BETWEEN PHYSICS AND PERCEPTION: SIGNAL MODELS FOR HIGH LEVEL AUDIO PROCESSING

Axel Röbel

IRCAM-CNRS-STMS,
Analysis-Synthesis Team,
Paris, France
axel(dot)roebel(at)ircam(dot)fr

ABSTRACT

The use of signal models is one of the key factors enabling us to establish high quality signal transformation algorithms with intuitive high level control parameters. In the present article we will discuss signal models, and the signal transformation algorithms that are based on these models, in relation to the physical properties of the sound source and the properties of human sound perception. We will argue that the implementation of perceptually intuitive high quality signal transformation algorithms requires strong links between the signal models and the perceptually relevant physical properties of the sound source. We will present an overview over the history of 2 sound models that are used for sound transformation and will show how the past and future evolution of sound transformation algorithms is driven by our understanding of the physical world.

1. INTRODUCTION

Signal transformation is one of the key topics of the DAFx conference and in the present article we will discuss signal transformation algorithms that can be controlled by means of high level controls. The term high level control will be used in the following for controls using terminology related to the everyday experience of ordinary people. For example making a sound longer is certainly an ambiguous specification but most people will have an intuitive understanding what they expect as result of this operation.

There are different classes of signal transformation algorithms and not in all situations high level controls are appropriate. Filters for example modify the energy distribution of the sound signals. The operators are relatively simple and have correspondence in the real life, but in many situations the control of filtering operations is best performed by means of direct specification of the modifications of the spectral content (e.g., increase amplitudes of high frequency content). High level control is required if the filtering operation is specified in terms of physical configurations of the filter as for example (change of room acoustics, change of resonator body of an instrument). For other more complex signal transformation algorithms, as for example ring modulation, there do not exist any real world operations that would allow to establish an intuitive control. In these cases intuitive control can only be achieved by means of training of the user[1]. Note that research activities that try to use sound descriptors to control sound transformation [1] will not be considered in the following. While signal descriptors like spectral centroid and spectral tilt extend the vocabulary for sound description, they do not introduce a different level of description by themselves.

There are many research topics that are related to high level control of signal transformation. Modification of the instrumentation for a given music signal, modification of the score or remixing of the instruments in a musical piece [2, 3, 4], modification of speaker characteristics (gender, age, complete identity, emotion) for a given speech signal [5, 6, 7, 8, 9, 10], manipulation of musical expression [11, 12], modification of the room acoustics or sound source position. The common factor of all these tasks is the reference to the physical world. The transformations are described in common language and it is expected that the algorithms finds one (of often many) possible transformations that are compatible with the description.

In the first part of the article some basic properties of the signal models that facilitate the introduction of this kind of high-level controls into signal transformation algorithms will be introduced. The second part of the article will discuss 2 important signal models that up to today have had a major impact on the construction of signal transformation algorithms. These models are the sinusoidal signal model [13, 14] and the source-filter model [15, 16]. The historical evolution of these 2 models will be reviewed and some recent trends that may allow improving and extending the signal transformation algorithms that are in use today are discussed.

To prevent the bibliography becoming too long only a few hopefully central references are given for most topics discussed. A general guideline was to prefer the selection of papers that have been presented at a DAFx conference

[1] While in the present article training is not considered as intuitive this does not exclude that after some time of experience a user may establish an intuitive understanding about every sound transformation algorithm he/she is working with.

whenever possible.

2. HIGH LEVEL CONTROL OF SIGNAL TRANSFORMATION

The discussion is started with an investigation into the properties of the sound representation that simplifies the high level control of signal transformation algorithms. The objective is to determine signal representations that represent the sound signal such that the original signal can be reproduced with no degradation and at the same time support the use of intuitive parameters to control sound signal transformation.

2.1. Intuitive sound transformation

Our intuition about how a signal should sound that is transformed using high level controls is related directly to our experience, our personal live in the physical world. As a simple examples one may consider the time stretching signal operator, one of the operators that has triggered major research activities. This operator is related to the physical concept of slow and fast playing style. One of the problems here is the ambiguity in the specification. There are infinitely many possibilities to play slower and there are even more possibilities to construct a signal operator that time stretches a given sound signal.

