauledat 2003)) we established an algorithm that did compare very favourably with existing algorithms like yin [N. Obin 2005] and that allowed controlling the type of errors between sub- and superharmonic errors. The algorithm was working in the spectral domain which, compared to the well known algorithm yin, provided the advantages that the noise outside the analysis frequency range does not effect the analysis results, that reverberation from previous notes had considerable less impact, and that the expected placement of the sinusoidal components can be modified to support inharmonic instruments as for example pitched percussive instruments like xylophone (LI7).
During our work with expressive speech signals (screaming) in the Respoken project[1] we found that the SWIPE algorithm [Camacho 2007; Camacho and Harris 2008] that showed very good performance even for signals that contain strong irregularities (sub harmonics) or noise. It is interesting that the algorithm achieves this performance without reducing the pitch range, which is required for our own algorithm. The possibility to avoid any strong F_0 priors is a very valuable feature for automated processing of speech and music signals, and therefore I investigated into the reasons of the high robustness of the algorithm. An important feature of the SWIPE algorithm seems to be the fact that it uses adaptive frequency resolution. In the master thesis of Laura Dale (2012) that dealt with the problem of note transcription for monophonic sources in the context of the Audio2Note project described below (RP4), we therefore started to investigate into the use of F0 estimation with adaptive resolution in noisy environments (instrument solo with drums).

8.2 Fundamental frequency estimation for polyphonic signals

Related projets

> (MA8) C. Yeh (2003). "Multiple fundamental frequency estimation". Rapport DEA Master ATIAM, supervision A. Roebel. MA thesis. Université Paris VI Pierre et Marie-Curie,
> (PhD2) C. Yeh (2008). "Multiple Fundamental Frequency Estimation of Polyphonic Recordings". Director X. Rodet, supervision A. Röbel. PhD thesis. Université Paris 6 (UPMC),
> (PhD3) W. C. Chang (2009). Visting PhD student at IRCAM in 2008, research on tracking of multiple fundamental frequency candidates in polyphonic music. Supervision A. Roebel, and C. Yeh. PhD thesis. National Cheng Kung University, Tainan, Taiwan,
> (MA13) R. Houzet (2010). "Formation de flux à partir d'une représentation objet de signaux musicaux polyphoniques". Intership Master 2 ATIAM, supervision C. Yeh, M. Lagrange, A. Roebel. MA thesis. Université Paris VI Pierre et Marie-Curie,
> (RP4) Chunghsin Yeh et al. (2010-2012). *Automatic midi annotation of polyphonic music*. Ableton. 2010-2012.

Given the very high performance that was achieved by the monophonic f_0 estimation algorithms around 2003, and further motivated by the good results obtained for example by Anssi Klapuri [Anssi Klapuri 2004] I started to investigate into the problem of fundamental frequency estimation for polyphonic sounds. The research related to this problem has been performed in the master thesis of Chunghsin Yeh (MA8) and later in his PhD thesis (PhD2) where we tried to establish a F_0 estimation algorithm based on the same principles that were used for the monophonic F_0 estimation algorithm by means of adapting the concepts to the polyphonic case. During the PhD thesis we were confronted with the question whether we

[1] The FEDER project Respoken was directed by Xavier Rodet and I contributed to this project with know how on speech transformation algorithms.

should continue to focus our efforts an estimator based on explicit domain knowledge as for example the algorithms in [Dressler 2012; Anssi Klapuri 2004, 2008; Saito et al. 2008] or whether we should follow the increasingly popular approach to focus on machine learning principles as for example [Bertin et al. 2009; Bertin et al. 2009; Grindlay and Daniel P.W. Ellis 2010; Kameoka et al. 2007; E. Vincent et al. 2010]. The methods based on machine learning principles try to estimate the F_0 using a given signal model or dictionary that is then adapted to the observed signal. Given the rather long runtime of the machine learning algorithms and given the very complex relations that are required especially to detect instruments playing in harmonic relations we have chosen to develop an estimator that integrates domain specific knowledge that itself was partly derived by means of machine learning techniques from databases of individual instrument sounds (C. Yeh et al. 2010).
The contributions that support multiple f_0 estimation algorithms are the following. A score function that evaluates f_0 candidates in polyphonic spectra (C. Yeh and A Röbel 2004a,b), the integration of a method for noise level estimation (C. Yeh and A. Röbel 2006a) that ponders the importance of a spectral peak that supports a given f_0 hypothesis. A two stage procedure that first determines a set of non-harmonically related f_0s before a refined search is done to select harmonically related f_0s (C. Yeh et al. 2010). A method for generation of synthetic polyphonic music with automatic ground truth annotation (C. Yeh et al. 2007). An algorithm that derives the expected amplitude of overlapping partials and an algorithm that allows to estimate the expected amplitude of an overlapping partial belonging to multiple harmonic sequences (Chunghsin Yeh and Axel Roebel 2009). A significant improvement of the fundamental frequency estimation solely on individual frames has been achieved by means of adding a subsequent tracking of the estimated fundamental frequencies candidates using as main constraint the continuity of the number of sources over time (W.-C. Chang et al. 2008). This work has been performed in the research scholarship of the PhD student Wei-Cheng Chang (PhD3).
The polyphonic fundamental frequency estimation has been evaluated in MIREX evaluations covering the years 2007-2011 (Chunghsin Yeh 2007; Chunghsin Yeh and Axel Roebel 2010, 2011; Chunghsin Yeh et al. 2008; Chungshin Yeh and Axel Roebel 2009). The method was ranked second in the category frame based multiple f0 estimation in MIREX 2007, and was ranked first for all MIREX evaluations from 2008 to 2011, which was the last MIREX we participated in. It is interesting to note that the algorithm ranked second over all evaluations so far is similar to ours based on explicit domain based knowledge [Dressler 2012]. The best algorithms based on machine learning techniques are currently still inferior by about 10% in accuracy [Bertin et al. 2009].
A topic that covers a subsequent problem of F_0 estimation for polyphonic files is the formation of instrument streams. This topic addresses the problem of classification of all the notes that were found by the F_0 estimation stage into the different instruments. This does not require that the instruments are recognized but that one can create a separate transcription for each instrument that is present. This problem was investigated in the master thesis of R. Houzet (MA13) using the task to cluster sequences of spectral envelopes of different notes of different instruments into groups each covering a single instrument. The performance was evaluated by means of the percentage of envelope frames that were not grouped into the correct class. The error depends on the number of instruments and was around 20% for 2 instruments and in the order of 50% for 6 instruments. This error is rather large if we consider that the spectral envelopes used were obtained from clean sources. The main problem for this kind of algorithm is the very strong variability of the spectra of music instrument sounds.

8.2.1 Audio2Note

Based on the success of the fundamental frequency estimation method (C. Yeh et al. 2010) we were able to attract the interest of the company Ableton to invest into a research partnership targeting a Audio2Midi transcription system (RP4). In this project we integrated our transient detection algorithm (A. Röbel 2003a,b) together with the polyphonic f_0 estimation algorithm (C. Yeh et al. 2010) into a system that describes a given polyphonic music in terms of a sequence of midi notes. After a successful integration of the algorithms by C. Yeh and later S. O'Leary the research collaboration was ended in Summer 2012 and the algorithm is now part of the software Ableton Live 9 that appeared in Spring 2013.

CHAPTER 9

SPOKEN AND SINGING VOICE

Analysis, synthesis and processing of the human voice is one of the major research activities of the analysis synthesis team (Pierre Lanchantin et al. 2011). I became involved in these activities after the development of the shape invariant processing option in the phase vocoder (see section 4.4) paved the way to voice transformation in the spectral domain. An important contribution to the research on speech processing was the development of the cepstrum based spectral envelope estimation technique *true envelope* that was able to reliably estimate near optimal spectral envelopes (see section 7.1) providing means for very high quality spectral envelope transformation. These results were essential for the high-level, semantic control of speech signal transformation using age, and gender parameters that has been developed in the VIVOS and Affective Avatar projects[1] (Snorre Farner et al. 2009; S. Farner et al. 2008; Xavier Rodet et al. 2009). I contributed to the development of the high-level control by means of work on the interface from the underlying speech processing functionality in SuperVP while the work on the control strategy was mainly performed by Snorre Farner and Xavier Rodet. All these results were important steps that later allowed us to develop the signal processing library for the professional audio plugin TRaX (LI12).
Besides the basic speech transformation technics discussed in section 4.4 my research on speech processing covers the areas: glottal source parameter estimation, voice conversion, parametric speech synthesis, and speech to singing conversion, that will be described in the following sections.

9.1 Glottal source parameter estimation and transformation

Related projects

> (PhD4) G. Degottex (2010). "Glottal source and vocal-tract separation". Director X. Rodet, supervision A. Röbel. PhD thesis. Université Paris 6 (UPMC),
> (PhD11) S. Huber (ongoing). "High Quality Voice Conversion by modelling and transformation of extended voice characteristics". Director X. Rodet, supervision A. Roebel. PhD thesis. Université Paris 6 (UPMC).
> (PhD7) W. C. Chien (2013). Visting PhD student at l'IRCAM from Aug. 2013 - Feb.2014, Estimating the source and filter from singing voice signals. Supervision A. Roebel. PhD thesis. National Taiwan University, Taiwan.

The transposition of speech signals is generally performed by means of changing the pitch and keeping the spectral envelope (A. Röbel and X. Rodet 2005a,b) if the speaker is preserved, or additionally transposing the envelope (Snorre Farner et al. 2009; S. Farner et al. 2008; Xavier Rodet et al. 2009) if the speaker identity (gender, age) is supposed to be changed. The underlying speech production model is the source filter model that assumes a white excitation source [Markel and Gray 1976]. In reality however, the excitation source is not white. Taking this into account yields an extended source-filter model that includes a specification of the glottal pulse signal, which have to obey certain characteristics to be accepted as natural voice [Rosenberg 1971]. One of the main parameters for the glottal pulse is its duration that is expressed relative to the fundamental period [Gunnar Fant et al. 1985]. In an extended source filter model pitch changes are accompanied with changes of the duration of the glottal pulse leading to a scaling of the excitation spectrum, which in turn will result in a modification of the spectral envelope even if the person (the vocal tract

[1] Both projects were directed by Xavier Rodet.

transfer function) remains the same. The possibility to take these modifications into account is expected to improve the quality of transposed speech. Manipulation of the speech source would open other interesting possibilities. Glottal source configurations have been shown to be important for characterization of affective states, emotions, speaking styles [Gobl and Chasaide 2003; Lorenzo-Trueba et al. 2012; Scherer et al. 1984]. Another interesting application is parametric speech synthesis for that today high-quality excitation algorithm are still not available [Alan W. Black et al. 2007; Zen et al. 2009]. Recently inverse glottal source filtering has been proposed as a new and promising approach [Raitio et al. 2011].

9.1.1 Estimation

An important precondition for the use of glottal source parameters is the availability of algorithms that estimate the glottal source signal or the glottal source parameters. Many approaches have been proposed: algorithms based on iterative adaptive inverse filtering (ITAIF) [Alku 1992], algorithms using AR models with excitation (ARX) for estimation [D. Vincent et al. 2005], algorithms based on closed phase vocal tract transfer function (VTF) estimation (CPVTF) and inverse filtering [Alku et al. 2009; Walker and P. Murphy 2005], and algorithms based on the decomposition into minimum and maximum phase components (ZZ) [B. Bozkurt and C. d'Alessandro 2012; Drugman et al. 2011].
All these algorithms have weak points that are hindering a robust application to real world signals. All but the last of these algorithms are based on the assumption that the VTF is all-pole, which is known to be incorrect due to the effect of the nasal tract. This assumption has posed many problems for speech transformation (see section 7.1) and there is no reason to expect that it should work reliably in the present context. ARX and ZZ require the correct detection of the glottal closure instant, which is a difficult problem in itself. CPVTF requires even the detection of the closed phase, which is even more difficult. All algorithms have problems with high pitch and breathy voices.
The potential benefits of the enhanced source filter model and the relatively weak performance of the existing algorithms has led us to initiate research into the estimation of the glottal source spectrum. This research has started with the PhD thesis of Gilles (PhD4). The algorithm developed aims to estimate the glottal pulse parameters of the Liljencrants-Fant (LF) glottal pulse model [Gunnar Fant et al. 1985]. While this model (like probably any other parametric model) cannot cover the complete set of possible glottal puls forms it certainly covers important characteristics of the glottal excitation that seem sufficient to synthesize expressive speech [Gobl and Chasaide 2003]. We investigated into an approach based on separation of minimum/maximum phase components (Gilles Degottex et al. 2009a,b, 2010, 2011). The approach estimates jointly the glottal closure time instants. Results obtained during the thesis suggested that the complete parameter space cannot be estimated robustly. First, the final closing phase of the LF model introduces minimum phase characteristics into the glottal pulse spectrum and can therefore not be detected using a method that distinguishes minimum and maximum phase signal components [B. Bozkurt and C. d'Alessandro 2012]. Second, the glottal pulse shape parameters open quotient and asymmetry can partly compensate each other [D. Vincent et al. 2005]. To address these ambiguities we have constrained the glottal shape parameters to the one dimensional subspace of the LF parameter set given by the R_d form parameter [G. Fant 1995].
A few improvements of the method have been developed in the context of the ongoing PhD thesis of Stefan Huber (PhD11) that tries to integrate glottal source parameters in our voice conversion algorithm (Stefan Huber et al. 2012) (see section 9.2). Notably, the range of the R_d regression model has been extended to consistently cover R_d values of very tense and relaxed excitation, a slightly more robust objective function has been developed (Stefan Huber et al. 2012), and a Viterbi smoothing post processing has been integrated to avoid occasional R_d jumps. The evaluation of our method using synthetic signals or real speech signals with EGG generally shows favourable results compared to state of the art methods [Gilles Degottex et al. 2011; Stefan Huber and Axel Roebel 2014; Stefan Huber et al. 2012], for many practical applications however, and especially for high pitched signals ($F_0 > 200Hz$) or relaxed excitation with very few partials ($R_d > 4$), the glottal pulse estimation remains still too unstable and will therefore remain one of the important research activities for the next years.
To finalize this section I note that recently we experimented with a completely new approach using machine learning technics to learn the relations between the open quotient of the glottal pulse and other relevant speech parameters (F_0, voiced/unvoiced frequency boundary, spectral envelope parameters) (Stefan Huber

