Low Bitrate Coding of Spot Audio Signals for Interactive and Immersive Audio Applications

Athanasios Mouchtaris, Christos Tzagkarakis, 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
{mouchtar, tzagarak, tsakalid} ics.forth.gr

Abstract. In the last few years, a revolution has occurred in the area of consumer audio. Similarly to the transition from analog to digital sound that took place during the 80s, we have been experiencing the transition from 2-channel stereophonic sound to multichannel sound (e.g., 5.1 systems). Future audiovisual systems will not make distinctions regarding whether the user will be watching a movie or listening to a music recording; they are envisioned to offer a realistic experience to the user who will be immersed into the content, implying that the user will be able to interact with the content according to his will. In this paper, an encoding procedure is proposed, focusing on spot microphone signals, which is necessary for providing interactivity between the user and the environment. A model is proposed which achieves high-quality audio reproduction with side information for each spot microphone signal in the order of 19 kbps.

1 Introduction

Similarly to the transition from analog to digital sound that took place during the 80s, these last years we have been experiencing the transition from 2-channel stereophonic sound to multichannel sound. This transition has shown the potential of multichannel audio to surround the listener with sound and offer a more realistic acoustic scene compared to 2-channel stereo. Current multichannel audio systems place 5 or 7 loudspeakers around the listener in pre-defined positions, and a loudspeaker for low-frequency sounds (5.1 [1] and 7.1 multichannel systems), and are utilised not only for film but also for audio-only content.

Multichannel audio offers the advantage of improved realism compared to 2-channel stereo sound at the expense of increased information concerning the storage and transmission of this medium. This is important in many network-based applications, such as Digital Radio and Internet audio. At a point where MPEG Surround (explained in the following paragraph) achieves coding rates

[*] This work has been funded by the Marie Curie TOK "ASPIRE" grant within the 6^{th} European Community Framework Program.

for 5.1 multichannel audio that are similar to MP3 coding rates for 2-channel stereo, it seems that the research in audio coding might have no future. However, this is far from the truth. Current multichannel audio formats will eventually be substituted by more advanced formats. Future audiovisual systems will not distinguish between whether the user will be watching a movie or listening to a music recording; audiovisual systems of the future are envisioned to offer a realistic experience to the user who will be immersed into the content. As opposed to listening and watching, the passive voice of immersed implies that the user's environment will be seamlessly transformed into the environment of his/her desire, the user being able to interact with the content according to his/her will. Using a large number of loudspeakers is useless if there is no increase in the content information. Immersive audio is largely based on enhanced audio content, which translates into using a large number of microphones (known as spot recordings) for obtaining a recording, containing as many sound sources as possible. These sources offer increased sound directions around the listener, but are also useful for providing interactivity between the user and the audio environment. The increase in audio content, combined with the strict requirements regarding the processing, network delays, and losses in the coding and transmission of immersive audio content, are issues that can be addressed based on the proposed methodology.

The proposed approach in this paper is an extension of multichannel audio coding. For 2-channel stereo sound, the importance of decreasing the bitrate in a music recording has been made apparent within the Internet audio domain with the proliferation of MP3 audio coding (MPEG-1 Layer III [2, 3]). MP3 audio coding allows for coding of stereo audio with rates as low as 128 Kbit/sec for high-quality audio (CD-like or transparent quality). Multichannel sound, as the successor of 2-channel stereo, has been in focus of all audio coding methods since the early 1990s. MPEG-2 Advanced Audio Coding (AAC) [4, 5] and Dolby AC-3 [6] were proposed among others and truly revolutionised the delivery of multichannel sound, allowing for bitrates as low as 320 Kbit/sec for 5.1 audio (transparent quality). These methods were soon adopted by all audio-related applications, such as newer versions of Internet music files (Apple's iTunes) and Digital Television (DTV).

