A Multichannel Sinusoidal Model Applied to Spot Microphone Signals for Immersive Audio

Christos Tzagkarakis, Athanasios Mouchtaris, *Member, IEEE*, and Panagiotis Tsakalides, *Member, IEEE*

***Abstract*—In this paper, a multichannel version of the sinusoids plus noise model (also known as deterministic plus stochastic decomposition) is proposed and applied to spot microphone signals of a music recording. These are the recordings captured by the various microphones placed in a venue, before the mixing process produces the final multichannel audio mix. Coding these microphone signals makes them available to the decoder, allowing for interactive audio reproduction which is a necessary component in immersive audio applications. The proposed model uses a single reference audio signal in order to derive a noise signal per spot microphone. This noise signal can significantly enhance the sinusoidal representation of the corresponding spot signal. The reference can be one of the spot signals or a downmix, depending on the application. Thus, for a collection of multiple spot signals, only the reference is fully encoded (e.g., as an MP3 monophonic signal). For the remaining spot signals, their sinusoidal parameters and corresponding noise spectral envelopes are retained and coded, resulting in bitrates for this side information in the order of 15 kb/s for perceptual performance above the 4.0 grade on the mean opinion score (MOS) scale.**

***Index Terms*—Deterministic plus stochastic decomposition, immersive audio, multichannel audio, noise transplantation, sinusoidal model.**

I. Introduction

In multichannel audio, multiple audio channels are used for audio reproduction with the objective to surround the listener with sound and offer a more realistic acoustic scene compared to two-channel stereo. Current multichannel audio systems place five or seven loudspeakers around the listener in predefined positions, and a further loudspeaker for low-frequency sounds (5.1 and 7.1 multichannel audio systems, respectively). Such systems are utilized not only for the reproduction of film, but also for audio-only content. Multichannel audio involves an equal number of audio recordings as the number of loudspeakers used, and this comes at the expense of increased storage and transmission requirements compared to two-channel stereo. This is important in many network-based applications, such as Digital Radio and Internet audio. Consequently, many compression techniques have been proposed in order to provide efficient solutions in several bitrate-constrained applications. Multichannel audio coding methods, such as MPEG Advanced Audio Coding (AAC) [1], [2] and Dolby AC-3 [3] achieve a significant coding gain but remain demanding for many low-bandwidth applications, such as streaming through the Internet and wireless channels.

An approach towards realizing higher compression ratios than those achieved by the aforementioned methods is to further exploit interchannel similarities. It must be mentioned that AAC and AC-3 both include algorithms that exploit interchannel redundancy, such as Mid/Side Coding [4] (usually applied in the lower frequency bands) and Intensity Stereo Coding [5] (usually applied in the higher frequency bands). In Mid/Side Coding, the sum and the difference signals of the stereo channels are quantized and coded instead of coding the actual channels separately. In Intensity Stereo Coding, only the sum signal of the channels is coded, as well as directional information as side information. A more recent method for better exploiting interchannel redundancy is described in [6], where the Karhunen–Loève Transform (KLT) is applied to multichannel audio signals within the AAC algorithm. Considering the 5.1 multichannel setting, Dolby AC-3 achieves at minimum a data rate of 320 kb/s for transparent audio coding (i.e., audio quality perceptually indistinguishable from the original uncompressed audio recording) [3], although a typical operating data rate for AC-3 is 384 kb/s. For the case of MPEG-2 AAC the minimum data rate of 320 kb/s for 5.1 channels with transparent quality has been reported in [7]. The method of [6] encodes multichannel audio signals at a data rate of 64 kb/s per channel.

Recently, MPEG Surround [8] has been introduced, achieving significant compression of multichannel audio recordings. MPEG Surround is based on the Spatial Audio Coding (SAC) concept. In SAC, only the spatial image of a multichannel audio signal is retained, by encoding one channel of audio (reference channel, which can be a downmix signal) and the parameters that capture the multichannel spatial image as side information. At the decoder, the original spatial image of the multichannel recording can be recreated, by applying the extracted spatial cues to the reference channel. MPEG Surround is based on the work on Binaural Cue Coding (BCC) [9], [10] and Parametric Stereo [11]. In MPEG Surround, it is possible to encode each channel with side information of only a few kb/s without significant loss in the spatial image. For MPEG Surround, rates as low as 48 kb/s for 5.1 multichannel audio of high-quality (not transparent though) have been recently reported [12]. It is noted that for backward compatibility with two-channel

Manuscript received July 23, 2008; revised April 09, 2009. First published April 28, 2009; current version published August 14, 2009. This work was supported by the Marie Curie Transfer of Knowledge Grant "ASPIRE" within the Sixth European Community Framework Program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gäel Richard.

The authors are with the Computer Science Department, University of Crete, Heraklion, Crete, Greece, GR-71409, and the Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH-ICS), Heraklion, Crete, Greece, GR-70013 (e-mail: tzagarak@ics.forth.gr; mouchtar@ics.forth.gr; tsakalid@ics.forth.gr).

Digital Object Identifier 10.1109/TASL.2009.2021716

stereo decoders, an alternative implementation of MPEG Surround suggests using a two-channel reference signal instead of using a monophonic downmix. Other recent improvements in multichannel audio coding include the improvement of AAC using Spectral Band Replication (SBR) [13], namely HE-AAC (High-Efficiency AAC) and HE-AAC v2 (version 2) which additionally incorporates the concepts from Parametric Stereo. In a recent report of the European Broadcasting Union (EBU) [14], several multichannel audio coding methods including Dolby AC-3, MPEG AAC, HE-AAC v2, MPEG Surround among others were tested using the MUSHRA subjective testing methodology. In this test, HE-AAC proved to be the most efficient, producing "Excellent" quality in the MUSHRA scale for 5.1 content at 160 kb/s on average for the various audio recordings used (however for some recordings the result was lower). MPEG Surround was found to be more useful for cases of low-bitrate applications (when nontransparent quality is acceptable).

In this paper, the focus is on immersive audio and more specifically on its application in music reproduction (especially live concert hall performances). In this case, immersive audio is largely based on enhanced audio content, which translates into using a large number of microphones (spot microphone signals) for obtaining a recording, containing as many sound sources as possible. These sources offer greater spatial fidelity to 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 methodology proposed in this paper. In this paper, the audio signals that are captured by the various microphones in a venue are encoded before they are mixed into the final multichannel mix. This is expected to offer a flexible reproduction of these signals at the decoder, which translates into allowing space for interactive applications at the client side. Examples of such interactive applications include virtual presence in a concert at a particular venue in real-time, remote collaboration of geographically distributed musicians, and so forth.

The process of mixing the multiple audio signals at the decoder can be termed remote mixing. 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. It is noted 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" (meta-data) that have been created by experts. In fact, recently proposed extensions on multichannel audio coding methods are becoming increasingly focused on the concept of interactivity in audio reproduction. The research in BCC [9], [10] includes the notion of "flexible rendering" for this purpose. The current MPEG work on Spatial Audio Object Coding (SAOC) [12] extends these concepts by spatially coding audio "objects," corresponding in many practical cases to the spot signals examined here. Of relevance is also the work on Spatial Audio Scene Coding (SASC) [15], which extends SAC in the sense that it is based on extracting the spatial cues of the audio channels for reproducing the perceived audio scene rather than the original channel configuration, and for this reason it is claimed to be more "universal" than SAC. The method proposed in this paper is also focused on offering flexible (i.e., interactive) audio rendering. However, in contrast to the aforementioned methods, our approach encodes the *actual content* of the spot audio signals, 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-based approaches. On the other hand, the proposed approach is more demanding in bitrates and less scalable than SAC-based approaches.

It is of interest to mention the work of [16], where a recorded scene is split into foreground and background components. Since the main interest in that work is to reconstruct a virtual 3-D audio scene based on field recordings it is considered as complementary to the work proposed here. Of relevance is also the work of [17], where the interest is to modify the source signals' spatial image when access to the original channels or field recordings is not possible. In that case, the authors propose using independent component analysis (ICA) in order to analyze the various audio sources, and the spatial image of these sources is subsequently obtained and modified. From the analysis of this and the previous paragraph it becomes clear that providing interactivity and an immersive audio experience to the listener is becoming indeed a central goal in current multichannel audio research.

In this paper, the Sinusoids plus Noise Model (henceforth referred to as SNM for brevity)—also known as deterministic plus stochastic decomposition—which has been used extensively for monophonic audio signals, is introduced in the context of low-bitrate coding for immersive audio applications. The proposed approach is to encode one audio signal only (which can be one of the spot signals or a downmix), while for the remaining spot signals only the parameters required for resynthesis of the content at the decoder are retained. These parameters are the sinusoidal parameters of each spot signal, as well as the short-time spectral envelope (estimated using Linear Predictive—LP—analysis) of the noise component of each spot signal. It is noted that the noise component is essentially the modeling error of the SNM. In contrast to these model parameters which can be encoded using a small amount of information, the true noise part of the SNM model is quite demanding with respect to coding rates. For this reason, the noise part of only the reference signal is retained. For resynthesis, each spot signal is reconstructed by adding its sinusoidal part to an estimated noise part. In turn, this noise part is synthesized by filtering the residual signal obtained from the reference channel with the time-varying noise envelope of each particular spot signal. This procedure, described in our recent work as *noise transplantation* [18], is based on the observation that the noise component of the spot signals of the same multichannel recording are very similar when the sinusoidal part has been captured by an appropriate number of sinusoids. For multiple spot signals without any similarities in their content, it is shown in this paper that the same transplantation procedure is valid if the noise component is obtained from a reference signal which is a downmix of the multiple audio recordings. The modeling and coding stages are described, and the bitrates that

the proposed system can achieve while retaining audio quality above 4.0 perceptual grade are experimentally found. Since in this paper there are two noise quantities described, the noise part of the SNM, as well as its whitened residual, at this point we clarify the terminology used in the remainder of this paper. Specifically, the following components are used:

- The *sinusoidal part* of the audio signal, which is given as a summation of a small number of sinusoids per signal segment.
- The *noise part* of the audio signal, which is the difference between the original audio signal and its sinusoidal part, and is modeled here as an autoregressive (AR) process.
- The *residual part* of the audio signal, which is the remainder of the noise part after its short-time spectral envelope has been extracted, or equivalently, the whitened version of the noise part.

