Audio Engineering Society

Convention Paper

Presented at the 130th Convention
2011 May 13–16 London, UK

On the Multichannel Sinusoidal Model for Coding Audio Object Signals

Toni Hirvonen[1]*and Athanasios Mouchtaris[1,2]

[1]*Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH-ICS) Heraklion, Crete, Greece.*

[2]*Department of Computer Science, University of Crete, Heraklion, Crete, Greece.*

Correspondence should be addressed to Athanasios Mouchtaris (mouchtar@ieee.org)

ABSTRACT

This paper presents two improvements on a recently proposed multichannel sinusoidal modeling system for coding multiple audio object signals. The system includes extracting the sinusoidal components and an LPC envelope for each object signal, as well as transform coding of the residuals' downmix. The contributions of this paper are: (a) a psychoacoustic model for enabling the system to scale well with multiple object signals, and (b) an improved method to encode the common residual, tailored to the "white" nature of this signal. As a result, sound quality of around 90% on the MUSHRA scale is obtained for 10 simultaneous object signals coded with a total rate of 150 kbit/s, while retaining their individual parametric representations

1. INTRODUCTION

Current multichannel audio coding systems are predominantly designed for mixed material, with the audio channels being typically highly spectro-temporarily correlated, such as in a 5.1 audio mix. Channel redundancy techniques such as parametric stereo [1] and Binaural Cue Coding (BCC) [2] (collectively referred to as Spatial Audio Coding (SAC) and standardized as MPEG Surround), have been proposed. These methods encode only the binaural cues necessary for delivering the spatial image of the initial multichannel recording, as well as a downmix signal of the multiple audio channels.

Coding of individual audio object signals with little correlation is starting to emerge as an important field of research, as evidenced by the standardization of Spatial Audio Object Coding (SAOC) [3]. If several object signals, such as different instruments of

*T. Hirvonen is now with Dolby Laboratories, Stockholm, Sweden.

a music performance, can be recovered for independent manipulation at the decoder, there are more possibilities for interactive and immersive audio applications than with pre-mixed material. Current applications of SAOC are based on applying channel redundancy techniques similar to BCC, such as coding object level differences and correlations at several frequency bands. In contrast, our approach in [4] encodes the actual content of the audio object signals (referred to as spot signals in that work), and not only the spatial image of the recording. In this sense, it can be claimed that the proposed methodology offers more freedom for flexible rendering applications compared to SAC-type approaches.

In this paper, improvements on the multichannel sinusoidal modeling system for object coding detailed in [4] are presented. Sinusoidal modeling [5] is a popular parametric audio modeling method for low bitrates, based on retaining only some spectral peaks of an audio signal which are deemed perceptually important. Sinusoidal representation allows for simple signal modifications, such as modifying the pitch/timbre or time duration, by manipulating the parameters directly. This is especially useful in an object coding framework.

An additional component which models the sinusoidal error signal can be included to the model. In [4], efficient sinusoidal error coding was accomplished by retaining only one spectrally whitened common signal (residual), as well as the whitening parameters (LPCs) for each object. Utilizing the common residual (downmix of the multiple residuals) was found to produce much better quality than low-bitrate residual modeling techniques, at the expense of fully encoding the downmix residual as one audio channel. The sinusoidal model had a constant number of components per object. As psychoacoustic masking was not utilized, the bitrate of the method increased proportionally to the number of signals.

In this paper we propose two main improvements to our previous work: (a) a psychoacoustic-based sinusoidal parameter estimation method, and (b) transform coding for the downmix residual, tailored to its "white" nature. We claim that both these approaches are not limited in scope to the particular system of [4], but are important to a wide class of audio coding methods. In particular, the proposed psychoacoustic analysis extends the existing matching pursuit (MP) methods to be utilized within a SAOC-type object coding framework. Our approach is applicable to other sinusoid-based multichannel audio coding redundancy techniques, such as [6, 7]. The second proposed method is using scalar quantization based on accurate modeling of the distribution of MDCT coefficients for the whitened residual. Given that this signal is the result of inverse LPC filtering of the sinusoidal error signal, *i.e.* a spectrally whitened signal, our approach is useful in any other audio coding method where a white residual needs to be separately encoded (single- or multi- channel sinusoidal audio coding, LPC-based audio coding, lossless audio coding).

