EFFICIENT REPRESENTATION OF SPATIAL AUDIO USING PERCEPTUAL PARAMETRIZATION

Christof Faller and Frank Baumgarte

Media Signal Processing Research, Agere Systems
600 Mountain Avenue, Murray Hill, NJ
{cfaller,fb}@agere.com

ABSTRACT

We introduce a new scheme for simultaneous placement of a number of sources in auditory space. The scheme is based on an assumption about the relevance of localization cues in different critical bands. Given the sum signal of a number of sources, i.e. a monophonic signal, and a set of parameters (side-information) the scheme is capable of generating a binaural signal by spatially placing the sources contained in the monophonic signal. Potential applications for the scheme are multi-talker desktop conferencing and audio coding. Preliminary experimental results suggest that the listener's ability to identify messages in a multi-talker environment significantly improves by enhancing a monophonic signal with the proposed scheme.

1. INTRODUCTION

Sound from a single freefield source reaches the two ears of a listener with an interaural level difference (ILD) and an interaural time difference (ITD). The ILD and ITD determine the perceived lateral position of a sound source in the horizontal plane. A more general description of localization cues for 3D audio is the direction-dependent transfer function of sound to the eardrum (head related transfer function HRTF [1]).

A monophonic recording of one sound source can be processed such that when reproduced over headphones the sound source is spatially placed by providing the sound localization cues (ILD, ITD, or HRTF) to the ear [2]. This process is shown in Fig. 1 and is called *binaural synthesis* (a binaural signal is defined as the two sound pressure signals at the eardrums of a listener).

Given a number of separated source signals, each source m ($1 \leq m \leq M$) can be spatially placed arbitrarily by appropriate binaural synthesis. Binaural synthesis is applied to each source m with spatial cues c_m chosen such that the source is placed at the desired location. The spatial cues are ILDs and ITDs, or HRTFs for 3D audio. If all the resulting binaural signals are mixed to one binaural signal, then the resulting synthesized signal consists of sound sources with individual spatial locations. In this paper we call a binaural signal generated as described above an *ideally synthesized binaural signal*. For example, the decoder of MPEG-4 Structured Audio [3] can spatially place sources in this way, using the separately decompressed or synthesized signals. The disadvantage of such a scheme is that it requires each of the source signals separately. Music signals usually consist of many sources which as separated source signals would require a large amount of storage. Therefore, such a scheme is of limited use for low-bitrate transmission and compression applications. Moreover, in many cases the separated source signals are not available (e.g. existing recordings).

In contrast to previous schemes [2, 3] the new scheme we are proposing does not need to transmit the separated source signals to the receiver for binaural synthesis. Instead, the new scheme depends only on one sum signal with additional side information (*spatial parameters*) as shown in Fig. 2. To achieve this, we approach the problem from the receiver perspective, i.e. we model only perceptually relevant spatial cues. We refer to the resulting synthesized binaural signal as a *perceptually synthesized binaural signal*. The perceived locations of sources are based on localization cues in the binaural signal. We make the following assumption about the relevance of localization cues in different critical bands:

> ASSUMPTION:
> The more the energy of a source in the sum signal dominates in a critical band the more perceptually relevant are the localization cues in that band. If several sources share the same localization cues they are treated as one source.

The consequence of the assumption is that the spatial locations of different sources in a binaural signal can be approximated by taking the sum signal of all source signals, i.e. a monophonic signal, and synthesizing the cues accurately only in the critical bands in which the energy from one source is dominant.

For example, in the case of spatially placing three sources, the dominant bands relative to their short-term power spectra at a given

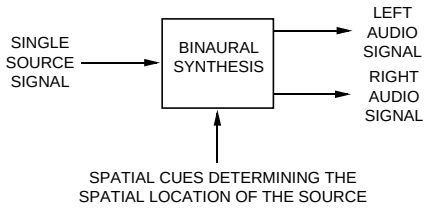

Figure 1: Synthesis of the spatial placement of one source.

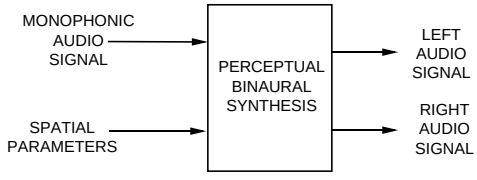

Figure 2: A scheme for the perceptual synthesis of a binaural signal given a monophonic signal and spatial parameters.

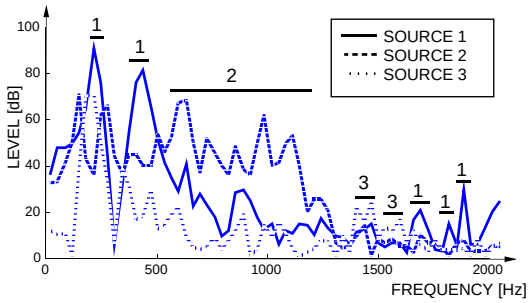

Figure 3: Power spectrum of 3 sources. Dominant bands are indicated with horizontal bars.