While the duration is the only factor that is unambiguously specified the duration of the transformed sound signal is certainly not the only criteria that will be used to evaluate the resulting sound. Additionally, the transformed sound should be perceptually close to the sound that could have been produced by the physical sound source when it is operated more slowly. One of the well known examples of this problem is related to time stretching of onsets [17, 18, 19, 20]. It is generally considered that time stretching fast sound onsets generates strong artifacts if the onset part is treated with the same algorithm that is applied to the stationary part. While the time scale modification is perfect in both cases, the separate treatment of the onsets will generally be preferred because it preserves the sound properties of the sound signals that are generated by the physical sound source when played more slowly. For a piano signal for example, playing slow does not change the duration of the attack. Accordingly the preservation of onset characteristics of time stretched piano signals is of crucial importance. For the case of a violin, however, onset times may or may not change with the playing tempo, and therefore, onset preservation has been considered much less an issue when violin sounds are transformed[2]. The more a given algorithm will be able to adapt its behavior to the signal that is treated, the more it will be considered to establish high level controls. From this perspective one may conclude that existing algorithms are generally not capable to establishment an advanced degree of high level controls.

[2]A special case are sound sources that do not change at all when tempo is modified. Playing a drum slowly will hardly change the sound itself, but will only separate the original drum beat signals. Accordingly, time stretching drum signals should ideally be performed by means of first separating the individual drum beat signals and then re-synthesizing displaced versions these individual drum beats.

2.2. Signal models for intuitive sound transformation

Having described the intuition that is the basis of high level sound transformation the properties of the signal models that allow implementing these intuitive controls can be addressed. Following the discussion above it appears that a requirement for an appropriate model would be the use of perceptually relevant components that have a simple relation with the physical properties of the sound sources. The simpler the relation between the perceptually relevant properties of the physical sound source and the signal model the easier it should be to provide controls that reflect our intuition that is built on physical interaction.

These kind of relations exist for example for models that are represented in terms of the vibration modes. Fortunately these individual modes can be represented in a rather simple manner as a sinusoid with time-varying amplitude and frequency. The relation between sinusoidal models and vibration modes is well know and has been used under the name *modal synthesis* for quite a while [21, 22]. In the case of analysis re-synthesis systems the modal representation is achieved by means of the sinusoidal model [13, 14, 16]. It can be concluded that the sinusoidal model establishes the desired link between the physical and the perceptual world. The vibrating modes, however, are generally not sufficient to describe a given sound signal. Noise sources are present in nearly all cases, for example as a side effect of the excitation. Because noise is generally perceived as an individual component we can add a noise component into our model without destroying the simplicity of the transformation.

A mathematical formulation of the sinusoids plus noise model that has been discussed is

$$\begin{array}{rcl} p_k(n) & = & a_k(n)\cos(\Theta_k(n)) \\ s(n) & = & \sum_k p_k(n) + r(n). \end{array} \quad (1)$$

Here p_k is a sinusoid with time varying amplitude $a_k(n)$ and phase $\Theta_k(n)$ and $s(n)$ is the signal that consists of a superposition of sinusoids and a noise component $r(n)$. Because all components of the sinusoidal model can independently vary amplitude and frequency over time, the model allows representing onsets, vibrato and other signal modulations. Accordingly, an additional transient component [23] is unnecessary at least from a conceptual point of view.