and Axel Roebel 2014). For training and validation we used the open quotient derived from EGG measurements available in the CMU Arctic speech database [Kominek and Alan W Black 2004]. Separately for each speaker the prediction of the open quotient was significantly better than what we can achieve with our R_d estimator. However, when we applied the predictor trained on two speakers to a third speaker we have got about the same prediction performance that we have with the estimation algorithms. While this result opens a new perspective to glottal pulse parameter estimation it is rather unclear whether the multi speaker prediction across gender and age can be performed reliably.

9.1.2 Transformation

One of the main reasons for investigation into the estimation of source excitation parameters is the new perspective to be able to manipulate the glottal pulse parameters and more generally the voice quality. Initial steps into expressive transformation of speech have been performed in [Grégory Beller 2009]. Later in the Respoken project[2] extensions for the VoiceForger library (SW5) have been developed that did allow experimentation with dynamic transformations of pitch, intensity, and duration as well as voice quality features like breathiness, glottal pulse parameters, and the spectral envelope. The result of this research demonstrated that by means of manually selected conversion strategies expressive signal transformations are feasible. The quickly changing transformations however pose problems because the signal transformation operators that are available today are not yet sufficiently robust to produce high quality speech with quickly changing parameters. On one hand the problems are related to insufficient robustness of the analysis algorithms, on the other hand are the signal transformations algorithms not sufficiently refined to be able to ensure a coherent signal after the transformation. One important problem is the coherence between spectral envelope and source signal after transformation. The transposed source may be such that noisy excitation falls into formant regions, which creates annoying noise in the transformed speech signal. Another problem is the fact that the signal transformation operator that have been developed so far take mainly into account the relations between different STFT frames, and treat the content of each STFT frame as locally stationary. These problems have been investigated partly in the project dealing with frequency domain transposition (see section 4.5), but further research would be required to resolve remaining issues. An alternative approach to speech signal transformation using an advanced analysis/synthesis scheme for speech signals has been developed in the thesis of G. Degottex ((PhD4), (G. Degottex et al. 2012)). This system analyses the speech signal in terms of voiced/unvoiced segments, fundamental frequency, glottal source parameters, noise and spectral envelope and vocal tract transfer function and provides means to transform all these parameters before re-synthesis. During re-synthesis the voiced part of signal is synthesized entirely from parameters which gives improved control such that a number of incoherencies can be avoided: e.g. when the voice is transformed into a more tense excitation the generation of additional sinusoids is required, which is rather straight forward when pulses are generated from scratch instead of transformed. The synthesis of pulses allows in principle to generate irregularities in the glottal pulse sequence that are characteristic for rough voices. An improved version of the algorithm is currently under development for experimentation in the context of voice conversion applications (PhD11).

9.2 Voice conversion

Related projects

(PhD5) Fernando Villavicencio (2010a). “Conversion de la voix de haute qualité”. Director X. Rodet, supervision A. Röbel. PhD thesis. Paris : Université Paris 6 (UPMC),
(PhD11) S. Huber (ongoing). “High Quality Voice Conversion by modelling and transformation of extended voice characteristics”. Director X. Rodet, supervision A. Roebel. PhD thesis. Université Paris 6 (UPMC),
(MF4) Axel Roebel et al. (2012). *Creation of the voice of Marilyn Monroe for the film "Marilyn" by P. Parreno*. Development and application of voice conversion algorithms,

[2]The research in the Respoken project was directed by Xavier Rodet. I contributed to this project with advice on speech signal transformation.

(MF5) Axel Roebel et al. (2012-2014). *Creation of voices of Philippe Petain, Léon Blum, Pierre Laval, Eduard Daladier, and Paul Reynaud, for the studio Maha Production*. Development and application of voice conversion algorithms

The term Voice Conversion is used for a subset of voice transformation tasks that aim to transform a given source voice into a specific target voice that generally is specified by means of a sound database of the target speaker. This problem is a natural extension of the high-level control of voice transformation algorithms that has been investigated in the VIVOS project. Accordingly, the research on voice conversion started after the VIVOS project with the PhD thesis of Fernando Villavicencio (PhD5). The research has later been continued by P. Lanchantin in the AngelStudio project[3] and is now continued in the PhD thesis of Stefan Huber (PhD11) in collaboration (CIFRE) with the company Acapela.

The first question to be solved when starting research on voice conversion is related to the features that will be used. Generally there are two classes of features that are considered for voice conversion: on one hand features related to speaking style and prosody, and on the other hand the features related to the vocal tract and excitation source characteristics [Kuwabara and Sagisaka 1995; Yannis Stylianou 2009]. Voice conversion research is in its majority investigating into the second set of features using the Gaussian mixture models to establish the feature conversion between source and target speaker [Kain 2001; Y. Stylianou 1996] and these models are trained generally on source and target speaker databases that contain the same phrases and are aligned (parallel), without using explicitly a phonetic annotation of the speech signals.

The main contribution of the thesis of F. Villavicencio was the investigation into the use of the new cepstrum based methods for spectral envelope estimation (see section 7.1) for voice conversion (F. Villavicencio et al. 2008, 2009). While the improved envelope estimation had beneficial effects on the converted voice similarity and quality (Fernando Villavicencio 2010b), the overall signal quality was still far from sufficient for professional projects (cinema). This was partly due to the problem of over-smoothing of the VTF after conversion, a problem that can be reduced by means of incorporating dynamic features into the conversion [T.Toda et al. 2007], as well as missing coherence between the converted envelope and the source excitation signal.

In the PhD thesis of Stefan Huber (PhD11), that is performed in collaboration with Acapela, the central objective was to evaluate strategies for the manipulation of the glottal source in the context of voice conversion systems. During his thesis, Stefan Huber has worked on improving the glottal pulse parameter estimation, and is currently investigating the conversion of excitation parameters in an analysis/synthesis system similar to the one proposed in (G. Degottex et al. 2012).

While the state of the art voice conversion systems are still far away from producing speech and conversion quality that would be sufficient for artistic production, we have recently received many requests from film and video production companies that ask for help with voice conversion projects, 2 in 2012 (MF4; MF5), and 3 in 2013 that are still under discussion. As these projects do not require real time conversion, we have implemented alternative offline strategies to achieve improved conversion and speech quality. The approach established in the context of the project (MF5) is particularly interesting because it allowed us to achieve very significant improvements of speech and conversion quality. As these new approaches are not yet published I will not give any more details here.

9.3 Singing synthesis

Related projects

(MF2) Axel Roebel and Joshua Fineberg (2006-2007). *Creation of voices for the opera* Lolita *of J. Fineberg*. Transformation of the voice of the main actor into girls singing voices,

(MA18) Luc Ardaillon (2013). "Singing synthesis". Stage Master 2 ATIAM, supervision A. Roebel. MA thesis. Université Paris VI Pierre et Marie-Curie

(RP9) *ANR project - Chanter – 2014-2017* (2014). Supervision of research on singing synthesis (WP2).

[3]The research in the FEDER project AngelStudio was directed by X. Rodet.

Singing synthesis is a research topic that was actively performed n the analysis synthesis team quite a while before I arrived at IRCAM. The work was mostly centred around the chant software program that performed synthesis by rules [Bennett and X. Rodet 1989; X. Rodet et al. 1984] and that produced results that still today are considered quite impressive[4], that however are restricted to synthesis of vowels.
Compared to speech synthesis there exist rather few research activities related to singing synthesis and most of these activities try to adapt technologies that were developed for speech synthesis. Current approaches cover the synthesis by rules mentioned above [Bennett and X. Rodet 1989; Berndtsson 1995; X. Rodet et al. 1984], concatenative synthesis [Bonada and Loscos 2003; Kenmochi and Ohshita 2007; Macon et al. 1997a,b], the conversion of speech into singing [Saitou et al. 2007], and singing synthesis based on HMM models [SynSY 2012]. Most of these systems have somewhat limited ambitions and do not take into account all the knowledge hat is available today. The most prominent system is certainly the singing synthesizer Vocaloid [Bonada and Loscos 2003; Kenmochi and Ohshita 2007] that is providing a complete system for singing synthesis.

9.3.1 Speech to singing conversion

My research related within this direction started with the project *Lolita* (MF2) of the composer Joshua Fineberg. For his imaginary Opera [J. Fineberg 2008] Joshua Fineberg wanted to create morphed voices between the main actor and the girl *Lolita* that was present only in his imagination. For this project I improved the shape invariant phase vocoder[5] and established means for gradual spectral envelope morphing (Axel Roebel and Joshua Fineberg 2007). The initial system did require manual adjustment of all parameter contours, but over time the system has been improved such that by today it is able to automatically transform speech into singing given an annotated input speech signal and a computer readable version of the target melody as well as vibrato control parameters. An experimental version of glottal pulse modification has been integrated demonstrating the importance of the modification of the source characteristics for expressive singing (A. Röbel et al. 2012). The modifications of the source are for the moment very ad hoc however, and a model of the singing voice source parameter contours relating vibrato, intensity, pitch and glottal source parameters eventually together with modifications of the spectral envelope is expected to create considerable improvements of the singing voice quality.

9.3.2 Singing voice synthesis

Motivated in part by the interesting results that have been obtained for speech to singing conversion I have recently engaged to revive the research activities related to the Chant project. The idea here is to develop an extended Chant synthesis program. The program is expected to be divided into a control part that generates the target contours for all speech signal parameters (F_0, intensity, ...) and that can be coupled with a small number of backends that will cover algorithms for speech to singing conversion, for singing synthesis based on concatenation and transformation of singing units in a database, and the F0F synthesis of the original Chant program.
Related research activities have started with the master thesis of Luc Ardaillon (MA18) in that an initial version of a singing synthesis backend based on concatenation and transformation has been developed. For this initial system the transformation did rely on signal transformation that were performed with the SuperVP phase vocoder using a new spectral domain phase vocoder based concatenation technique. In the future a more refined approach using the speech oriented analysis synthesis techniques (G. Degottex et al. 2012) discussed in section 9.1 and section 9.2 should be integrated. These topics will be adressed and the research will be continued in the ANR project Chanter (RP9) that starts in January 2014.

[4] see example at: http://anasynth.ircam.fr/home/media/singing-synthesis-chant-program
[5] see section 4.4

CHAPTER 10

DEVELOPMENT ACTIVITIES

As was already mentioned in the first chapter, there is a strong interest at IRCAM to transform research results into applications that can be used inside (by composers or the musical assistents at IRCAM) or outside of IRCAM (by the members of the IRCAM Forum or by industrial clients). Therefore, I have invested considerable effort into the development of C++ libraries for signal processing that will be summarized shortly in the following sections. Frédéric Cornu, who joined the analysis synthesis team in 2007, is participating in the development of all the C++ libraries mentioned below. He has contributed many important extensions and corrections, and since 2010 he is the main developer for all the signal processing libraries of the team.

