SINUSOIDAL SPATIAL AUDIO CODING FOR LOW-BITRATE BINAURAL REPRODUCTION

Toni Hirvonen and Athanasios Mouchtaris

Department of Computer Science, University of Crete, and
Institute of Computer Science
Foundation for Research and Technology - Hellas (FORTH-ICS)
Heraklion, Crete, Greece

ABSTRACT

A binaural audio synthesis system based on sinusoidal modeling is proposed for spatial, low-bitrate audio coding utilized for example in teleconference applications. The system transmits monaural sinusoidal parameters of a downmix signal, from which the left and right binaural signals are synthesized according to the directional metadata at the receiver. Typical sinusoidal synthesis methods, as well as the effectiveness of a monaural frequency masking model, are evaluated in binaural context. Furthermore, a method for binaural noise residual synthesis and efficiency improvements for HRTF parameter acquisition are suggested. Tests utilizing speech signals indicate that sinusoidal modeling is an attractive technique for applications such as the proposed one.

Index Terms— Binaural synthesis, sinusoidal coding, spatial audio coding.

1. INTRODUCTION

Systems for coding of spatial audio, such as MPEG Surround, have emerged as a prominent part of the new media landscape [1, 2]. Initially used to compress/downmix an existing (static) multichannel audio mix, these methods have been extended to extracting and transmitting the directional qualities of actual acoustic situations, such as musical performances or real-time teleconferences [3]. Spatial audio coding systems usually transmit a monophonic or stereophonic channel of audio (downmix of several audio channels), along with metadata containing the spatial information as a function of time and frequency. In this manner, these systems achieve high compression ratios while being capable of excellent reproduction of spatial quality at the receiving listening setup.

The teleconference version of Directional Audio Coding (DirAC) is a system for recording and reproducing spatial audio when transmitting one monophonic audio channel and directional metadata [3]. These metadata include the spatial information based on the analysis at the transmitter; the rate of metadata can be as low as 3-6 kbit/s [4]. At the receiver, all required audio channel signals are created. DirAC utilizes STFT-based transform coding for each overlap-windowed time frame. This allows for easy manipulation of the gains of different frequency bins according to the directional metadata. Teleconferencing [4, 5] and binaural synthesis [6] with DirAC have been investigated. Another system similar to DirAC is Spatial Audio Scene Coding (SASC), which utilizes efficient frequency-domain manipulation and synthesis [7]. The use of sinusoidal coding in the SASC context was described in [8]. However, this study focused on exploiting channel redundancy in static multichannel audio mixes and did not investigate the binaural synthesis quality or psychoacoustic-motivated selection of the sinusoidal frequencies.

This paper is concerned with binaural headphone synthesis in the context of spatial audio coding, based on the sinusoidal model. In the case examined here, both ear signals are constructed at the receiver, by directionally processing the downmix signal in order to form the left and right channels, using the transmitted directional cues. This approach is more efficient in terms of bitrate than creating the binaural signals at the transmitter, and is suitable for spatial audio coding applications where the directional information is encoded separately (as in SASC or DirAC). The fact that the proposed methodology is based on the sinusoidal model, implies that lower bitrates for coding can be obtained compared to non-parametric spatial audio coding, at the expense of lower audio quality. Thus, the proposed approach is suitable for low-bitrate applications such as Voice-over-IP (VoIP) teleconference applications. The benefit of spatial coding in teleconferencing is notable as *e.g.* simultaneous speakers from different directions can be reproduced to be spatially separated, thus greatly increasing intelligibility and naturalness in the reproduction [9]. The need for the study of low-bitrate applications is significant, as present spatial audio systems have been mostly used at higher bitrates.

The general goal of this paper is to propose and evaluate sinusoidal modeling for binaural synthesis within the context of spatial audio coding. Specifically, the following issues are examined: (a) We investigate windowed overlap-add (OLA)- and interpolation-based synthesis methods in their ability to preserve spatial quality and remove frame discontinuities for varying spatial directions. (b) The effect of applying a monaural frequency masking model in binaural synthesis is evaluated. The possible quality reduction caused by binaural effects is investigated by comparing the use of a masking model for sinusoidal analysis, with the case where no masking model is applied. (c) The effect of binaurally adding the residual noise part, which is utilized in sinusoidal plus noise modeling, is examined. Tests are performed to investigate whether artifacts are created by the binaural residual. (d) Additionally, we propose an efficiency improvement for binaural synthesis of sinusoidally modeled audio based on the Goertzel algorithm.