Based on this description, the parameters of the model that are retained are the sinusoidal parameters and the spectral envelopes of the noise part. The residual part, which is costly to encode due to its noise-like nature, is not retained. In the decoder, an estimate of this signal is obtained by the reference signal used in our model, which is one of the spot signals or their dowmix, depending on whether the spot signals contain common information or not, respectively.

To our knowledge, this is the first attempt to tailor and apply the sinusoidal model to high-quality multichannel and immersive audio applications. In principle, the proposed method attempts to model each microphone signal with respect to a reference audio signal, so in this sense it follows the SAC philosophy. However, as explained, the proposed method offers more flexibility regarding interactivity compared to SAC-based approaches, which is essential in many immersive applications. It is pertinent to refer the reader to the work in [19], where the sinusoidal model was also applied in the context of multichannel audio coding. In that work, the sinusoidal parameters of the various channels of a multichannel recording are estimated simultaneously using a multichannel matching pursuit. As an application, the author proposed using this model as a front-end for SAC, where the spatial parameters are estimated based on the sinusoidal model, and the reference signal can be obtained by a summation of the sinusoidal parameters of the various channels. However, no treatment for the noise part is provided, and there is no indication of the resulting audio quality of the method.

It is noted that the approach proposed in this paper is based on our previous work in [20]. There, a multiresolution source/filter model was applied for coding of the spot recordings and resynthesizing them from a reference signal. In that work, a significant issue was the leakage, from the reference channel to the resynthesized recordings, of information that was not initially part of the recorded spot signals. This issue is usually described as crosstalk. In this current work, the objective has been to alleviate this issue by the use of the sinusoidal model, additionally to the source/filter model. It is shown here that indeed the sinusoidal model can achieve this objective, at the expense of higher bitrates for coding.

The remainder of this paper is organized as follows. In Section II, a brief overview is given of how the multichannel recordings are obtained in concert hall performances, which are the main concern in this paper. In Section III, a description is provided of the sinusoids plus noise model, which is the basis of our modeling method. In Section IV, the noise transplantation procedure is described, which is the central idea of this paper. Section V focuses on the coding of the model parameters, and is divided in two subsections: Section V-A, which describes the background theory for the quantization of the sinusoidal parameters, and Section V-B which explains how the noise spectral envelopes can be encoded. In Section VI, subjective results are provided for both the modeling and coding parts of the proposed algorithm. Emphasis is given on the modeling rather than the coding side, since the coded parameters include sinusoidal and spectral envelope parameters, which have been considered extensively in the past. The paper closes with concluding remarks in Section VII.

II. Microphone Signals of a Multichannel Recording

In this section, we briefly describe the procedure of creating a multichannel audio recording using multiple microphones placed in a recording venue, such as a concert hall. In this paper, we mainly focus on live concert hall performances. A number of microphones is used to capture several characteristics of the recording 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 of reference microphone signals (or even only one, which could be a sum signal). The result would be a significant reduction in transmission requirements, while enabling interactivity at the receiver. For achieving high quality resynthesis, we propose the use of some additional information for each microphone with the constraint that this additional information requires minimal data rates 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 depend primarily 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 time and frequency. Reverberant microphones are the microphones placed far from the sound source, which mainly capture the reverberation information of the venue. Here, we focus on the recordings made by spot microphones. Modeling of reverberant microphones has been considered in our earlier work, where linear time-invariant filters were proposed for transforming a reference signal into a given reverberant signal [21].

III. Sinusoids Plus Noise Model

The sinusoidal model was initially applied towards the analysis/synthesis of speech [22]. Under this model, a signal $s(n)$

is represented as the sum of a small number of sinusoids with time-varying amplitudes and frequencies. This can be written as

$$s(n) = \sum_{\ell=1}^{L} \alpha_\ell(n) \cos(\theta_\ell(n)) \quad (1)$$

where $\alpha_\ell(n)$ and $\theta_\ell(n)$ is the instantaneous amplitude and phase of the ℓth sinusoid, respectively. To estimate the parameters of the model, one needs to segment the signal into a number of short-time frames and compute a short-time frequency representation for each frame. Subsequently, the prominent spectral peaks are identified using a peak detection algorithm (possibly enhanced by perceptual-based criteria, e.g., [23]). Interpolation methods and tracking of sinusoids, such as [24] and [25], can be used to increase the accuracy of the parameter estimation. Each peak at the qth frame is represented as a triad of the form $\{\alpha_\ell^q, \omega_\ell^q, \phi_\ell^q\}$ (amplitude, frequency, phase), corresponding to the ℓth sinewave. In our implementation we apply a peak continuation algorithm in order to assign each peak to a frequency trajectory by matching the peaks of the previous frame to the current frame, using linear amplitude interpolation and cubic phase interpolation. However we mention that Overlap-Add (OLA) approaches in sinusoidal analysis/synthesis, such as [26]–[30], can also be employed within the context of the proposed research. The latter methods are attractive in practice since peak continuation is not required.

The sinusoidal model of (1) is not appropriate for a high-quality manipulation of audio signals, because these signals contain stochastic (i.e., unstructured, noise-like) components as well. A more accurate representation of audio signals is achieved when a stochastic component is included in the model. Although such a model is described under various names in the literature, it is described generally here as the *sinusoids plus noise model*. We briefly refer to this model as SNM. It is noted that SNM does not correspond to a particular implementation but rather to the general concept as described next.

In SNM, the signal representation is obtained by including in the model a noise part $e(n)$ as well, i.e., for each short-time frame the signal can be represented as

$$s(n) = \sum_{\ell=1}^{L} \alpha_\ell(n) \cos(\theta_\ell(n)) + e(n). \quad (2)$$

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

Several variations of the sinusoids 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 part), and 3) representing signal transients. Problem 1) has been extensively treated for speech signals, e.g., [22], [31], and variations of these approaches have been extended to wideband audio. In music, a sinusoids plus noise model was first proposed in [24]. For addressing problem 3), use of damped sinusoids and AM modulated sinusoids (instead of constant amplitude sinusoids) has been proposed (e.g., [32], [33]). A multiresolution analysis method [34] has also been proposed for better estimating the sinusoidal parameters by passing the signal through an octave-spaced filterbank, combined with transform coding of transient components. The noise is modeled based on a Bark-band noise model, as in [35].

In this paper, the focus is on the problem of noise representation, concentrating on the multiple signal (multichannel) case. In the first SNM derivation for audio signals [24], 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). Popular methods for modeling the noise part have been described in [35] and [36]. In the former approach, the spectrum of the noise signal is divided into critical bands and the spectral envelope is estimated by retaining the energy in each band. Then, the piecewise constant envelope is added to the sinusoidal part, in the frequency domain, where the phase spectrum of the noise part is estimated using a uniform random phase. The approximated signal is finally computed by taking the inverse Fourier transform of the aforementioned spectral sum. In the latter approach, the envelope of the noise part is estimated based on a perceptually motivated LPC computation, based on the masking threshold [37]. The noise part is reconstructed by filtering white noise by the estimated LPC filter.

Based on the above description, single-channel SNM methods focus on modeling the noise part 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 worse than the quality of the original audio signal: subjective results with average grades below 4.0 in a 5-grade scale have been reported for the aforementioned methods [36]. This is also true for the HILN method of [38] (Harmonic and Individual Lines plus Noise) which is part of MPEG-4, and models the noise part by retaining only the LPC envelope and filtering white noise for resynthesis.

In this paper, we are interested in low-bitrate high-quality audio modeling, which we show is feasible when more than one signals are to be encoded simultaneously. In this paper, the SNM is applied for simultaneously modeling more than one audio signals, i.e., in a multichannel manner.

In our case, it is desirable to achieve a high-quality result, i.e., grades above 4.0 in a 5-grade perceptual scale. We focus on the sinusoids plus noise model of [24]. Although more recent models offer improved modeling using fewer sinusoidal components [23], our objective is to provide a proof of concept for the noise transplantation procedure that is described in Section IV. In other words, we are interested to show that indeed our approach results in good audio quality compared both to the sinusoids-only model and to the original recording as well. This is regardless of the specific implementation of the SNM method that is used, as long as the spectral envelope of the noise is retained in some form. In this paper, we model the noise part $e(n)$ of the sinusoidal model as the result of filtering a residual part with an AR filter that models the noise part spectral envelope. The choice of AR filters, as opposed to using general ARMA filters, is due to the efficiency of linear predictive (LP) analysis

