16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

MODELING AND CODING OF SPOT MICROPHONE SIGNALS FOR IMMERSIVE AUDIO BASED ON THE SINUSOIDAL MODEL

Christos Tzagkarakis, Athanasios Mouchtaris, and Panagiotis Tsakalides

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

ABSTRACT

In this paper, the Sinusoids plus Noise Model (briefly SNM) is applied in a novel manner, in order to efficiently encode spot audio signals. These are the microphone recordings of a performance, before obtaining the multichannel mix, and are important for immersive audio applications since they can be used to provide interactivity. The SNM, as well as the SNM error spectral envelope, are extracted from each spot signal, providing a low-quality version of the signals. The main contribution of the paper corresponds to the use of a single audio reference signal which significantly enhances the quality of all the modeled spot signals. Reproduction of good quality and without loss of image width can be achieved using the proposed approach (above 4.0 perceptual grade for modeling and coding), by encoding a single audio (reference) signal, with side information per spot signal on the order of 19 kbps.

1. INTRODUCTION

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

Multichannel audio offers the advantage of improved realism compared to 2-channel stereo sound at the expense of increased storage and transmission requirements. 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 [1, 2], achieve a significant coding gain but remain demanding for many low-bandwidth applications, such as streaming through the Internet and wireless channels. Recently, MPEG Surround [3] has been introduced, achieving significant compression of multichannel audio recordings. MPEG Surround is based on the Spatial Audio Coding (SAC) concept; SAC captures the spatial image of a multichannel audio signal by encoding only 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. For each channel (excluding the reference), these spatial cues can be encoded with rates as low as 5 kbps. MPEG Surround is based on the work on Binaural Cue Coding [4] and Parametric Stereo [5].

This work has been funded by the Marie Curie TOK "ASPIRE" grant within the 6th European Community Framework Program.

Our objective is to derive a low bitrate coding method for immersive audio applications. Immersive audio as opposed to multichannel audio, implies that the listener's environment is seamlessly transformed into the environment of his/her desire, and that the listener is able to interact with the content according to his/her will. Immersive audio is largely based on enhanced audio content, which translates into using a large number of microphones for obtaining a recording, containing as many sound sources as possible. These sources offer increased sound directions around the listener during reproduction, but are also useful for providing interactivity between the user and the audio environment. Examples include collaboration of geographically distributed musicians [6], or tele-presence in a concert hall performance where the user can "move" around the venue. Consequently, emphasis is on encoding the multiple microphone recordings of a given performance before those are mixed into the final multichannel mix. These microphone signals, also referred to as spot signals, are the signals that are captured *e.g.* by the various microphones that are placed inside a concert hall.

In this paper, the Sinusoids plus Noise Model (henceforth referred to as SNM for brevity), which has been used extensively for monophonic audio signals, is introduced in the context of low-bitrate coding for *Immersive* audio. As in the SAC method for low bitrate *multichannel* audio coding, our approach is to encode one audio channel only (which can be one of the spot signals or a downmix), while for the remaining spot signals we retain only the parameters required for resynthesis of the content at the decoder. These parameters are the sinusoidal parameters (harmonic part) of each spot signal, as well as the short-time spectral envelope (estimated using Linear Predictive – LP – analysis) of the sinusoidal noise component of each spot signal. These parameters are not as demanding—with respect to coding rates—as the true noise part of the SNM model. For this reason, the noise part of only the reference signal is retained. For resynthesis, each spot signal is reconstructed by adding its harmonic part to an estimated noise part. In turn, this noise part is synthesized by filtering the noise residual 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* [7], is based on the observation that the noise component of the spot signals of the same multichannel recording are very similar when the harmonic part has been captured with an appropriate number of sinusoids. Here, the modeling and codings stages are described, and the bitrates that our proposed system can achieve while retaining audio quality above 4.0 perceptual grade are experimentally found.