2.2.1. Signal transformation with the sinusoids plus noise model

The sinusoids plus noise model discussed up to now does contain the information that is required for high level signal transformation, but the information is provided in a way that is still too far away from the physical structure of sound sources for intuitive controls to be established. When changing the note on a guitar, for example, the resonator part stays approximately the same, but the excitation oscillator changes its properties. Accordingly, it would be beneficial if one could separate the modal structure of the source oscillator from the transfer function of the source resonator filter. This type of separation is generally achieved by means of the source/filter model [16] that adds a separate resonator filter into the model given by eq. (1). If the signal spectrum is represented in terms of a short time Fourier transform (STFT) a model including the source filter model can be written in the following form

$$S(w,n) = \sum_k E(w,n)P'_k(w,n) + N(w,n)R'(w,n). \quad (2)$$

Here $P'_k(w,n)$ is the STFT of the sinusoidal component k that represents only the amplitude and phase evolution that is due to the excitation source. The effect of the resonator and radiation is summarized in $E(w,n)$. This component may be time varying (e.g. if sound source position changes). For the noise component a similar split into the noise source component $R'(w,n)$ and a potentially time varying noise filter $N(w,n)$ is used. Note, that the distinction of sinusoidal resonator and radiation filter and noise resonator and radiation filter is an extension of the standard source filter model. It can be justified physically in many situations (wind noise not passing through the flute body) and the effects of these separate filters have been observed in a number of recent studies related to transformation of expressive playing styles at IRCAM. These results will be published elsewhere.

The sinusoidal model is especially well suited for high level control of signal transformations because most physical manipulations require simply a modification of the amplitude and frequency of the vibration modes which could in principle be implemented rather perfectly by means of manipulating the amplitude and frequency parameter trajectories of the model. For some sound sources, notably voice, the sinusoidal components have to be modified following an additional constraint (shape invariant signal modification [24, 25, 26]). An important precondition to be able to keep the parameter trajectory mapping simple is the fact that the sinusoidal components representing the stable resonating modes are individually resolved in the signal representation. If this is not the case time stretching for example will introduce modification of the beating pattern of the sinusoidal components and parameter trajectory mapping becomes extremely complicated.

2.3. Model implementation

Before the discussion of the history and future evolution of the sinusoids plus noise model and the source filter model a short note on the implementation seems appropriate. While many implementations of the sinusoids plus noise model use the explicit formulation of the model eq. (1) [18, 27, 28, 29] other implementations are possible and often beneficial. An important example is the phase vocoder [30, 31] that provides an implicit sinusoidal model that achieves a very efficient (in calculation time and quality) signal representation.

3. PAST, PRESENT AND FUTURE

The following section contains a short history of the evolution of the 2 model components that have been discussed, including especially the key steps that have been taken to bring the signal transformation algorithms closer to the physical reality.

3.1. Sinusoids plus noise model

The sinusoidal models have their origin in the vocoder developed by Dudley in 1939 [32]. The ideas of Dudley evolved with the invention of computers and digital signal processing into early versions of the phase vocoder [33]. These phase vocoders used very low number of bands (30 bands with 100Hz bandwidth) such that the resolution of the individual sinusoids could not be guaranteed. With further increasing computing capacities and the use of FFT algorithms the number of bands (today bins) increased and as a next step explicit harmonic sinusoidal models were developed [34]. The use of the sinusoidal modeling techniques for musical applications also started with the early phase vocoder [35] and evolved into an explicit sinusoidal model [36]. The main advantage of the explicit sinusoidal model compared to the phase vocoder was the peak picking that was part of the analysis for the explicit sinusoidal models. The peak picking and subsequent parameter estimation did allow to increase frequency resolution and improved the tracking of time varying sinusoids. As a next step the sinusoidal model was extended by means of a dedicated noise model [13] so that the sinusoidal model present in eq. (1) was completed. After the introduction of the intra-sinusoidal phase synchronization [31] the phase vocoder has evolved into an implicit implementation of a sinusoidal model that generally is computationally more efficient than the explicit sinusoidal model. Due to the fact that the phase vocoder representation achieves a better representation of potential structure in the aperiodic (noise) component it often achieves better quality than the explicit sinusoidal model.

The main problem with the sinusoids plus noise model is related to finding the model parameters from the original

signal. This problem has triggered numerous research efforts over the last decades and despite the many interesting and powerful methods that have been developed and that extended the boundaries of the signal representation that can be obtained using this model there remain problems that are still considered unsolved by today.