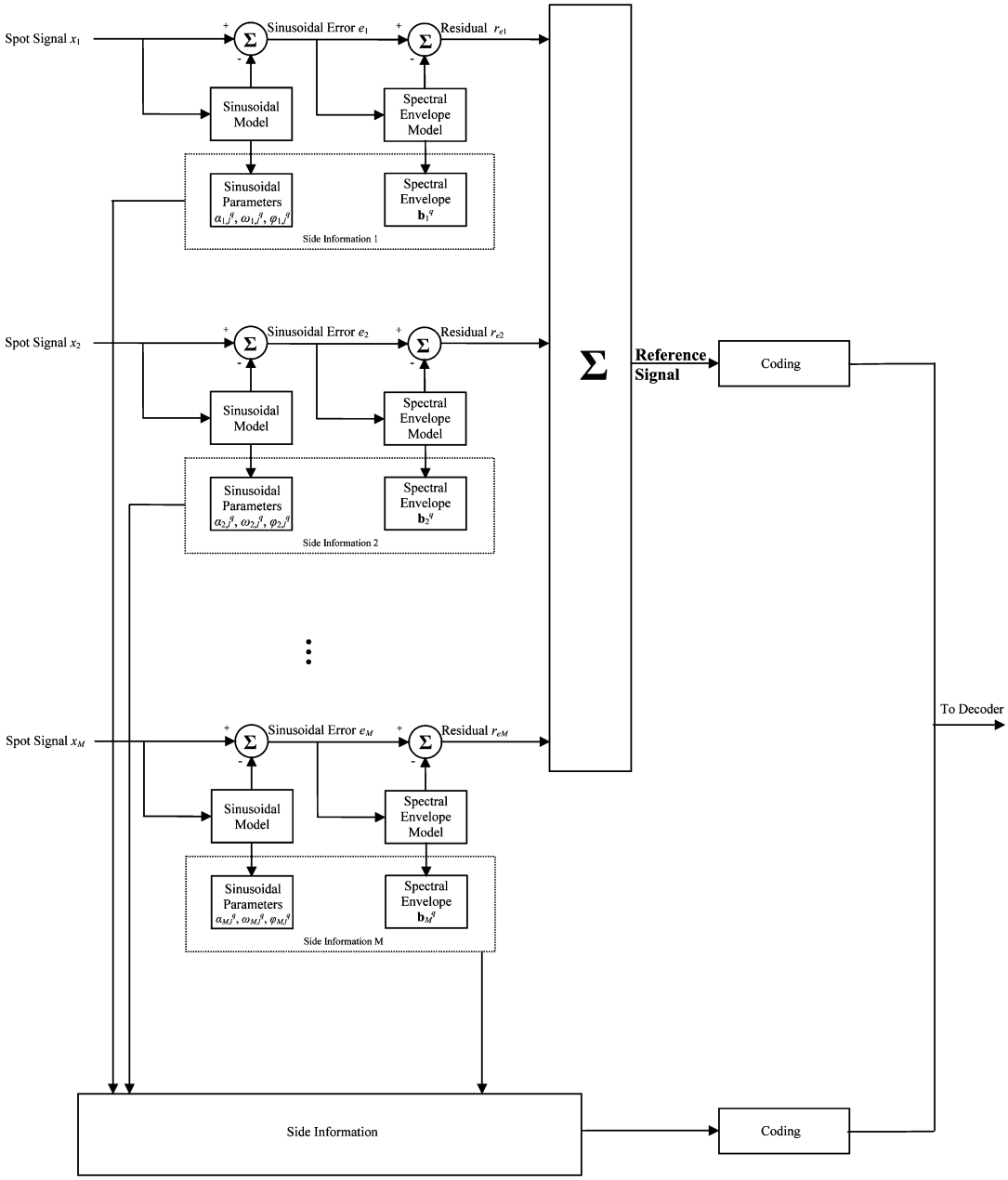

Fig. 1. Diagram of the proposed analysis approach. The per-frame sinusoidal and noise spectral envelope parameters are retained as side information and coded. The reference signal is created as a downmix of the residuals of the various spot signals, and coded using a monophonic audio coder. Alternatively, if the various spot signals contain a high degree of common information, the reference signal can be one of the spot signals.

for estimating the AR filter coefficients. Thus, we assume the following equation for the noise part of the sinusoidal model

$$e(n) = \sum_{p=1}^{P} b(p)e(n-p) + r_e(n). \tag{3}$$

The quantity $e(n)$ is the noise part, while $r_e(n)$ is the residual part and P is the AR filter order. The $P+1$th-dimensional vector $\mathbf{b}^T = [1, -b_1, -b_2, \ldots, -b_P]$ represents the coefficients of the LP filter, directly related to the spectral envelope of the noise part $e(n)$. In the frequency domain, (3) becomes

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

where $S_e(e^{j\omega})$ and $S_{r_e}(e^{j\omega})$ are the power spectra of $e(n)$ and $r_e(n)$, respectively, while

$$B(e^{j\omega}) = 1 - \sum_{p=1}^{P} b(p)e^{-j\omega p} \tag{5}$$

is the frequency response of the LP filter $\mathbf{b}$. As explained in Section I, in this paper there are two noise quantities introduced, i.e., the noise part $e(n)$ and its whitened version $r_e(n)$ which is the residual part.

IV. Noise Transplantation

In this section, we describe the main novelty of our proposed approach, namely noise transplantation. Consider a collection

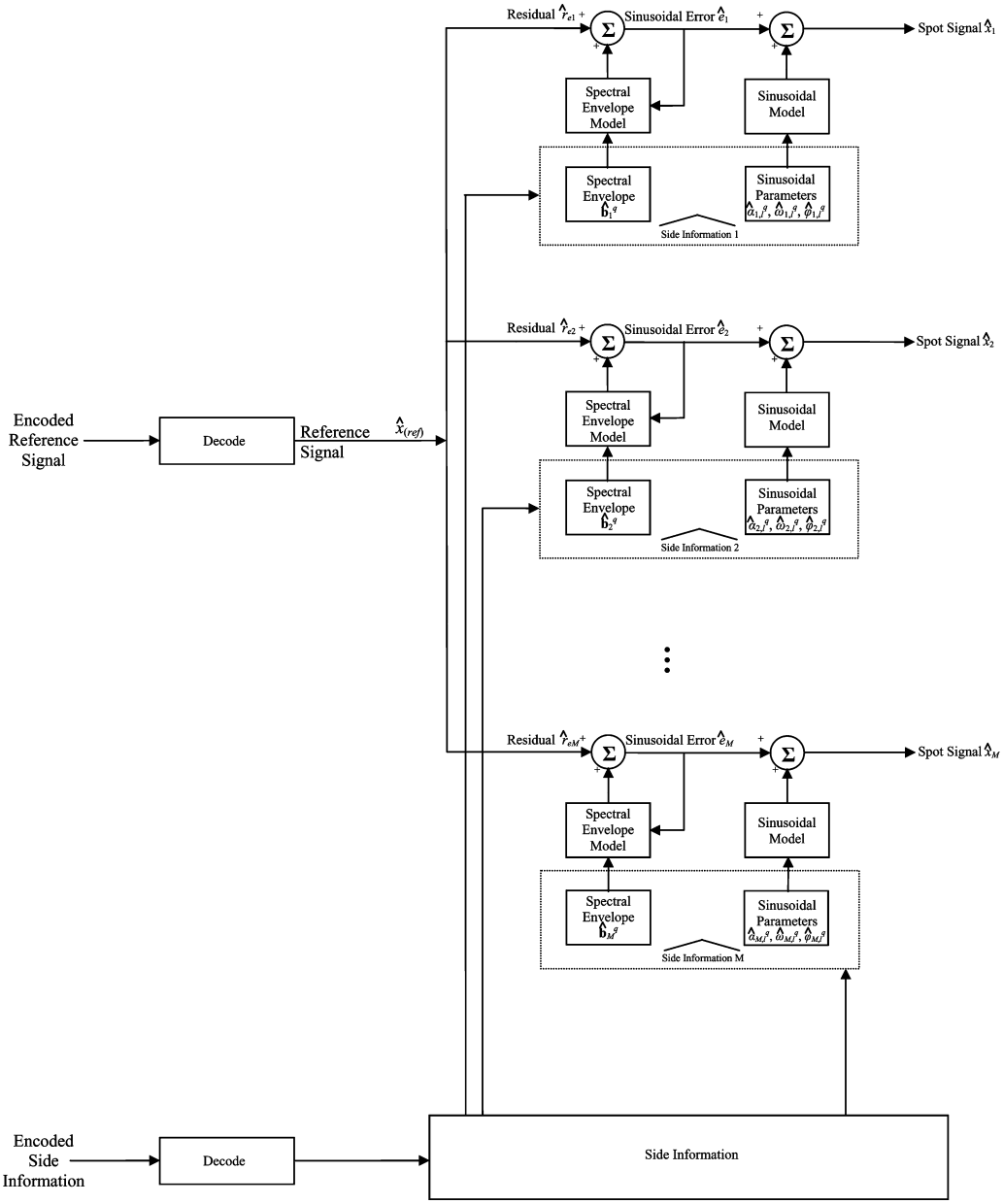

Fig. 2. Diagram of the proposed resynthesis approach. The side information parameters are decoded back into the original sinusoidal and spectral envelope parameters, with some quantization error. The reference signal is first decoded, and then filtered by the spectral envelope of each spot signal to recreate the sinusoidal error signal. This is then added to the corresponding sinusoidal parameters to resynthesize the estimate of the original spot signal.

of M spot microphone signals that need not necessarily have similar acoustical content. The proposed modeling approach is divided in two parts, the analysis part, which is followed by the coding of the model parameters at the transmitter, and the resynthesis part, which follows the decoding of the model parameters at the receiver. In this section, the analysis/synthesis model is described. In Section V, the method of encoding/decoding the model parameters is described.

A. Analysis Model

We encode as one audio channel the reference signal. At this point, we clarify the fact that the residual can be obtained using one of the following alternative methods, depending on the application.

- *Case 1*: The reference signal can be a downmix of the residual parts of the various spot signals. This is the case examined in the remainder of this section, and is the case applied for obtaining the results in Figs. 4–6, in Section VI.
- *Case 2*: The reference signal can be a downmix of the various spot signals.
- *Case 3*: The reference signal can be one of the spot signals (or its residual). This is the case applied for obtaining the results in Figs. 3 and 7, in Section VI.

In all cases, the reference signal is used for obtaining the residual part of the spot signals in the decoder. Thus, the advantage in Case 1 is that this residual is equal to the reference signal, while in Cases 2 and 3 the residual must be obtained

by LP analysis of the reference signal at the decoder, after the sinusoidal parts of all spot signals have been subtracted. One issue, though, with Case 1 is that the reference signal is proposed here to be encoded using a monophonic encoder, such as MP3. This is not optimal for the particular reference signal which is a sum of residual signals (instead of recorded audio signals) and thus could be more efficiently coded by exploiting this fact. This issue was not further examined in this paper. In this sense, the proposed coding approach for the reference signal is more suitable for Case 2. Also, another advantage for Case 2 is that, instead of encoding a monophonic downmix of the spot signals, it could be possible to encode a two-channel dowmix so as to allow for backward compatibility with stereophonic decoders, as in some SAC schemes. Finally, Case 3 is advantageous for concert hall recordings, where usually two of the multiple spot signals are used as the stereo channels for stereophonic reproduction. In that case, one or both of these signals could be used for obtaining the residuals at the decoder, so that stereo compatibility is achieved by using as reference the *actual* stereophonic channels.