The results presented in this paper also illustrate the resulting quality and image width obtained using the proposed model in a stereophonic playback setting. It is of interest to examine whether the proposed approach results in introducing correlation among the various spot signals, which in turn would result in loss of spatial image width in the resulting stereophonic recording. The coding of the sinusoidal parameters is based on the scheme of [8], while the encoding process of the noise envelopes is based on [9].

2. MODELING OF SPOT SIGNALS

2.1 Sinusoids Plus Noise Model

The sinusoidal model for harmonic signals was initially proposed for speech signals in [11]. The sinusoids plus noise model (SNM) extends the sinusoidal model by representing a signal $s(n)$ with harmonic nature, as the sum of a predefined number of sinusoids (harmonic part) and a noise term (stochastic part) $e(n)$ (for each short-time analysis frame). A popular implementation of SNM for audio signals can be found in [10]. Thus, the SNM can be written in the following form:

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

where L denotes the number of sinusoids, $\{\alpha_l\,,\,\omega_l\,,\,\phi_l\}_{l=1}^{L}$ are the constant amplitudes, frequencies and phases respectively and N is the length (in samples) of the analysis short-time frame of the signal. The noise component is also needed for representing the noise-like part of audio signals which is audible and is necessary for high-quality resynthesis. The noise component can be computed by subtracting the harmonic component from the original signal.

Modeling the noise component is a challenging task. We follow the popular approach of modeling $e(n)$ as the result of filtering a residual noise component with an autoregressive (AR) filter that models the noise spectral envelope, *i.e.*,

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

where $r_e(n)$ is the residual of the noise, p is the AR filter order, and vector $\vec{b} = (1, -b(1), -b(2), ..., -b(p))^T$ represents the spectral envelope of the noise component $e(n)$ which can be obtained by LP analysis. In the frequency domain (2) becomes

$$S_e(e^{j\omega}) = \left|\frac{1}{B(e^{j\omega})}\right|^2 S_{r_e}(e^{j\omega}), \qquad (3)$$

where $S_e(e^{j\omega})$ and $S_{r_e}(e^{j\omega})$ is the power spectrum of $e(n)$ and $r_e(n)$, respectively, and $B(e^{j\omega})$ is the frequency response of the LP filter $\vec{b}$. In the remainder of the paper, we refer to $e(n)$ as the (sinusoidal) *noise* signal, and to $r_e(n)$ as the *residual* (noise) of $e(n)$.

2.2 Noise Transplantation

In this section, we describe the main novelty of our proposed approach, namely noise transplantation. Consider a collection of M microphone signals that correspond to the same multichannel recording and thus have similar acoustical content. We model and encode only one of the signals as a full audio channel (alternatively it can be a downmix, *e.g.* a sum signal), which is the reference signal. The remaining (side) signals are modeled by the SNM, as explained in the previous sub-section, retaining their sinusoidal components, and the noise spectral envelope (filter $\vec{b}$ in (2)).

In order to reconstruct the spot side signals, the residual signals are needed. In our approach, only one such signal is used for all spot signals. In fact, this signal is the residual noise of the reference signal. It is obtained by extracting first the sinusoidal noise of the reference signal and from it the residual noise. In other words, the residual noise of the reference signal is used in order to reconstruct all spot signals during decoding. First this noise is filtered by each of the LP spectral envelopes (one for each spot signal), and then the derived signal is added to the corresponding harmonic part in order to recreate the high-quality resynthesized spot signals.

In this manner, we avoid encoding the residual of each of the side signals. This is important, as the noise signals in general are of highly stochastic nature, and cannot be adequately represented using a small number of parameters (thus, it is highly demanding in bitrates for accurate encoding). We note that modeling this signal with parametric models results in low-quality audio resynthesis; in [7] we showed that our noise transplantation method can result in significantly better quality audio modeling compared to parametric models for the residual signal. We obtained subjective scores around 4.0 using as low as 10 sinusoids, which is very important for low bitrate coding.