10.1 MatMTL

The research in the analysis/synthesis team is performed in MATLAB, and more recently as well in Python, so that porting to C/C++ is necessary if the algorithms should be distributed to end users. Very soon after arriving at IRCAM I investigated into efficient porting of MATLAB implementations into C++. In 2001 and in collaboration with Patrice Tisserand, I started the development of the Matrix Mathematics Template library (MatMTL) (SW4) with the objective to combine the efficiency that was achieved by modern c++ template libraries (e.g. [ES1, Blitz], [ES2, MTL], and more recently [ES7, Eigen]) by means of modern c++ programming technics like meta programming and expression templates [Veldhuizen 1995] with a high-level programming interface that was as close as possible to MATLAB.
Today MatMTL has grown to support all element-wise numerical expressions of the MATLAB programming language. The use of expression templates allows computing complex sequences of element-wise operations without temporary variables, and the compiler automatically transforms many kinds of vector and matrix expressions into vectorized SIMD code. SIMD implementations of mathematical functions (log, exp, sin, cos) have been developed in the Sample Orchestrator project (RP2) by Frédéric Cornu. MatMTL is cross platform supporting Linux, MacOSX, Windows, and iOS and it is used for nearly all developments for industrial projects in the team. A MATLAB to MatMTL compiler (Mat2MTL) has been developed during the master thesis of B. Pratz (MA11), but this compiler has been discontinued due to the difficulties to support MATLAB's user defined classes.

10.2 SuperVP

SuperVP (SW2) is a library and stand alone program for STFT based signal processing. It incorporates a number of signal analysis algorithms (F0, spectral envelope, voiced/unvoiced frequency boundary) as well as an extended phase vocoder algorithm for signal transformation. Research and development for this software started in 1995 and was initially supervised by P. Depalle, who left IRCAM in 1999. As part of my research activities on phase vocoder based signal processing I became responsible as well for the development of SuperVP.
Over the last ten years the SuperVP signal processing library has been extended by new functionalities that have been described partly in the present document[1]. To support real-time applications a memory based API has been added to the SuperVP interface. A modular analysis/synthesis API has been developed in the

[1] see chapter 4, and section 7.1

Sample Orchestrator project (RP2) and later used for the development of the SVPX MaxMSP objects by Jean-Philippe Lambert in collaboration with Cycling'74.
Within IRCAM SuperVP is used in many different forms: The integration of the command line program in to the graphical user interface AudioSculpt (N. Bogaards et al. 2004) allows controlling the sound analysis and transformation engine by means of a graphical user interface. Another use is the integration in Max/MSP by means of the collection of Max/MSP objects developed by Norbert Schnell from the IMTR team [ES4, SuperVP for Max/MSP]. Since many years the command line version of SuperVP can be controlled from within OpenMusic thanks to the OM-SuperVP module by Jean Bresson and Jean Lochard ([ES6], [Bresson 2006]). The multitude of incarnations (AudioSculpt, Max/MSP, OpenMusic) and the high audio quality delivered by the library have generated a lot of interest in SuperVP from the artistic users at IRCAM. According to Arshia Cont, responsible for the IRCAM Forum and the interaction between the artistic and research departments, there are *very few artistic projects that do not use SuperVP in one of its different incarnations (AudioSculpt, Max/MSP, OM-SuperVP, command line) in one of the phases of the project*.
Besides the use at IRCAM the SuperVP library has been the object of numerous industrial projects and licences (LI1; MF1; LI2; LI3; LI4; MF3; LI5; LI6; LI7; LI8; LI9; LI10; LI11; LI12; RP4; LI13).

10.2.1 VoiceForger

VoiceForger is a library for voice transformation that is based on SuperVP speech processing facilities that simplifies the user interface due to the possibility to use high-level control of voice characteristics (Snorre Farner et al. 2009; S. Farner et al. 2008; Xavier Rodet et al. 2009). The development of the C++ version of the library voice transformation began early in 2009 under the Affective Avatars project where I supervised the development activities of S. Farner. In this project we implemented the basic architecture for the management of analysis and synthesis of the SuperVP library using the modular analysis/synthesis interface developed in the Sample Orchestrator project (RP2) that allows processing and reusing of various spectral domain analysis in the VoiceForger library. The library performs completely automatic reconfiguration of the chain of treatments based on the high-level specifications of source and target voices. The available high-level controls cover age and gender of the speaker, the ambitus (extend) of the fundamental frequency contour, as well as the breathiness of the voice. The programming interface of the VoiceForger library is organized in layers, such that alternatively to the high-level interface, many control parameters of the low-level interface of the underlying SuperVP library are exposed as well. The VoiceForger library constitutes the programming interface that is used in the IrcamTOOLS-TRaX plugin (LI12).

10.3 Pm2

Pm2 is a command line software, and c++ library for sound analysis/synthesis based on the sinusoidal signal model. Pm2 is a reimplementation of the earlier Pm software integrating recent state of the art algorithms for partial analysis that improve notably the estimation errors for the analysis of non-stationary sinusoids (A. Röbel 2007a,b, 2008). Pm2 supports sinusoidal analysis of harmonic and inharmonic sounds, and different modes of partial chord analysis. Pm2 is integrated as sinusoidal analysis/synthesis kernel in AudioSculpt and via the OM-Pm2 module in OpenMusic [ES8]. Moreover it performs sinusoidal based analysis of the ircamDescriptor software.

10.4 A2N

The development of the A2N library (SW6) has been started in the context of the industrial project Audio2Note (RP4) that was aiming to enhance the Ableton Live software by means of providing automatic midi transcription of arbitrary polyphonic music material. This library integrates research results described in section 8.2, as well as the onset detection algorithm described in section 4.2 that is implemented in the SuperVP library. A first version of the library was finalized in 2012 and has then been integrated in Ableton

Live Version 9, which has been released in early 2013. In the future we plan to integrate the A2N library or an enhanced version supporting signal separation into AudioSculpt.

10.5 Scientific Python

Before I conclude this chapter on development activities, I would like to discuss an important trend in my approach to scientific computing. After having used MATLAB for about 20 years as the main tool for my research activities I have recently switched to use python ([ES5, NumPy]/[ES3, SciPy]). This decision was due to the increasing costs for MATLAB software extensions (Parallelization Toolbox) and the difficulty to create command line executables from of MATLAB implementations. I started to investigate into the potential of scientific computing with Python in the speech to singing conversion project subsection 9.3.1, and later in the master thesis of Laura Dale (MA17). The conclusion of these tests were very satisfying showing many benefits of scientific computing with Python that are especially important if many different command line tools are used for the research, and if parallelization and interfacing with external libraries is needed. As a conclusion I have continuously migrated my research activities to Python/NumPY/SciPY, and this summer I have released within IRCAM a python toolbox for scientific computing.

CHAPTER 11

FUTURE RESEARCH DIRECTIONS

The topic of this last chapter is a vision of the next 5 years of research. It is clear that for none of the problems discussed until now I have found a final solution, and all of them may be the object of additional research, especially if I am confronted with a practical problem where the solutions proposed until now do not provide a satisfying solution. These basic consideration set aside, there is a direction in that the research topics will evolve and the present chapter will discuss 4 directions that seem to represent new approaches to signal transformation.

11.1 Transformation and synthesis of expressive sounds

The high quality obtained when transforming music and speech sounds with constant or slowly varying transformations has now led us to investigate into expressive signal transformation and signal synthesis. Under this term I gather transformation and synthesis problems that try to achieve at the same time high quality, realism and expressiveness that resemble natural sounds. For music signal synthesis and transformation this includes the synthesis and transformation of notes played by musical instruments. Here the work initiated in the SOR2 project (RP3) and the thesis of H. Hahn (PhD8) has to be continued and it has to be integrated with models that represent the style of interpretation of individual instrumentalists. Accordingly analysis and representation of musical ornamentation is essential (PhD10).

For speech and singing, this will concern the analysis, modelling and representation of the voice source parameters. As the voice source is not independent of the vocal tract filter modifications of the vocal tract filter may be required when changing the voice source. One of the central results in this work is the analysis of the parameters of the glottal source of the voice (Gilles Degottex et al. 2011). This analysis must be improved, particularly for high fundamental frequency sounds ($f_0 > 200$ Hz) for which the current algorithm is not yet working reliably. Another problem are signal segments with relaxed excitation for that only very few sinusoids are available to estimate the source characteristics. Here again the available analysis algorithms do not provide sufficient accuracy.

The increasing robustness and performance of the analysis of the speech source characteristics will open means to subsequently use the results in the context of processing and expressive synthesis singing and speech. One of the target applications is learning models of singing style from the signal models that are subsequently used for transformation and synthesis of singing. Voice conversion, that is the transformation of the speaker identity is a subject of ongoing research at IRCAM since 2006 (PhD5), for which there is still no satisfactory solution. An extension to be dealt with is cross language voice conversion. In the PhD thesis of S. Huber (PhD11) the research on voice conversion is continued including the management of glottal source and the prosody characteristics of the speakers. More advanced speech conversion algorithms that allow the manipulation of aperiodic voice source signals and emotional expressions are long-term objectives.

One of the key projects in this area will be the ANR project Chanter (RP9) that will start in early 2014 and deals with expressive singing synthesis. This project will investigate into new solutions for singing synthesis, but also into representation and synthesis of the singing style (voice type, vibrato, note articulation).

11.2 Analysis, separation and transformation of polyphonic sounds

The processing of the individual sound sources within polyphonic sounds is a signal processing problem that is only in its early stages. With respect to the transcription problem, we have developed a very competitive algorithm (C. Yeh 2008; C. Yeh et al. 2010) that achieves results that are rather satisfactory when compared to the state of the art. Nevertheless it is clear that a performance that achieves only 60% of note accuracy with very approximate note onset and offset locations is not sufficient for many applications.
Considering the transformation of the individual sound sources, there are already a few commercial products appearing that allow the manipulation of individual notes in polyphonic music[1]. Modification of individual notes in a polyphonic signal is closely linked to the signal separation problem. Today, the perceptual quality of the separated sounds is generally far from sufficient for using these signals for anything but mixing it into the original polyphonic sound at the original time position.
With respect to the separation of the individual sound sources for further processing or analysis the current focus of the research is clearly oriented towards nonnegative factorization methods. A rather promising approach to solve underdetermined separation of individual notes from polyphonic music has been presented recently in [FitzGerald et al. 2008].
Given the many interesting applications in this domain the research on polyphonic signal transcription and polyphonic music will constitute one of the central research topics for many years to come and the research will continue in several directions: The note transcription errors should be reduced. This concerns especially the errors related to notes played in harmonic relations. Here additional means should be established to take into account the musical context (harmony constraint) and the musical instruments that are present (acoustic constraints either given a priori or derived through the analysis or the musical performance). The note transcription should be accompanied by the information of the instrument playing the note, and an additional transcription of the part played by the drum kit. Finally, as mentioned before, the extraction (de-mixing) of the individual notes is an important problem to consider. On one hand it leads to many new applications, as for example remixing and/or re-orchestration, but on the other hand the separation can also provide information about the individual sources (e.g. the spectral envelope) that can then be inserted back into the transcription process. Means to use the partly redundant information that is present in the different channels of stereo and multi-track sounds similar as the approaches being currently developed in the 3DTVs project (RP5) have to be developed.
An example of the research to be performed in this area is the master thesis of Jordi Pons Puig (MA20) where we will try to integrate the audio 2 note algorithm that provides an approximate score representation of the music signal with the NTD based signal separation algorithms [FitzGerald et al. 2005, 2008]. The algorithms developed in the 3DTVs project for gunshot detection seem particularly interesting for drum detection and separation.

