SINUSOIDAL MODELING OF SPOT MICROPHONE SIGNALS BASED ON NOISE TRANSPLANTATION FOR MULTICHANNEL AUDIO CODING

Christos Tzagkarakis, Athanasios Mouchtaris, and Panagiotis Tsakalides

Department of Computer Science, University of Crete and
Institute of Computer Science (FORTH-ICS)
Foundation for Research and Technology - Hellas
Heraklion, Crete, Greece
{tzagarak, mouchtar, tsakalid}@ics.forth.gr

ABSTRACT

This paper focuses on high-fidelity multichannel audio modeling based on an enhanced adaptation of the well-known sinusoidal plus noise model (SNM). Sinusoids cannot be used per se for high-quality audio modeling because they do not represent all the audible information of a recording. The noise part has also to be treated to avoid an artificial sounding resynthesis of the audio signal. Generally, the encoding process needs much higher bitrates for the noise part than the sinusoidal one. Our objective is to model spot microphone signals using the SNM, by taking advantage of the interchannel similarities to achieve low bitrates. We demonstrate that for a given multichannel audio recording, the noise part for each spot microphone signal (microphone signals before the mixing stage) can be obtained from the noise part of one of the signals (reference, which is fully encoded), by transforming the reference noise part using the noise envelope of each of the remaining signals.

1. INTRODUCTION

Multichannel audio offers significant advantages to audio reproduction when compared to stereophonic sound[1]. Multichannel audio "surrounds" the listener with sound, and immerses the listener with a realistic acoustic scene. The problem with adding many channels is that there is an increased demand of higher datarates for transmission and storage. While many coding methods have been proposed for multichannel audio (MPEG AAC, Dolby AC-3, *etc.*) that achieve a significant coding gain, they still remain demanding for many low-bandwidth applications, such as streaming through the Internet and wireless channels. In this paper, we focus on the problem of further exploiting interchannel similarities. We mention that multichannel audio coding methods that exploit interchannel redundancy have been proposed, such as Mid/Side Coding [1] (for frequencies below 2 kHz), Intensity Stereo Coding [2] (for frequencies above 2 kHz), and

KLT-based methods (for example [3]). In this paper, we are interested on modeling the multiple microphone signals that are captured by the various microphones in a venue, *before* those are mixed into the multiple channels of the multichannel recording. These signals are usually referred to as spot microphone signals. Our purpose is to exploit the similarities among the various spot signals of the same multichannel recording, with the final objective of low bitrate coding of these spot signals.

Towards the direction of low bitrate multichannel audio coding with emphasis on exploiting interchannel redundancies, Spatial Audio Coding (SAC) was introduced in recent years. SAC captures the spatial image of a multichannel audio signal with a compact set of parameters; by encoding only one channel of audio (reference channel) and these parameters as side information, the objective is to resynthesize the original multichannel spatial image at the decoder. Conceptually, SAC can be seen as an extension of the Binaural Cue Coding (BCC) scheme [4] towards using more than one transmitted downmix channel. Particularly, BCC encodes as additional information the subband interchannel level difference, time difference, and correlation of each channel with respect to the reference audio channel, achieving bitrates in the order of few KBits/sec for the side information of each channel. Parametric Stereo (PS) [5] also operates based on a similar philosophy. SAC is currently under a standardization process within the MPEG community (known as MPEG Surround [6]).

A large number of microphones in a venue is used, to create a multichannel audio recording. These are then mixed in order to produce the final multichannel audio recording. It would be desirable to transmit the multiple microphone signals of a performance, before those are mixed into the (usually much smaller number of) channels of the multichannel recording. This would allow for interactive applications that are of immense interest for immersive audio environments, such as remote mixing of the multichannel recording and remote collaboration of geographically distributed musicians [7]. For these applications, the number of audio channels to be encoded is higher than in multichannel recordings, and low bitrate encoding for each channel is important.

In this paper, we introduce and validate a novel sinusoidal

[1] In this paper the term stereophonic sound refers to 2-channel stereo.

plus noise transplantation approach for encoding the multiple microphone signals of a music performance with moderate datarate requirements. This would allow for transmission through low bandwidth channels such as the current Internet infrastructure, and for broadcasting over wireless networks. Our method focuses on the microphone signals of a performance *before they are mixed*, and thus can be applied to applications such as remote mixing and distributed performances. In principle, our method attempts to model each microphone signal with respect to a reference audio signal, so in this sense it follows the SAC philosophy. However, while in SAC the objective is to retain the *spatial image* of the original (before the coding stage) multichannel recording, our objective is to retain the *content* of each of the spot microphone signals. In both cases, audio quality is of central importance.