The spot signals are modeled by the SNM, as explained in the previous section, retaining their sinusoidal components, and the noise spectral envelope (filter $\mathbf{b}$ in (3)). This procedure is the modeling part of the proposed algorithm, which is then followed by the encoding of the reference signal and the side information (the sinusoidal parameters and the noise spectral envelopes).

The analysis procedure is depicted in Fig. 1. Denoting the M originally recorded spot signals as $x_k(n)$, $k = 1, \ldots, M$, and following the notation of the previous section, the analysis procedure can be summarized as follows.

Step 1) Obtain the qth frame sinusoidal parameters of spot signal $x_k(n)$, denoted as $\{\alpha^q_{k,\ell}, \omega^q_{k,\ell}, \phi^q_{k,\ell}\}$, $\ell = 1, \ldots, L$, where L denotes the user-defined number of sinusoids per frame.

Step 2) Obtain the noise part of spot signal $x_k(n)$ denoted as $e_k(n)$, which is given by the following relation for the qth frame

$$e^q_k(n) = x^q_k(n) - \sum_{\ell=1}^{L} \alpha^q_{k,\ell}(n) \cos\left(\omega^q_{k,\ell} n + \phi^q_{k,\ell}\right). \quad (6)$$

Step 3) Obtain the qth frame LPC vector of the noise part $e^q_k(n)$, denoted as $\mathbf{b}^q_k$.

Step 4) Obtain the qth frame residual part $e^q_k(n)$, denoted as $r^q_{e_k}(n)$, using the relation

$$r^q_{e_k}(n) = e^q_k(n) - \sum_{p=1}^{P} b^q_k(p) e^q_k(n-p). \quad (7)$$

The above relation can be written in the frequency domain (power spectral densities) as

$$S_{r^q_{e_k}}(e^{j\omega}) = \left|1 - \sum_{p=1}^{P} b^q_k(p) e^{-j\omega p}\right|^2 S_{e^q_k}(e^{j\omega}). \quad (8)$$

The per frame energy of the residual part must also be retained, as will be explained in the description of the resynthesis procedure. This coefficient is denoted as $c_{r^q_{e_k}} = \sum_n |r^q_{e_k}(n)|^2$. This coefficient is scalar quantized in this work, using 16-bit uniform resolution, since the required bitrate for transmission even without any further coding is negligible.

Step 5) Obtain the reference signal $x_{(\mathrm{ref})}(n)$ as a downmix of the M residual signals $r_{e_k}(n), k = 1, \cdots, M$, i.e.,

$$x_{(\mathrm{ref})}(n) = \sum_{k=1}^{M} r_{e_k}(n). \quad (9)$$

The downmix signal can be obtained in a framewise manner, using an overlap-add process. Different weights for each residual signal can be used when creating the downmix. As explained, the original spot signals can be alternatively used for creating the reference signal. In that case, the residual part of the reference signal must be obtained in the decoder, by estimating its sinusoidal model. In cases when significant leakage exists among the spot signals (e.g., when these are obtained from a concert hall recording), one of the spot signals can be used as a reference signal.

The model parameters, i.e., the per-frame sinusoidal parameters $\{\alpha^q_{k,\ell}, \omega^q_{k,\ell}, \phi^q_{k,\ell}\}$ and LPC vectors $\mathbf{b}^q_k$, are encoded using the method described in Section V and transmitted to the receiver. The reference signal is also encoded using a monophonic audio encoder such as MP3, and transmitted.

B. Resynthesis Model

In order to reconstruct the spot signals, the residual parts are needed. Under the proposed approach, the reference signal is used as the residual part in order to resynthesize all spot signals in the following manner. First this noise is filtered by each of the LP spectral envelopes (one for each spot signal). Then, the derived signal is added to the corresponding sinusoidal part in order to recreate the high-quality resynthesized spot signals. This procedure is the resynthesis phase of our method, i.e., after decoding the encoded reference signal and side information parameters.

The resynthesis procedure is depicted in Fig. 2. The decoded model parameters are first obtained as will be explained in Section V. Since the parameter encoding/decoding procedure will introduce quantization error, the model parameters at the decoder side are denoted as $\{\hat{\alpha}^q_{k,\ell}, \hat{\omega}^q_{k,\ell}, \hat{\phi}^q_{k,\ell}\}$ and vector $\hat{\mathbf{b}}^q_k$. Similarly, the reference signal at the decoder is denoted as $\hat{x}_{(\mathrm{ref})}$. The resynthesis process can be summarized as follows.

Step 1) Obtain the qth frame estimate of the residual part $e^q_k(n)$, denoted as $\hat{r}^q_{e_k}(n)$, using the reference signal, i.e.,

$$\hat{r}^q_{e_k}(n) = \sqrt{\frac{c_{r^q_{e_k}}}{\sum_n \left|x^q_{(\mathrm{ref})}(n)\right|^2}}\, \hat{x}^q_{(\mathrm{ref})}(n). \quad (10)$$

Step 2) Obtain the estimate of the noise part of spot signal $x_k(n)$, denoted as $\hat{e}_k(n)$, which is given by the following AR relation for the qth frame

$$\hat{e}^q_k(n) = \sum_{p=1}^{P} \hat{b}^q_k(p) \hat{e}^q_k(n-p) + \hat{r}^q_{e_k}(n). \quad (11)$$

In the frequency domain (power spectral densities), the above relation can be written as follows:

$$S_{\hat{e}_k}(e^{j\omega}) = \left| \frac{1}{1 - \sum_{p=1}^{P} \hat{b}_k(p) e^{-j\omega p}} \right|^2 S_{\hat{r}_{e_k}}(e^{j\omega}). \quad (12)$$

It is noted that (12) produces an approximation of the noise part since the reference signal is used instead of the residual part of the corresponding spot signal. The exact relation that produces the noise part from its corresponding residual part based on the LPC model is in fact (4), which can be written for spot signal $x_k(n)$ as

$$S_{e_k}(e^{j\omega}) = \left| \frac{1}{1 - \sum_{p=1}^{P} b_k(p) e^{-j\omega p}} \right|^2 S_{r_{e_k}}(e^{j\omega}). \quad (13)$$

Since for each frame the reference signal $x_{(\mathrm{ref})}$ is used instead of the actual residual signal r_{e_k}, the energy normalization of (10) is needed. Consequently, the per-frame energy of the correct residual signal r_{e_k} for each spot signal must be retained, along with the corresponding LPC vector, as mentioned in the analysis model description. The audio improvement of this normalization was perceptible in preliminary listening tests we conducted.

Step 3) Obtain the qth frame estimate of spot signal $x_k(n)$, denoted as $\hat{x}_k(n)$, using the estimated noise part from Step 2 and the decoded sinusoidal parameters

$$\hat{x}_k^q(n) = \sum_{\ell=1}^{L} \hat{\alpha}_{k,\ell}^q(n) \cos\left(\hat{\omega}_{k,\ell}^q n + \hat{\phi}_{k,\ell}^q\right) + \hat{e}_k^q(n). \quad (14)$$

The final spot signal estimate $\hat{x}_k(n)$ can be obtained in a framewise manner using overlap-add. Clearly, the objective of the proposed approach is to avoid encoding the residual part of each of the spot signals. This is important, as the residual part (as well as the noise part) is in general of highly stochastic nature, and cannot be adequately represented using a small number of parameters. Consequently, this part of each spot signal is costly to be fully coded (e.g., using MP3) for low bitrate applications. We note that modeling such signals with parametric models results in low-quality audio resynthesis. It is shown, later in this paper, that the noise transplantation method can result in significantly better quality audio modeling compared to parametric models for the residual signals. This is true even when using as low as ten sinusoids, which is very important for low bitrate coding, since less parameters must be encoded.

V. Coding of Spot Signals

The model parameters to be encoded at the transmitter consist of the sinusoidal parameters and the LP parameters per spot microphone signal. Coding of such parameters has been considered extensively in the literature. Thus, in this section two representative methods (one for the sinusoidal and one for the LP parameters) are briefly described. The performance of these methods in the particular problem examined in this paper is evaluated in Section VI-C.

A. Coding of the Sinusoidal Parameters

We adopt the coding scheme of [39], developed for jointly optimal quantization of sinusoidal frequencies, amplitudes and phases. The sinusoidal parameters are quantized in polar form, assuming a dependence of the frequency quantization on the amplitude, and a dependence of the phase quantization on the amplitude and the frequency. This scheme is called Unrestricted Polar Quantization (UPQ) and represents a combination of three scalar quantizers, based on high-rate quantization. High-rate quantizers are formulated in terms of quantization point density functions, which are defined as the inverse of the quantizer step size Δ_i, where i denotes the ith cell. The quantization scheme is entropy constrained, and each reconstruction level is transmitted with variable codeword length.