It is noted that the system proposed in [4] and extended here, in essence provides a sinusoidal noise modeling method for the case that multiple audio signals are encoded (and consequently rendered) simultaneously. In [4] we showed that the proposed noise modeling approach outperforms state-of-art noise modeling methods based on retaining only the noise spectral envelope (such as [8, 9]), for this particular multiple signal encoding case. Thus, the proposed system can be applied for extending state-of-art single-channel sinusoidal audio coding systems (such as MPEG-4 HILN [10], MPEG-4 SSC, the systems described in [11], in [12], and in [13]) so that they can be applied to multiple audio signal encoding for flexible rendering applications.

2. ANALYSIS AND SYNTHESIS MODEL

2.1. Brief Description of the System in [4]

Fig. 1 presents a block-diagram of the encoding and decoding parts of the system examined in this paper. The main assumption, as is usual in audio object coding, is that k object signals are rendered simultaneously, and can thus be jointly analyzed. The system consists of independent blocks of sinusoidal modeling, LPC modeling of the sinusoidal error signal, residual extraction per object signal, and transform coding of the common residual (downmix of the residuals corresponding to all object signals). For clarity, the *sinusoidal error* signal is the signal obtained by subtracting the sinusoidally-modeled signal from the original object signal (*i.e.* the sinusoidal model error signal). Furthermore, the term *residual* in this paper corresponds to the signal which remains

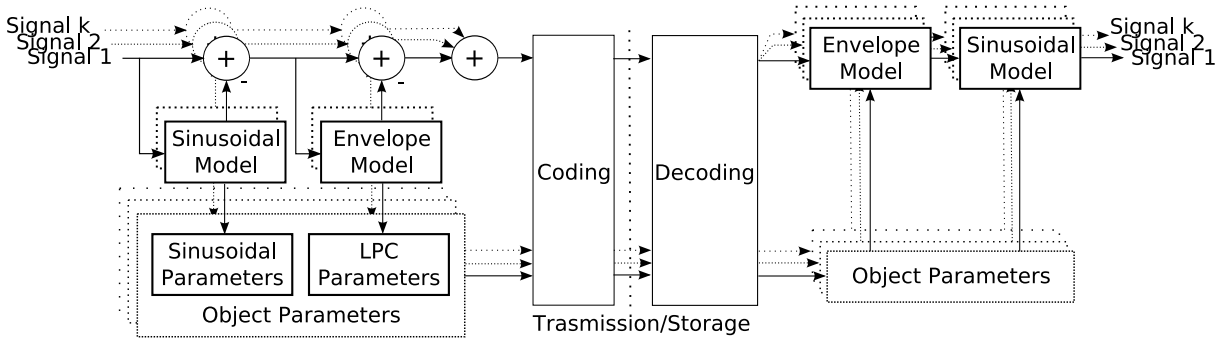

Fig. 1: Block diagram of the audio object encoding and decoding chains.

after the sinusoidal error signal is whitened by filtering with the inverse LPC filter. The sinusoidal analysis and LPC analysis are performed using all separate object signals, while at the last stage, all whitened residual signals are summed to form the common residual. The encoded parameters are: (i) the sinusoidal parameters per object signal, (ii) the LPC parameters per object signal, and (iii) the common residual, which has the same number of samples as any one of the object signals. This procedure is performed by processing the audio signals in short-time frames. Below, specific improvements to that work are proposed.

2.2. Psychoacoustic Analysis

Sinusoidal modeling results in retaining only the spectral peaks of an audio signal that are perceptually important related to the masking threshold. The current state-of-the-art methods for sinusoidal modeling utilize greedy matching pursuit algorithms. In [14], an improved frequency masking model was combined with Psychoacoustic Matching Pursuit (PAMP). At each iteration i, PAMP estimates the i^{th} sinusoidal component by minimizing the perceptual distortion measure

$$D_i = \sum_f \hat{a}_i(f)|\hat{r}_i(f)|^2, \qquad (1)$$

where $\hat{}$ indicates the frequency-domain signal, f is the frequency index, r_i indicates the sinusoidal error signal at the i^{th} iteration, and $\hat{a}(f)$ is a weighting function usually set as the reciprocal of the masking energy. Thus, updating the sinusoidal error magnitude spectrum and the masking curve is required at each iteration.

If $\hat{a}(f)$ is positive and real for all f, the distortion measure (1) defines a norm as $||x||^2 = \sum_f \hat{a}_i(f)|\hat{x}_i(f)|^2$ on the underlying signal space. The norm is induced by the inner product

$$\langle x, y\rangle = \sum_f \hat{a}(f)\hat{x}(f)\hat{y}^*(f), \qquad (2)$$

utilizing the distortion measure in the sinusoidal component selection of the matching pursuit algorithm.