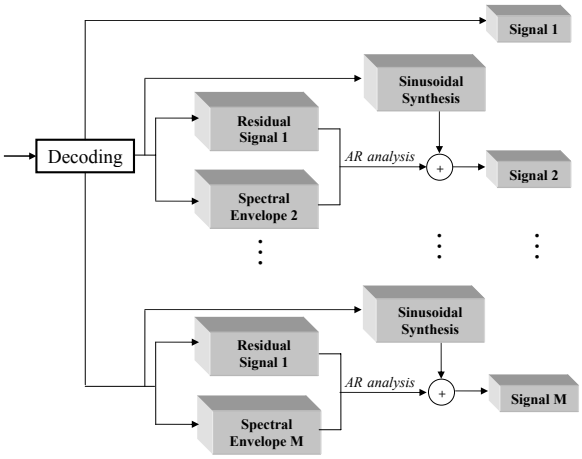

Figure 1: Diagram of the proposed decoding approach.

The relation for the resynthesis of one of the *side* microphone signals x_k (using the reference signal $x_{(ref)}$) is (see also Fig. 1)

$$\hat{x}_k(n) = \sum_{l=1}^{L} \alpha_{k,l}\, cos(\omega_{k,l}\, n + \phi_{k,l}) + \hat{e}_k(n)\ ,k = 1,...,M, \qquad (4)$$

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

$$S_{\hat{e}_k}(e^{j\omega}) = \left|\frac{1}{1 - \sum_{i=1}^{p} b_k(i)\, e^{-j\omega i}}\right|^2 S_{r_{e_{(ref)}}}(e^{j\omega}). \qquad (5)$$

In the relations above, $\{\alpha_{k,l}\,,\,\omega_{k,l}\,,\,\phi_{k,l}\}$ are the sinusoidal parameters of side signal x_k and $\{b_k\}$ is the signal's LP noise shaping filter. The approximated noise component for signal x_k, $\hat{e}_k(n)$, is computed by filtering the reference signal's residual noise with the noise shaping filter $\{b_k\}$ of the corresponding side signal. The power spectrum of the residual of the reference signal can be computed by the following expression

$$S_{r_{e_{(ref)}}}(e^{j\omega}) = \left|1 - \sum_{i=1}^{p} b_{(ref)}(i)\, e^{-j\omega i}\right|^2 S_{e_{(ref)}}(e^{j\omega}), \qquad (6)$$

where $e_{(ref)}$ is the sinusoidal noise of the reference signal.

3. CODING OF SPOT SIGNALS

The second part of our method is the coding procedure. It can be divided into two tasks; the quantization of the sinusoidal parameters and the quantization of the noise spectral envelopes for each side signal (for each short-time frame). In Fig. 1 we can see the decoding process, where the reference signal (Signal 1) is fully encoded (*e.g.* using an MP3 encoder at 64 kbps), while the remaining $M-1$ signals are reconstructed using the quantized sinusoidal and LP parameters, and the LP residual obtained from the reference signal. It must be noted that the side information for each spot signal consists of the sinusoidal parameters and the LP filter parameters. Thus, the coding procedure proposed follows previously proposed methods for coding such parameters, namely [8] (sinusoidal parameters) and [9] (LP parameters). The description of these methods is mainly given here for completeness.

3.1 Coding of the Sinusoidal Parameters

According to the coding scheme of [8], the sinusoidal parameters are quantized in polar form, assuming a dependence of the fre-

quency quantization on the amplitude, and a dependence of the phase quantization on the amplitude and the frequency. This scheme is called Unrestricted Polar Quantization and represents a combination of three scalar quantizers, based on high-rate quantization.