We employ the sinusoidal plus noise model for each spot microphone signal, and we model the signal with the sinusoidal parameters (harmonic part) and the short-time spectral envelope of the noise (modeling noise part). For resynthesis of each microphone signal, we add the harmonic part that was fully encoded, to the noise part which is recreated by using the corresponding noise envelope with the noise residual obtained from the reference signal. This procedure, which we term as *noise transplantation*, is based on the observation that the noise signals of the various spot signals of the same multichannel recording are very similar when the harmonic part has been captured with an appropriate number of sinusoids. The sinusoidal plus noise model has been applied to audio (music) signals under various approaches (more in Section 2.2). To our knowledge, this is the first attempt to tailor and apply this model to the specific case of multichannel audio, for low bitrate high-fidelity multichannel audio coding.

2. BACKGROUND INFORMATION

2.1. Microphone signals of a Multichannel Recording

A brief description is given below, of how the multiple microphone signals for multichannel rendering are recorded. In this paper, we mainly focus on live concert hall performances. A number of microphones is used to capture several characteristics of the venue, resulting in an equal number of microphone signals (stem recordings). Our main goal is to design a system that is able to recreate at the receiving end all of the target microphone signals from a smaller set (or even only one, which could be a sum signal) of reference microphone signals. The result would be a significant reduction in transmission requirements, while enabling interactivity at the receiver.

In order to achieve high quality resynthesis, we propose the use of some additional information for each microphone with the constraint that this additional information requires minimal datarates for transmission. By examining the acoustical characteristics of the various stem recordings, the distinction of microphones is made into reverberant and spot microphones.

Spot microphones are microphones that are placed close to the sound source. The recordings of these microphones heavily depend on the instruments that are near the microphone and not so much on the hall acoustics; these recordings recreate the sense that the sound source is not a point source but rather distributed such as in an orchestra. Hence, resynthesizing the signals captured by these microphones involves enhancing certain instruments and diminishing others, which in most cases overlap in the time and frequency domains. Reverberant microphones are the microphones placed far from the sound source, that mainly capture the reverberation information of the venue. Here, we focus on the recordings made by spot microphones since modeling their spectral properties is more challenging compared to reverberant microphone signals. Modeling of the latter signals has been considered in our earlier work, where linear time-invariant filters were proposed for transforming a reference signal into a given reverberant signal [8].

2.2. Sinusoidal Plus Noise Model

The sinusoidal model represents a harmonic signal $s(n)$ as the sum of a small number of sinusoids with time-varying amplitudes and frequencies

$$s(n) = \sum_{l=1}^{L} A_l(n) \cos(\theta_l(n)) \qquad (1)$$

where $A_l(n)$ and $\theta_l(n)$ is the instantaneous amplitude and phase, respectively.

A more accurate representation of audio signals, such as speech or music, is achieved with the so-called *sinusoidal plus noise model*. The signal representation is obtained by restricting the sinusoids to modeling only the deterministic part of the audio signal, leaving the rest of the spectral information in the sinusoidal noise component $e(n)$, *i.e.*, for each short-time frame the signal can be represented as

$$s(n) = \sum_{l=1}^{L} A_l(n) \cos(\theta_l(n)) + e(n). \qquad (2)$$

Practically, after the sinusoidal parameters are estimated, the noise component is computed by subtracting the harmonic component from the original signal.

Several variations of the sinusoidal plus noise model have been proposed for applications such as signal modifications and low bitrate coding, focusing on three different problems: (1) accurately estimating the sinusoidal parameters from the original spectrum, (2) representing the modeling error (noise component), and (3) representing signal transients. Problem (1) has been extensively treated for speech signals, *e.g.* [9, 10], and variations of these approaches have been extended to wideband audio. Here, as explained, we focus on the problem of noise representation. In music, a harmonic plus noise model was first proposed in [11], where the noise part was modeled based on a piecewise-linear approximation of its short-time spectral envelope or alternatively its Linear Predictive Coding (LPC) envelope (assuming white noise excitation during synthesis).