11.3 Sound textures

Work on time frequency representation of sound textures has recently begun in the thesis of Wei-Hsiang Liao (PhD12) and in the ANR project PHYSIS (RP8) that began in mid-2012. The objectives of the research is the modelling and transformation of sound textures and noise (see section 5.2). A first objective of this research is to establish signal transformation (time scaling) algorithms that preserve the perceptually relevant statistical characteristics of the sound signal.
An interesting question that arises in this context is related to the question to what extend the principles derived for texture modification can be used for time scaling music and speech signals. It is obviously straightforward to apply the same principles that are used for texture preservation to non texture signals. This has been done for example in [Dubnov et al. 2002; O'Regan and Kokaram 2007]. The fundamental difference with standard approaches to time scaling is the fact that the sound characteristics preserved when interpreting the sound as a texture are related to time averages, while standard algorithms exclusively take into account instantaneous signal characteristics (e.g. amplitude and frequency of sinusoids). It clear that the results of time scaling with texture based algorithms will strongly depend on the time scale that is used to derive the averages. This can be explained easily with an example. Assume that we have perfect

[1] Melodyne DNA of the German company Celemony, `http://www.celemony.com/dna`.

algorithms for time stretching not limited by any analysis problems and that we are working with a drum loop signal containing some measures with a few articulations changes in the hi-hat pattern.
Time stretching this drum loop with standard algorithms will preserve the number of measures in the loop when stretched resulting in a different tempo. When using texture stretching algorithms using the complete drum loop for the estimation of the signal statistics the texture time scaling algorithms will preserve the duration of the individual measures and add more measures at the end of the signal preserving the tempo of the original loop. The variations present in the hi-hat signal would be reproduced such that the statistical parameters remain unchanged which however does not require that the complete loop is periodically repeated. If on the other hand we would use texture characteristics estimated over individual measures and would time scale each of the regions individually we would expect to create drum loops that repeat the individual measures each with its characteristic hi-hat pattern.
Assuming further that we are able to interpolate the texture statistics we could try to synthesize with interpolated texture statistics. In this case we could expect to obtain some sort of interpolation between the different versions of the hi-hat patterns present in the original example. If we assume we would be able to estimate the texture statistics from a single analysis window, then preserving the evolution of the texture statistics would result in a time scaled signal very similar to that produced by the standard algorithm. That means in that case the tempo of the drum loop would change and the number of measures would be preserved.
Besides time scaling there are other texture transformations that should be investigated. High-level control of texture properties related to physical scene parameters (rain strength, wind strength and turbulence) should allow efficient texture synthesis control. The target values of the statistical texture characteristics would have to be established from analysis of target textures and the required sampling of the texture statistics (number of intermediate texture examples) is an open question.
Finally it would be very interesting to investigate whether the texture transformation algorithms can be integrated transparently into traditional time scaling algorithms to allow improved quality of the time scaling of the aperiodic signal components for example in speech signals. Again there are many open questions here, notably related to the separation of sinusoidal and aperiodic signal components and the time scale that is necessary to estimate the relevant signal statistics with sufficient precision.

11.4 Adaptive parameter selection for analysis and transformation

One of the problems remaining for all signal transformation algorithms is the fact that results depend on the parameters chosen for these algorithms. In section section 8.1 we discussed the algorithm SWIPE that uses a multi-resolution representation to achieve highly robust single source F_0 estimation. Our initial studies of adaptive time-frequency representations seem to indicate that the most sensitive parameter for STFT based signal transformation: the size of the analysis window determining the time-frequency resolution of the signal representation, can be selected by means of adaptive procedures.
These are two examples that could be understood as precursors of a new strategy to design algorithms based on the use of multiple representations with different resolutions from that the most appropriate resolution will then be selected automatically. To achieve this goal there is a long way to go. The intelligence of the signal processing expert will have to be built into the algorithms, which renders the development of the algorithms considerably more complex. Still, the benefit for the use of these algorithms could be very significant.
The objectives for coming years will be to continue the research on adaptive representations, and the integration of these representations in the algorithms for analysis and transformation audio signals. As a first example, I will continue the investigation of an adaptive F_0 estimation algorithm. Such an algorithm would significantly improve existing applications as for example the TRaX plugin (LI12) that today requires rough user presets to match the F_0 range to the user. A considerably more complex objective is the integration of the adaptive principles into signal transformation algorithms. As a first step we will soon integrate adaptive time varying time-frequency resolution (time varying window size) into AudioSculpt. In a longer future we will investigate into the adaptation of the sound processing algorithms to representations having adaptive the time-frequency resolution in time and frequency. An additional long-term goal would be use of adaptive time-frequency resolution for sound source separation.

CHAPTER 12

REFERENCES

Personal publications

Behles, G., S. Starke, and A. Röbel (1998). "Quasi-Synchronous and Pitch-Synchonous Granular Sound Processing with Stampede II". In: *Computer Music Journal* 22.2, pp. 44–51.

Bogaards, N., A. Röbel, and X. Rodet (2004). "Sound Analysis and Processing with AudioSculpt 2". In: *Proc. Int. Computer Music Conference (ICMC)*.

Burred, J. J., A. Röbel, and X. Rodet (2006). "An Accurate Timbre Model for Musical Instruments and its Application to Classification". In: *Proc. 1st Workshop on Learning the Semantics of Audio Signals (LSAS'06)*, pp.

Burred, J. J., A. Röbel, and T. Sikora (2010). "Dynamic Spectral Envelope Modeling for the Analysis of Musical Instrument Sounds". In: *IEEE Transactions on Audio, Speech and Language Processing* 18.3, pp. 663–674.

Burred, Juan José and Axel Roebel (2010). "A Segmental Spectro-Temporal Model of Musical Timbre". In: *Int. Conf. on Digital Audio Effects (DAFx)*.

Burred, Juan José, Axel Roebel, and Thomas Sikora (2009). "Polyphonic Musical Instrument Recognition Based on a Dynamic Model of the Spectral Envelope". In: *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 173 –176.

Chang, W.-C., A. W. Y. Su, C. Yeh, A. Röbel, and X. Rodet (2008). "Multiple-F0 tracking based on a high-order HMM model". In: *Proc. of the 11th Int. Conf. on Digital Audio Effects (DAFx'08)*, pp. 379–386.

Degottex, Gilles, Axel Roebel, and Xavier Rodet (2009a). "Glottal Closure Instant detection from a glottal shape estimate". In: *International Conference on Speech and Computer, SPECOM*. St-Petersbourg, Russie, pp. 226–231.

— (2009b). "Shape parameter estimate for a glottal model without time position". In: *International Conference on Speech and Computer, SPECOM*. St-Petersbourg, Russie, pp. 345–349.

— (2010). "Joint estimate of shape and time-synchronization of a glottal source model by phase flatness". In: *ICASSP*, pp. 5058–5061.

— (2011). "Phase minimization for glottal model estimation". In: *IEEE Transactions on Acoustics, Speech and Language Processing* 19.5, pp. 1080–1090.

Degottex, G., P. Lanchantin, A. Roebel, and X. Rodet (2012). "Mixed source model and its adapted vocal-tract filter estimate for voice transformation and synthesis". In: *Speech Communication* 55.2, pp. 278–294.

Farner, Snorre, Axel Roebel, and Xavier Rodet (2009). "Natural transformation of type and nature of the voice for extending vocal repertoire in high-fidelity applications". In: *The 35th International AES Conference (Audio for Games)*. CD proceedings only.

Farner, S., A. Roebel, C. Veaux, G. Beller, X. Rodet, and L. Ach (June 2008). *Voice transformation and speech synthesis for video games*. Paris Game Developers Conference. Paris, France.

Hahn, Henrik and Axel Roebel (2013). "Extended Source-Filter Model for Harmonic Instruments for Expressive Control of Sound Synthesis and Transformation". In: *Proc. 16th Int. Conf. on Digital Audio Effects (DAFx)*.

Hahn, H. and A. Roebel (2012). "Enhanced Source-Filter Model of Quasi-Harmonic Instruments for Sound Synthesis, Transformation and Interpolation". In: *Proc. Sound and Music Computing Conference (SMC)*.

Huber, Stefan and Axel Roebel (2014). "On the use of voice descriptors for glottal source shape parameter estimation". In: *Computer Speech and Language*. Accepted for publication.

Huber, Stefan, Axel Roebel, and Gilles Degottex (2012). "Glottal source shape parameter estimation using phase minimization variants". In: *Interspeech*.

Lanchantin, Pierre, Snorre Farner, Christophe Veaux, Gilles Degottex, Nicolas Obin, Gr?©gory Beller, and Stefan Huber (2011). "Vivos Voco: A survey of recent research on voice transformation at IRCAM". In: *International Conf on Digital Audio Effects*, pp. 277–285. URL: http://articles.ircam.fr/textes/Lanchantin11c/.

Liao, Wei-Hsiang, Axel Röbel, and Alvin W.Y. Su (2013). "On the Modeling of Sound Textures Based on the STFT Representation". In: *Proc. of 16th Int. Conf on Digital Audio Effects (DAFx)*. accepted for publication.

Liao, W.-H., A. Röbel, and A. W.-Y. Su (2012). "On stretching Gaussian noises with the phase vocoder". In: *Proc. Int. Conf. on Digital Audio Effects (DAFx)*. accepted for publication.

Lithaud, A., N. Bogaards, A. Röbel, and A. Gerszo (2008). *Audiosculpt Sound Examples*. http://support.ircam.fr/forum-ol-doc/audiosculpt/2.9.2-exe-son-FR-EN-1.0/co/exe-son-EN.html.

Liuni, Marco, Axel Roebel, Ewa Matusiak, Marco Romito, and Xavier Rodet (2013). "Automatic Adaptation of the Time-Frequency Resolution for Sound Analysis and Re-Synthesis". In: *IEEE Transactions on Audio, Speech, and Language Processing*. accepted for publication. URL: http://articles.ircam.fr/textes/Liuni13a/.

Mitsufuji, Y. and A. Röbel (2013). "Sound Source Separation Based on Non-Negative Tensor Factorization Incorporating Spatial Cue as Prior Knowledge". In: *ICASSP*.

Rigaud, Francois, Mathieu Lagrange, Axel Roebel, and Geoffroy Peeters (2011). "Drum Extraction From Polyphonic Music Based on a Spectro-Temporal Model of Percussive Sounds". In: *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 381 –384.

Röbel, A. (1993). "Neural Models of Nonlinear Dynamical Systems and their Application to Musical Signals". In German. PhD thesis. Technical University of Berlin.

— (1994a). "Dynamic pattern selection: Effectively training backpropagation neural networks". In: *Proceedings of the Intern. Conference on Artificial Neural Networks, ICANN'94*.

— (1994b). "Dynamic pattern selection for faster learning and improved generalization of neural networks". In: *Proceedings of the European Symposium on Artificial Neural Networks, ESANN'94*, pp. 187–193.

— (1995a). "Neural models for estimating Lyapunov exponents and embedding dimension from time series of nonlinear dynamical systems". In: *Proceedings of the Intern. Conference on Artificial Neural Networks, ICANN'95, Vol. II*. Paris, pp. 533–538.

— (1995b). "Neural networks for modeling time series of musical instruments". In: *Proc. of the International Computer Music Conference, ICMC*. Banff, Canada, pp. 424–428.

— (1995c). "Using neural models for analyzing time series of nonlinear dynamical systems". In: *Proceedings of the 5th International IMACS-Symposium on System Analysis and Simulation*.

— (1996). "Scaling Properties of Neural Networks for the Prediction of Time Series". In: *Proceedings of the 1996 IEEE Workshop on Neural Networks for Signal Processing VI*, pp. 190–199.

— (1997). "Neural Network Modeling of Speech and Music Signals". In: *Neural Information Processing Systems 9 (NIPS 96)*, pp. 779–785.

— (1998a). "Morphing Dynamical Sound Models". In: *Proceedings of the 1998 IEEE Workshop on Neural Networks for Signal Processing VIII*, pp. 409–418.

— (1998b). "Morphing Sound Dynamics". In: *Proceedings of the 1998 International Conference on Neural Networks and Brain*.

— (1999a). "Adaptive Additive Synthesis of Sound". In: *Proc. Int. Computer Music Conference, (ICMC'99)*, pp. 256–259.