3.1.1. Sinusoids and noise model estimation

The first parameter estimators that have been developed imposed very strong constraints on the parameters of the sinusoids, notably that the parameter changes were very slow [36]. A consequence of this constraint is the fact that the parameter trajectories representing fast onsets of sinusoidal components could not be correctly estimated. To improve the representation of fast onsets in the sinusoidal models significant research efforts were undertaken [37, 38, 18, 39, 40] most of which make use of the time reassignment operators [41] that significantly reduces onset smearing but unfortunately will systematically cut the sinusoidal onsets [39, 40]. Another approach to onset representation that has been proposed is to extend the sinusoidal plus noise model by means of a dedicated transient component [23]. The main drawback of this approach is the fact that transient and non transient parts of sinusoidal components need to be processed independently ensuring a smooth connection between these 2 parts after processing. For the phase vocoder implementation of the sinusoids plus noise model onset preservation methods have been proposed in [19, 20]. A rather different approach that achieves increased time and frequency resolution by means of using stronger constraints on the sinusoidal components are high resolution methods [42].

The strong bias that exists for most of the sinusoidal parameter estimators whenever the sinusoidal parameters vary over time has recently let to the investigation of parameter estimation methods with reduced bias [43, 44, 45]. The use of these bias reduced estimators significantly extends the range of the signals that may be represented reliably.

Up to now only the parameter estimation for the sinusoidal components has been considered. However, a sinusoids plus noise model contains a noise component as well. The noise components have initially simply collected the residual part of the signal after subtracting all sinusoids [13]. With the rather high quality sinusoidal parameter estimators that exist today, more effort has been invested as well into the estimation and representation of the noise components. This shift of interest is especially visible in the increasing number of algorithms that address the problem to separate sinusoidal and noise components even before the analysis has been started [46, 47, 48, 49, 50, 51]. In our research this separation has proven to be of major importance for the solution of a number of research problems, notably the estimation of multiple fundamental frequencies in polyphonic audio [52, 53] and the shape invariant phase vocoder implementation [26].

3.1.2. Outlook

The discussion of existing approaches in the previous section reveals that the sinusoidal and transient components have received a significant part of the research effort that has been invested into the related signal models. The noise components seem to be somewhat neglected and given that rather high quality transformation algorithms for the sinusoidal and transient components are available today one can expect that the appropriate representation and transformation of the wide variety of structure that the aperiodic signal component may exhibit [54] receives more interest. The implicit sinusoids plus noise model that is used in the phase vocoder often leads to an improved transformation of noise components that preserves some of the structure of the noise components [26], but one can expect that more research will be needed. Some research activities that try to capture the complex time frequency structure that characterizes the aperiodic components have already been started [55] and it can be expected that these results will be refined and hopefully at some point be integrated into the signal transformation algorithms.

Another open research problem is the appropriate selection of the time and frequency resolution of the analysis. For a long time the research and applications have used standard short time Fourier transform algorithms to derive the spectral domain representation of the signal. One can expect that Wavelets providing a frequency dependent frequency and time resolution will not be of major importance for signal transformation algorithms. Our main argument here would be the fact that sinusoidal components of harmonic sounds do not require any adaptation of the frequency resolution as a function of the frequency. They could benefit from an adaptation of the frequency resolution as a function of the pitch and it can be expected that the research into signal representation with time varying signal adaptive time frequency resolution [56, 57] will be a key to resolve the sometimes contradicting demands a signal may pose on today's transformation algorithms.

As another central point of research appears the work on polyphonic signal separation and editing. While there exist numerous proposals to use non parametric decomposition algorithms [58, 59, 60] for signal decomposition the first commercially available system, Melodyne DNA [4], is based on a sinusoidal and noise model. One could argue that the strong, physically and perceptually relevant constraints that are imposed by the sinusoidal model are essential to achieve a sufficiently high quality for the signal separation phase of the DNA system. The strong activity in the area of signal separation will certainly continue and deliver new audio processing tools.