In order to derive the quantizers, the goal is to minimize, in a frame-by-frame basis, the average weighted mean squared error (WMSE) for L sinusoids

$$D = \frac{1}{L} \sum_{\ell=1}^{L} w_\ell D_\ell \quad (15)$$

under the entropy constraint

$$H = \frac{1}{L} \sum_{\ell=1}^{L} (H(I_{\alpha_\ell}) + H(I_{\omega_\ell}|I_{\alpha_\ell}) + H(I_{\phi_\ell}|I_{\alpha_\ell})). \quad (16)$$

The given total entropy per sinusoid (amplitude, frequency, and phase) is denoted by H. The entropies $H(I_{\alpha_\ell}), H(I_{\omega_\ell}|I_{\alpha_\ell})$ and $H(I_{\phi_\ell}|I_{\alpha_\ell})$ express the entropies of the individual quantization parameters. The mean squared error (MSE) D_ℓ introduced by the quantization of the ℓth sinusoid is assigned a perceptual weight w_ℓ, which is defined as $w_\ell = 1/m_\ell, \ell = 1, \ldots L$, where m_ℓ is the masking threshold at the frequency of the corresponding sinusoid [40]. The MSE D_l over a frame of length N, can be expressed as

$$D_\ell = E\left\{ \frac{1}{N} \sum_{n=-(N-1)/2}^{(N-1)/2} (\alpha_\ell \cos(\omega_\ell n + \phi_\ell) - \hat{\alpha}_\ell \cos(\hat{\omega}_\ell n + \hat{\phi}_\ell))^2 \right\} \quad (17)$$

where $\{\alpha_\ell, \omega_\ell, \phi_\ell\}$ and $\{\hat{\alpha}_\ell, \hat{\omega}_\ell, \hat{\phi}_\ell\}$ are the non-quantized and quantized sinusoidal parameters, respectively, and $E\{\cdot\}$ denotes the expectation operation. Thus, the optimization problem is to minimize the WMSE in (15) under the constraint expressed in (16). This constrained minimization problem can be solved using the method of Lagrange multipliers. The evaluation of the Euler–Lagrange equations with respect to the point densities $g_A(\alpha), g_\Omega(\omega)$ and $g_\Phi(\phi)$ (corresponding to amplitude, frequency, and phase, respectively) give the optimum quantization point densities [39]

$$g_A(\alpha) = g_A = \frac{w_\alpha^{\frac{1}{6}} 2^{\frac{1}{3}\tilde{H} - \frac{2}{3}b(A)}}{w_g^{\frac{1}{6}} \left(\frac{N^2}{12}\right)^{\frac{1}{6}}} \quad (18)$$

$$g_\Omega(\omega,\alpha)=g_\Omega(\alpha)=\frac{\alpha w_\alpha^{\frac{1}{6}}\left(\frac{N^2}{12}\right)^{\frac{1}{3}}2^{\frac{1}{3}\tilde{H}-\frac{2}{3}b(A)}}{w_g^{\frac{1}{6}}} \quad (19)$$

$$g_\Phi(\phi,\alpha,w_\ell)=g_\Phi(\alpha,w_\ell)=\frac{\alpha w_\ell^{\frac{1}{2}}2^{\frac{1}{3}\tilde{H}-\frac{2}{3}b(A)}}{w_\alpha^{\frac{1}{3}}w_g^{\frac{1}{6}}\left(\frac{N^2}{12}\right)^{\frac{1}{6}}} \quad (20)$$

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.

B. 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 noise part. We follow the quantization scheme of [41]. The LP coefficients of each spot signal that model the noise spectral envelope are transformed to line spectral frequencies (LSFs) which are modeled by means of a Gaussian mixture model (GMM), defined as

$$g(\boldsymbol{x})=\sum_{i=1}^{C}p_iN(\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i). \quad (21)$$

In the equation above, $N(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the normal multivariate distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, p_i is the prior probability of Gaussian class i, and C is the number of classes. The covariance matrix of each class can be diagonalized using eigenvalue decomposition as

$$\boldsymbol{\Sigma}_i=\mathbf{Q}_i\boldsymbol{\Lambda}_i\mathbf{Q}_i^T \quad (22)$$

where $i=1,\ldots,C$. The matrix $\boldsymbol{\Lambda}_i$ is diagonal and contains the corresponding eigenvalues of $\boldsymbol{\Sigma}_i$, while $\mathbf{Q}_i$ is the matrix containing the corresponding set of orthogonal eigenvectors of $\boldsymbol{\Sigma}_i$, for the ith Gaussian class of the model. Then, the Karhunen–Loève Transform (KLT) substitutes each LSF vector for time segment k, $\boldsymbol{z}_k$, with another decorrelated vector $\boldsymbol{w}_k$, where $\boldsymbol{w}_k=\mathbf{Q}_i^T(\boldsymbol{z}_k-\boldsymbol{\mu}_i)$. Afterwards, the components of the vector $\boldsymbol{w}_k$ can be independently quantized by a nonuniform quantizer, i.e., through a compressor, a uniform quantizer and an expander. Each LSF vector is classified to only one of the C Gaussians, so that the above scheme can be applied. 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, and the vector is classified to the class associated with the minimal LSD, which is defined as

$$\mathrm{LSD}(i)=\left(\frac{1}{F_s}\int_0^{F_s}\left[10\log_{10}\left(\frac{S(f)}{\hat{S}^{(i)}(f)}\right)\right]^2df\right)^{\frac{1}{2}} \quad (23)$$

where F_s is the sampling rate, $S(f),\hat{S}^{(i)}(f)$ are, respectively, the LP power spectra corresponding to the original vector $\boldsymbol{z}_k$, and the quantized vector $\hat{\boldsymbol{z}}_k^{(i)}$, for each class $i=1,\ldots,C$. In the decoder side of the quantization procedure, the correlated version of the quantized vector is reconstructed by left multiplying of the reconstructed $\boldsymbol{w}_k$ with the matrix $\mathbf{Q}_i$. Finally, the class mean $\boldsymbol{\mu}_i$ is added to obtain the quantized value of $\boldsymbol{z}_k$, denoted as $\hat{\boldsymbol{z}}_k$.

VI. Performance Evaluation

In this section, we examine the modeling as well as the coding performance of our proposed system, with respect to the resulting audio quality. For this purpose, several listening tests were performed, both in monophonic as well as in stereophonic settings, evaluating first the modeling approach, and subsequently the coding of the model parameters.

For the results of this section, a 30-ms analysis/synthesis frame was used for the sinusoidal model with 50% overlap (with overlap-add synthesis) for all sinusoidal model implementations. For the noise modeling using our proposed approach, we used tenth-order LPC analysis, with a window of 23 ms and 75% overlap. The sampling rate was 44.1 kHz for the music waveforms and 22 kHz for the speech waveforms.

A. Modeling Performance Using Monophonic Test Signals

For the results of this section we used two monophonic microphone signals of a multichannel recording of a concert hall performance.[1] These signals are actual spot microphone signals. One of the microphones captures mainly the male voices of the orchestra's chorus and is used here as the spot signal, while the other one mainly captures the female voices and is used as the reference signal. This two-signal example can be easily extended to an arbitrary number of recordings. In this test, the objective is to resynthesize the male chorus signal using its sinusoidal parameters and noise part spectral envelopes, using the residual part of the female chorus. The two recordings used here were chosen based on the fact that they have been used in our previous experiments with other modeling methods [20].

Twelve listeners participated in the listening tests individually, under the same environmental conditions (i.e., a quiet office space), using high-quality headphones (Sennheiser HD-650). From the two concert hall recordings, we chose three different parts of the performance with 10-s 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 listening tests, which were performed following the ITU-R BS.1116 [42] recommendations (no anchor signals were used). The grades characterize the *quality* of the resynthesized signal in relation to the original recording, "5" corresponding to "not perceived" (difference in quality), "4" to "perceived but not annoying," "3" to "slightly annoying," "2" to "annoying," and "1" to "very annoying."

In Fig. 3, we plot the average subjective grades for each of the three test signals. Each of the two figures corresponds to a different choice of sinusoidal parameters per frame. The upper plot corresponds to 40 sinusoids and the lower plot to 10 sinusoids per frame. A graphical representation of the 95% confidence interval is indicated by the two horizontal lines above and below the mean value. In each of the two different plots of Fig. 3, the following results are depicted. The squares correspond to the sinusoidal model without retaining the noise envelope (denoted as "sin" in the figure). The stars correspond to our proposed

[1]Provided by Prof. C. Kyriakakis of the University of Southern California.

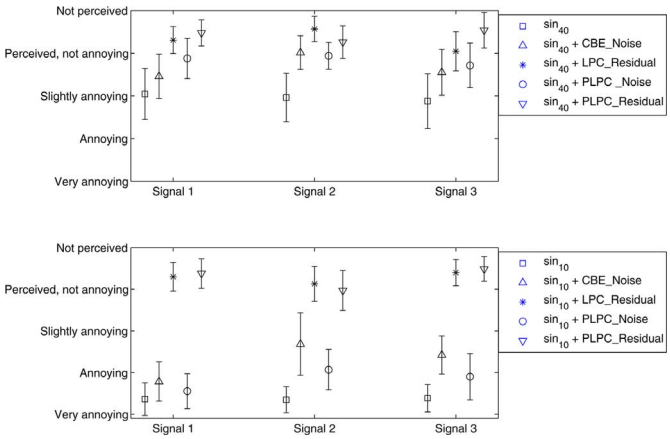

Fig. 3. Results from the quality rating monophonic listening tests corresponding to sinusoidal modeling with (a) 40 sinusoids per frame (middle), and (b) ten sinusoids per frame (lower). The proposed method ("sin + LPC_Residual") is tested against the sinusoids-only model ("sin"), the CBE noise model of [35] ("sin + CBE_Noise"), and the PLPC noise model of [36] ("sin + PLPC_Noise"). A version of our proposed method using the PLPC noise envelope estimation ("sin + PLPC_residual") instead of using the LPC method is also tested.