More recent is the work in [12], where multiresolution analysis was applied for better estimating the sinusoidal parameters by passing the signal through an octave-spaced filterbank which was designed for avoiding aliasing problems. Regarding the noise part, it was not parametrically modeled for best audio quality. Previous studies [13, 14] have focused in the noise part modeling, where in the former approach the envelope of the noise component is estimated by retaining the signal energy in each critical band. Then, the envelope is

added to the sinusoidal part in the frequency domain, and the original signal is approximated by taking the Inverse Fourier Transform of this spectral sum. In the latter approach, the envelope of the noise part is estimated based on a perceptually motivated LPC computation. In particular, LPC analysis is applied to the noise component, which first has been transformed in the perceptual domain. The noise part is reconstructed by filtering white noise through the inverse LPC filter and then, by transforming the filtered white noise back to the initial (time) domain. Thus, more recent methods focus on modeling the noise component using only its perceptually relevant (short-time) spectral envelope. While these methods offer the advantage of low bitrate coding for the noise part, the resulting audio quality is usually worse than the quality of the original audio signal (subjective results with average grades around 3.0 in a 5-grade scale have been reported).

In our case, we are interested in high-quality audio modeling (achieving a grade around 4.0 is desirable). We focus on the sinusoidal plus noise model of [11]. Although more recent models offer improved modeling using fewer sinusoidal components, our objective is to provide a proof of concept for the noise transplantation procedure. In other words, we are interested to show that indeed this method results in good audio quality compared not only to the sinusoids-only model but also compared with the original recording.

In this paper, we model the noise component of the sinusoidal model as the result of filtering a residual noise component with an autoregressive (AR) filter that models the noise spectral envelope. Linear Predictive (LP) analysis is applied to estimate the spectral envelope of the sinusoidal noise. In other words, we assume the following equation for the noise component of the sinusoidal model

$$e(n) = \sum_{i=1}^{p} \alpha(i)\, e(n-i) + r_e(n). \quad (3)$$

The quantity $e(n)$ is the sinusoidal noise component, while $r_e(n)$ is the residual of the noise and p is the AR filter order. The $p+1^{th}$-dimensional vector $\vec{\alpha}^T = (1, -\alpha_1, -\alpha_2, ..., -\alpha_p)$ represents the spectral envelope of the noise component $e(n)$. In the frequency domain (3) becomes

$$S_e(e^{j\omega}) = \left| \frac{1}{A(e^{j\omega})} \right|^2 S_{r_e}(e^{j\omega}) \quad (4)$$

where $S_e(e^{j\omega})$ and $S_{r_e}(e^{j\omega})$ is the power spectrum of $e(n)$ and $r_e(n)$, respectively, while $A(e^{j\omega})$ is the frequency response of the LP filter $\vec{\alpha}$. Since in this paper there are two noise quantities introduced, i.e., the sinusoidal model noise e and its whitened version r_e, we will refer to e as the (sinusoidal) *noise* signal and to r_e as the *residual* (noise) of e. For convenience, we refer to this model as the Sinusoidal plus Noise Model (SNM).

3. NOISE TRANSPLANTATION

We start by considering two spot microphone signals of a music performance, in which the two microphones are placed close to two distinct groups of instruments of the orchestra. The first microphone signal is denoted by $x_L(n)$ (for simplicity we refer to this signal as the left channel, which should not be confused with the channels of the multichannel mix),
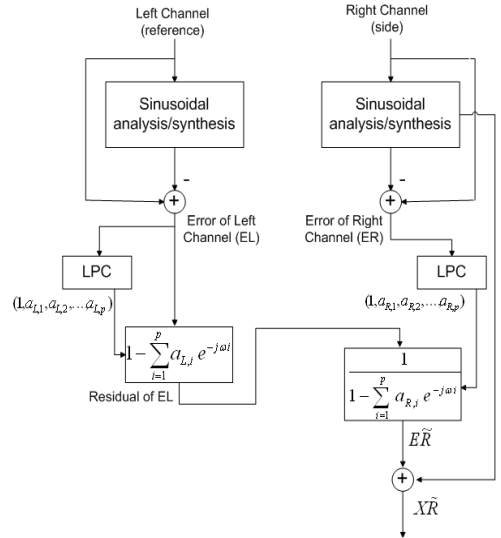

Fig. 1: Noise transplantation. The LPC residual of the reference signal's noise component is filtered by the side signal's noise envelope and added to its sinusoidal component.

while the second one is denoted by $x_R(n)$ (referred to as the right channel). Each of these microphone signals mainly captures the sound from the closest group of instruments, but also captures the sound from all the other instruments of the orchestra (this is especially true for live concert hall performances). Thus, the two recordings are similar in content, and this is apparent in most multichannel recordings in such settings. Alternatively, one of the channels (the reference signal) could be a sum signal of all the spot recordings.