In the following, we present a modified matching pursuit method to be used in the analysis of multiple object signals. First, we assume that we have access to k object signals $s_k(n)$, and these are reproduced simultaneously at the receiver. This implies that it is meaningful to analyze the sum of these signals $s_t(n) = \sum_k g_k s_k(n - \delta_k)$, for performing the psychoacoustic analysis. This holds if we have knowledge of the relative signal gains g_k, as well as relative arrival delays δ_k in the reproduction position. In SAOC-type framework g_k typically remain at unity, and δ_k are equal. In real-time transmission, interactivity at the decoder implies that g_k and/or δ_k change, thus feedback of these values to the transmitter is required.

Our recent work indicates that the perceptual weighting $\hat{a}(f)$ can be calculated from $s_t(n)$ in cases, where the reproduced signals do not have large interaural differences [15]. Furthermore, because of the use of an estimated residual signal of the sinusoidal error during decoding, the signal energy of

each analysis frame is well reproduced in synthesis. In this light, it is reasonable to use the masking curve of the whole sum signal $\hat{a}_t(f)$ instead of iterating it at each step. In addition to being faster, our recent work showed that this approach produces better quality with added residual [15].

The selection of the perceptually important sinusoidal components is performed based on the Matching Pursuit (MP) approach that was explained. Since multiple audio signals are modeled at the same time, the MP algorithm is performed simultaneously on these signals, by picking at each iteration the sinusoidal component which minimizes the sum of the distortions for all object signals. In this manner, a pre-defined total number of sinusoidal components for all object signals can be given depending on the desired total bitrate or application. An important issue related to the matching pursuit approach for the case of object coding that is addressed here is that the phases and amplitudes of the chosen sinusoidal components should match the parameters of the best-fitting object signal at the sinusoid's frequency. The phases of $s_k(n)$ and $s_t(n)$ may differ, and consequently using a norm where only absolute values are taken into account may lead to biased component choices. Thus, we suggest using a distortion measure for MP object coding that is similar to what is used in conjugate subspace MP:

$$D_{ki} = \sum_f \hat{a}_t(f)\Re(\hat{r}_{ti}(f)\hat{r}^*_{ki}(f)), \quad (3)$$

where $r_{ti}(f)$ is the sinusoidal error signal corresponding to the sum signal and $r_{ki}(f)$ the sinusoidal error signal of object signal k — at iteration i of one analysis frame —, $\Re()$ denotes the real part, and * denotes the complex conjugate. The perceptual weighting function $\hat{a}_t(f)$ is calculated only once for each frame using the masker energy of $s_t(n)$. Utilizing the real part of the product takes into account the possible phase difference between $s_k(n)$ and $s_t(n)$ and ensures that no component that adds to the distortion is actually chosen.

The distortion measure (3) does not define a norm, since it can yield non-positive results. However, it can be used by choosing the frequency γ that, at each iteration i, maximizes

$$\gamma_i = \arg\sup_{\gamma,k} \sum_f \hat{a}_t(f)\Re(\hat{r}_{ti}(f)\hat{r}^*_{ki}(\gamma)\hat{w}(f-\gamma)), \quad (4)$$

where w is the window function used for analysis frames. Amplitude and phase are obtained from $\hat{r}_{ki}(\gamma_i)$.

If a distortion measure yielding positive values is desired, an approximation of (3) can be used

$$\tilde{D_{ki}} = \sum_f \hat{a}_t(f)|\hat{r}_{ti}(f)\hat{r}^*_{ki}(f)|. \quad (5)$$

This measure defines a norm $\langle x, y\rangle = \sum_f \hat{a}_i(f)|\hat{x}_i(f)\hat{y}^*_i(f)|$ as in (2). Measure (5) can in some cases lead to selecting a suboptimal component but was nevertheless deemed mostly appropriate. The tests of this paper utilize measure (3). If $\hat{r}_{ti}(f)$ and $\hat{r}_{ki}(f)$ have the same phase, (3) reduces to (5).

The method additionally requires sending side information describing the object signal that each transmitted sinusoid was obtained from. This information is in this paper represented with 4 bits per component and is Huffmann coded. Some low-level object signals may obtain no sinusoidal components in one frame. For the quantization of the sinusoidal parameters, we have utilized the method of [16] since it does not require transmitting the masking curve of the psychoacoustic model.