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

$$D = \frac{1}{L}\sum_{l=1}^{L} w_l D_l \tag{7}$$

under the entropy constraint

$$H = \frac{1}{L}\sum_{l=1}^{L}(H(I_{\alpha l}) + H(I_{\omega l}|I_{\alpha l}) + H(I_{\phi l}|I_{\alpha l})). \tag{8}$$

The given total entropy per sinusoid (amplitude, frequency, and phase) is denoted by H. The entropies $H(I_{\alpha l})$, $H(I_{\omega l}|I_{\alpha l})$ and $H(I_{\phi l}|I_{\alpha l})$ express the entropies of the individual quantization parameters. The mean squared error (MSE) D_l introduced by the quantization of the l^{th} sinusoid is assigned a perceptual weight w_l, which is defined as $w_l = 1/m_{th,l}$, $l = 1,\ldots L$, where $m_{th,l}$ is the masking threshold at the frequency of the corresponding sinusoid [12].

The MSE D_l over a segment of length N, can be expressed as

$$D_l = E\left\{\frac{1}{N}\sum_{n=-(N-1)/2}^{(N-1)/2}(\alpha_l\cos(\omega_l n+\phi_l) - \hat{\alpha}_l\cos(\hat{\omega}_l n+\hat{\phi}_l))^2\right\}, \tag{9}$$

where $\{\alpha_l, \omega_l, \phi_l\}$ and $\{\hat{\alpha}_l, \hat{\omega}_l, \hat{\phi}_l\}$ 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 (7) under the constraint expressed in (8). 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

$$g_A(\alpha) = g_A = \frac{w_\alpha^{\frac{1}{6}} 2^{\frac{1}{3}\tilde{H}-\frac{2}{3}b(A)}}{w_g^{\frac{1}{6}}(\frac{N^2}{12})^{\frac{1}{6}}} \tag{10}$$

$$g_\Omega(\omega,\alpha) = g_\Omega(\alpha) = \frac{\alpha w_\alpha^{\frac{1}{6}}\left(\frac{N^2}{12}\right)^{\frac{1}{3}} 2^{\frac{1}{3}\tilde{H}-\frac{2}{3}b(A)}}{w_g^{\frac{1}{6}}} \tag{11}$$

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

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

3.2 Coding of the Spectral Envelopes

The second group of parameters for each spot signal that need to be encoded are the spectral envelopes of the sinusoidal noise. We follow the quantization scheme of [9]. The LP coefficients of each spot signal that model the noise spectral envelope are transformed to LSF's (Line Spectral Frequencies) which are modeled by a Gaussian Mixture Model (GMM), defined as

$$g(x) = \sum_{i=1}^{C} p_i N(x;\mu_i,\mathbf{\Sigma}_i). \tag{13}$$

In the equation above, $N(x;\mu,\mathbf{\Sigma})$ is the normal multivariate distribution with mean vector μ and covariance matrix $\mathbf{\Sigma}$, p_i is the prior probability that the observation x has been generated by cluster i and C is the number of clusters. The covariance matrix of each cluster can be diagonalized using eigenvalue decomposition as

$$\mathbf{\Sigma}_i = \mathbf{Q}_i\mathbf{\Lambda}_i\mathbf{Q}_i^T, \tag{14}$$

where $i = 1,\ldots,C$. The matrix $\mathbf{\Lambda}_i$ is diagonal and contains the corresponding eigenvalues of $\mathbf{\Sigma}_i$, while $\mathbf{Q}_i$ is the matrix containing the corresponding set of orthogonal eigenvectors of $\mathbf{\Sigma}_i$, for the i^{th} Gaussian class of the model. Then, the Karhunen Loève Transform (KLT) substitutes each LSF vector for time segment k, z_k, with another decorrelated vector w_k, where $w_k = \mathbf{Q}_i^T(z_k - \mu_i)$. Consequently, the components of the vector w_k can be independently quantized by a non-uniform quantizer, *i.e.*, through a compressor, a uniform quantizer and an expander.