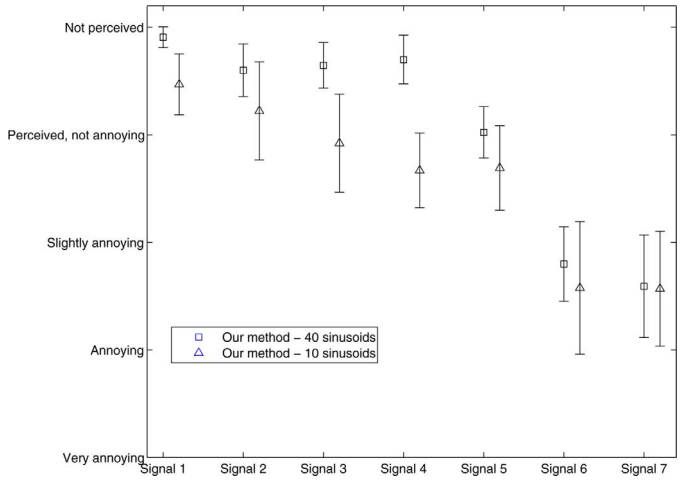

Fig. 4. Results from the monophonic quality rating listening tests for the downmix case, corresponding to sinusoidal modeling with (a) 40 sinusoids per frame (squares), and (b) ten sinusoids per frame (triangles).

model ("sin + LPC_Residual"). The triangles correspond to enhancing the sinusoidal model using the Critical Band Energy (CBE) method of [35], which models the noise part by retaining only its energy in each critical band ("sin + CBE_Noise"). The circles correspond to enhancing the sinusoidal model with the Perceptually-motivated LPC estimation (PLPC noise modeling method of [36], "sin + PLPC_Noise"). In the latter method, a perceptually estimated spectral envelope of the noise part is retained. It must be noted that the CBE and PLPC methods are single-channel noise modeling methods for the SNM model, and the residual part is obtained by a random number generator (uniformly random phase for CBE, white noise for PLPC). An alternative version of our noise transplantation method using PLPC estimation of the noise envelopes (instead of LPC) was implemented and tested in this test ("sin + PLPC_Residual," denoted in the figure using the inverted triangles). The novelty in our approach is the use of the residual part from the reference signal, which can be adapted for any other model where a residual part is needed (including PLPC), and is only possible in the joint modeling scenario that is examined here. For all five cases examined (sinusoidal model only, our approach, our approach using PLPC, CBE, and PLPC filtering white noise), the sinusoidal parameters are the same, for straightforward comparison of the results.

The results of Fig. 3 indicate that all noise modeling methods are superior in comparison to the model based on sinusoidal parameters only. Clearly, the noise part of the sinusoidal model must be treated to achieve high-quality resynthesis. Both CBE and PLPC approaches achieve an improvement over the sinusoids-only model. Their resulting audio quality, though, remains lower than our proposed method using LPC or PLPC envelope estimation. Especially for the ten sinusoids case, our noise transplantation method retains a grade around 4.0 (as in the case of 40 sinusoids), while the two other noise modeling methods achieve a grade below 3.0. This can be attributed to the fact that the PLPC and CBE methods treat the envelope of the noise part only, while our method provides a residual part as well, as explained, at the expense of higher bitrate (corresponding to encoding the reference signal). The fact that the audio quality in our method remains high even for a low number of sinusoids per frame, directly translates into a significant coding benefit, since less sinusoidal parameters must be encoded per frame. This is especially important given that in previously reported applications of the sinusoidal model (mainly in speech coding), 14–20 bits per sinusoid were needed for high-quality results (e.g., [39], [43], [44]). It also noted that the LPC and PLPC envelope estimation methods under our noise transplantation approach achieve very similar audio quality. Given that LPC is simpler to implement, this was the method used for envelope estimation in our approach for the remainder of Section VI.

We also examine here the resynthesis quality of various spot signals from a downmix signal. This is important in cases where spot signals do not contain similar audio content, which is often the case in studio recordings. We used seven audio signals for the test. Each reference sound file used contained a sum of the residual parts of two original recordings, and more specifically the following reference signals were created: 1) bass singer plus soprano singer, 2) electric guitar plus rock singer, 3) harpsichord plus violin, 4) female plus male speech, 5) trumpet plus violin, 6) violin plus guitar, and 7) violin plus harpsichord. These seven signals correspond one-to-one to Signals 1–7 in the subjective results depicted in Fig. 4 (again the ITU-R BS.1116 methodology with no anchors was followed). The recordings of the bass singer, soprano singer, harpsichord, trumpet, and violin are excerpts from the EBU SQAM (Sound Quality Assessment Material) test disc.[2] These are stereo recordings, and only one of the two channels was used in this experiment. The recordings of the electric guitar and the rock singer are spot recordings which are a courtesy of rock band "Orange Moon." The speech signals were obtained from the VOICES corpus,[3] available by OGI's CSLU [45]. The instrument that is referred first in the above list is the instrument (side signal) that we wanted to resynthesize

[2]http://sound.media.mit.edu/mpeg4/audio/sqam/

[3]http://www.cslu.ogi.edu/corpora/voices/

from the sum signal (reference signal). In this experiment 12 volunteers participated.

The results depicted in Fig. 4 include the average subjective results (along with the 95% confidence intervals) for all seven signals modeled by our method, using 40 sinusoids (squares) and ten sinusoids (triangles) per time frame. From the results of this monophonic downmix subjective test we can notice that in the 40 sinusoids case, Signals 1–5 achieve a grade above 4.0, while Signals 6–7 achieve a grade below 3.0 because the percussive sounds cannot be adequately modeled by the SNM, and significant information remains in the residual. In the 10–sinusoids case, the performance further deteriorates for Signals 3–4 as well.

It was apparent in these tests that the main source of degradation was due to leakage from the reference recording to the resynthesized spot signal. In other words, parts of the reference signal which were not originally present in the originally recorded spot signal (i.e., from the other spot signals), were included in the resynthesized spot signal. This fact is an undesired effect of the transplantation procedure, and can be termed as leakage or crosstalk effect. It is not possible in practice to avoid this crosstalk, given that the model parameters cannot capture all the microphone-specific information and completely "whiten" the residual part of the reference signal. At the same time, it was clear from our tests that, apart from this interference, the quality of the resynthesized spot signals was not severely affected. These observations are important since the proposed model is designed for applications when all modeled signals are rendered simultaneously, possibly after a mixing process at the decoder. Thus, more important than the perceived quality of the individual recordings is the perceived quality when these are rendered simultaneously. Given that the only degradation of the modeled signals is the leakage among the several recordings, this should appear in the stereophonic or multichannel setup as an image width distortion rather than a quality distortion.

The above observations led us to test the same audio files evaluated in this section under a stereophonic (two-channel) setting. In this test, the two spot signals that were used to create the reference signal were presented simultaneously to a listener using headphones. The results are described in the following section.

B. Modeling Performance Using Stereophonic Test Signals

In the previous section, listening tests were performed under a monophonic setting approach. In other words, the proposed model was used in order to derive one microphone recording (monophonic signal) from the reference signal, and this was presented separately to each listener using headphones. In this section, we are interested in testing the assumptions of the previous paragraph, i.e., that the leakage introduced by our model does not affect the audio quality (if the spot signals are rendered simultaneously), and that the amount of this leakage is small and does not severely affect the image width of the (mixed or unmixed) spot signals. For this reason, the following two listening tests were designed. The first test was designed in order to test the quality of two modeled signals when rendered simultaneously, excluding the image width distortion. The second test was designed to test only the image width distortion, excluding quality distortion. Both tests were performed following the ITU-R BS.1116 [42] recommendations. In both tests 12 volunteers participated, who were trained in the beginning of the session so that they could distinguish among the types of distortion examined (using the same headphones as in the previous tests). Separate monophonic spot recordings were modeled by the proposed approach for deriving the sound files used in the listening tests. In other words, from each spot signal its sinusoidal part (using ten sinusoids per frame as explained later) and the spectral envelope of its noise part were retained. Also, a reference signal was created as a downmix of the spot signals. This procedure was employed due to the fact that under the proposed scheme the actual stereophonic or multichannel recordings are mixed *after* decoding. The proposed algorithm is designed so as to recreate the content of each spot signal separately, and is not designed to specifically retain the relative amplitude and time differences between the audio channels such as in SAC for example. The relative amplitude and time differences are closely related with the spatial image of a multichannel recording, consequently in the proposed method the spatial image of an already mixed recording may be distorted. It is for this reason that the our algorithm is proposed to be used for monophonic spot signals and not mixed multichannel recordings, and the mixing process in our method to be created after decoding.

The following monophonic recordings were used, each containing a separate instrument recording (the duration of each audio clip was around 10 s): i) bass singer, ii) soprano, iii) trumpet, iv) harpsichord, v) violin, vi) rock singer, vii) rock guitar, viii) male speech, ix) female speech, x) male chorus, xi) female chorus. These signals are the same used in Section VI-A. More specifically, signals i)–ix) are the signals used in the downmix test (for creating the seven downmix signals), while signals x)–xi) are the concert hall performance spot signals, used in the test of Fig. 3.

Using these recordings, stereophonic signals were created by mixing two monophonic signals at a time, with a relative level difference of ± 14 dB for the left and right channel (amplitude panning). More specifically the following signals were created: 1) bass plus soprano, 2) guitar plus rock singer, 3) harpsichord plus violin, 4) female plus male speech, 5) trumpet plus violin, 6) violin plus guitar, 7)- 9) male plus female chorus (three different parts of the recording). The reader can view these nine signals as the stereophonic equivalent of the signals of Section VI-A. More specifically, Signals 1)–6) are similar to those used for the results of Fig. 4 (violin plus harpsichord and harpsichord plus violin is the same signal for the stereophonic case examined here). Similarly, Signals 7)–9) correspond to the three signals used for the results in Fig. 3.