2.3. LPC modeling

After removing the selected sinusoidal components, each object signal is further whitened by LPC inverse filtering. In this paper we used 20th order filters. The LPC coefficients for each object are transformed to line spectral frequencies (LSF) and quantized using the method of [17].

2.4. Residual Quantization

In [4], it was proposed to use a proprietary audio coder for the residual signal (*e.g.* MP3 monophonic encoding). However this is not an optimal solution given the specific "white" nature of the residual signal. In this paper, we utilize uniform scalar quantization (SQ) followed by entropy (Huffman) coding. Uniform SQ was chosen because of computational speed, and the theoretical near Gaussian-noise nature of the residual signal. We also assume that the LPC curve approximates the masking curve, and whitening roughly equalizes the perceptual importance of the different frequencies. Thus, no further

weighting/bit allocation between frequency bands is performed in the quantization.

The time-domain residual was transformed using the Modified Discrete Cosine Transform (MDTC). As common with MDCT coding, one frame of 1024 samples was processed as either one long, or four short frames, depending on transient analysis. A transient was set to occur if the energy of the 5 kHz high-passed 256-sample frame exceeded the energy of the previous frame sixfold. Being approximately zero-mean and symmetric, the probability density function (pdf) of the transformed coefficients is modeled by Generalized Gaussian distribution (GGD). GGD adds a shape parameter to the normal distribution, so that the family encompasses various distributions from Laplace to uniform distribution. As in [18], the estimated shape parameter $\hat{\alpha}$ for a random variable can be obtained as a solution to equation

$$\frac{\hat{\sigma}^2}{\hat{\mu}^2} = \frac{\Gamma(1/\alpha)\Gamma(3/\alpha)}{\Gamma^2(2/\alpha)}, \tag{6}$$

where Γ denotes the Gamma function, $\hat{\sigma}^2$ indicates the sample variance and $\hat{\mu}$ the first absolute moment. Here, a single shape parameter $\hat{\alpha}$ is estimated for the whole MDCT coefficient vector of one frame. The fit of the GGD model to the whitened residuals was examined heuristically and deemed to be good with $\hat{\alpha}$ generally close to 1. To obtain the optimal step size for the quantizer of this frame, we use the method in [18], where Lagrangian techniques are applied to minimize distortion. It was shown that the optimal step size is

$$q = \sqrt{\frac{6\lambda_{opt}}{ln(2)}}, \tag{7}$$

where λ_{opt} is the optimal Lagrange multiplier that depends on the target bitrate per sample.

2.5. Bit allocation

Fig. 2 shows the Objective Difference Grade (ODG) scores as a function of total bitrate with different amounts of sinusoidal components (total number for all object signals). The ODG scores were obtained using the PEAQ algorithm (Perceptual Evaluation of Audio Quality), based on recommendation ITU-R BS.1387. In the ODG scale, grades are given in a scale between -4 to 0, where -4 is the minimum grade and corresponds to the decoded signal being

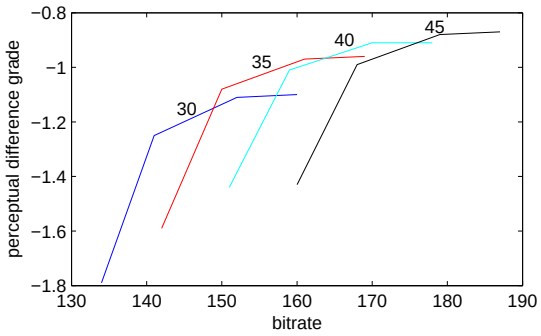

Fig. 2: Perceptual difference grades as a function of total bitrate with different (total) amounts of sinusoidal components.

of "very annoying" quality compared to the original recording; similarly, 0 is the maximum grade and corresponds to the decoded signal being of "imperceptible" quality compared to the original signal. The test was done by comparing the original sum signal $s_t(n)$ to the synthesized sum signal using the "Classical 1" test sample as discussed in the following section. The total bitrate was adjusted by varying the residual quantization rate between 1-2 bits per sample, without using further bit reduction techniques that replicate high frequency spectral part. The results may vary for different frame length parameters.