Each LSF vector is classified to only one of the C GMM's, 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 cluster associated with the minimal LSD, which is defined as

$$LSD(i) = \left(\frac{1}{F_s}\int_0^{F_s}\left[10\log_{10}\left(\frac{S(f)}{\hat{S}^{(i)}(f)}\right)\right]^2 df\right)^{\frac{1}{2}}, \tag{15}$$

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

4. PERFORMANCE EVALUATION

In this section, we examine the modeling and coding performance of our proposed system, regarding the resulting audio quality. For this purpose, several listening tests were performed. The parameter choices used in the following experiments were 10 sinusoids per frame for the sinusoidal model, 10^{th} LP order for the noise spectral envelope, 30 msec window with 50% overlapping for the sinusoidal model, and 23 msec and 75% overlapping for the LP model. The audio signals were sampled at 44.1 kHz unless otherwise stated.

4.1 Modeling Performance

In our previous work [7], listening tests were performed in a monophonic setting. In other words, the proposed model was used to derive one microphone recording (monophonic signal) from the reference signal, and this was presented separately to each listener using headphones. The results of this test indicated that the perceived quality was good when using 40 sinusoids in the model, but the results were lower when only 10 sinusoids were used. However, it was apparent that the main source of degradation was due to the fact that parts of the reference recording which were not originally present in the originally recorded spot signal, were included in the resynthesized spot signal. This fact is an undesired effect of the transplantation procedure, the model parameters cannot capture all the microphone-specific information and completely "whiten" the reference residual. As a result, it is not possible in practice to avoid leakage of the reference signal into the resynthesized signal. At the same time, it was clear 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. Thus, if the only degradation of the modeled signals is the leakage among the several recordings, this will appear in the stereophonic or multichannel setup as an image width—and not as a quality—distortion.

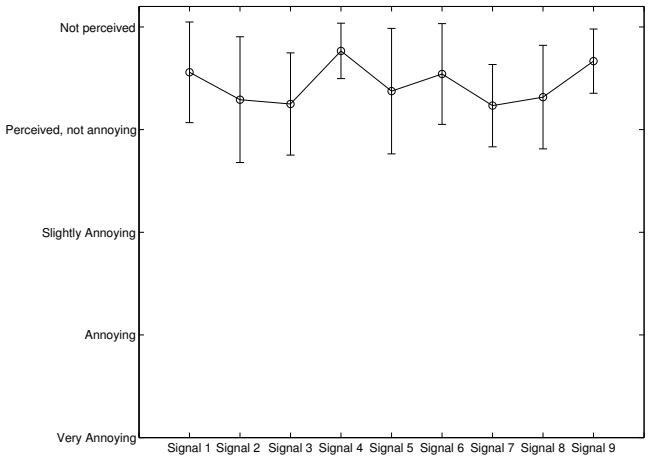

Figure 2: Results from the quality rating listening test. Only quality was rated, and listeners were asked to ignore image width distortion.

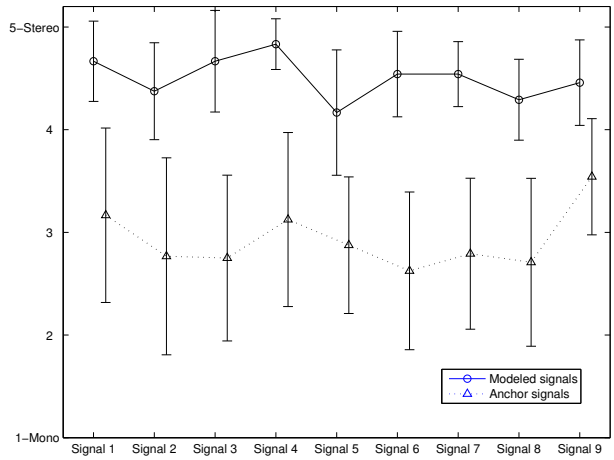

Figure 3: Results from the image width rating listening tests (ignoring quality distortion).