Most of the sound files that were used in the listening tests can be found in our web site.[4] The sound examples that can be found there correspond to the above cases 1)–7), and more specifically: i) the original monophonic recordings; ii) their sinusoids-only representation using ten sinusoids per frame; iii) the improved sinusoidal representation using our proposed approach; iv) the synthesized stereophonic recording using the original monophonic pairs (with $\pm$14-dB panning); and v) the synthesized

[4]http://www.ics.forth.gr/~mouchtar/snm/

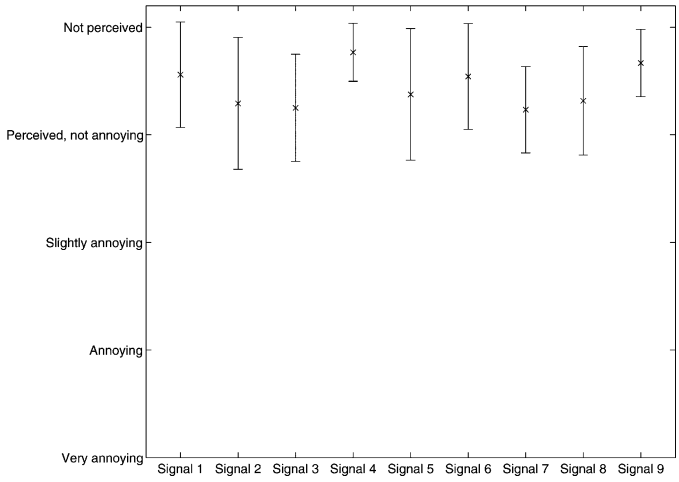

Fig. 5. Results from the stereophonic quality rating listening test. Only quality was rated, and listeners were asked to ignore image width distortion.

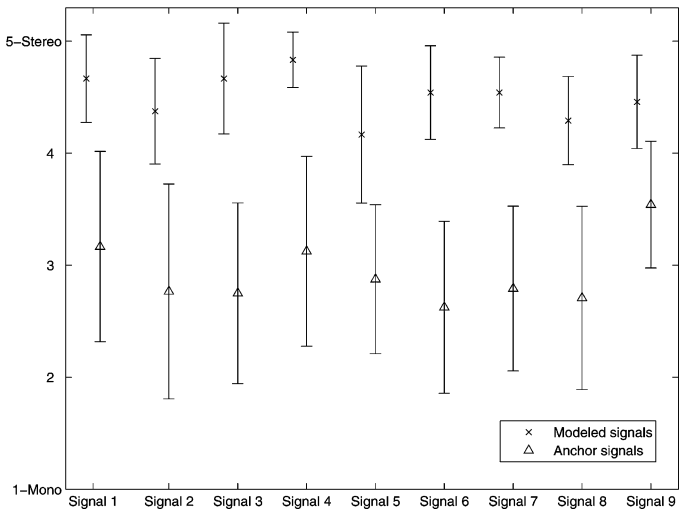

Fig. 6. Results from the image width rating stereophonic listening tests (ignoring quality distortion).

stereophonic recording using the modeled pairs with our transplantation approach (with $\pm$14-dB panning). Clearly, the sound files of case i) are mixed to synthesize the files of case iv), and similarly, the sound files of case iii) are mixed to synthesize the files of case v). Also, the sinusoidal example sound files of case ii) are the same that are enhanced under the proposed method in case iii). In these examples, the small amount of leakage from one spot signal to another can be perceived, especially in percussive-like sounds. At the same time, the fact that the audio quality in the resynthesized signals remains high, excluding the leakage effect, is also demonstrated. It is noted that these signals were used for the results both of Figs. 5 and 6 of this section.

The nine signals described correspond to the Signals 1–9 in the figures depicting the results of the listening tests of this section. It is important to mention that excluding the chorus signals, the remaining monophonic signals do not contain any common information (crosstalk). In such cases, the proposed model can result in high quality resynthesis if the reference signal is derived as the summation (downmix) of the various monophonic signals or their residuals, and this was the approach followed in the tests of this section.

In the first listening test, the quality of the modeled signals in a stereophonic setting was assessed. The listeners were asked to grade quality while ignoring any possibly noticeable image width distortion. Following the ITU-R [42] methodology, the modeled signals were compared against the originally recorded signals, mixed with the same $\pm$14-dB factors. A 5-scale grading system (from 1-"very annoying" audio quality compared to the original, to 5-"not perceived" difference in quality), was employed. No anchor signals were used. The results of this test are shown in Fig. 5, where the 95% confidence intervals are shown. It is clear from this image that the quality for all samples remains well above the 4.0 grade, even for the more complex chorus signals.

In the second listening test, the resulting image width was evaluated against the originally recorded signals. The procedure was similar to the procedure of the first test, but now grading referred to the resulting image width compared to the original stereo recording. Thus, a grade of 1.0 corresponded to a fully monophonic perception of the recording, while 5.0 corresponded to the image width of the original. Listeners were instructed to ignore quality distortion. An anchor signal was designed for this test, which was created by mixing the original signals with level differences of $\pm$2.5 dB instead of $\pm$14 dB. The results of this test are shown in Fig. 6, and it is clear that the proposed approach (denoted using "x" in the figure) introduces only a small degree of image width distortion for all nine testing signals. At the same time, the test results for the anchor signals (triangles in the figure) indicates that the subjects were able to correctly perceive image width distortion in the audio clips.

Overall, the results of this section justify our claim that high-quality resynthesis can be obtained even when using a small number of sinusoids and LP order in each frame, as long as the audio signals are rendered simultaneously. At the same time, the leakage between the signals which is introduced by our model, results only in a small degradation of the image width of the original stereophonic recording. In this sense, it is claimed that the results of the monophonic downmix tests of Fig. 4 are somewhat misleading, since the main source of degradation is due to the leakage, and in that monophonic setting the listeners did not tolerate any effect of leakage. In practice, even in cases when the user of an interactive audio reproduction system wishes to "move" closer to a particular instrument he/she will tolerate some instruments being audible slightly in the background. In other words, the results of Fig. 4 are valid only in cases when perfect separation of the spot signals in the decoder is desired, and in most practical scenarios the results of Figs. 5 and 6 more accurately represent the modeling performance of the proposed method.

C. Coding Performance

In this section, the perceived quality of the audio signals after the proposed modeling and coding procedure is evaluated. For this purpose we performed subjective (listening) tests by employing the ITU-R BS.1116 methodology (no anchor signals were used). In this test, listeners graded the coded versus the original signals using the aforementioned 5-scale grading. For

our listening tests, we used three signals, referred to as Signals 1—3, which are the *monophonic* concert hall signals of Section V-A, using ten sinusoids for our method. The female chorus signals were used in our experiments as the modeled (spot) signals, and the male chorus signals as the reference signals. Thus, the objective is to test whether the spot signal can be accurately reproduced when using the residual part derived from the reference signal. In this section our objective is to examine the lower limit in bitrates which can be achieved by our system without degradation of audio quality below the 4.0 level. Only the chorus signals were used for the results of this section since they contain more complex information compared to single instrument recordings, and thus quality distortions are easier to notice using these signals.

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 14 kb/s and 10 kb/s, respectively. It is noted that an analysis/synthesis window of 40 ms was used for the results of this section, since this allowed for a decreased bitrate compared to 30 ms, without noticeable degradation in the audio quality. It is noted that by applying more recent methods in sinusoidal audio coding such as [43] and [44], a lower bitrate in the order of 20% for the sinusoidal part may be obtained, for similar audio quality as the one given in our tests.

Regarding the coding of the LP parameters (noise spectral envelope), 28 bits were used per LSF vector which corresponds to 4.8 kb/s for the noise envelopes. Thus, the resulting bitrates that were tested are 18.8 kb/s and 14.8 kb/s (adding the bitrate of the sinusoidal parameters and the noise envelopes). This is in fact the target (theoretical) bitrate. By measuring the bit allocation in the actual experiments, the corresponding practical bitrates become on average 15.04 kb/s and 12.06 kb/s for the sinusoidal part, and 5.73 kb/s for the noise envelopes. This results in average practical rates of 20.77 kb/s and 17.79 kb/s for the target rates of 18.8 kb/s and 14.8 kb/s, respectively. It is informative to mention the maximum practical rates, which are 16.15 kb/s and 12.15 kb/s for the sinusoidal part, and 5.78 kb/s for the noise envelopes. This results in total maximum practical rates of 21.93 kb/s and 17.93 kb/s for the target rates of 18.8 kb/s and 14.8 kb/s, respectively.

A training audio dataset of about 100 000 LSF vectors (approximately 9.5 min of audio) was used to estimate the parameters of a 64-class GMM. The training database consisted of recordings of the classical music performance (corresponding to a different part of the same recording). Details about the coding procedure for the LP parameters can be found in our earlier work [20].

Twelve volunteers participated in these listening tests using headphones. The results of the tests are depicted in Fig. 7 for the cases of coding with a bitrate of 18.8 kb/s (squares) and of 14.8 kb/s (triangles), along with the modeling results without coding the parameters (circles). The results without coding are the same as in Fig. 3, corresponding to the label "$\sin_{10}$ +LPC_residual," and are given here again for reference. The results of the figure verify that the quality of the coded audio signals is good (above 4.0 in average), and that this quality can be maintained at as low as 14.8 kb/s 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 kb/s (monophonic case).

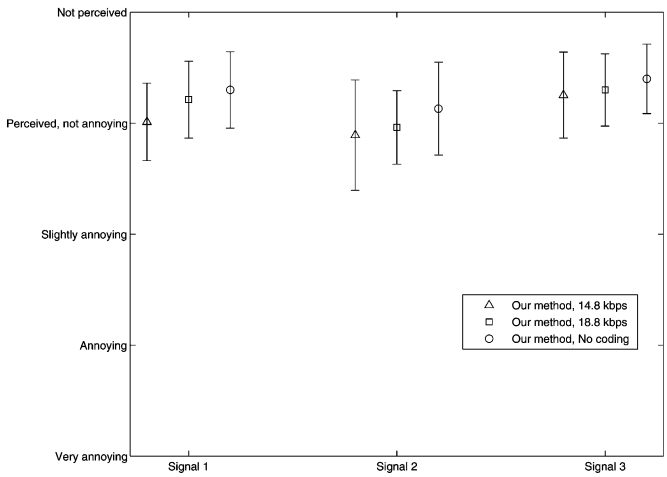

Fig. 7. Results from the monophonic quality rating listening tests, corresponding to using ten sinusoids and coding with (a) 14.8 kb/s (triangles), (b) 18.8 kb/s (squares), (c) no coding (only modeling, circles).