In the audio coding methods mentioned in the previous paragraph, the concept of perceptual audio coding has been of central importance. Perceptual audio coding refers to colouring the coding noise in the frequency domain, so that it will be inaudible by the human auditory system. However, early on it was apparent that coding methods that exploit interchannel (for 2-channel or multichannel audio) were necessary for achieving best coding results. In MPEG-1 and MPEG-2 audio coding, Mid/Side [7] and Intensity Stereo Coding [8] were employed. The former operated on the audio channels in an approximate Karhunen-Loeve-type approach for decorrelation of the channel samples, while the latter was applied to higher frequency bands by exploiting the fact that the auditory image in these bands can be retained by only using the energy envelope of each channel at each short-time audio segment. In early 2007, a new standard for very low bitrate coding of multichannel audio became an International Standard under the name MPEG Surround [9]. MPEG Surround allows for coding of multichannel audio content with rates as low as 64 Kbit/sec for transparent quality. It is based on Binaural Cue Coding (BCC) [10] and Parametric Stereo (PS) [11]. Both methods operate on the same philosophy, which is to capture (at the encoder) and re-synthesise (at the decoder) the cues needed for sound localisation by the human auditory system. In this manner, it is possible to recreate the original spatial image of the multichannel recording by encoding only one monophonic audio downmix signal (the sum of the various audio channels of a particular recording), as well as the binaural cues which constitute only a small amount of additional (side) information. MPEG Surround and (related) AAC+ are expected to replace the current MP3 and AAC formats for Internet audio, and to dominate in broadcasting applications.

Immersive audio, as opposed to multichannel audio, is based on providing the listener the option to interact with the sound environment. This translates, as explained later in this paper, into different objectives in the content to be encoded and transmitted, which cannot be fulfilled by current multichannel audio coding approaches. Our goal is to introduce mathematical models specifically directed towards immersive audio, for compressing the content and allowing model-based reconstruction of lost or delayed information. Our aspirations are towards finally implementing long-proposed ideas in the audio community, such as (network-based) telepresence of a user in a concert hall performance in real-time, implying interaction with the environment, *e.g.*, being able to move around in the hall and appreciate the hall acoustics; virtual music performances, where the musicians are located all around the world; collaborative environments for the production of music; and so forth.

In this paper, the sinusoids plus noise model (henceforth denoted as SNM for brevity), which has been used extensively for monophonic audio signals, is introduced in the context of low-bitrate coding for *Immersive* audio. As in the SAC method for low bitrate *multichannel* audio coding, our approach is to encode one audio channel only (which can be one of the spot signals or a downmix), while for the remaining spot signals we retain only the parameters that allow for resynthesis of the content at the decoder. These parameters are the sinusoidal parameters (harmonic part) of each spot signal, as well as the short-time spectral envelope (estimated using Linear Predictive – LP – analysis) of the sinusoidal noise component of each spot signal. These parameters are not as demanding in coding rates, as the true noise part of the SNM model. For this reason, the noise part of only the reference signal is retained; during the resynthesis of each spot signal, its harmonic part is added to the noise part, which is recreated by using the corresponding noise envelope with the noise residual obtained from the reference channel. This procedure, has been described in our recent work as *noise transplantation* [12], and is based on the observation that the noise component of the spot signals of the same multichannel recording are very similar when the harmonic part has been captured with an appropriate number of sinusoids. In this paper, we focus on describing the coding stage of the model parameters, and defining the lower limits in terms of bitrate that our proposed system can achieve. The coding of the sinusoidal parameters is based on the high-rate quantization scheme of [13], while the encoding process of the noise envelope is based on the vector quantization method described in [14].

2 Modeling Methodology

Initially, we briefly explain how interactivity can be achieved using the multiple microphone recordings (spot microphone signals) of a particular multichannel recording. The number of these multiple microphone signals is usually higher than the available loudspeakers, thus a mixing process is needed when producing a multichannel audio recording. We place emphasis on the mixing of the multi-microphone audio recordings on the decoder side. Remote mixing is imperative for immersive audio applications, since it offers the amount of freedom for the creation of the content that is needed for interactivity. Thus, in immersive audio applications, current multichannel audio coding methods. This is due to the fact that, for audio mixing (remote or not), not only the spatial image but the content of each microphone recording must be encoded, so that the audio engineer will have full control of the content. We note that remote mixing, when the user is not an experienced audio engineer, can be accomplished in practice by storing at the decoder a number of predefined mixing "files" that have been created by experts for each specific recording. The limitations of transmitting the microphone recordings through a low-bandwidth medium (*e.g.*, the Internet or wireless channels) are due to: (i) the increase in the audio channels, which translates into the need of high transmission rates which are not available, and (ii) network delays and losses which are unacceptable in high-quality real-time audio applications. In order to address these problems, we propose using the source/filter and sinusoidal models.