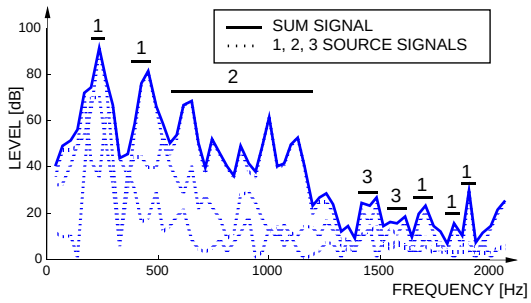

Figure 4: Power spectrum of the sum of the three sources of Fig. 3.

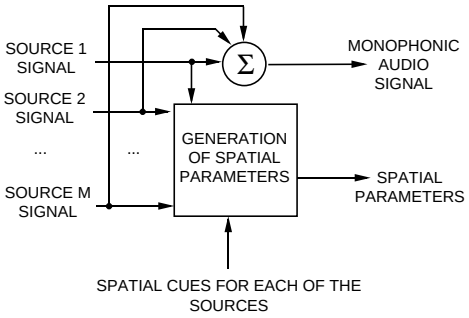

Figure 5: A scheme for the generation of the input of the perceptual synthesis scheme of Fig. 2 (the monophonic signal and the spatial parameters).

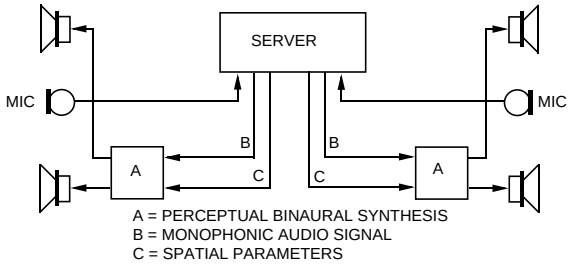

Figure 6: Desktop conferencing is an application for the perceptual synthesis of binaural signals.

time are shown in Fig. 3 (1, 2, 3). The short-term power spectrum of the sum of the three source signals is shown in Fig. 4. In this case, the binaural signal is synthesized by taking the sum signal and introducing in each critical band an ILD or ITD corresponding to the localization cues c_m of the source m which is dominant in that band (1, 2, and 3 in Fig. 4). In the case of auralization with HRTFs the binaural signal is synthesized by filtering the sum signal with left and right composite HRTFs. The left composite HRTF frequency response is computed by selecting for each critical band the left HRTF of the dominant source m. The right composite HRTF is constructed analogously. The resulting spatial locations of the sources are assumed to be perceived as being approximately the same as for an ideally synthesized binaural signal with the same corresponding cues but applied over the whole spectrum of each separated source. The scheme for generating the monophonic signal and the spatial parameters is shown in Fig. 5. An algorithm for obtaining spatial parameters is described in Section 2.2.

We also implemented a scheme to obtain the spatial parameters from a binaural signal without the necessity of having given the separated source signals. By examination of the cross-correlation function for each critical band, it is possible to obtain the spatial parameters. It is out of the scope of this paper to describe this scheme. Preliminary experiments suggest that the performance of this analysis scheme is nearly as good as when the separated source signals are given.

Figure 6 shows an application for the schemes of Figs. 2 and 5 for desktop conferencing. For convenience, the scheme is shown for only two clients, but it can be easily extended to more clients. Each client transmits an audio signal to the server. The server consists of schemes as shown in Fig. 5 to generate for each client the sum signal of all other clients and spatial parameters. The server generates the spatial parameters such that each participant is located in auditory space as desired. The spatial synthesis scheme of each client generates the left and right signals. As opposed to previous approaches [4], only a monophonic signal and a small number of spatial parameters need to be transmitted from the server to the participants instead of a binaural signal. In addition, each client can optionally place the participants individually by modifying the spatial parameters. The results presented in Section 3 suggest that for such a system a listener's ability to identify messages in a multi-talker system is significantly improved over the case of only monophonic playback.

In this paper we specifically address the spatial placement of sound sources by synthesis of a binaural signal using headphones. But similar techniques can most likely be applied for the synthesis of stereo or multi-channel signals for loudspeaker playback.

In Section 2 we describe in detail how we perceptually synthesize a binaural signal given a monophonic signal and spatial parameters. In Section 3 we compare a listener's ability to identify messages in a multi-talker communication environment scenario of diotic signals, perceptually synthesized binaural signals, and ideally synthesized binaural signals. Some conclusions are drawn in Section 4.

2. PERCEPTUAL SYNTHESIS OF BINAURAL SIGNALS

The scheme for perceptually synthesizing binaural signals is shown in Fig. 7. The given monophonic audio signal is first converted to the spectral domain. From the monophonic signal, the spectral coefficients of the binaural signal are obtained by modification of the monophonic spectrum. To obtain the perceptually synthesized bin-

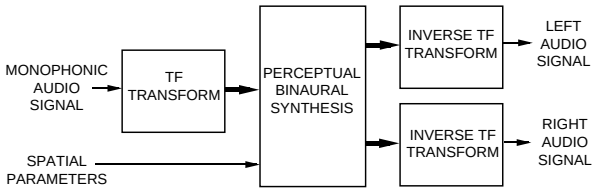

Figure 7: The scheme for the perceptual synthesis of binaural signals.

aural signal these spectra are transformed back to the time domain.