VII. Conclusion

A novel multichannel sinusoidal model was proposed for jointly coding multiple monophonic audio signals. The focus has been on spot audio signals, since these must be available at the decoder when interactivity between the listener and the acoustic environment is needed, as in truly immersive environments. The proposed approach is based on enhancing the sinusoidal model with the noise signal extracted from a reference, which can be one of the spot signals or a downmix. The proposed approach offers the advantage of employing the very flexible sinusoidal model into low bitrate multichannel audio coding. It was shown that the proposed method allows for high-quality audio modeling with only a negligible loss regarding the perceived audio image width in a stereophonic setting. It was also shown that the model parameters can be coded with rates as low as 15 kb/s per spot signal, which can be considered as a very encouraging result. Compared to SAC schemes this bitrate is higher. On the other hand, the proposed method encodes the actual content of the spot signals instead of only their spatial image, and in this sense it offers more flexibility for interactivity at the decoder and immersive audio applications.

Acknowledgment

The authors would like to thank the listening tests volunteers, 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 concert hall spot

signals, and A. Holzapfel and his band "Orange Moon" for providing the rock music recordings.

REFERENCES

[1] *Generic Coding of Moving Pictures and Associated Audio: Advanced Audio Coding*, ISO/IEC JTC1/SC29/WG11 (MPEG) International Standard ISO/IEC 13818-7, 1997.

[2] K. Brandenburg and M. Bosi, "ISO/IEC MPEG-2 advanced audio coding: Overview and applications," in *Proc. 103rd Conv. Audio Eng. Soc. (AES)*, Sep. 1997, Preprint No. 4641.

[3] M. Davis, "The AC-3 multichannel coder," in *Proc. 95th Conv. Audio Eng. Soc. (AES)*, Oct. 1993, Preprint No. 3774.

[4] J. D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 1992, pp. 569–572.

[5] J. Herre, K. Brandenburg, and D. Lederer, "Intensity stereo coding," in *Proc. 96th Conv. Audio Eng. Soc. (AES)*, Feb. 1994, Preprint No. 3799.

[6] D. Yang, H. Ai, C. Kyriakakis, and C.-C. J. Kuo, "High-fidelity multichannel audio coding with Karhunen–Loeve transform," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 365–380, Jul. 2003.

[7] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 100–120, Apr. 2000.

[8] J. Breebaart *et al.*, "MPEG spatial audio coding/MPEG surround: Overview and current status," in *Proc. AES 119th Conv.*, Oct. 2005, Paper 6599.

[9] F. Baumgarte and C. Faller, "Binaural cue coding—Part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 509–519, Nov. 2003.

[10] C. Faller and F. Baumgarte, "Binaural cue coding—Part II: Schemes and applications," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 520–531, Nov. 2003.

[11] 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.

[12] J. Herre and S. Disch, "New concepts in parametric coding of spatial audio: From SAC to SAOC," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2007, pp. 1894–1897.

[13] M. Wolters, K. Kjorling, D. Homm, and H. Purnhagen, "A closer look into MPEG-4 high efficiency AAC," in *Proc. 115th Conv. Audio Eng. Soc. (AES)*, Oct. 2003, Preprint No. 5871.

[14] "EBU evaluations of multichannel audio codecs," Geneva, Switzerland, Tech. Rep. 3324 EBU, Sep. 2007.

[15] M. M. Goodwin and J.-M. Jot, "A frequency domain framework for spatial audio coding based on universal spatial cues," in *Proc. 120th Conv. Audio Eng. Soc. (AES)*, May 2006, Preprint No. 6751.

[16] E. Gallo and N. Tsingos, "Extracting and re-rendering structured auditory scenes from field recordings," in *Proc. 30th Int. Conf. Intell. Audio Environ. Audio Eng. Soc. (AES)*, Mar. 2007.

[17] Y. Haraguchi, S. Miyabe, H. Saruwatari, K. Shikano, and T. Nomura, "Source-oriented localization control of stereo audio signals based on blind source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2008, pp. 177–180.

[18] C. Tzagkarakis, A. Mouchtaris, and P. Tsakalides, "Modeling spot microphone signals using the sinusoidal plus noise approach," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2007, pp. 183–186.

[19] M. Goodwin, "Multichannel matching pursuit and applications to spatial audio coding," in *Proc. 40th Annu. Asilomar Conf. Signals, Syst. Comput.*, Oct.–Nov. 2006, pp. 1114–1118.

[20] A. Mouchtaris, K. Karadimou, and P. Tsakalides, "Multiresolution source/filter model for low bitrate coding of spot microphone signals," *EURASIP J. Audio, Speech, Music Process.*, 2008, doi:10.1155/2008/624321.

[21] A. Mouchtaris, S. S. Narayanan, and C. Kyriakakis, "Virtual microphones for multichannel audio resynthesis," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 10, pp. 968–979, Sep. 2003.

[22] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

[23] R. Heusdens, R. Vafin, and W. B. Kleijn, "Sinusoidal modeling using psychoacoustic-adaptive matching pursuits," *IEEE Signal Process. Lett.*, vol. 9, no. 8, pp. 262–265, Aug. 2002.

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

[25] M. Lagrange, S. Marchand, and J.-B. Rault, "Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1625–1634, Jul. 2007.

[26] X. Rodet and P. Depalle, "Spectral envelopes and inverse FFT synthesis," in *Proc. 93rd Conv. Audio Eng. Soc. (AES)*, Oct. 1992.

[27] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 389–406, Sep. 1997.

[28] K. Vos, R. Vafin, R. Heusdens, and W. B. Kleijn, "High-quality consistent analysis-synthesis in sinusoidal coding," in *Proc. 17th Int. Conf. High Quality Audio Coding Audio Eng. Soc. (AES)*, Sep. 1999.

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

[30] M. M. Goodwin, "Multiscale overlap-add sinusoidal modeling using matching pursuit and refinements," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2001, pp. 207–210.

[31] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 1, pp. 21–29, Jan. 2001.

[32] J. Jensen, R. Heusdens, and S. H. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 121–132, Mar. 2004.

[33] M. G. Christensen, A. Jakobsson, S. V. Andersen, and S. H. Jensen, "Linear AM decomposition for sinusoidal audio coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2005, vol. 3, pp. 165–168.

[34] S. N. Levine, T. S. Verma, and J. O. Smith, "Multiresolution sinusoidal modeling for wideband audio with modifications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 1998, vol. 6, pp. 3585–3588.

[35] M. Goodwin, "Residual modeling in music analysis-synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 1996, vol. 2, pp. 1005–1008.

[36] R. C. Hendriks, R. Heusdens, and J. Jensen, "Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2004, vol. 4, pp. 189–192.

[37] *Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 mbit/s—Part 3: Audio*, ISO/IEC IS 11172-3, 1992.

[38] H. Purnhagen and N. Meine, "HILN—The MPEG-4 parametric audio coding tools," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2000, pp. 201–204.

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

[40] R. Vafin, S. V. Andersen, and W. B. Kleijn, "Exploiting time and frequency masking in consistent sinusoidal analysis-synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Jun. 2000, vol. 2, pp. 901–904.

[41] A. D. Subramaniam and B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 365–380, Mar. 2003.

[42] "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," ITU-R, 1997.

[43] R. Vafin and W. B. Kleijn, "Jointly optimal quantization of parameters in sinusoidal audio coding," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2005, pp. 247–250.

[44] P. Korten, J. Jensen, and R. Heusdens, "High resolution spherical quantization of sinusoidal parameters," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 966–981, 2007.

[45] A. Kain, "High Resolution Voice Transformation," Ph.D. dissertation, OGI School of Sci. and Eng., Oregon Health and Sci. Univ., Hillsboro, OR, Oct. 2001.

Christos Tzagkarakis received the B.Sc. and M.Sc. degrees in computer science from the University of Crete, Heraklion, Greece, in 2005, and 2007, respectively. He is currently pursuing the Ph.D. degree in the Computer Science Department, University of Crete, in the area of audio signal processing.

His research interests include signal processing for immersive audio environments, audio modeling and coding, and music information retrieval.

Athanasios Mouchtaris (S'02–M'04) received the Diploma degree in electrical engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1997 and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, in 1999 and 2003, respectively.

He is currently an Assistant Professor of Computer Science at the University of Crete, Heraklion, Greece, and a Researcher with the Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH-ICS), Heraklion. From 2004 to 2007, he was a Postdoctoral Researcher at FORTH-ICS, and a Visiting Assistant Professor at the Computer Science Department of the University of Crete. From 2003 to 2004, he was a Postdoctoral Researcher at the Electrical and Systems Engineering Department, University of Pennsylvania, Philadelphia. His research interests include signal processing for immersive audio environments, spatial audio rendering, multichannel audio modeling, speech synthesis with emphasis on voice conversion, and speech enhancement. He has coauthored more than 40 journal and conference papers in these areas.

Dr. Mouchtaris is a member of Eta Kappa Nu.

Panagiotis Tsakalides (M'95) received the Diploma in electrical engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1990, and the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 1995.

He is an Associate Professor of Computer Science at the University of Crete, and a Researcher with the Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH-ICS), Greece. From 2004 to 2006, he served as the Department Chairman. From 1999 to 2002, he was with the Department of Electrical Engineering, University of Patras, Patras, Greece. From 1996 to 1998, he was a Research Assistant Professor with the Signal and Image Processing Institute, USC, and he consulted for the U.S. Navy and Air Force. His research interests lie in the field of statistical signal processing with emphasis in non-Gaussian estimation and detection theory, and applications in sensor networks, audio, imaging, and multimedia systems. He has coauthored over 90 technical publications in these areas, including 25 journal papers.

Dr. Tsakalides was awarded the IEE's A. H. Reeve Premium in 2002 (with coauthors P. Reveliotis and C. L. Nikias) for the paper "Scalar quantization of heavy tailed signals" published in the October 2000 issue of the *IEE Proceedings-Vision, Image, and Signal Processing*. He is a member of the ERCIM Network of Innovation/Technology and Knowledge Transfer Experts (I-Board).