Another topic that will become increasingly important is research on modeling and transformation of expressive per-

formance. The research presented in [19, 11] is only a first step. Much more work is required to be able to transform the signal in a physically coherent manner when modulation parameters are changed. At IRCAM the Sample Orchestrator 2 research project that will address these questions has just started.

3.2. Source-Filter model

The source-filter model is another important signal model that is widely used for signal transformation algorithms. It has the same origins as the sinusoids plus noise model [32]. In the first application of this model the excitation source had been represented by either a impulse train parametrized by the fundamental frequency, or by means of white noise. In both cases the filter part has been achieved by means of modulating the energy of the excitation signal in bands of constant bandwidth (=250Hz). This basic setting is still in use today. The band wise filtering will in most cases be replaced by a continuous filter function that is called the spectral envelope [61, 16].

The source filter model has many applications for signal transformations. Cross synthesis for example can be achieved by means of using the excitation signal (source) from one signal and the resonator (filter) from another. Other applications are transformations that require independent transposition of pitch and formant structure.

3.2.1. Spectral envelope estimation

An important precondition for the source filter model is that the distinct source and filter parts can be estimated from the original signal. A short summary of the existing approaches will be given in the present section.

One of the first techniques that has been used for spectral envelope estimation is linear prediction (LPC) [62]. This method assumes an autoregressive filter function and has been used especially for speech signals for which the autoregressive filter model has a physical justification at least for some configurations of the vocal tract [61]. A problem of the LPC estimate is the fact that it is strongly biased if the excitation spectrum contains sinusoids. This problem has been addressed in the discrete all pole model [63]. Alternative spectral envelope estimators use the cepstral representation to derive the spectral envelope. An early rather costly and complex method is the discrete cepstrum [64]. Later a more efficient method has been developed [65, 66] that is using the same envelope representation but makes use of a rediscovered proposal of an iterative cepstral envelope estimator [67]. The method is referred to as *True Envelope Estimator*. It has proven to provide nearly optimal estimates given that the spectral envelope can only be observed in a strongly sub sampled version that is produced by the sinusoidal components sampling the filter transfer function [66]. At IRCAM this method has been applied especially in the context of speech notably for voice conversion [68, 10] and it has been found to produce very good results.

The estimation of the noise envelopes of the background noise that is part of a complex sound consisting of sinusoidal and noise components is a problem that has received relatively few interest. The estimation of the sinusoidal parameters and estimation of the noise level from the residual is a possible procedure, but if the sinusoidal components are superimposed as for example in polyphonic music this procedure will not provide robust results. There exist only a few methods that allow to establish a background noise estimate for complex polyphonic sounds [69, 48]. As discussed in section **3.1.1** such method can sometimes be extremely helpful.

3.2.2. Outlook

One of the main problems of the source filter model that has been used so far is the assumption that the noise or sinusoidal excitation spectrum is white and that the spectral color is completely provided by the filter. It is rather simple to demonstrate that this assumption is far away from the physical reality. The glottal pulse for example, that is exciting the vocal tract filter during voiced speech, is a smooth signal and therefore does not have a white spectrum [70]. The same is true for all musical instruments. Vibrating chords, e.g., do not produce a white excitation. A number of recent research projects address the problem to jointly estimate the glottal pulse parameters and the vocal tract filter parameters[71, 72]. Such estimators would allow to establish a significant step towards a physically correct separation of source and filter components in speech signals. By consequence new high level signal transformations become possible like transformations of voice quality from soft into tense voice (or the inverse). Because voice quality is partly determined by the emotional state of the speaker, such transformation operators may help to provide expressive speech transformations (transforming a neutral voice into an angry expression).

For music signals one can expect that the possibility to estimate a physically reasonable non-white sound source jointly with a corresponding resonator filter opens new possibilities for a number of applications. With respect to instrument recognition one may expect that this kind of physically more reasonable source filter separation will help to find consistent instrument features which significantly reduced variance [73]. With respect to sound transformation it can be expected that a physically more reasonable resonator filter will provide new intuitive means for signal transformation as for example a modification of the material of the exciting chords in a given guitar signal. In the Analysis/Synthesis team at IRCAM we have just started to investigate these new possibilities [74] and we expect a number of interesting results in the near future.