2. BACKGROUND

2.1. Sinusoidal Modeling

Sinusoidal modeling is a prominent technique for low-bitrate coding, especially for speech signals [10]. Current state-of-the-art methods employ perceptual matching pursuit algorithms to determine the sinusoidal parameters of each frame. Psychoacoustical

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

Matching Pursuit (PAMP) [11] minimizes the perceptual distortion measure D given by:

$$D = \int_{\omega} A(\omega)|R(\omega)|^2, \tag{1}$$

where $R(\omega)$ is the Fourier transform of the residual after each matching pursuit iteration, and $A(\omega)$ is a frequency weighting function usually set as the reciprocal of the current masking threshold.

Given the sinusoidal model parameters, common methods for synthesizing the final time-domain signal include windowed overlap-add (OLA) and interpolation-based methods. The former is simpler in that every windowed synthesized component is simply summed to form the frame signal. In the latter, samples for each frame are created by interpolating between the corresponding frequency and phase parameters according to determined sinusoidal tracks [10]. Although it is more computationally expensive, we chose to also study the interpolation method because of the increased synthesis accuracy it provides.

In general, audio signals modeled by the sinusoidal model sound artificial, since the stochastic nature of audio signals is not captured. In order to address this issue, a noise residual based on the modeling error between the original and sinusoidally modeled signals can be introduced. It has been shown that good results can be obtained with *e.g.* the Perceptual LPC (PLPC) method [12]. The residual creates "comfort noise" that increases naturalness. In this paper, the PLPC method is employed for the modeled signals which include a noise part for the sinusoidal model.

2.2. Binaural Synthesis

Binaural synthesis usually corresponds to the process of filtering the monophonic signal with the Head Related Transfer Functions (HRTFs) for a specific direction in order to create the left and right ear signals [13]. HRTF filtering implements the interaural time difference (ITD) and level difference (ILD) spatial cues of the desired spatial direction, which are frequency- and location-dependent due to sound filtering by the human upper body and pinnae.

By introducing different HRTF information for different frequencies within a single time frame, simultaneous sound sources from different directions can be reproduced in the context of spatial audio coding methods such as DirAC and SASC. Binaural synthesis for SASC has been suggested in [14]. The HRTF filtering can be performed in the frequency domain by manipulating the amplitudes and phases of the frequency bins, *i.e.*

$$Y(\omega) = X(\omega)|H(\omega)|e^{j\phi(\omega)}. \tag{2}$$

In the above equation, $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of the monophonic signal and the synthesis result, respectively, for either left or right channel at a given time segment. The term $|H(\omega)|$ represents the directional magnitude response that can be obtained from a single HRTF in case of constant direction for all frequencies, or as is more common, combined from several HRTFs. This procedure implements filtering due to the ILD spatial cues. Similarly, the phase manipulation according to $\phi(\omega)$ implements the ITD cues. The ITD can be approximated to be constant for all frequencies, and obtained with simple formulae [6, 14]. This is equivalent to the minimum-phase HRTF approximation and does not notably degrade reproduction [15].

An important practical issue examined in this paper is the avoidance of audible discontinuities between frames in HRTF synthesis. These are related to the smooth combination of the changing HRTF information. In the case of one constant sound source without echoes, the HRTF filtering can be realized by rectangularly-windowed OLA FFT-filtering. However, with multiple or moving sounds, the required directions for different frequencies usually change very rapidly within and between frames. With DirAC, methods were presented to overcome the problem [3, 6]. These included signal windowing in both analysis and synthesis, as well as slowing the HRTF gain changes. In SASC, the discontinuities were reportedly avoided without gain slowing and using only single analysis window [14]. Based on our simulations, the reason for this is the use of parametric IDFT synthesis with minimum-phase HRTFs.

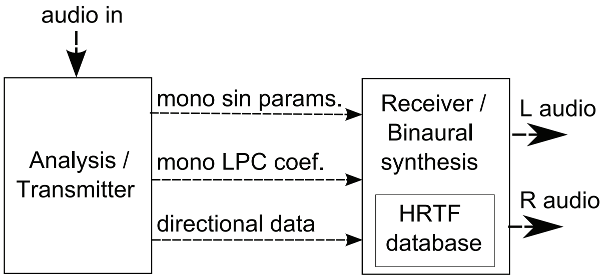

Fig. 1. Proposed system for audio transmission and binaural synthesis. The monophonic downmix signal is modeled by the sinusoidal model and transmitted along with the spatial cues. In the transmitter the binaural signals are synthesized and rendered.