To test these assumptions, two listening tests were designed in a stereophonic setup using good-quality headphones (Sennheiser HD-650). The first test was designed in order to test the quality of two modeled signals when rendered simultaneously, ignoring the image width distortion. The second test was designed to test only the image width distortion, ignoring the quality distortion. Both tests were performed following the ITU-R BS.1116 [14] recommendations. In both tests 12 volunteers participated, who were trained so that they could distinguish among the types of distortion examined.

Separate *monophonic* spot recordings were modeled by the proposed approach for deriving the sound files used in the listening tests. This is due to the fact that, under the proposed scheme, the actual stereophonic or multichannel recordings are mixed *after* decoding. The proposed algorithm cannot retain relative amplitude and time differences between the audio channels, so the spatial image of an already mixed recording would be distorted by our method. The following monophonic recordings were used, each containing a separate instrument recording (the duration of each audio clip was around 10 sec): (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. Signals (i)-(v) are excerpts from the EBU SQAM (Sound Quality Assessment Material) test disc[1]. These are stereo recordings, and only one of the 2 channels was used in our experiments. Signals (vi)-(vii) are a courtesy of rock band "Orange Moon". Signals (viii)-(ix) were obtained from the VOICES corpus[2], available by OGI's CSLU [13]. Signals (x)-(xi) are actual spot signals from a concert hall performance [3]. The speech signals were sampled at 22 kHz.

Using the above mentioned 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). These nine signals correspond to the Signals 1-9 in the figures depicting the results of the listening tests of this section (modeling results). It is important to mention that apart from 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, and this was the approach followed in the results of this section.

In the first listening test, the listeners were asked to grade quality while ignoring any possibly noticeable image width distortion. Following the ITU-R [14] methodology, the modeled signals were compared against the originally recorded signals, mixed with the same ± 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. 2, where the vertical lines indicate the 95% confidence limits. It is clear from this image that the quality for all samples remains well above 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 the same as in the first test, but in this case the resulting image width compared to the original stereo recording was graded. 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 ± 2.5 dB instead of ± 14 dB of the original stereo signals. The results of this test are shown in Fig. 3, and it is clear that the proposed approach (solid line 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 (dashed line in the figure) indicate 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—inaudible in this case—leakage between the signals, which is introduced by our model, results only in a small degradation of the image width of the original recording (after the monophonic signals are mixed in order to form the final stereo or multichannel recording).

4.2 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 Degradation Category Rating (DCR) test. In this test, listeners graded the coded *vs.* the original waveform using the aforementioned 5-scale grading. For our listening tests, we used three signals, referred to as Signals 1-3, which correspond to the *monophonic* chorus signals of the previous section (*i.e.* before they were

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

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

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

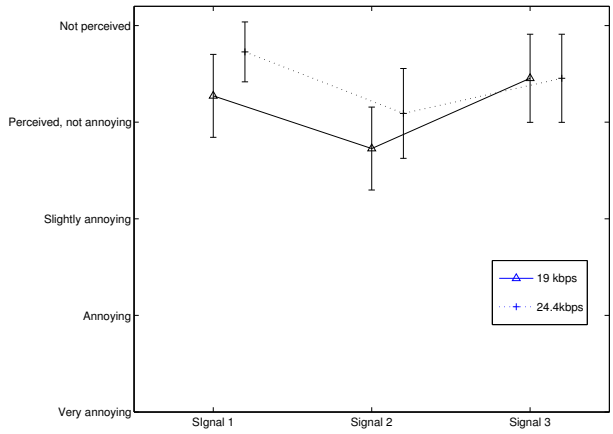

Figure 4: Results from the quality rating DCR listening tests, corresponding to coding with (a) 24.4 kbps (dotted), (b) 19 kbps (solid).

mixed). 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 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 loss of audio quality below a 4.0 grade. 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 19.6 kbps and 14.2 kbps respectively. Regarding the coding of the LP parameters (noise spectral envelope), 28 bits were used per LSF vector which corresponds to 4.8 kbps for the noise envelopes. Thus, the resulting bitrates that were tested are 24.4 kbps and 19 kbps (adding the bitrate of the sinusoidal parameters and the noise envelopes). A training audio dataset of about 100,000 LSF vectors (approx. 9.5 min of audio) was used to estimate the parameters of a 16-class GMM. The training database consisted of recordings of the classical music performance (corresponding to a different part of the same recording). Details about the coding procedure for the LP parameters can be found in our earlier work [15].