Sinusoidal models capture the harmonics of the original audio signal well if the number of harmonics used is carefully chosen. However, especially for music signals, the harmonic component is not sufficient for high-quality synthesis; its structured nature and the lack of "randomness" in the signal is audible even if a high number of sinusoids is used. The noise signal e, which contains the spectral information which is considered of random nature, is necessary for high-quality audio synthesis. It mostly contains higher-frequency information, and adds the acoustically needed "randomness" to the sinusoidal component. In coding applications, the noise signal will require a much higher degree in terms of datarates compared to the sinusoidal component, exactly due to its quasi-random nature. Thus, we are interested here to propose a model that is based on the sinusoidal component of the audio signal, but can result in high-quality audio synthesis at the decoder.

In order to achieve this objective, we propose a scheme that is similar to the Spatial Audio Coding philosophy. In other words, we propose that given a collection of microphone signals that correspond to the same multichannel recording (and thus have similar content), we encode as a full audio channel only one of the signals (reference signal). We model the remaining signals with the SNM model, retaining their sinusoidal components and the noise spectral envelope (filter $\vec{\alpha}$ in (3)). For resynthesis, we model the reference signal with the SNM in order to obtain its noise signal e, and

from it we obtain the LP residual r_e using LPC analysis. Finally, we reconstruct each microphone signal using its sinusoidal component and its noise LP filter; its sinusoidal component is added to the noise component that we obtain by filtering with the signals LP noise shaping filter the LPC residual of the sinusoidal noise from the reference signal. The assumption is that, as the harmonics capture most of the important information for each microphone signal and the LP coefficients capture most of the channel-specific noise characteristics, the residual noise part that remains will be similar for all the microphone signals. This assumption is in fact verified in Section 4. By taking the reference residual (whitened sinusoidal noise) and filtering it with the correct noise envelope (the envelope of side channel k, where the reference and side signals must be time-aligned), we can obtain a noise signal with very similar spectral properties to the initial noise component of the side channel k. This procedure is depicted in the diagram of Fig. 1.

To formalize the previous discussion, considering a multichannel recording with M microphone signals, we introduce the general relation for the resynthesis of one of the *side* microphone signals x_k (as opposed to the *reference* signal $x_{(ref)}$),

$$\hat{x}_k(n) = \sum_{l=1}^{L} A_{k,l}(n) \cos(\theta_{k,l}(n)) + \hat{e}_k(n) , k = 1, ..., M, \quad (5)$$

where $\hat{e}_k(n)$ is represented in the frequency domain as

$$\hat{S}_{e_k}(e^{j\omega}) = \left| \frac{1}{1 - \sum_{i=1}^{p} \alpha_k(i) e^{-j\omega i}} \right|^2 S_{r_{e_{(ref)}}}(e^{j\omega}). \quad (6)$$

In the equations above, $A_{k,l}(t)$ and $\theta_{k,l}(t)$ are the estimated sinusoidal parameters of microphone signal k, $\{\alpha_k\}$ is the signal's LP noise shaping filter, while $\hat{e}_k(n)$ is the estimated noise component using the noise transplantation procedure described. The residual of the noise component of the reference signal can be found as

$$S_{r_{e_{(ref)}}}(e^{j\omega}) = \left| 1 - \sum_{i=1}^{p} \alpha_{(ref)}(i) e^{-j\omega i} \right|^2 S_{e_{(ref)}}(e^{j\omega}). \quad (7)$$

Thus, $S_{r_{e_{(ref)}}}(e^{j\omega})$ is the power spectrum of the reference signal noise residual (AR modeling error of the sinusoidal noise), and $e_{(ref)}$ is the sinusoidal noise obtained from the reference.