The source/filter model [15] segments the signal in short (around 30 ms) segments, and the spectral envelope of each segment is modelled (*e.g.*, by linear prediction) using a small number of coefficients (filter part). The remaining modelling error has the same number of samples as the initial segment (source part), and contains important spectral information. For speech signals, the source part theoretically contains the integer multiples of the pitch, so it can be modelled using a small number of coefficients. Many speech compression methods are based on this concept. However, for audio signals, methods for reducing the dimensionality of the source signal and retaining high quality have not yet been derived. We have recently found that multiresolution estimation of the filter parameters can greatly improve the modelling performance of the filter model. We, then, were able to show that the source/filter model can separate the spot microphone signals of multimicrophone recording into a part that is specific to each microphone (filter) and a part which can be considered common to all signals (source) [16]. Thus, for each spot recording we can only encode its filter part (using around 10 Kbit/sec), while one reference audio signal (can be a dowmix) must be encoded, *e.g.*, using MP3.

Our aforementioned method introduces an amount of correlation between the recordings (crosstalk), and is not suitable for some audio signals (*e.g.*, transients). These problems can be overcome by additional use of the sinusoidal model [17–19]. It has been applied to speech and audio signals and is based on retaining (for each segment) only the prominent spectral peaks. The sinusoidal parameters alone cannot model audio signals with enough accuracy. Representing the modelling error is an important problem for enhancing the low audio quality of the sinusoids-only model. It has been proposed that the error signal can be modelled by only retaining its spectral envelope (*e.g.*, [19–21]). The sinusoids plus noise model (SNM) represents a signal $s(n)$, with harmonic nature, as the sum of a predefined number of sinusoids (harmonic part) and a noise term (stochastic part) $e(n)$ (for each short-time analysis frame)

$$s(n) = \sum_{l=1}^{L} \alpha_l \, \cos(\omega_l \, n + \phi_l) + e(n) \, , \; n = 0, \ldots, N-1, \tag{1}$$

where L denotes the number of sinusoids, $\{\alpha_l \, , \, \omega_l \, , \, \phi_l\}_{l=1}^{L}$ are the constant amplitudes, frequencies and phases respectively and N is the length (in samples) of the analysis short-time frame of the signal. The noise component is also needed for representing the noise-like part of audio signals which is audible and is necessary for high-quality resynthesis. The noise component can be computed by subtracting the harmonic component from the original signal. Modeling the noise component is a challenging task. We follow the popular approach of modeling $e(n)$ as the result of filtering a residual noise component with an autoregressive (AR) filter that models the noise spectral envelope, *i.e.*,

$$e(n) = \sum_{i=1}^{p} b(i) \, e(n-i) + r_e(n), \tag{2}$$

where $r_e(n)$ is the residual of the noise, and p is the AR filter order, while vector $\boldsymbol{b} = (1, -b(1), -b(2), ..., -b(p))^T$ represents the spectral envelope of the noise component $e(n)$ and can be obtained by LP analysis. In the remainder of the paper, we refer to $e(n)$ as the (sinusoidal) *noise* signal, and to $r_e(n)$ as the *residual* (noise) of $e(n)$.

Fully parametric models under the SNM degrade audio quality, since the residual of the original audio signals is discarded and replaced by (filtered) white noise or parametrically generated. Thus, so far the sinusoidal (as the source/filter) model is considered useful (for audio) only in low-bitrate low-quality applications (*e.g.*, scalable audio coding in MPEG-4). The idea in our research is to apply our findings of the source/filter model not to the actual audio signal but to the sinusoidal error signal. Our preliminary efforts have shown that this “noise transplantation” procedure is indeed valid and can overcome the problems of crosstalk and transient sounds, since even only a few sinusoidal coefficients can capture the significant components of an audio signal. In fact, by using our approach, the number of sinusoidal coefficients can be greatly decreased compared to current sinusoidal models, due to the improved accuracy in the noise modelling of the proposed multiresolution source/filter model.