Eleven volunteers participated in the DCR tests using headphones. The results of the DCR tests are depicted in Fig. 4. The solid line shows the results for the case of coding with a bitrate of 19 kbps, while the dotted line shows the results for the 24.4 kbps case. The results of the figure verify that the quality of the coded audio signals is good, and that this quality can be maintained at as low as 19 kbps per side signal. We note that the reference signal was PCM coded with 16 bits per sample, however similar results were obtained for the side signals when the reference signal was MP3 coded at 64 kbps (monophonic case).

5. CONCLUSIONS

In this paper, we presented a complete system for low bitrate coding of spot microphone signals for multichannel audio applications. Spot signals were treated here since preserving their content and quality is important when interactivity between the listener and the acoustic environment is needed, as in truly immersive environments. Our proposed approach is based on the sinusoidal model, as well as the newly introduced concept of noise transplantation which exploits the interchannel similarities of a given multichannel recording. It was shown that the proposed method allows for good-quality audio modeling without a significant loss of 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 19 kbps per spot signal, which can be considered as a very encouraging result. Since this research is at an early stage, we are confident that the required rates of the proposed system can be further reduced.

6. ACKNOWLEDGMENTS

The authors wish to thank the listening tests volunteers, as well as Prof. Y. Stylianou for his insightful suggestions and for his help with the implementation of the sinusoidal model algorithm.

REFERENCES

[1] K. Brandenburg, and F. Bosi, "ISO/IEC MPEG-2 advanced audio coding: Overview and applications," *Proc.* 103^{rd} *Convention of the Audio Engineering Society (AES)*, preprint No. 4641, (New York, NY), Sep. 1997.

[2] M. Davis, "The AC-3 multichannel coder," *in Proc.* 95^{th} *Convention of the Audio Engineering Society (AES)*, preprint No. 3774, (New York, NY), Oct. 1993.

[3] J. Breebaart et al., "MPEG Spatial Audio Coding / MPEG Surround: Overview and Current Status," *Proc. AES* 119^{th} *Convention*, Paper 6599, Oct. 2005.

[4] F. Baumgarte, and C. Faller, "Binaural Cue Coding - Part I: Psychoacoustic Fundamentals and Design Principles," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, pp. 509–519, Nov. 2003.

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

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

[7] C. Tzagkarakis, A. Mouchtaris, and P. Tsakalides, "Modeling spot microphone signals using the sinusoidal plus noise approach," *in Proc. Workshop on Appl. of Signal Proc. to Audio and Acoust.*, Oct. 2007.

[8] R. Vafin, D. Prakash, and W. B. Kleijn, "On Frequency Quantization in Sinusoidal Audio Coding," *IEEE Signal Proc. Letters*, vol. 12, no. 3, pp. 210–213, Mar. 2005.

[9] A. D. Subramaniam, and B. D. Rao, "PDF optimized parametric vector quantization of speech line spectral frequencies," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, pp. 365–380, Mar. 2003.

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

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

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

[13] A. Kain, "High Resolution Voice Transformation," PhD Thesis, OGI School of Science and Engineering at Oregon Health and Science University, Oct. 2001.

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

[15] K. Karadimou, A. Mouchtaris, and P. Tsakalides, "Multichannel Audio Modeling and Coding Using a Multiband Source/Filter Model," *Conf. Record of the* 39^{th} *Asilomar Conf. Signals, Systems and Computers*, pp. 907–911, 2005.