At this point, we note that the proposed method presents scalability with regard to the quality. It can be controlled through the number of sinusoids, the LP order and the percentage of frame overlapping. By changing these values, different datarates can be achieved. It is also important to clarify that for the reference signal, the SNM model is applied only for obtaining the noise residual. This signal is assumed to be encoded and transmitted as a monophonic audio signal (*e.g.* MP3) to the receiver. Also, it is possible that more than one reference signals might be necessary for the method to perform well in practice, depending on the nature of the multiple signals of a particular multichannel recording or when backwards compatibility with stereo decoders is required.

4. EXPERIMENTAL RESULTS

In this section, we are interested in illustrating the validity of our claims regarding the estimation of the noise part of a side microphone signal from the reference signal. The results given in this section are from subjective (listening) tests, given the fact that the importance of the noise part in sinusoidal modeling of music can mostly be quantified subjectively.

For the results of this section we used two microphone signals of a multichannel recording of a concert hall performance. One of the microphones captures mainly the female voices of the orchestra's chorus and is used here as the side channel, while the other one mainly captures the male voices and is used as the reference signal. It is important to mention that this two-channel example can be easily extended to an arbitrary number of recordings. The example examined here is proposed based on the fact that our goal is to resynthesize each microphone signal independently of the others, with the use of only one reference signal and the model parameters (sinusoids and LP filter) that characterize the side microphone signal. The two recordings used here were chosen based on the fact that they have been used in our previous experiments with other modeling methods [15].

The particular implementation in this section is based on a 20 msec analysis/synthesis frame for the sinusoidal model with 50% overlapping (with overlap-add synthesis). The LP order for the AR noise shaping filters is 25. The sampling rate for the recordings used is 44.1 kHz. Sixteen listeners participated in the listening tests individually (the authors are not included), under the same environmental conditions, using high-quality headphones. From the two concert hall recordings, we chose three different parts of the performance of about 10 sec duration each (referred to as Signals 1-3). In order to compare the quality of the resynthesized (side) signal with respect to the original microphone recording, we conducted three different DCR-based (Degradation Category Rating) listening tests, using a five-grade scale from 1 (very annoying perceived quality) to 5 (not perceived difference in quality) [16].

In Fig. 2, we plot the average DCR tests for each of the 3 testing signals. Each of the three figures corresponds to a different choice of sinusoidal parameters per frame. The upper plot corresponds to 80 sinusoids, the middle plot to 40 sinusoids, and the lower plot to 10 sinusoids per frame. In each plot, the solid line corresponds to the sinusoidal model resynthesis ("sin"), the dotted line to our proposed model ("sin plus LPC noise"), the dashed-dotted line corresponds to adding to the sinusoidal part of the side signal the noise of the side signal (with perceptual LPC (PLPC) [14] noise shaping model, "sin plus PLPC noise"), while the dashed line corresponds to adding the noise part (obtained with the Critical Band Energy (CBE) approach [13]) of the side signal to the sinusoidal part of the side signal ("sin plus CBE noise"). A graphical representation of the 95% confidence interval is also given by the two horizontal lines above and below the mean value.

From Fig. 2 it can be seen that the three noise-based methods are superior in comparison to the model based on sinusoidal parameters only. Thus, it is apparent that the noise has to be treated to achieve high-quality resynthesis. Both CBE and PLPC approaches give slightly worse results compared to our method, in the case of high enough number of sinusoids (40 and 80 sinusoids in total). However, in the 10 sinusoids case, our LPC-based method still achieves a grade around 4.0, while the two other noise-based methods achieve a grade be-
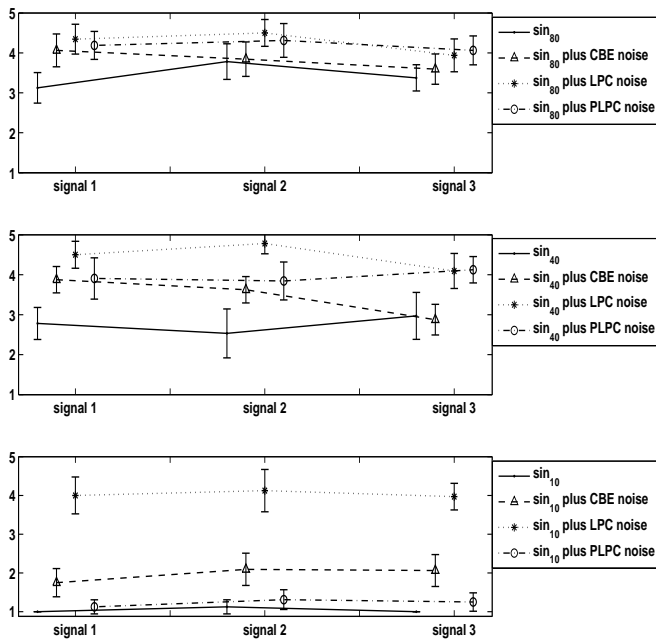
Fig. 2: Results from the quality rating DCR listening tests corresponding to sinusoidal modeling with (a) 80 sinusoids per frame (upper), (b) 40 sinusoids per frame (middle), and (c) 10 sinusoids per frame (lower).