It can be seen that for each choice of the number of sinusoids, the optimal MDCT-rate is similar, around 1.4 bits per sample, and that the quality improvement stagnates with larger rate. This is a property caused by utilizing the same sum residual for all objects in synthesis. Thus, we use here constant bitrate of 1.4 bits per sample in MDCT-quantization. This allows for adjusting the total bitrate only by the number of sinusoids. We used 25 bits per sinusoidal component, which is relatively high; optimized quantizers report good results even with 16-20 bits. The bitrate for LPC coefficient transmission was kept constant as 20 bits per frame per object signal. All audio signals in this paper were sampled at 44.1 kHz.

3. SUBJECTIVE TESTING

In this section, the objective is to evaluate the performance of the proposed method using listening

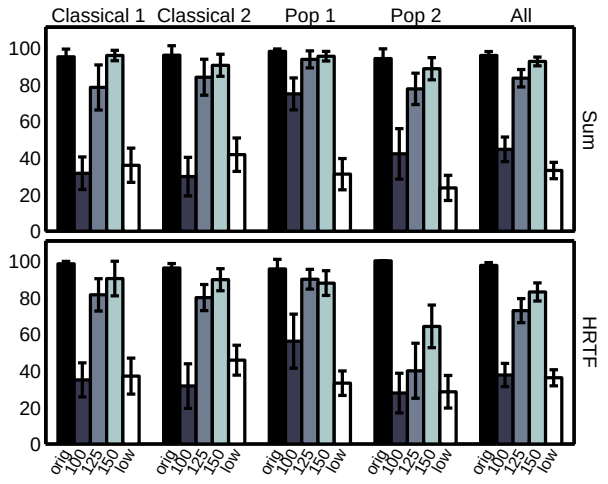

Fig. 3: MUSHRA listening test results for different signals and bitrates. Bar values indicate grade means and whiskers 95% confidence intervals. Upper panel: object signals summed. Lower panel: object signals spread equally in separate directions with HRTFs.

tests. As explained, the current work proposes two important improvements compared to our previous work in [4] (*i.e.* the psychoacoustic model and the residual coding). Our goal in this paper is to examine the total bitrate that our algorithm must operate for sound quality at approximately 90% in the MUSHRA scale, when these improvements are included in the implementation. It must be noted that the main feature of the method is to provide a means for deriving a sinusoidal error signal which can lead to good quality audio resynthesis (noise modeling); in this sense, it was compared and found to be superior (for the specific case of multiple audio signal encoding) than state-of-art sinusoid noise modeling methods in [4]. Thus, in this section the objective is to examine the bitrates obtained with the proposed improvements and not to compare our method with state-of-the-art.

A double-blind MUSHRA listening test was utilized to evaluate the overall quality of the method at various bitrates. Four sets of individual instrument tracks from musical performances (2 classical, 2 pop/rock) were used to create four samples. The classical music samples consisted of ten different object signals each (some doubling instruments were mixed into one object). The pop/rock samples consisted of eight object signals. Material for classical recordings was obtained from [19], and pop/rock tracks from the material made publicly available by the band "Nine Inch Nails". It is noted that these signals cover a variety of sound events which are important to examine in audio coding, including transient sounds in many parts of the recordings.

Two test cases were created for each of the four samples: a single-channel sum signal of all synthesized objects (monophonic listened diotically), as well as an HRTF-synthesized binaural version, created by placing each of the object signals in separate, evenly spaced, directions. The purpose of the binaural material was to examine whether the psychoacoustic analysis performed on the sum signal $s_t(n)$ of all objects is appropriate if the spatial properties of the objects are separately manipulated. In the binaural tests, subjects could also evaluate spatial attributes in addition to sound quality.

The test included a total of eight cases, each of which was repeated in one run. For each case, the MUSHRA evaluation included five items: original hidden reference, the coding method applied at 100, 125, and 150 kbit/s, as well as a 3.4 kHz low-pass filtered version of the original as an anchor. All signals were played back using Sennheiser HD 650 headphones, whose response was also compensated in the HRTF-cases. Nine subjects (non-expert volunteers who had previous participation in listening tests) participated in the tests.

Results can be seen in Fig. 3. Examining the monophonic case first (upper plot, denoted as "sum" in the figure), the quality can be seen to be similar to the reference with 150 kbit/s and 125 kbit/s signals. A difference was deemed significant or insignificant based on the confidence intervals confirmed by Tukey's HSD tests. Analysis over all samples indicated that 150 kbit/s signals' grades did not differ from the original signals' grades, while the 125 kbit/s signals were found to be slightly different compared to the original. In the binaural case (lower plot in Fig. 3, denoted as "HRTF"), the quality was also found to be similar, with the exception of sample "Pop 2". In this latter sample, the quality can be seen to be significantly reduced compared to the monophonic case. This is attributed to the fact that

this particular sample contained a distorted high-frequency instrument that stood out when spatially separating the object signals, which can be considered as a relatively rare occurrence.