The use of psychoacoustic masking models in binaural synthesis is also examined here. This issue becomes important in our work since monaural perceptual masking is used in sinusoidal modeling; only perceptually relevant spectral components are considered for deriving the sinusoidal model parameters. Given, however, that the modeled mono channel is subsequently filtered to create the binaural left and right channels, it is important to examine whether sound colour- and/or Binaural Masking Level Difference (BMLD) effects [13] lead to reduced audio quality due to the use of the monaural masker analysis. In transform coding such as MP3, some quantization noise components that are masked in monaural listening, may be audible (masking release) when binaural processing is applied. However, it is not clear how the use of sinusoidal modeling induces binaural quality differences. Furthermore, we are in this paper not considering parameter *quantization*, but rather wish to examine whether the parameter selection method itself, based on monaural masking model, is suitable for binaural synthesis in that it can produce natural sound timbre and not create a sensation of "missing components" due to BMLD. The BMLD effects possibly caused by sinusoidal parameter quantization are hypothesized not as prominent as in transform coding because of the reduced masking energy between spectral components, and are left to future studies.

3. PROPOSED SYSTEM

The proposed system is described in Fig. 1. The LPC coefficients implement the noise residual needed in the sinusoidal plus noise model. The effect of the noise part in binaural synthesis is discussed in Section 3.3. Directional data can be obtained with *e.g.* a microphone grid that also records the audio input as in DirAC, or by assigning spatial properties to audio signals if they are separately available before downmixing. Although not tested here, simultaneous input from multiple sites can also be utilized. In these situations, the audio signal from every site would be synthesized separately, and the receiver would perform a downmix for the listener, who could also interact with the spatial cues for each input if desired.

Spatial coding methods for high-quality audio in DirAC and

SASC include additional metadata describing the energy of the diffuse sound as a function of frequency. However, as discussed in [4, 5], this is commonly not beneficial for low-bitrate applications, and is not included here. Note that the noise residual technique described here should not be confused with the synthesis of diffuse sound.

3.1. Binaural synthesis

This study utilizes HRTFs measured from the first author's ears (open ear canal) with miniature microphones in an anechoic room [6]. The angular resolution in the horizontal plane is about 5° on average, which is close to the minimum audible angle resolution of human hearing.

To facilitate the binaural synthesis, we extend the frequency domain binaural synthesis as suggested in [14], by directly modifying the sinusoidal amplitude and phase parameters according to the desired direction at that frequency prior to sinusoidal synthesis:

$$\begin{aligned} A_{L,R}(k) &= A(k)|H_{L,R}(\theta, f_k)|, & (3)\\ \phi_L(k) &= \phi(k) - \pi f_k \Delta(\theta), & (4)\\ \phi_R(k) &= \phi(k) + \pi f_k \Delta(\theta). & (5) \end{aligned}$$

In the above equations, $A_L(k)$, $A_R(k)$, $\phi_L(k)$, and $\phi_R(k)$ indicate the k^{th} binaural left and right channel amplitude and phase parameters, $A(k)$ and $\phi(k)$ the monaural amplitude and phase, f_k the parameter frequency, θ the direction obtained from the metadata, and $\Delta(\theta)$ the ITD value.

Typically, the magnitudes of the HRTF responses $|H_L(\theta, f_k)|$ $|H_R(\theta, f_k)|$, *i.e.* the ILD cues must be evaluated for each sinusoidal component individually, as the directional metadata varies rapidly with frequency even for a non-moving source. This requires at least twice as many FFT operations as there are components. We suggest an efficiency improvement when using sinusoidal model: the HRTF magnitudes are here obtained with the Goertzel algorithm, a well known method for calculating a single DFT bin. Calculating a single bin for each sinusoidal component this way is beneficial, as the Goertzel algorithm is more efficient than FFT when the number of bins is less than $log_2(N)$, where N is the frame size.

When the desired direction does not exactly match any of the measured HRTF directions, the amplitude values are interpolated linearly from the two closest available HRTFs. The ITD is approximated as constant for all frequencies and modeled with a smooth function which varies according to the average measured group delays of the HRTFs, as in [15]. Furthermore, the HRTFs are assumed symmetric and right-side measurements are mirrored to the left side.