Röbel, A. (1999b). "Morphing Sound Attractors". In: *Proc. of the 3rd. World Multiconference on Systemics, Cybernetics and Informatics (SCI'99) and the 5th. Int. Conference on Information Systems Analysis and Synthesis (ISAS'99), Vol 6*, pp. 340–347.
— (2001a). "Adaptive additive synthesis using spline based parameter trajectory models". In: *Proc. Int. Computer Music Conference (ICMC)*, pp. 369–372.
— (2001b). "Forschungsbericht: Adaptive Additive Synthese". Dfg research report (in German). URL: http://recherche.ircam.fr/equipes/analyse-synthese/roebel/publications/dfgrepshort.ps.
— (2001c). "Synthesizing natural sounds using dynamical models of sound attractors". In: *Computer Music Journal* 25.2, pp. 46–61.
— (2003a). "A new approach to transient processing in the phase vocoder". In: *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*, pp. 344–349.
— (2003b). "Transient detection and preservation in the phase vocoder". In: *Proc. Int. Computer Music Conference (ICMC)*, pp. 247–250.
— (2006a). "Adaptive additive modeling with continuous parameter trajectories". In: *IEEE Transactions on Speech and Audio Processing* 14.4, pp. 1440–1453.
— (2006b). "Estimation of partial parameters for non stationary sinusoids". In: *Proc. Int. Computer Music Conference (ICMC)*, pp. 167–170.
— (2006c). *Lecture Analysis, modeling and transformation of audio signals*. http://recherche.ircam.fr/equipes/analyse-synthese/roebel/amt_lecture.html. Material for Edgar-Var?®se guest professor lecture at TU Berlin.
— (2007a). "Frequency Slope Estimation and its Application for Non-Stationary Sinusoidal Parameter Estimation". In: *Proc. of the 10th Int. Conf. on Digital Audio Effects (DAFx'07)*, pp. 77–84.
— (2007b). "Parameter Estimation for linear AM/FM sinusoids using frequency domain demodulation". In: *Proc. of the 9th IASTED Int. Conf. on Signal and Image Processing (SIP2007)*.
— (2008). "Frequency-Slope-Estimation and Its Application to Parameter Estimation for Non-Stationary Sinusoids". In: *Computer Music Journal* 32.2, pp. 68–79.
— (2010a). "Between Physics and Perception: Signal Models for High Level Audio Processing". In: *Int. Conf on Digital Audio Effects (DAFx)*.
— (2010b). "Shape-invariant speech transformation with the phase vocoder". In: *Proc. International Conf. on Spoken Language Processing (InterSpeech)*, pp. 2146–2149.
Röbel, A., S. Huber, X. Rodet, and G. Degottex (2012). "Analysis and modification of excitation source characteristics for singing voice synthesis". In: *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5381–5384.
Röbel, A. and X. Rodet (2005a). "Efficient Spectral Envelope Estimation and its application to pitch shifting and envelope preservation". In: *Proc. of the 8th Int. Conf. on Digital Audio Effects (DAFx05)*, pp. 30–35.
— (2005b). "Real time signal transposition with envelope preservation in the phase vocoder". In: *Proc. Int. Computer Music Conference (ICMC)*, pp. 672–675.
Röbel, A., F. Villavicencio, and X. Rodet (2007). "On Cepstral and All-Pole based Spectral Envelope Modeling with unknown Model order". In: *Pattern Recognition Letters, Special issue on Advances in Pattern Recognition for Speech and Audio Processing* 28.6, pp. 1343–1350.
Röbel, A., M. Zivanovic, and X. Rodet (2004). "Signal decomposition by means of classification of spectral peaks". In: *Proc. Int. Computer Music Conference (ICMC)*, pp. 446–449.
Rodet, Xavier, Grégory Beller, Niels Bogaards, Gilles Degottex, Snorre Farner, Pierre Lanchantin, Nicolas Obin, Axel Roebel, Christophe Veaux, and Fernando Villavicencio (2009). "Transformation et synthèse de la voix parlée et de la voix chantée". In: *Parole et Musique*. Ed. by Stanislas Dehaene et Christine Petit. Odile Jacob.
Roebel, Axel and Joshua Fineberg (2007). *Speech to chant transformation with the phase vocoder*. URL: http://articles.ircam.fr/textes/Roebel07d/index.pdf.
van Praagh, Kay (1995). "Neuronale Modelle zur Vorhersage von Ozonimmissionen im Stadtgebiet von Berlin". Studienarbeit. Technische Universit"at Berlin.
Villavicencio, Fernando (2010b). "Conversion de la voix de haute qualité". In English. PhD thesis. Paris : Université Paris 6 (UPMC).

Villavicencio, F., A. Röbel, and X. Rodet (Apr. 2008). "Extending efficient spectral envelope modeling to mel-frequency based representation". In: *ICASSP*. Las Vegas, pp. 1625–1628. URL: http://mediatheque.ircam.fr/articles/textes/Villavicencio08a/.

— (Apr. 2009). "Applying improved spectral modeling for High Quality voice conversion". In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Taipei, Taiwan, pp. 4285–4288.

Vinet, H., G. Assayag, J. Burred, G. Carpentier, N. Misdariis, G. Peeters, A. Roebel, N. Schnell, D. Schwarz, and D. Tardieu (2011). "Sample Orchestrator : gestion par le contenu d'échantillons sonores". In: *Traitement du Signal* 28.3, pp. 417–468.

Wright, M., M. J. Beauchamp, K. Fitz, X. Rodet, A. Röbel, X. Serra, and G. Wakefield (2000). "Analysis/Synthesis Comparison". In: *Organised Sound* 5.3, pp. 173–189.

Yeh, C., N. Bogaards, and A. Röbel (2007). "Synthesized Polyphonic Music Database with Verifiable Ground Truth for Multiple F0 Estimation". In: *Proc. of the 8th Int. Conf. Music Information Retrieval (ISMIR 07)*, pp. 393–398.

Yeh, Chunghsin (2007). *Multiple F0 estimation for MIREX 2007*. URL: http://www.music-ir.org/mirex/abstracts/2007/F0_yeh.pdf.

Yeh, Chunghsin and Axel Roebel (2009). "The expected amplitude of overlapping partials of harmonic sounds". In: *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3169 –3172.

— (2010). *Multiple-F0 Estimation For MIREX 2010*. URL: http://articles.ircam.fr/textes/Yeh10b/index.pdf.

— (2011). *Multiple-F0 estimation for MIREX 2011*. URL: http://articles.ircam.fr/textes/Yeh11a/index.pdf.

Yeh, Chunghsin, Axel Roebel, and Wei-Chen Chang (2008). *Multiple F0 estimation for MIREX 2008*. URL: http://www.music-ir.org/mirex/abstracts/2008/mirex08_yeh.pdf.

Yeh, Chungshin and Axel Roebel (2009). *Multiple-F0 Estimation For MIREX 2009*. URL: http://articles.ircam.fr/textes/Yeh09b/index.pdf.

Yeh, C. and A Röbel (2004a). "A new score function for joint evaluation of multiple F0 hypothesis". In: *Proc. of the 7th Int. Conf. on Digital Audio Effects (DAFx'04)*, pp. 234–239.

— (2004b). "Physical principles driven joint evaluation of multiple F0 hypotheses". In: *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA'04)*.

— (2006a). "Adaptive noise level estimation". In: *Proc. of the 9th Int. Conf. on Digital Audio Effects (DAFx'06)*, pp. 145–148.

— (2006b). "Adaptive noise level estimation". In: *Workshop on Computer Music and Audio Technology (WOCMAT'06)*.

Yeh, C., A. Roebel, and X. Rodet (2010). "Multiple Fundamental Frequency Estimation and Polyphony Inference of Polyphonic Music Signals". In: *IEEE Transactions on Audio, Speech and Language Processing* 18.6, pp. 1116–1126. URL: http://articles.ircam.fr/textes/Yeh10a/.

Zivanovic, M., A. Röbel, and X. Rodet (2004). "A new approach to spectral peak classification". In: *Proc. of the 12th European Signal Processing Conference (EUSIPCO)*, pp. 1277–1280.

— (2007). "Adaptive Threshold Determination for Spectral Peak Classification". In: *Proc. of the 10th Int. Conf. on Digital Audio Effects (DAFx'07)*.

— (2008). "Adaptive Threshold Determination for Spectral Peak Classification". In: *Computer Music Journal* 32.2, pp. 57–67.

PostDoc researchers supervised or co-supervised

[PD1] CanadasQuesada, Francisco (2012). *Signal separation and signal transcription applied to singing voice*. Visting PostDoc Researcher at IRCAM, July - Sep. 2012.

[PD2] Özbek, Erdal (2011). *Sound reconstruction for clipped or missing audio samples*. Visting PostDoc Researcher at IRCAM, Oct 2011- Sep. 2012.

[PD3] Zivanovic, Miroslav (2003). *Detection, estimation and extraction of non-stationary sinusoids in noise: application to musical signals*. Visting PostDoc Researcher at IRCAM, Jan. 2003 - Juin. 2003.

[PD4] Zivanovic, Miroslav (2006). *Improving state of the art strategies for automatic detection and classification of signal components into sinusoidal and noise components*. Visting PostDoc Researcher at IRCAM, Sep. 2006 - Fev. 2007.

PhD Thesis supervised or co-supervised

[PhD1] Burred, J. J. (2008). "From Sparse Models to Timbre Learning: New Methods for Musical Source Separation". Visiting PhD student at IRCAM in 2006, Research on modeling of spectral envelopes of musical instruments for musical source separation, supervision A. Roebel. PhD thesis. Communication Systems Group, Technical University of Berlin.

[PhD2] Yeh, C. (2008). "Multiple Fundamental Frequency Estimation of Polyphonic Recordings". Director X. Rodet, supervision A. Röbel. PhD thesis. Université Paris 6 (UPMC).

[PhD3] Chang, W. C. (2009). Visting PhD student at IRCAM in 2008, research on tracking of multiple fundamental frequency candidates in polyphonic music. Supervision A. Roebel, and C. Yeh. PhD thesis. National Cheng Kung University, Tainan, Taiwan.

[PhD4] Degottex, G. (2010). "Glottal source and vocal-tract separation". Director X. Rodet, supervision A. Röbel. PhD thesis. Université Paris 6 (UPMC).

[PhD5] Villavicencio, Fernando (2010a). "Conversion de la voix de haute qualité". Director X. Rodet, supervision A. Röbel. PhD thesis. Paris : Université Paris 6 (UPMC).

[PhD6] Liuni, M. (2012). "Automatic Adaptation of Sound Analysis and Synthesis". Directors X. Rodet, et M. Romito, supervision A. Röbel. PhD thesis. Università di Firenze, Italie/Université Paris 6 (UPMC), France.

[PhD7] Chien, W. C. (2013). Visting PhD student at l'IRCAM from Aug. 2013 - Feb.2014, Estimating the source and filter from singing voice signals. Supervision A. Roebel. PhD thesis. National Taiwan University, Taiwan.

[PhD8] Hahn, H. (ongoing). "Synthèse et transformation des sons basés sur des modèles de type source-filtre étendu pour les instruments de musique". Director X. Rodet, supervision A. Roebel. PhD thesis. Université Paris 6 (UPMC).

[PhD9] Wang, T. M. (ongoing). Visting PhD student at IRCAM in 2011, Research on separation of drum signals from polyphonic music, supervision A. Roebel, C. Yeh, M. Lagrange. PhD thesis. National Cheng Kung University, Tainan, Taiwan.

[PhD10] Coler, H. v. (ongoing). "Expressive Sample Based Sound Synthesis". Director S. Weinzierl, supervision A. Roebel. PhD thesis. Technical University of Berlin, Allemagne.

[PhD11] Huber, S. (ongoing). "High Quality Voice Conversion by modelling and transformation of extended voice characteristics". Director X. Rodet, supervision A. Roebel. PhD thesis. Université Paris 6 (UPMC).

[PhD12] Liao, W. H. (ongoing). "Modelling and transformation of sound textures and environmental sounds". Directors X. Rodet et A. Su, supervision A. Roebel. PhD thesis. Université Paris 6 (UPMC) and National Cheng Kung University, Tainan, Taiwan.

Master thesis supervised or co-supervised

[MA1] Ehrke, J. (1993). "Dynamische Entwicklung der Topologie Neuronaler Netze. Erprobung von Verfahren auf der Basis des Backpropagation Algorithmus". Diplomarbeit (Master), supervision A. Roebel. MA thesis. Technische Universität Berlin, Allemagne.

[MA2] Assimakopoulos, T. (1994). "Modellierung nichtautonomer dynamischer Systeme durch verdeckt gesteuerte Neuronale Netze". Diplomarbeit (Master), supervision A. Roebel. MA thesis. Technische Universität Berlin, Allemagne.

[MA3] Behles, G. (1997). "Entwurf und Implementierung einer Echtzeitsoftware zur musikalischen Gestaltung auf der Basis von granularen und PSOLA Klangverarbeitungsverfahren". Diplomarbeit (Master), supervision A. Roebel. MA thesis. Technische Universität Berlin, Allemagne.

[MA4] Flohrer, T. (1999). “Anwendung von Algorithmen der blinden Signalverarbeitung zur verlustlosen Kompression von Audiosignalen”. Diplomarbeit (Master) (Master), supervision A. Roebel. MA thesis. Technische Universität Berlin, Allemagne.

[MA5] Starke, S. (2002). “Bestimmung der Grundfrequenz mit Hilfe des Algorithmus zur Maximierung der Likelihood der harmonischen Auswahl von Maxima”. Diplomarbeit (Master), supervision A. Roebel. MA thesis. Technische Universität Berlin, Allemagne.

[MA6] Schulz, S. (2002). “Bestimmung der Grundfrequenz mit Hilfe wahrscheinlichkeitstheoretischer Bewertung von Signalspektren unter Verwendung spektraler Energiedichtemodelle”. Diplomarbeit (Master), supervision A. Roebel. MA thesis. Technische Universität Berlin, Allemagne.

[MA7] Krauledat, M. (2003). “Fundamental frequency estimation”. Leonardo Internship, director X. Rodet, supervision A. Roebel. MA thesis. Westfälische Wilhelmsuniversität Münster.

[MA8] Yeh, C. (2003). “Multiple fundamental frequency estimation”. Rapport DEA Master ATIAM, supervision A. Roebel. MA thesis. Université Paris VI Pierre et Marie-Curie.