The results of this section indicate that good audio quality (on average 90% on the MUSHRA scale) can be generally achieved using 12.5-15 kbit/s per object signal. It is of interest to note that Binaural Masking Level Difference effects are largely, but not completely, accounted for by the approximation of analyzing $s_t(n)$, instead of considering the final reproduction configuration of the objects. However, our work in [15] indicates that such considerations can be integrated into the psychoacoustic analysis by utilizing individual weighting functions $\hat{a}_k(f)$ for each object. These can then be manipulated according to masking release effects.

4. CONCLUSIONS

We have presented two improvements on a previously proposed method [4] for audio object coding. The method focuses on encoding the individual audio object signals based on sinusoidal modeling, and sinusoidal error synthesis from individual LPC parameters and common whitened residual. A matching pursuit method for sinusoidal object coding was presented, and scalar quantization-based MDCT-coding was utilized for the whitened residual. The presented psychoacoustic analysis allowed for efficient coding of a large number of object signals, and can be exploited in any system where sinusoidal audio coding of multiple audio signals is performed. Subjective tests, using individual instrument recordings as objects, showed that good quality audio (90% on the MUSHRA scale) can be obtained with 150 kbit/s when using 10 object signals, which is a significant improvement compared to the work in [4]. Object-specific parametric representation allows convinient spatial and signal processing manipulations that are to be studied in the future.

5. ACKNOWLEDGEMENTS

This work has been funded in part by the Marie Curie TOK "ASPIRE" grant, and in part by the PEOPLE-IAPP "AVID-MODE" grant, within the 6th and 7th European Community Framework Programs respectively.

6. REFERENCES

[1] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *EURASIP J. Appl. Signal Process.*, vol. 9, pp. 1305–1322, 2005.

[2] F. Baumgarte and C. Faller, "Binaural cue coding - part I: psychoacoustic fundamentals and design principles," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 509–519, 2003.

[3] J. Breebaart et al., "Spatial audio object coding (SAOC) - the upcoming MPEG standard on parametric object based audio coding," *Presented at 124th AES Convention*, May 2008.

[4] C. Tzagkarakis, A. Mouchtaris, and P. Tsakalides, "A multichannel sinusoidal model applied to spot microphone signals for immersive audio," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 8, pp. 1483–1497, Nov. 2009.

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

[6] M. Goodwin, "Multichannel matching pursuit and applications to spatial audio coding," *Asilomar Conf. Signals, Syst., Computers*, October 2006.

[7] A. Härmä and C. Faller, "Spatial decomposition of time-frequency regions: subands or sinusoids," *Presented at 126th AES Convention*, May 2004.

[8] M. Goodwin, "Residual modeling in music analysis-synthesis," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 1996, pp. 1005–1008.

[9] R. C. Hendriks, R. Heusdens, and J. Jensen, "Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2004, pp. 189–192.

[10] H. Purnhagen and N. Meine, "HILN - the MPEG-4 parametric audio coding tools," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2000, pp. 201–204.

[11] B. Edler et al., "ASAS analysis/synthesis codec for very low bit rates," *Presented at 100th AES Convention*, 1996.

[12] S. N. Levine and J. O. Smith III, "Improvements to the switched parametric and transform audio coder," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 1999, pp. 43–46.

[13] T. Verma and T. Meng, "A 6 kbps to 85 kbps scalable audio coder," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, June 2000, pp. 877–880.

[14] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 1, pp. 1292–1304, 2005.

[15] T. Hirvonen and A. Mouchtaris, "Top-down strategies in parameter selection of sinusoidal modeling of audio," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2010, pp. 273–276.

[16] R. Vafin, D. Prakash, and W. B. Kleijn, "On frequency quantization in sinusoidal audio coding," *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 210–213, 2005.

[17] A. D. Subramaniam and B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 130–142, 2003.

[18] M. Oger, S. Ragot, and M. Antonini, "Transform audio coding with arithmetic-coded scalar quantization and model-based bit allocation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2007, pp. 545–548.

[19] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica united with Acustica*, vol. 94, no. 6, November/December 2008.