low 2.0. This can be attributed to the fact that PLPC and CBE methods treat the envelope of the noise part only, while our method provides a residual signal for the noise as well (based on similarities among the various spot microphone signals). Finally, it is important to note that, since the subjective audio quality remains high even for a very small number of sinusoids, we can achieve the final objective of increased coding performance (low datarate), since it translates into decreasing the bitrate needed for encoding the sinusoidal component.

5. CONCLUSIONS

In this paper, we presented a sinusoidal plus noise model that is specifically tailored for multichannel audio, with the objective of low bitrate coding by taking advantage of the similarities among the various spot microphone signals (recordings before the mixing process). Our approach offers the possibility of employing the flexible sinusoidal model into low bitrate multichannel audio coding, following a similar philosophy to spatial audio coding. At the same time, by focusing on the spot signals before those are mixed into the final multichannel mix, our method allows for many applications that are not feasible if the spot signals are not available to the decoder. In the future, we intend to examine the bitrates that are possible to achieve with this model and to improve the system by better modeling of the transient signals and by using multiresolution analysis.

7. REFERENCES

[1] J. D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding," in *ICASSP*, pp. 569–572, 1992.

[2] J. Herre, K. Brandenburg, and D. Lederer, "Intensity stereo coding," in *Proc. 96th Convention of the Audio Eng. Soc. (AES), preprint No. 3799*, 1994.

[3] D. Yang, H. Ai, C. Kyriakakis, and C.-C. J. Kuo, "High-fidelity multichannel audio coding with karhunen-loeve transform," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, pp. 365–380, July 2003.

[4] F. Baumgarte and C. Faller, "Binaural cue coding - Part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. Speech and Audio Proc.*, vol. 11, pp. 509–519, Nov. 2003.

[5] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *EURASIP Journal on Applied Signal Proc.*, pp. 1305–1322, 2005:9.

[6] J. Breebaart et al., "MPEG Spatial Audio Coding / MPEG Surround: Overview and current status," in *Proc. AES 119th Convention, Paper 6599*, Oct. 2005.

[7] A. Sawchuk, E. Chew, R. Zimmermann, C. Papadopoulos, and C. Kyriakakis, "From remote media immersion to distributed immersive performance," *in Proc. ACM SIGMM Workshop on Experiential Telepresence (ETP)*, Nov. 2003.

[8] A. Mouchtaris, S. S. Narayanan, and C. Kyriakakis, "Virutal microphones for multichannel audio resynthesis," *EURASIP Journal on Applied Signal Proc.*, vol. 2003:10, pp. 968–979, Sep. 2003.

[9] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. 34(4), pp. 744–754, Aug. 1986.

[10] Y. Stylianou, "Applying the harmonic plus noise model in concatinative speech synthesis," *IEEE Trans. Speech and Audio Proc.*, vol. 9(1), pp. 21–29, 2001.

[11] X. Serra and J. O. Smith, "Spectral modeling sythesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14(4), pp. 12–24, Winter 1990.

[12] S. N. Levine, T. S. Verma, and J. O. Smith, "Multiresolution sinusoidal modeling for wideband audio with modifications," *ICASSP*, 1998.

[13] M. Goodwin, "Residual modeling in music analysis-synthesis," in *ICASSP*, May 1996.

[14] R. C. Hendriks, R. Heusdens, and J. Jensen, "Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding," in *ICASSP*, May 2004.

[15] K. Karadimou, A. Mouchtaris, and P. Tsakalides, "Multichannel audio modeling and coding using a multiband source/filter model," in *Conf. Record of the Thirty-Ninth Asilomar Conf. Signals, Systems and Computers*, (Pacific Grove, CA), pp. 907–911, Nov. 2005.

[16] W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis*. Elsevier Science, 1995.