[MA9] Champion, G. (2004). “Application du modele additif shape invariant pour la transformation de la voix”. Rapport DEA Master ATIAM, supervision A. Roebel. MA thesis. Université Paris VI Pierre et Marie-Curie.

[MA10] Hamnane, S. (2006). “Traitement et représentation temps-fréquence des sons avec résolution adaptative”. Report DEA Master ATIAM, supervision A. Roebel et R. Gribonval (INRIA/IRISA). MA thesis. Université Paris VI Pierre et Marie-Curie.

[MA11] Pratz, B. (2008). “Projet compilateur MATLAB vers C++ – Mat2MTL”. Master Informatique (M1), supervision A. Roebel. MA thesis. University d’Orsay Paris XI.

[MA12] Contreras, J. (2009). “Transformation des modulations et des gestes ornementaux dans les sons musicaux”. Report Master 2 ATIAM, supervision A. Roebel. MA thesis. Université Paris VI Pierre et Marie-Curie.

[MA13] Houzet, R. (2010). “Formation de flux à partir d’une représentation objet de signaux musicaux polyphoniques”. Intership Master 2 ATIAM, supervision C. Yeh, M. Lagrange, A. Roebel. MA thesis. Université Paris VI Pierre et Marie-Curie.

[MA14] Rigaud, F. (2010). “Séparation de la partie percussive d’un morceau de musique”. Stage Master 2 ATIAM, supervision M. Lagrange, A. Roebel and G. Peeters. MA thesis. University Paris VI – Pierre et Marie-Curie.

[MA15] Hahn, H. (2010). “Generalisierte, grundtonabhängige Modelle für quasi-harmonische Instrumente”. Magisterarbeit (Master), supervision A. Roebel. MA thesis. Technische Universität Berlin, Allemagne.

[MA16] Bonnefoy, A. (2012). “Transcription de la partie percussive d’un morceau de musique”. Stage Master 2 ATIAM, supervision M. Lagrange, A. Roebel and G. Peeters. MA thesis. Université Paris VI Pierre et Marie-Curie.

[MA17] Dale, L. (2012). “Automatic Note Detection in Monopohonic Sound Files”. Sciences de l’Ingénieur, Master Mécanique (M1), orientation Acoustique, supervision A. Roebel. MA thesis. University Paris VI – Pierre et Marie-Curie.

[MA18] Ardaillon, Luc (2013). “Singing synthesis”. Stage Master 2 ATIAM, supervision A. Roebel. MA thesis. Université Paris VI Pierre et Marie-Curie.

[MA19] Saulnier, Hugo (2013). “Synthesis of Sound Textures”. Stage Master 2 ATIAM, supervision A. Roebel and S. O’Leary. MA thesis. Université Paris VI Pierre et Marie-Curie.

[MA20] Puig, Jordi Pons (2013). “Source separation for music signals”. MA thesis. Universitat Politècnica de Catalunya · BarcelonaTech (UPC).

Research projects supervised or performed

[RP1] Roebel, A. (2000). *Adaptive additive synthesis of non-stationary sounds*. Research scholarship at CCRMA, DFG project, Ref RO2277/1-1.

[RP2] *Projet ANR - Sample Orchestrator – 2006-2009* (2006). Task 3.1: Enhanced phase vocoder analysis and transformations in real time applications.

[RP3] *ANR project - Sample Orchestrator II – 2010-2013* (2010). Supervision of research on modeling timbre spaces of musical instruments by means of extended parameterized source filter models (WP2).

[RP4] Yeh, Chunghsin, Sèan O'Leary, and Axel Roebel (2010-2012). *Automatic midi annotation of polyphonic music*. Ableton. 2010-2012.

[RP5] *Projet FP7-ICT-2011, 3DTVS, 3DTV Content Search – 2011-2014* (2011). Direction des travaux sur 3D Audio & Multi Modal Content Analysis and Description (WP4).

[RP6] Lambert, Jean-Philippe and Axel Roebel (2011). *Development of an interface to SuperVP proccessing library in MaxMSP*. Cycling'74. 2011-2012.

[RP7] Mitsufuji, Yuki and Axel Roebel (2011-2012). *Collaborative research project with Sony Japan*. Source separation in multichannel audio recordings.

[RP8] *Projet ANR - PHYSIS, Physically informed and semantically controllable interactive sound synthesis – 2012-2015* (2012). Direction des travaux sur low level sound representation (WP3).

[RP9] *ANR project - Chanter – 2014-2017* (2014). Supervision of research on singing synthesis (WP2).

Industrial software licenses of research results

[LI1] Roebel, A. and X. Rodet (2002). *Mobistation*. Library for real time voice effects and voice transformation.

[LI2] — (2004). *MakeMusic*. Library for time stretching and pitch shifting with transient preservation for music signals.

[LI3] Roebel, A., Xavier Rodet, and Norbert Schnell (2005). *Voxler*. Library for voice transformation.

[LI4] Roebel, A. (2005). *Roni Music*. Library for time scaling and pitch shifting with transient preservation for music signals.

[LI5] Roebel, Axel (2008). *NeoCraft*. Library for transposition and time-scaling with transient preservation for music signals.

[LI6] Roebel, Axel, Snorre Farner, and Xavier Rodet (2008). *Xtranormal*. Library for voice transformation.

[LI7] Roebel, A. and X. Rodet (2008). *MakeMusic*. Library for fundamental frequency estimation for monophonic instrumental sounds.

[LI8] Roebel, Axel (2009a). *MXP4*. Library for time scaling and pitch shifting with transient preservation for music signals.

[LI9] — (2009b). *UniversSons - MachFive 3*. Library for time scaling and pitch shifting with transient preservation for music signals.

[LI10] Roebel, Axel, Snorre Farner, and Xavier Rodet (2010). *Xtranormal*. Library for voice transformation with high level control.

[LI11] Roebel, Axel (2010a). *OhmForce*. Use of supervp library for sample precise time scaling with transient preservation for music signals.

[LI12] — (2010b). *IRCAMTools-TRaX*. Development of a professional audio plugin for music and voice transformation with high level controls and high sound quality, in collaboration with the French software development company Flux in Orléans.

[LI13] Lochard, Jean (2013). *IRCAM MAX Collection : SuperVP*. Use of supervp - max objects for sound effects patches in Max/MSP.

Software development

[SW1] Röbel, A. and Frédéric Cornu. *PM2: command line tool and c++ library for analysis/synthesis with sinusoidal models*. A. Roebel: Scientific direction and software development since 2000, F. Cornu: software development since 2007.

[SW2] Roebel, A., F. Cornu, and P. Depalle (2000). *SuperVP: command line tool and c++ library for audio treatment in real time or non real time based on an extended phase vocoder*. A. Roebel: Scientific direction and software development since 2000, F. Cornu: software development since 2007, P. Depalle: initial version before 2000.

[SW3] Röbel, A., Patrice Tisserand, Fabien Tisserand, and Diemo Schwarz. *EaSDIF: c++ library for high-level manipulation of SDIF files*. F. Tisserand : initial development, P. Tisserand : software development 2002 - 2005, D. Schwarz : initial concepts, A. Roebel : software development 2002 -. URL: `http://sourceforge.net/projects/sdif/files/Easdif/`.

[SW4] Röbel, A. and Frédéric Cornu. *MatMTL: c++ template library implementation for matrix and vector operations allowing efficient implementation of Matlab code in C++*. A. Roebel : initial concept, and software development since 2003, F. Cornu : software development since 2007.

[SW5] Farner, S., X. Rodet, A. Roebel, and F. Cornu (2009). *VoiceForger: C++ library for real time speech and music conversion with high-level control based on SuperVP*. X. Rodet: Scientific direction, S. Farner: Research and Software development until 2009, A. Roebel: Scientific direction and Software development since 2010, F. Cornu: software development since 2010.

[SW6] Yeh, C., S. O'Leary, and A. Röbel. *Audio2Note transcription library*. A. Roebel: Scientific direction and software development since 2003, C. Yeh: research and software development since 2008-2011, S. O'Leary: research and software development in 2012.

Music and film projects

[MF1] Rodet, Xavier, Axel Roebel, and Alain Lithaud (2003). *Creation of voices for* Tiresia *from B. Bonello*. Transformation of the voice of a female actor into a male voice.

[MF2] Roebel, Axel and Joshua Fineberg (2006-2007). *Creation of voices for the opera* Lolita *of J. Fineberg*. Transformation of the voice of the main actor into girls singing voices.

[MF3] Rodet, Xavier, Snorre Farner, Axel Roebel, and Alain Lithaud (2007). *Creation of voices for* Les Amours d'Astrée et de Céladon *from E. Rohmer*. Transformation of the voice of the actor playing Céladon into a femal voice.

[MF4] Roebel, Axel, Nicolas Obin, and Stefan Huber (2012). *Creation of the voice of Marilyn Monroe for the film "Marilyn" by P. Parreno*. Development and application of voice conversion algorithms.

[MF5] Roebel, Axel, Nicolas Obin, Stefan Huber, and Marco Liuni (2012-2014). *Creation of voices of Philippe Petain, Léon Blum, Pierre Laval, Eduard Daladier, and Paul Reynaud, for the studio Maha Production*. Development and application of voice conversion algorithms.

External References

Schulz, A. (1998). "Entwicklung einer Simulationsumgebung zur Analyse von Sigma Delta AD/DA-Wandlern". Magisterarbeit (Master), supervision A. Roebel. MA thesis. Technische Universität Berlin, Allemagne.

Abe, M. and J. O. Smith (2004). *Design Criteria for the Quadratically Interpolated FFT Method (I): Bias due to Interpolation*. Tech. rep. STAN-M-117. available at `http://ccrma.stanford.edu/STANM/stanms/stanm114/index.html`. Stanford University, Department of Music.

— (2005). "AM/FM rate estimation for time-varying sinusoidal modeling". In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 201–204 (Vol. III).

Alku, Paavo (1992). "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering". In: *Speech Communication* 11.2, pp. 109–118.

Alku, Paavo, Carlo Magi, Santeri Yrttiaho, Tom Bäckström, and Brad Story (2009). "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering". In: *the Journal of the Acoustical Society of America* 125, pp. 3289–3305.

Amatriain, X., J. Bonada, A. Loscos, and X. Serra (2002). "Spectral Processing". In: *Digital Audio Effects*. Ed. by U. Zölzer. John Wiley & Sons. Chap. 10, pp. 373–438.

Athineos, M. and D. Ellis (2003). “Sound texture modeling with linear prediction in both time and frequency domains”. In: *ICASSP*. Vol. 5, pp. 648–651.

Auger, F. and P. Flandrin (1995). “Improving the readability of time-frequency and time-scale representations by the reassignment method”. In: *IEEE Trans. on Signal Processing* 43.5, pp. 1068–1089.

Beller, Grégory (2009). “Analyse et Modèle Génératif de l’Expressivité, Application à la Parole et à l’Interprétation Musicale”. PhD thesis. University Paris VI - Pierre and Marie Curie.

Bello, J. P., L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler (2005). “A tutorial on onset detection in music signals”. In: *IEEE Transactions on Speech and Audio Processing* 13.5 Part 2, pp. 1035–1047.

Bello, J. P., C. Duxbury, M. Davies, and M. Sandler (2004). “On the use of phase and energy for musical onset detection in the complex domain”. In: *IEEE Signal Processing Letters* 11.6, pp. 553–556.

Bennett, G. and X. Rodet (1989). “Synthesis of the Singing Voice”. In: *Current Directions in Computer Music Research*. MIT Press, pp. 19–44.

Berndtsson, G. (1995). “The KTH rule system for singing synthesis”. In: *STL-QPSR*. Vol. 36. 1. KTH, pp. 1–22.

Bertin, Nancy, Roland Badeau, and Emmanuel Vincent (2009). “Enforcing Harmonicity and Smoothness in Bayesian Non-negative Matrix Factorization Applied to Polyphonic Music Transcription.” In: *IEEE Trans. on Audio, Speech and Language Processing* 18.3, pp. 538–549.

Bertin, Nancy, Emmanuel Vincent, and Roland Badeau (2009). *Fast Bayesian constrained NMF for polyphonic pitch transcription*. URL: http://www.music-ir.org/mirex/abstracts/2009/BVB.pdf.

Betser, Michaël (2009). “Sinusoidal polynomial parameter estimation using the distribution derivative”. In: *IEEE Transactions on Signal Processing* 57.12, pp. 4633–4645.

Bishop, Christopher M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Black, Alan W., Heiga Zen, and Keiichi Tokuda (2007). “Statistical parametric speech synthesis”. In: *ICASSP*. Vol. IV, pp. 1229–1232.