4. REFERENCES

[1] G. Coleman and J. Bonada, "Sound transformation by descriptor using an analytic domain," in *Proc. Int. Conf. on Digital Audio Effects (DAFx 08)*, Espoo, Finland, 01/09/2008 2008.

[2] J. Woodruff, B. Pardo, and R. Dannenberg, "Remixing stereo music with score-informed source separation," in *Proc. Int. Soc. for Music Information Retrieval Conf (ISMIR)*, 2006, pp. 314–319.

[3] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive componets in music signals," in *Proc. of the 9th Int. Soc. for Music Information Retrieval Conf (ISMIR)*, 2008, pp. 139–144.

[4] M. Senior, "Celemony Melodyne DNA Editor," *Sound on Sound*, , no. 12, 2009.

[5] R. Lawlor and A.D. Fagan, "A novel efficient algorithm for voice gender conversion," in *Proc 14. Congress of Phonetic Sciences (ICPS)*, 1999.

[6] B. P. Nguyen and M. Akagi, "Spectral modification for voice gender conversion using temporal decomposition," *Journal of Signal Processing*, vol. 11, no. 4, pp. 333–336, 2007.

[7] A. Kain, *High resolution voice transformation*, Ph.D. thesis, OGI School of Science and Engineering at Oregon Health and Science University, 2001.

[8] A. Kain and M.W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, 2001, vol. 2, pp. 813–816.

[9] S. Farner, C. Veaux, G. Beller, X. Rodet, and L. Ach, "Voice transformation and speech synthesis for video games," in *Paris Game Developers Conference*, Paris, France, Juin 2008.

[10] F. Villavicencio, A. Röbel, and X. Rodet, "Applying improved spectral modeling for high quality voice conversion," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Avril 2009, pp. 4285–4288.

[11] E. Lindemann, "Music synthesis with reconstructive phrase modeling," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 80–91, 2007.

[12] P. Herrera and J. Bonada, "Vibrato extraction and parameterization in the spectral modeling synthesis framework," in *Proc. COST-G6 Workshop on Digital Audio Effects (DAFx)*, 1998.

[13] X. J. Serra and J. O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition.," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[14] X. Serra, *Musical signal processing*, chapter Musical Sound Modeling with Sinusoids and Noise, pp. 91–122, Studies on New Music Research. Swets & Zeitlinger B. V., 1997.

[15] G. Fant, "The source filter concept in voice production," *QPSR, Dept of speech, music and hearing, KTH*, vol. 22, no. 1, pp. 21–37, 1981.

[16] X. Rodet and D. Schwarz, "Spectral envelopes and additive+residual analysis-synthesis," in *Analysis, Synthesis, and Perception of Musical Sounds: The Sound of Music*, James W. Beauchamp, Ed., pp. 175–227. Springer, New York, USA, 2007.

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

[18] K. Fitz, L. Haken, and P. Christensen, "Transient preservation under transformation in an additive sound model," in *Proc. of the Int. Computer Music Conference (ICMC))*, 2000, pp. 392–395.

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

[20] A. Röbel, "A new approach to transient processing in the phase vocoder," in *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*, 2003, pp. 344–349.

[21] J.D. Morrison and J.M. Adrien, "Mosaic: A framework for modal synthesis," *Computer Music Journal*, vol. 17, no. 1, pp. 45–56, 1993.

[22] P. Djoharian, "Generating models for modal synthesis," *Computer Music Journal*, vol. 17, no. 1, pp. 57–65, 1993.

[23] S. Levine and J. O. Smith, "A sines+transients+noise audio representation for data compression and time/pitch-scale modifications," in *105th AES Convention*, 1998, Preprint 4781.

[24] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, no. 3, pp. 497–510, 1992.

[25] J. Laroche, "Frequency-domain techniques for high-quality voice modification," in *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*, 2003.