In more detail, consider a collection of M microphone signals that correspond to the same multichannel recording and thus have similar acoustical content. We model and encode as a full audio channel only one of the signals (alternatively it can be a downmix, *e.g.* a sum signal), which is the reference signal. The remaining (side) signals are modeled by the SNM, retaining their sinusoidal components and the noise spectral envelope (filter $\boldsymbol{b}$ in (2)). In order to reconstruct the side signals, we obtain the LP residual of the reference channel's noise signal. Each side microphone signal is reconstructed using its sinusoidal (harmonic) component and its noise LP filter. In specific, its harmonic component is added to the noise component that it is obtained by filtering, with the signal's LP noise shaping filter, the LP residual of the sinusoidal noise from the reference signal. In this manner, we avoid encoding the residual of each of the side signals. This is important, since this signal is of highly stochastic nature, and cannot be adequately represented using a small number of parameters (thus, it is highly demanding in bitrates for accurate encoding). We note that modeling this signal with parametric models results in low-quality audio resynthesis; in our previous work [12] we have shown that our noise transplantation method can result in significantly better quality audio modeling compared to parametric models for the residual signal. We obtained subjective scores around 4.0 using as low as 10 sinusoids, which is very important for low bitrate coding.

For decoding, the proposed model operates as follows. The reference signal (Signal 1) is fully encoded (*e.g.* using an MP3 encoder at 64 kbps), while the remaining microphone signals are reconstructed using the quantized sinusoidal and LP parameters, using the LP residual obtained from the reference channel.

3 Coding Methodology

The second part of our method is the coding procedure. It can be divided into two tasks; the quantization of the sinusoidal parameters and the quantization of the noise spectral envelopes for each side signal (for each short-time frame).

3.1 Coding of the Sinusoidal Parameters

We adopt the coding scheme of [13], developed for jointly optimal quantization of sinusoidal frequencies, amplitudes and phases. Due to space limitations, and since the details of the coding can be found in [13], we only provide here the final equations for the coding. More specifically, the quantizations point densities $g_A(\alpha)$, $g_\Omega(\omega)$ and $g_\Phi(\phi)$ (corresponding to amplitude, frequency, and phase, respectively) are given by the following equations:

$$g_A(\alpha) = g_A = \frac{w_\alpha^{\frac{1}{6}} 2^{\frac{1}{3}\tilde{H} - \frac{2}{3}b(A)}}{w_g^{\frac{1}{6}} \left(\frac{N^2}{12}\right)^{\frac{1}{6}}}, \tag{3}$$

$$g_\Omega(\omega, \alpha) = g_\Omega(\alpha) = \frac{\alpha w_\alpha^{\frac{1}{6}} \left(\frac{N^2}{12}\right)^{\frac{1}{3}} 2^{\frac{1}{3}\tilde{H} - \frac{2}{3}b(A)}}{w_g^{\frac{1}{6}}}, \tag{4}$$

$$g_{\Phi}(\phi, \alpha, w_l) = g_{\Phi}(\alpha, w_l) = \frac{\alpha w_l^{\frac{1}{2}} 2^{\frac{1}{3}\tilde{H} - \frac{2}{3}b(A)}}{w_{\alpha}^{\frac{1}{3}} w_g^{\frac{1}{6}} (\frac{N^2}{12})^{\frac{1}{6}}}, \tag{5}$$

where w_α and w_g are the arithmetic and geometric mean of the perceptual weights of the L sinusoids, respectively, $\tilde{H} = H - h(A) - h(\Omega) - h(\Phi)$ and $b(A) = \int f_A(\alpha) \log_2(\alpha)\, d\alpha$. The quantities $h(A)$, $h(\Omega)$ and $h(\Phi)$ are the differential entropies of the amplitude, frequency and phase variables, respectively, while $f_A(\alpha)$ denotes the marginal pdf of the amplitude variable.

3.2 Coding of the Spectral Envelopes

The second group of parameters for each spot signal that need to be encoded are the spectral envelopes of the sinusoidal noise. We follow the quantization scheme of [14]. The LP coefficients of each spot signal that model the noise spectral envelope are transformed to LSF's (Line Spectral Frequencies) which are modeled by means of a Gaussian Mixture Model (GMM). Then, the Karhunen Loève Transform (KLT) decorrelates each LSF vector for each time segment. The decorrelated components can be independently quantized by a non-uniform quantizer (compressor, uniform quantizer and expander). Each LSF vector is classified to only one of the GMM clusters. This classification is performed in an analysis-by-synthesis manner. For each LSF vector, the Log Spectral Distortion (LSD) is computed for each GMM class (the distortion among the spectral envelopes obtained by the original and the quantized LSF vectors), and the vector is classified to the cluster associated with the minimal LSD.