Boeck, S., F. Krebs, and M. Schedl (2012). “Evaluating the online capabilities of onset detection methods”. In: *Proceedings ISMIR*.

Bonada, J. (2000). “Automatic technique in frequency domain for near-lossless time-scale modification of audio”. In: *Proc. of the Int. Computer Music Conference (ICMC)*, pp. 396–399.

Bonada, J. and A. Loscos (2003). “Sample-based Singing-voice Synthesizer by Spectral Concatenation”. In: *Proc. of Stockholm Music Acoustics Conf.* Pp. 439–442.

Bozkurt, B. and T. Dutoit C. d’Alessandro B. Doval (2012). “Zeros of Z-Transform Representation With Application to Source-Filter Separation in Speech”. In: *IEEE Signal Porcessing Letters* 12.4, pp. 344–347.

Bresson, Jean (2006). “Sound processing in OpenMusic”. In: *Proceedings of the International Conference on Digital Audio Effects (DAFx-06)*.

Camacho, Arturo (2007). “SWIPE: A sawtooth waveform inspired pitch estimator for speech and music”. PhD thesis. University of Florida. URL: http://www.kerwa.ucr.ac.cr:8080/handle/10669/536.

Camacho, Arturo and John G Harris (2008). “A sawtooth waveform inspired pitch estimator for speech and music”. In: *The Journal of the Acoustical Society of America* 124, p. 1638. URL: http://link.aip.org/link/?JASMAN/124/1638/1.

Cardoso, J. F. (1997). “Statistical Principles of Source Separation”. In: *Proc. Of 11th IFAC Symposium on system identification (SYSID?Äô97)*, pp. 1837–1845.

Cohen, L. (1995). *Time-frequency analysis*. Signal Processing Series. Prentice Hall.

Cotton, Courtenay V. and Daniel PW Ellis (2011). “Spectral vs. spectro-temporal features for acoustic event detection”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 69–72.

Day, M. and S. Godsill (2002). “Bayesian Harmonic Models for Musical Signal Analysis”. In: *Proc. Seventh Valencia International meeting (Bayesian Statistics 7)*.

de Boor, C. (1987). “Multivariate Approximation”. In: In *The State of the Art in Numerical Analysis*. A. Iserles and M. J. Powell (Eds.),Clarendon Press, Oxford, pp. 87–109.

de Cheveigné, A. and H. Kawahara (2002). "YIN, a fundamental frequency estimator for speech and music". In: *Journal Acoust. Soc. Am.* 111.4, pp. 1917–1930.

Depalle, P. and X. Rodet (1995). "Sound synthesis process". US5401897. URL: http://www.google.com/patents/US5401897.

Dessein, A., A. Cont, and G. Lemaitre (2012). "Real-Time detection of overlapping sound events with non-negative matrix factorization". In: *Matrix Information Geometry*. Ed. by F. Nielsen and R. Bhatia. Springer, pp. 341–372.

Ding, Y. and Q. Qian (1997). "Processing of Musical Tones Using a Combined Quadratic Polynomial-Phase Sinusoid and Residual (QUASAR) Signal Model". In: *Journal of the Audio Engineering Society* 45.7/8, pp. 571–585.

Dixon, S. (2006). "Onset detection revisted". In: *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx)*, pp. 133–137.

Dolson, M. (1986). "The phase vocoder: A tutorial". In: *Computer Music Journal* 10.4, pp. 14–27.

Dörfler, M. (2011). "Quilted Gabor frames: A new concept for adaptive time-frequency representation". In: *Advances in Applied Mathematics* 47.4, pp. 668–687.

Doval, B. and X. Rodet (1991). "Estimation of fundamental frequency of musical sound signals". In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 3657–3660 (Vol. V).

— (1993). "Fundamental Frequency Estimation and Tracking using Maximum Likelihodd Harmonic Matching and HMMs". In: *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'93)*, pp. 221–224.

Dressler, Karin (2012). *Multiple fundamental frequency extraction for MIREX 2012*. URL: http://www.music-ir.org/mirex/abstracts/2012/KD3.pdf.

Drugman, Thomas, Baris Bozkurt, and Thierry Dutoit (2011). "Causal–anticausal decomposition of speech using complex cepstrum for glottal source estimation". In: *Speech Communication* 53.6, pp. 855–866.

Duan, Z., G. Mysore, and P. Smaragdis (2012). "Online PLCA for real-time semi-supervised source separation". In: *Latent Variable Analysis and Signal Separation*, pp. 34–41.

Dubnov, S., Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman (2002). "Synthesizing sound textures through wavelet tree learning". In: *IEEE Computer Graphics and Applications* 22.4, pp. 38–48.

Dudas, R. (2002). "Spectral Envelope Correction for Real-Time Transposition: Proposal of a "Floating-Formant" Method". In: *Proc Int. Conf on Computer Music (ICMC)*, pp. 126–129.

Duxbury, C., M. Davies, and M. Sandler (2001). "Separation of transient information in musical audio using multiresolution analysis techniques". In: *Proceedings of the Digital Audio Effects (DAFx)*, pp. 1–4.

— (2002). "Improved time-scaling of musical audio using phase locking at transients". In: *112th AES Convention*. Convention Paper 5530.

Evangelista, G., M. Dörfler, and E. Matusiak (2012). "Phase Vocoders with arbitrary frequency band selection". In: *Proceedings of the 9th Sound and Music Computing Conference (SMC)*.

Fant, G. (1995). "The LF-model revisited. Transformations and frequency domain analysis." In: *Quarterly Progress and Status Report, Dept of speech, music and hearing, KTH* 36.2-3, pp. 119–156.

Fant, Gunnar, Johan Liljencrants, and Qi-guang Lin (1985). "A four-parameter model of glottal flow". In: *STL-QPSR, Dept of speech, music and hearing, KTH* 4.1985, pp. 1–13.

Févotte, C. and A. Ozerov (2010). "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation". In: *IEEE Trans. on Acoustics, Speech, and Language Processing* 18.3, pp. 550–563.

Fineberg, J. (2008). *Lolita*. http://brahms.ircam.fr/works/work/18304/.

FitzGerald, D., M. Cranitch, and E. Coyle (2005). "Non-negative tensor factorisation for sound source separation". In: *Proc. of the Irish Signals and Systems Conf. (ISCC) 2005*.

— (2008). "Extended nonnegative tensor factorisation models for musical sound source separation". In: *Computational Intelligence and Neuroscience* 2008.

Fitz, K., L. Haken, and P. Christensen (2000). "Transient Preservation under Transformation in an Additive Sound Model". In: *Proc. of the Int. Computer Music Conference (ICMC))*, pp. 392–395.

Flanagan, J. L. and R. M. Golden (1966). "Phase Vocoder". In: *Bell System Technical Journal* 45, pp. 1493–1509.

Freed, A., X. Rodet, and P. Depalle (1992). “Synthesis and Control of Hundreds of Sinusoidal Partials on a Desktop Computer without Custom Hardware”. In: *Int. Conf. on Signal Processing Applications and Technology*. Available at http://mediatheque.ircam.fr/articles/textes/Rodet92/note.html.

Gobl, Christer and Ailbhe Ní Chasaide (2003). “The role of voice quality in communicating emotion, mood and attitude”. In: *Speech Communication* 40.1, pp. 189–212.

Goodwin, M. and X. Rodet (1994). “Efficient Fourier Synthesis of Nonstationary Sinusoids”. In: *Proceedings of the International Computer Music Conference (ICMC)*, pp. 333–334.

Grindlay, Graham and Daniel P.W. Ellis (2010). “A probabilistic subspace model for multi-instrument polyphonic transcrption”. In: *Proc. of Intern. Soc. for Music Information Retrieval. Conf (ISMIR 2010)*.

Grofit, S. and Y. Lavner (2008). “Time-Scale Modification of Audio Signals Using Enhanced WSOLA With Management of Transients”. In: *IEEE Transactions on Audio, Speech & Language Processing* 16.1, pp. 106–115.

Hainsworth, S. and M. Macleod (2003). “Onset detection in musical audio signals”. In: *Proc. Int. Computer Music Conference (ICMC)*, pp. 163–166.

Hainsworth, S. W., M. D. Macleod, and P. J. Wolfe (2001). “Analysis of reassigned spectrograms for musical transcription”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 23–26.

Hélie, Thomas and David Roze (2008). “Sound synthesis of a nonlinear string using Volterra series.” In: *Journal of Sound and Vibration* 314, pp. 275–306.

Hélie, Thomas and V. Smet (2008). “Simulation of the weakly nonlinear propagation in a straight pipe: application to a real-time brassy audio effect.” In: *Mediterranean Conference on Control and Automation.* Vol. 16, pp. 1580–1585.

Imai, S. and Y. Abe (1979). “Spectral envelope extraction by improved cepstral method”. In: *Electron. and Commun. in Japan* 62-A.4. in Japanese, pp. 10–17.

El-Jaroudi, A. and J. Makhoul (1991). “Discrete All-Pole Modeling”. In: *IEEE Transactions on Signal Processing* 39.2, pp. 411–423.

Jean Laroche, Santa Cruz CA and Ben Lomond CA Mark Dolson (Apr. 15, 2003). “Phase-vocoder pitch-shifting”. Patent US 6549884 (US). URL: `http://www.patentlens.net/patentlens/patent/US_6549884/en/`.

Kain, A. (2001). “High resolution voice transformation”. PhD thesis. OGI School of Science, Engineering at Oregon Health, and Science University.

Kameoka, Hirokazu, Takuya Nishimoto, and Shigeki Sagayama (2007). “A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering”. In: *IEEE Trans. on Audio, Speech and Language Processing* 15.3, pp. 982–994.

Kawahara, H. (1997). “Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited”. In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. Vol. 2, pp. 1303–1306.

Kenmochi, H. and H. Ohshita (2007). “VOCALOID-commercial singing synthesizer based on sample concatenation”. In: *Proc InterSpeech*, pp. 409–410.

Klapuri, A. (1999). “Sound onset detection by applying psychoacoustic knowledge”. In: *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’99)*. Vol. 6, pp. 3089–3092.

Klapuri, Anssi (2004). “Signal Processing Methods for the Automatic Transcription of Music”. PhD thesis. Tampere University of Technology.

— (2008). “Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model”. In: *IEEE Trans. on Audio, Speech and Language Processing* 16.2, pp. 255–266.

Kominek, John and Alan W Black (2004). “The CMU Arctic speech databases”. In: *Proc. of the 5th ISCA Workshop on Speech Synthesis*, pp. 223–224. URL: `http://www.isca-speech.org/archive_open/ssw5/ssw5_223.html`.

Kuwabara, H. and Yoshinori Sagisaka (1995). “Acoustic charcteristics of pseaker individuality: Control and conversion”. In: *Speech Communication* 16.2, pp. 165–173.

Lacoste, Alexandre and Douglas Eck (2007). “A Supervised Classification Algorithm for Note Onset Detection”. In: *EURASIP J. Appl. Signal Process.* 1, pp. 153–166.

Lagrange, Mathieu, Sylvain Marchand, and Jean-Bernard Rault (2002). “Sinusoidal parameter extraction and component selection in a non stationary model”. In: *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx)*, pp. 59–64.

Laroche, J. (2003). “Frequency-domain techniques for high-quality voice modification”. In: *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*.

Laroche, J. and M. Dolson (1999a). “Improved Phase Vocoder Time-Scale Modification of Audio”. In: *IEEE Transactions on Speech and Audio Processing* 7.3, pp. 323–332.

— (1999b). “New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing and other exotic audio modifications”. In: *Journal of the AES* 47.11, pp. 928–936.

Lee, D. D. and H. Sebastian Seung (2000). “Algorithms for non-negative matrix factorization”. In: *Advances in Neural Information Processing Systems*. MIT Press, pp. 556–562.

Levine, S. (1998). “Audio representations for data compression and compressed domain processing”. PhD thesis. Department of Electrical Engineering, CCRMA, Stanford University.

Levine, S. and J. O. Smith (1998). “A Sines+Transients+Noise Audio Representation for Data Compression and Time/Pitch-Scale Modifications”. In: *105th AES Convention*. Preprint 4781.

Lewicki, Michael S (2002). “Efficient coding of natural sounds”. In: *Nature neuroscience* 5.4, pp. 356–363. URL: http://www.nature.com/neuro/journal/v5/n4/abs/nn831.html.

Lorenzo-Trueba, J., Roberto Barra-Chicote, Tuomo Raitio, Nicolas Obin, Paavo Alku, Junichi Yamagishi, and Juan M. Montero (2012). “Towards Glottal Source Controllability in Expressive Speech Synthesis”. In: *Proc. of Interspeech*.

Lu, Lie, Liu Wenyin, and Hong-Jiang Zhang (2004). “Audio textures: theory and applications”. In: *IEEE Transactions on Speech and Audio Processing* 12.2, pp. 156–167.