[26] A. Röbel, "A shape-invariant phase vocoder for speech transformation," in *Proc. 13th int. Conf. on Digital Audio Effects (DAFx)*, 2010.

[27] K. Fitz, L. Haken, and P. Christensen, "A new algorithm for bandwidth association in bandwidth-enhanced additive sound modeling," in *Proc. of the Int. Computer Music Conference (ICMC)*, 2000, pp. 384–387.

[28] X. Amatriain, J. Bonada, A. Loscos, and X. Serra, "Spectral processing," in *Digital Audio Effects*, U. Zölzer, Ed., chapter 10, pp. 373–438. John Wiley & Sons, 2002.

[29] M. Klingbeil, "Software for spectral analysis, editing, and synthesis," in *Proc. of the Int. Computer Music Conference (ICMC)*. Citeseer, 2005, pp. 107–110.

[30] M. Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[31] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

[32] Homer Dudley, "Remaking speech," *Journal of the Acoustical Society of America (JASA)*, vol. 11, no. 2, pp. 169–177, 1939.

[33] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, pp. 1493–1509, 1966.

[34] R. J. McAulay and T. F. Quatieri, "Speech analysis-synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[35] J. A. Moorer, "The use of the phase vocoder in computer music applications," *Journal of the Audio Engineering Society*, vol. 26, no. 1/2, pp. 42–45, 1978.

[36] J. O. Smith and X. Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *Proc. Int. Computer Music Conference (ICMC)*, 1987, pp. 290–297.

[37] T. Verma and T. H. Y. Meng, "An analysis / synthesis tool for transient signals that allows a flexible sines + transients + noise model for audio," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.

[38] G. Peeters and X. Rodet, "SINOLA: A new analysis/synthesis method using spectrum peak shape distortion, phase and reassigned spectrum," in *Proc. Int. Computer Music Conference*, 1999, pp. 153–156.

[39] A. S. Master and K. Lee, "Explicit onset modeling of sinusoids using time reassignment," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[40] A. Röbel, "Adaptive additive modeling with continuous parameter trajectories," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 4, pp. 1440–1453, 2006.

[41] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Trans. on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, 1995.

[42] R. Badeau, B. David, and G Richard, "High resolution spectral analysis of mixtures of complex exponentials modulated by polynomials," *IEEE Transactions on Signal Processing*, vol. 54, no. 4, pp. 1341–1350, 2006.

[43] M. Abe and J. O. Smith, "AM/FM rate estimation for time-varying sinusoidal modeling," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2005, pp. 201–204 (Vol. III).

[44] A. Röbel, "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)*, 2007.

[45] A. Röbel, "Frequency-slopeÂestimation and its application to parameter estimation for non-stationary sinusoids," *Computer Music Journal*, vol. 32, no. 2, pp. 68–79, 2008.

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

[47] A. Röbel, M. Zivanovic, and X. Rodet, "Signal decomposition by means of classification of spectral peaks," in *Proc. Int. Computer Music Conference (ICMC)*, 2004, pp. 446–449.

[48] C. Yeh and A. Röbel, "Adaptive noise level estimation," in *Proc. of the 9th Int. Conf. on Digital Audio Effects (DAFx'06)*, 2006, pp. 145–148.

[49] M. Zivanovic, A. Röbel, and X. Rodet, "Adaptive threshold determination for spectral peak classification," in *Proc. of the 10th Int. Conf. on Digital Audio Effects (DAFx'07)*, 2007.

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

[51] J. J. Wells and D. T. Murphy, "A comparative evaluation of techniques for single-frame discrimination of nonstationary sinusoids," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 18, no. 3, pp. 498–508, 2010.

[52] W.-C. Chang, A. W. Y. Su, C. Yeh, A. Röbel, and X. Rodet, "Multiple-f0 tracking based on a high-order hmm model," in *Proc. of the 11th Int. Conf. on Digital Audio Effects (DAFx'08)*, 2008, pp. 379–386.

[53] C. Yeh, A. Roebel, and X. Rodet, "Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals," *IEEE Transactions on Audio, Speech and Language Processing*, 2010, accepted for publication.