3.2. Sinusoidal Parameter Selection

In the sinusoidal analysis, the local maxima, *i.e.* the true peaks of the frame magnitude spectrum were selected as the initial dictionary of the matching pursuit. For the reference signal that did not utilize psychoacoustic masking, all sinusoids of this dictionary were chosen, resulting in approximately 80 sinusoids per frame. This relatively high number of components was selected in order to achieve quality close to reference so that the audio model quality affected the listening test as less as possible.

To compare the effect of the monaural masking model to the sinusoidally modeled reference, PAMP was utilized with the previous initial dictionary, and a maximum of 50 sinusoidal components per frame. These were usually the components above the final masking curve of the perceptual model. When listening monaurally, the quality of these two signals was the same. The Hanning-windowed analysis frame was 20 ms with a 10 ms hopsize. As discussed in Section 2.2, no quantization was included.

3.3. Proposed Binaural Noise Residual Method

As the phase is randomized in a noise residual, a natural approach is to multiply the magnitude response of the the synthesized monaural residual by the HRTF magnitudes $H_L(\omega)$ and $H_R(\omega)$:

$$N_{L,R}(\omega) = |H_{L,R}(\omega)|N(\omega) \qquad (6)$$

where $N(\omega)$ is the Fourier transform of the monaural residual and $N_L(\omega)$ and $N_R(\omega)$ are the binaural left and right residuals. After inverse transform, the resulting left and right residual time domain signals are summed to the synthesized sinusoidal signals. The correct energy relationships are also preserved. Due to the human hearing resolution, the spectral magnitudes of the HRTFs can be averaged if desired. Here, this is done at critical band resolution.

4. LISTENING TESTS

4.1. Samples

Two samples of approximately 6 seconds were used, sampled at 22050 Hz. To obtain the directional data, we used a synthetic method where the directions of the sound objects (the individual speaker signals) were known a priori. For the first sample, a female speaker was simulated to move circularly 180° by interpolating between -90° and 90° so that the speaker moved smoothly. For the second sample, two different speakers (one male and one female speaker uttering different content) were reproduced at constant azimuth angles simultaneously. In this case, the sources did not move, but the direction varied in frequency within one frame among two values in the following manner. Prior to downmix, the relative energies of the two speaker samples were compared at each frequency bin of each frame. An azimuth value was assigned to the bin according to the energy relation as in [3], with the speakers assumed at -45° and 45°. The directional data was also averaged across each critical frequency band. The former sample represents the case of *e.g.* gaming applications, and had more pauses between words making it easier to perceive quality differences. Simulating the directional information for the former sample is a rather simple task, since this information is applied to all frequencies of the non-directional monophonic signal, and is only varying in time. The latter sample represents a situation in a teleconference, since the two speakers overlapped densely without many pauses.

As seen in Fig. 2, nine test items were created from each of the two samples. For each sample, the following signals were created. A high-quality *reference* signal was synthesized from the original speech recording using the SASC method [14] (indicated as "Reference" in Fig. 2). SASC was also applied to the non-masked sinusoidal case ("No mask SASC"). Both OLA and interpolation sinusoidal synthesis were applied ("int", "ola" in Fig. 2) with and without masking. For all the sinusoidally modeled signals so far, 80 components were used for the non-masked analysis, and 50 components were used for the masked analysis. To test the effect of the additive residual, items with 25 sinusoidal components per frame were interpolation-synthesized ("25" in Fig. 2), as the effect of this residual in the resulting quality is more evident when a lower number of sinusoids is used. For synthesizing this residual, the PLPC method with 20 prediction coefficients was used for creating the monaural version of the residual signal; this was then binaurally modified (see Section 3.3). As a low-quality reference (anchor signal), the interpolation-synthesized signal using 25 sinusoids was filtered with rectangular-window OLA using the desired HRTFs ("Anchor"). Although suitable for cases with constant filters, here it resulted in notable frame discontinuities due to the changing HRTF information.

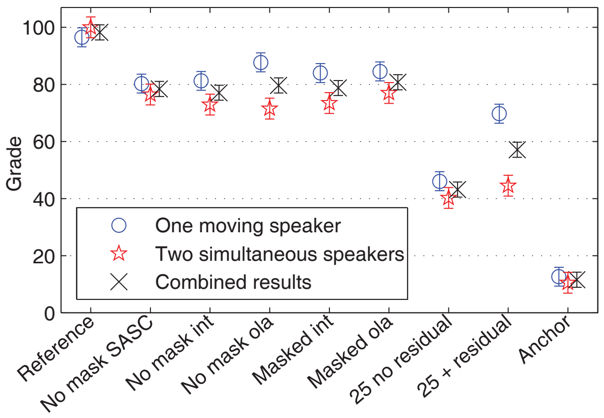

Fig. 2. Results from the MUSHRA listening test.