Macon, M., L. Jensen-Link, J. Oliverio, M. Clements, and E. B. George (1997a). “A system for singing voice synthesis based on sinusoidal modeling”. In: *ICASSP*. Vol. I, pp. 435–438.

— (1997b). “Concatenation-based MIDI-to-Singing Voice Synthesis”. In: *Proc. of the 103rd Meeting of Audio Engineering Society*.

Marchand, Sylvain and Philippe Depalle (2008). “Generalization of the derivative analysis method to non-stationary sinusoidal modeling”. In: *Proceedings of the Digital Audio Effects (DAFx) Conference*, pp. 281–288. URL: http://hal.archives-ouvertes.fr/hal-00351950/.

Mark Dolson, Ben Lomond CA (Aug. 29, 2000). *System for fourier transform-based modification of audio*. Patent. US 6112169. URL: http://www.patentlens.net/patentlens/patent/US_6112169/en/.

Markel, J. D. and A. H. Gray (1976). *Linear Prediction of Speech*. Springer Verlag.

Marques, J. S. and L. B. Almeida (1986). “A background for sinusoid based representation of voiced speech”. In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1233–1236 (Vol. II).

— (1989). “Frequency-Varying Sinusoidal Modeling of Speech”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37.5, pp. 763–765.

Masri, P. and A. Bateman (1996). “Improved modelling of attack transients in music analysis-resynthesis”. In: *Proceedings of the International Computer Music Conference (ICMC)*, pp. 100–103.

McAulay, R. J. and T. F. Quatieri (1986). “Speech analysis-synthesis based on a sinusoidal representation”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 34.4, pp. 744–754.

McDermott, J.H., A.J. Oxenham, and E.P. Simoncelli (2009). “Sound texture synthesis via filter statistics”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 5, pp. 297–300.

McDermott, Josh H. and Eero P. Simoncelli (2011). “Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis”. In: *Neuron* 71.5, pp. 926–940.

Mettin, R. and G. Mayer-Kress (1996). “Chaotic attractors from homotopic mixing of vector fields”. In: *International Journal of Bifurcation and Chaos* 6.2, pp. 395–408.

IMIRSEL (Sept. 2005). *MIREX 2005:Audio Onset Detection Results*. http://www.music-ir.org/mirex/wiki/2005:Audio_Onset_Detection_Results. ISMIR 2005, London, Great Britain.

— (Oct. 2006). *MIREX 2006:Audio Onset Detection Results*. http://www.music-ir.org/mirex/wiki/2006:Audio_Onset_Detection_Results. ISMIR 2006, Victoria, Canada.

IMIRSEL (Sept. 2007). *MIREX 2007:Audio Onset Detection Results*. http://www.music-ir.org/mirex/wiki/2007:Audio_Onset_Detection_Results. ISMIR 2007, Vienna, Austria.
— (2009). *MIREX 2009:Audio Onset Detection Results*. http://www.music-ir.org/mirex/wiki/2009:Audio_Onset_Detection_Results.
— (2010). *MIREX 2010:Audio Onset Detection Results*. http://www.music-ir.org/mirex/wiki/2010:Audio_Onset_Detection_Results.
— (2011). *MIREX 2011:Audio Onset Detection Results*. http://www.music-ir.org/mirex/wiki/2011:Audio_Onset_Detection_Results.
M. J. Wright (2008). "The shape of an instant: Measuring and modeling perceptual attack time with probability density functions". PhD thesis. Department of Music, Stanford University.
Moulines, Eric and Francis Charpentier (1990). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones". In: *Speech Communication* 9.5, pp. 453–467.
Moulines, Eric and Jean Laroche (1995). "Non-parametric techniques for pitch-scale and time-scale modification of speech". In: *Speech Communication* 16.2, pp. 175–205.
Obin, N. (2005). "Evaluation des algorithmes d'estimation de la fréquence fondamentale mono-pitch". MA thesis. University of California Berkeley.
O'Regan, Deirdre and Anil Kokaram (2007). "Multi-resolution sound texture synthesis using the dual-tree complex wavelet transform". In: *Proc. European Signal Processing Conference (EUSIPCO)*.
Ozerov, A. and C. Févotte (2010). "Notes on Nonnegative Tensor Factorization of the Spectrogram for Audio Source Separation: Statistical Insights and Towards Self-Clustering of the Spatial Cues". In: *7th International Symposium on Computer Music Modeling and Retrieval*.
Parker, T. S. and L. O. Chua (1987). "Chaos: A Tutorial for Engineers". In: *Proc. of the IEEE* 75.8, pp. 982–1008.
Peeters, G. and X. Rodet (1999). "SINOLA: A new analysis/synthesis method using spectrum peak shape distortion, phase and reassigned spectrum". In: *Proc. Int. Computer Music Conference*, pp. 153–156.
Peleg, S. and B. Porat (1991). "Linear FM signal parameter estimation from discrete-timeobservations". In: *IEEE transactions on aerospace and electronic systems* 27.4, pp. 607–616.
Puckette, M. S. (1995). "Phase-locked vocoder". In: *Proceedings of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 222–225.
Quatieri, T. F. and R. J. McAulay (1992). "Shape invariant time-scale and pitch modification of speech". In: *IEEE Transactions on Signal Processing* 40.3, pp. 497–510.
Raitio, Tuomo, Antti Suni, Junichi Yamagishi, Hannu Pulakka, Jani Nurminen, Martti Vainio, and Paavo Alku (2011). "HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering". In: *ITASLP* 19.1, pp. 153–165.
Refenes, A.-P., ed. (1995). *Neural Netorks in the Captial Markets*. John Wiley & Sons.
Rodet, Xavier (1998). "Musical sound signals analysis/synthesis: Sinusoidal+ residual and elementary waveform models". In: *Applied Signal Processing* 4.3, pp. 131–141.
Rodet, X. and P. Depalle (1992). "A new additive synthesis method using inverse Fourier transform and spectral envelopes". In: *Proceedings of the International Computer Music Conference (ICMC)*, pp. 410–411.
Rodet, X. and F. Jaillet (2001). "Detection and modeling of fast attack transients". In: *Proc. Int. Computer Music Conference (ICMC)*, pp. 30–33.
Rodet, X., Y. Potard, and J.-B. Barriere (1984). "The CHANT project : from synthesis of the singing voice to synthesis in general". In: *Computer Music Journal* 8.3, pp. 15–31.
Rosenberg, A. E. (1971). "Effect of Glottal Pulse Shape on the Quality of Natural Vowels". In: *jasa* 49.2, pp. 583–590.
Roucos, S. and A. Wilgus (1985). "High Quality Time-Scale Modification for Speech". In: *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 493–496.
Saint-Arnaud, Nicolas (1995). "Classification of sound textures". MA thesis. Massachusetts Institute of Technology.
Saint-Arnaud, Nicolas and Kris Popat (1998). "Analysis and synthesis of sound textures". In: *Computational Auditory Scene Analysis*. Ed. by D.F. Rosenthal and H.G. Okuno. L. Erlbaum Associates Inc, pp. 293–308.

Saito, Shoichiro, Hirokazu Kameoka, Keigo Takahashi, Takuya Nishimoto, and Shigeki Sagayama (2008). "Specmurt Analysis of Polyphonic Music Signals". In: *IEEE Trans. on Audio, Speech and Language Processing* 16.3, pp. 639–650.

Saitou, T., M. Goto, M. Unoki, and M. Akagi (2007). "Speech-to-Singing synthesis : converting speaking voices to singing voices by controlling acoustical features unique to singing voices". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 215 –218.

Scherer, Klaus R., D. Robert Ladd, and Kim A .E. Silverman (1984). "Vocal cues to speaker affect: Testing two models". In: *jasa* 76.5, pp. 1346–1356.

Schwarz, Diemo and Norbert Schnell (2010). "Descriptor-based sound texture sampling". In: *Proc. 7th Sound and Music Computing Conference*.

Serra, X. J. and J. O. Smith (1990). "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition." In: *Computer Music Journal* 14.4, pp. 12–24.

Smola, A. J. and B. Schölkopf (2004). "A Tutorial on Support Vector Regression". In: *Statistics and Computing* 14, pp. 199–222.

Strobl, Gerda, Gerhard Eckel, Davide Rocchesso, and S le Grazie (2006). "Sound texture modeling: A survey". In: *Proceedings of the 2006 Sound and Music Computing (SMC) International Conference*, pp. 61–65.

Stylianou, Y. (1996). "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification". PhD thesis. Ecole Nationale Supèrieure des Télécommunications, Paris, France.

— (2001). "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis". In: *IEEE Transactions on Speech and Audio Processing* 9.1, pp. 21–29.

Stylianou, Yannis (2009). "Voice transformation: a survey". In: *ICASSP*, pp. 3585–3588.

SynSY (2012). *HMM Based Signing Voice Synthesis System*. `http://www.sinsy.jp/`.

Takens, F. (1981). "Detecting Strange Attractors in Turbulence". In: vol. 898. Lecture Notes in Mathematics (Dynamical Systems and Turbulence, Warwick 1980). D. A. Rand and L. S. Young, Eds. Berlin: Springer, pp. 366–381.

Thomson, D. J. (1982). "Spectrum Estimation and Harmonic Analysis". In: *Proceedings of the IEEE* 70.9, pp. 1055–1096.

T.Toda, A. Black, and K. Tokuda (2007). "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory". In: *ITASLP* 15.8, pp. 2222 –2235.

Veldhuizen, Todd (1995). "Expression Templates". In: *C++ Report* 7, pp. 26–31.

Verron, Charles (2010). "Synthèse immersive de sons d'environnement". PhD thesis. Université Aix-Marseille I.

Vincent, Damien, Olivier Rosec, and Thierry Chonavel (2005). "Estimation of LF glottal source parameters based on an ARX model". In: *Proc. of Interspeech*.

Vincent, Emmanuel, Nancy Bertin, and Roland Badeau (2010). "Adaptive Harmonic Spectral Decomposition for Multiple Pitch Estimation". In: *ITASLP* 18.3, pp. 528–537.

Virtanen, T. (2007). "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria". In: *IEEE Trans. on Audio, Speech, and Language Processing* 15.3, pp. 1066–1074.

Walker, Jacqueline and Peter Murphy (2005). "Advanced methods for glottal wave extraction". In: *Nonlinear analyses and algorithms for speech processing*. Springer, pp. 139–149.

Weigend, A. S. and N. A. Gershenfeld (1993). *Time Series Prediction: Forecasting the Future and Understanding the Past*. Addison-Wesley Pub. Comp.

Wells, J. J. and D. T. Murphy (2007). "Single-frame discrimination of non-stationary sinusoids". In: *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 94–97.

— (2010). "A comparative evaluation of techniques for single-frame discrimination of nonstationary sinusoids". In: *IEEE Trans. Audio, Speech and Lang. Proc.* 18.3, pp. 498–508. DOI: `http://dx.doi.org/10.1109/TASL.2009.2039088`.

Wen, Xue and Mark Sandler (2009). "Notes on model-based non-stationary sinusoid estimation methods using derivatives". In: *Proc. Digital Audio Effects (DAFx)*.

Zen, Heiga, Keiichi Tokuda, and Alan W. Black (2009). "Statistical parametric speech synthesis". In: *Speech Communication* 51.11, pp. 1039–1064.

Zhu, Xinglei and Lonce Wyse (2004). “Sound texture modeling and time-frequency LPC”. In: *Proc. 7th Int. Conf. on Digital Audio Effects (DAFx)*.

External Software

[ES1] Veldhuizen, Todd, Patrik Jonsson, and Julian Cummings (1995). *Blitz++: a C++ class library for scientific computing*. `https://sourceforge.net/projects/blitz/`.

[ES2] University, OSL - Indiana (2001). *The Matrix Template Library 2*. `http://osl.iu.edu/research/mtl/mtl2.php3`.

[ES3] Oliphant, Trevis, Eric Jones, Pearu Peterson, and Community (2001). *SciPy: Scientific Computing in Python*. `http://www.scipy.org/`.

[ES4] Schnell, N. (2005). *SuperVP objects in Max MSP*. Development since 2005.

[ES5] Oliphant, Trevis and Community (2005). *NumPy: Numeric Computing in Python*. `http://www.numpy.org/`.

[ES6] Bresson, Jean and Jean Lochard (2006). *SuperVP for OpenMusic*. Development since 2006.

[ES7] Guennebaud, Gaël, Benoît Jacob, et al. (2010). *Eigen v3*. `http://eigen.tuxfamily.org`.

[ES8] Bresson, Jean (2010). *Pm2 for OpenMusic*. Development since 2010.