4 Results

In this section, we are interested to examine the coding performance of our proposed system, with respect to the resulting audio quality. For this purpose we performed subjective (listening) tests. We employed the Degradation Category Rating (DCR) test, in which listeners grade the coded *vs* the original waveform using a 5-scale grading system (from 1-"very annoying" audio quality compared to the original, to 5-"not perceived" difference in quality). For our listening tests, we used three signals, referred to as Signals 1-3. These signals are parts of a multichannel recording of a concert hall performance. We used the recordings from two different microphones, one of which captured mainly the female voices of the orchestra chorus, while the second one captured mainly the male voices. The former was used in our experiments as the side channel, and the latter as the reference signal. Thus, the objective is to test whether the side signal can be accurately reproduced when using the residual from the reference signal. We note that in our previous work [12], we showed that the proposed noise transplantation approach results in very good quality (around 4.0 grade in DCR tests in most cases) for various music signals, with the number of sinusoids per frame as low as 10. Thus, in this section our objective is to examine the lower limit in bitrates which can be achieved by our system without loss of audio quality below the grade achieved by modeling alone (*i.e.* 4.0 grade for the three signals tested here).

Regarding the parameters used for deriving the waveforms used in the tests, the sampling rate for the audio data was 44.1 kHz and the LP order for the AR noise shaping filters was 10. The analysis/synthesis frame for the implementation of the sinusoidal model is 30 msec with 50% overlapping between successive frames. The coding efficiency for the sinusoidal parameters was tested for a given (target) entropy of 28 and 20 bits per sinusoid (amplitudes, frequencies and phases in total), which gives a bitrate of 19.6 kbps and 14.2 kbps respectively. Regarding the coding of the LP parameters (noise spectral envelope), 28 bits were used per LSF vector. With 23 msec frame and 75 % overlapping, this corresponds to 4.8 kbps for the noise envelopes. Thus, the resulting bitrates that were tested are 24.4 kbps and 19 kbps (adding the bitrate of the sinusoidal parameters and the noise envelopes). A training audio dataset of about 100,000 LSF vectors (approximately 9.5 min of audio) was used to estimate the parameters of a 16-class GMM. The training database consisted of recordings of the classical music performance (corresponding to the recording from which Signals 1-3 originated, but a different part of the recording than the one used for testing). Details about the implementation of the coding procedure for the LP parameters can be found in our earlier work [16].

Eleven volunteers participated in the DCR tests, using high-quality headphones. The results of the DCR tests are depicted in Fig. 1, where the 95% confidence interval are shown (the vertical lines indicate the confidence limits). The solid line shows the results for the case of coding with a bitrate of 19 kbps, while the dotted line shows the results for the 24.4 kbps case. The results of the figure verify that the quality of the coded audio signals is good and the proposed algorithm offers an encouraging performance, and that this quality can be maintained at as low as 19 kbps per side signal. We note that the reference signal was PCM coded with 16 bits per sample, however similar results were obtained for the side signals when the reference signal was MP3 coded at 64 kbps (monophonic case).

5 Conclusions

In this paper a novel modeling approach, namely noise transplantation, was proposed for achieving interactive and immersive audio applications of high-quality audio at low bitrates. The approach was based on applying the sinusoidal model at spot microphone signals, *i.e.* the multiple audio recordings before performing the mixing process which produces the final multichannel mix. It was shown that these signals can be encoded collectively using a bitrate as low as 19 kbps per spot signal. Further research efforts are necessary in order to achieve even lower bitrates while preserving high audio quality, while a more detailed testing procedure using subjective methods is currently underway.

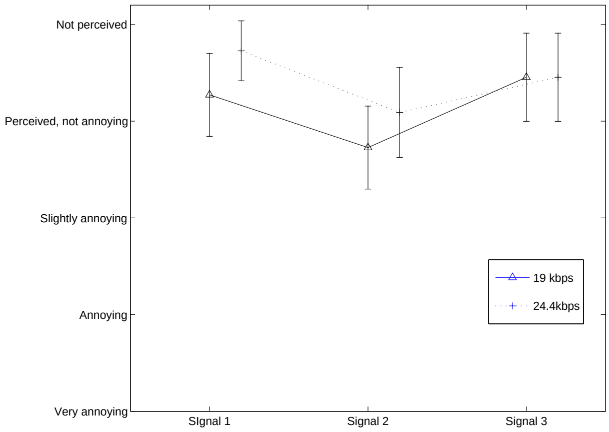

Fig. 1. Results from the quality rating DCR listening tests, corresponding to coding with (a) 24.4 kbps (dotted), (b) 19 kbps (solid). Each frame is modeled with 10 sinusoids and 10 LP parameters.