4.2. Procedure

The listening tests were implemented using the MUSHRA methodology. The Sennheiser HD650 headphones were used, and their response was compensated. Eight volunteers (authors not included) participated in the test. The listening task was to evaluate the quality of each item compared to the reference signal in overall quality and in spatial quality. Each subject performed two repetitions of the two test cases, and these results were averaged.

4.3. Results

In Fig. 2, the mean scores and the standard errors of the means for all test items of both samples and combined results are given. The reported significant differences between the mean values are based on multiple-comparison hypothesis tests using Tukey's HSC criterion.

As expected, good quality (reference) and low quality (anchor) signals occupy the opposite ends of the scale. It can be seen that both sinusoidal synthesis methods perform similarly to SASC STFT synthesis. Surprisingly, OLA synthesis has some variation between the results of the two samples. Although not statistically significant, this variance is probably caused by the fact that this method smooths the signal more than the interpolation synthesis. We plan to study the interpolation synthesis further without frame overlap to decrease bitrate, as its quality is shown to be good here.

The effect of the monaural masking model can be seen to not introduce any perceivable degradation in the binaural synthesis. Similar to the monaural case, the masked results do not differ significantly in quality from the non-masked results with either synthesis method.

The overall effect of the residual is a significant quality improvement in the low-component modeling case of the moving listener, although the effect is not significant in the sample with two simultaneous speakers. This is deemed to be caused by the dense nature of the sound in this sample that makes it difficult to notice small distortions. Nevertheless, the residual is valuable for most practical situations when a low number of sinusoidal components are used for modeling speech signals.

5. CONCLUSIONS

In this paper, the use of sinusoidal modeling in binaural synthesis for spatial audio coding was examined. The proposed system is mainly useful for speech-based applications, such as teleconferencing, thus speech signals were used for testing. Results show that the sinusoidal model can achieve good quality reproduction free of artifacts, and improve synthesis efficiency when using HRTFs. The monaural masking model for the sinusoidal analysis of the downmix signal was found to not introduce any degradation in binaural synthesis. Finally, the proposed sinusoidal plus noise model proved to be useful in binaural synthesis.

6. REFERENCES

[1] J. Herre et al, "Spatial audio coding: Next-generation efficient and compatible coding of multi-channel audio," Presented at the AES 117th Convention, San Francisco, CA, USA, 2004.

[2] ISO/IEC 23003-1:2007, "Information technology – MPEG audio technologies – part 1: Mpeg surround," Tech. Rep., International Organization for Standardization, Geneva, Switzerland, 2007.

[3] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, June 2007.

[4] T. Hirvonen, J. Ahonen, and V. Pulkki, "Perceptual compression methods for metadata in directional audio coding applied to audiovisual teleconference," Presented at the AES 126th Convention, Munich, Germany, 2009.

[5] J. Ahonen, V. Pulkki, and T. Lokki, "Teleconference application and b-format microphone array for directional audio coding," in *AES 30th International Conference: Intelligent Audio Environments*, 2007.

[6] M. V. Laitinen, "Binaural reproduction for directional audio coding," M.S. thesis, Helsinki University of Technology, 2008.

[7] M. Goodwin and J. M. Jot, "Spatial audio scene coding," Presented at the AES 125th Convention, San Francisco, CA, USA, 2008.

[8] M. Goodwin, "Multichannel matching pursuit and applications to spatial audio coding," *Asilomar Conf. on Signals, Systems, and Computers*, 2006.

[9] A. Ihlefeld and B. G. Shinn-Cunningham, "Spatial release from energetic and informational masking in a divided speech identification task," *J. Acoust. Soc. Am.*, vol. 123, no. 6, 2008.

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

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

[12] R. C. Hendriks, R. Heusdens, and J. Jensen, "Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, (ICASSP '04)*, 2004.

[13] J. Blauert, *Spatial Hearing*, The MIT Press, Cambridge, MA, USA, revised edition, 1997.

[14] M. Goodwin and J. M. Jot, "Binaural 3-D audio rendering based on spatial audio scene coding," Presented at the AES 123th Convention, New York, NY, USA, 2007.

[15] D.J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1637–1647, 1991.