[54] G. Richard and C. Alessandro, "Analysis/synthesis and modification of the speech aperiodic components," *Speech Communication*, vol. 19, no. 3, pp. 221–244, 1996.

[55] P. Hanna and M. Desainte-Catherine, "Time scale modification of noises using a spectral and statistical model," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2003, pp. 181–184 (Vol. 6).

[56] M. Linui, A. Röbel, M. Romito, and X. Rodet, "A reduced multiple Gabor frame for local time adaptation of the spectrogram," in *Proc. 13th int. Conf. on Digital Audio Effects (DAFx)*, 2010.

[57] F. Jaillet, , P. Balazs, M. Dörfler, and N. Engelputzeder, "nonstationary gabor frames," in *proc. 8th Int. Conf. on Sampling Theory and Applications (SAMPTA)*, 2009.

[58] N. Mitianoudis and M.E. Davies, "Audio source separation of convolutive mixtures," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 489–497, 2003.

[59] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 91–98, 2006.

[60] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[61] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, Springer Verlag, 1976.

[62] John Makhoul, "Linear prediction: A turorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[63] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 411–423, 1991.

[64] O. Cappé and E. Moulines, "Regularization techniques for discrete cepstrum estimation," *IEEE Signal Processing Letters*, vol. 3, no. 4, pp. 100–102, 1996.

[65] A. Röbel and X. Rodet, "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)*, 2005, pp. 30–35.

[66] A. Röbel, F. Villavicencio, and X. Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order," *Pattern Recognition Letters, Special issue on Advances in Pattern Recognition for Speech and Audio Processing*, pp. 1343–1350, 2007.

[67] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *Electron. and Commun. in Japan*, vol. 62-A, no. 4, pp. 10–17, 1979, in Japanese.

[68] F. Villaviciencio, A. Röbel, and X. Rodet, "Improving LPC spectral envelope extraction of voiced speech by true envelope estimation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2006, pp. 869–872 (Vol. I).

[69] Guillaume Meurisse, Pierre Hanna, and Sylvain Marchand, "A new analysis method for sinusoids+noise spectral models," in *Proc. of the 9th Int. Conf. on Digital Audio Effects (DAFx'06)*, 2006, pp. 139–144.

[70] G. Fant, J. Liljencrants, and Q.-G. Lin, "A four-parameter model of glottal flow," *QPSR, Dept of speech, music and hearing, KTH*, vol. 26, no. 4, pp. 1–13, 1985.

[71] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filer model optimization," *IEEE Trans Speech and Audio Processing*, vol. 14, no. 2, pp. 492–501, 2006.

[72] G. Degottex, A. Roebel, and X. Rodet, "Joint estimate of shape and time-synchronization of a glottal source model by phase flatness," in *ICASSP*, 2010.

[73] A. Klapuri, "Analysis of musical instrument sounds by source-filter-decay model," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, pp. 53–56 (Vol. I).

[74] H. Hahn, A. Röbel, J. J. Burred, and S. Weinzierl, "A source-filter model for quasi-harmonic instruments," in *Proc. 13th int. Conf. on Digital Audio Effects (DAFx)*, 2010.

BIOGRAPHY

Axel Roebel received the Diploma in electrical engineering from Hannover University in 1990 and the Ph.D. degree (summa cum laude) in computer science from the Technical University of Berlin in 1993. In 1994 he joined the German National Research Center for Information Technology (GMD-First) in Berlin where he continued his research on adaptive modeling of time series of nonlinear dynamical systems. In 1996 he became assistant professor for digital signal processing in the communication science department of the Technical University of Berlin. In 2000 he was visiting researcher at CCRMA Standford University, where he worked on adaptive sinusoidal modeling. In the same year he joined the IRCAM to work on sound analysis, synthesis and transformation algorithms. In summer 2006 he was Edgar-Varese guest professor for computer music at the Electronic studio of the Technical University of Berlin and currently he is head of the analysis-synthesis team at IRCAM. His present research interests are related to music and speech signal transformation.