Acknowledgments

The authors wish to thank Prof. Y. Stylianou for his insightful suggestions and for his help with the implementation of the sinusoidal model algorithm, Prof. C. Kyriakakis for providing the audio recordings used in the experiments, as well as the listening tests volunteers.

References

1. ITU-R BS.1116, “Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems,” 1994. International Telecommunications Union, Geneva, Switzerland.
2. ISO/IEC JTC1/SC29/WG11 (MPEG) International Standard ISO/IEC 11172-3, “Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s,” 1992.
3. K. Brandenburg, “MP3 and AAC explained,” *in Proc. 17th International Conference on High Quality Audio Coding of the Audio Engineering Society (AES)*, September 1999.
4. ISO/IEC JTC1/SC29/WG11 (MPEG) International Standard ISO/IEC 13818-7, “Generic coding of moving pictures and associated audio: Advanced audio coding,” 1997.
5. M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, “ISO/IEC MPEG-2 advanced audio coding,” *in Proc. 101st Convention of the Audio Engineering Society (AES)*, preprint No. 4382, (Los Angeles, CA), November 1996.
6. M. Davis, “The AC-3 multichannel coder,” *in Proc.* 95th *Convention of the Audio Engineering Society (AES)*, preprint No. 3774, (New York, NY), Oct. 1993.

7. J. D. Johnston and A. J. Ferreira, “Sum-difference stereo transform coding,” *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 569-572, 1992.
8. J. Herre, K. Brandenburg, and D. Lederer, “Intensity stereo coding,” *in Proc. 96th Convention of the Audio Engineering Society (AES)*, preprint No. 3799, February 1994.
9. J. Breebaart, J. Herre, C. Faller, J. Roden, F. Myburg, S. Disch, H. Purnhagen, G. Hotho, M. Neusinger, K. Kjorling, and W. Oomen, “MPEG Spatial Audio Coding / MPEG Surround: Overview and current status,” *in Proc. AES 119^{th} Convention*, Paper 6599, (New York, NY), October 2005.
10. F. Baumgarte, and C. Faller, “Binaural Cue Coding - Part I: Psychoacoustic Fundamentals and Design Principles,” *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, pp. 509–519, Nov. 2003.
11. J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, “Parametric coding of stereo audio,” *EURASIP Journal on Applied Signal Processing*, pp. 1305-1322, 2005:9.
12. C. Tzagkarakis, A. Mouchtaris, and P. Tsakalides, “Modeling spot microphone signals using the sinusoidal plus noise approach,” *in Proc. Workshop on Appl. of Signal Proc. to Audio and Acoust.*, Oct. 2007.
13. R. Vafin, D. Prakash, and W. B. Kleijn, “On Frequency Quantization in Sinusoidal Audio Coding,” *IEEE Signal Proc. Letters*, vol. 12, no. 3, pp. 210–213, Mar. 2005.
14. A. D. Subramaniam, and B. D. Rao, “PDF optimized parametric vector quantization of speech line spectral frequencies,” *IEEE Trans. on Speech and Audio Proc.*, vol. 11, pp. 365–380, Mar. 2003.
15. L. Rabiner and B.-H. Juang. Fundamentals of Speech Recognition. Prentice Hall,1993.
16. K. Karadimou, A. Mouchtaris, and P. Tsakalides, “Multichannel Audio Modeling and Coding Using a Multiband Source/Filter Model,” *Conf. Record of the Thirty-Ninth Asilomar Conf. Signals, Systems and Computers*, pp. 907–911, 2005.
17. R. J. McAulay and T. F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 34(4), pp. 744-754, August 1986.
18. Y. Stylianou, “Applying the harmonic plus noise model in concatenative speech synthesis,” *IEEE Trans. Speech and Audio Process.*, vol. 9(1), pp. 21-29, 2001.
19. X. 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(4), pp. 12-24, Winter 1990.
20. M. Goodwin, “Residual modeling in music analysis-synthesis,” *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 1005-1008, May 1996.
21. R. C. Hendriks, R. Heusdens, and J. Jensen, “Perceptual linear predictive noise modeling for sinusoid-plus-noise audio coding,” *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 189-192, May 2004.