2.1. The Time-Frequency Transform

The signal is transformed to the spectral domain frame-wise since we aim for a system suitable for real-time applications. We would like to be able to introduce for each frequency band n at each time k (frame number) a level difference ΔL_n or time difference τ_n into the underlying audio signal, or implement HRTF filtering. For that purpose we use a DFT based transform. We choose the transform based on the desire of synthesizing frequency-dependent and time-adaptive time differences τ_n. It is shown that the same transform can be readily used for the synthesis of frequency-dependent and time-adaptive level differences ΔL_n and HRTF filtering.

When W samples $s_0, \ldots, s_{W-1}$ are converted to a complex spectral domain $S_0, \ldots, S_{W-1}$ with a DFT, a circular time shift of d time domain samples can be obtained by modifying the W spectral values by $\hat{S}_n = S_n e^{-\frac{2\pi n d}{W}}$. However we are not interested in a circular time shift within the frame. To achieve a time shift without time aliasing due to the circular shift we pad the samples $s_0, \ldots, s_{W-1}$ with Z zeros at the beginning and the end of each frame and then use a DFT of size $N = 2Z + W$. By modifying the resulting spectral coefficients a time shift within the range $d \in [-Z, Z]$ can be implemented by modifying the resulting N spectral coefficients according to $\hat{S}_n = S_n e^{-\frac{2\pi n d}{N}}$.

The described scheme works perfectly as long as the time shift d is not varied over time. Since the desired d usually varies over time we have to smooth the transitions by using overlapping windows for the analysis transform. A frame of N samples is multiplied with the analysis window before an N-point DFT is applied. We use the following analysis window which includes zero padding at the beginning and at the end,

$$w_a[i] = \begin{cases} 0 & \text{for} \quad i < Z \vee i \geq Z + W \\ \frac{1}{4}\sin^2(\frac{(i-Z)\pi}{W}) & \text{for} \quad Z \leq i < Z + W \end{cases}, \tag{1}$$

where Z is the width of the zero region before and after the window. The non-zero window span is W and the size of the transform is $N = 2Z + W$. Adjacent windows are overlapping and shifted by $W/8$ samples. We chose such a high overlap to increase oversampling and thus reduce frequency domain aliasing which occurs when modifying the complex spectrum. The window was chosen such that the the overlapping windows add up to a constant value of one. For our experiments we chose $W = 256$, $Z = 128$, and $N = 512$ for a sample rate of 32 kHz.

The zero padding of the window we use (1) allows the implementation of filtering with HRTFs as simple multiplications in the frequency domain. Therefore, the transform is also suitable for applying HRTFs and level differences. For a discussion of such time-frequency transforms the reader is referred to [6].

2.2. Obtaining the Spatial Parameters

For separated sources, the spatial parameters can be derived as follows. The spatial cues c_m associated with each source m are interaural level differences ΔL_m, interaural time differences τ_m, or HRTFs. A source m is considered to be dominant in a critical band if its power is at least larger by $T\ dB$ (e.g. $T = 3$ dB) than the power of the second strongest source. For each critical band b the spatial cues of the dominant source m are chosen,

$$C_b = c_m \text{ with } m = \arg \max_{1 \leq i \leq M} \{P_{bi}\}, \tag{2}$$

where P_{bi} is the power of source i in the critical band b. If none of the sources is dominant, spatial cues are chosen for a location in the middle between left and right, $C_b = c_{center}$. The spatial cues which are applied to each spectral coefficient n are obtained by interpolation over frequency of the spatial cues C_b between the center frequencies of adjacent critical bands. For each band n this results in an individual interaural level difference ΔL_n, interaural time difference τ_n, and in the case of HRTFs in a complex value for left and right, H_n^L and H_n^R, which is composed of the values of the HRTFs of the dominant sources. The presented algorithm for obtaining spatial parameters has low complexity and works well.

2.3. Obtaining the Binaural Signal

Given the spectral coefficients of the monophonic signal $\{S_n\}$ the level differences ΔL_n are applied such that the perceived loudness of the synthesized binaural signal is (approximately) independent of ΔL_n. The time differences τ_n are applied symmetrically by shifting by $\tau_n/2$ and $-\tau_n/2$ to obtain the left and right spectra:

$$\begin{aligned} S_n^L &= \frac{10^{\frac{\Delta L_n}{20}}}{\sqrt{1 + 10^{\frac{\Delta L_n}{10}}}} S_n e^{-\frac{2\pi n \tau_n}{2N}} \\ S_n^R &= \frac{1}{\sqrt{1 + 10^{\frac{\Delta L_n}{10}}}} S_n e^{\frac{2\pi n \tau_n}{2N}}. \end{aligned} \tag{3}$$

$\{S_n^L\}$ and $\{S_n^R\}$ are the spectral coefficients of the resulting binaural signal. The level differences $\{\Delta L_n\}$ are expressed in dB and the time differences $\{\tau_n\}$ in sampling intervals. For the perceptual synthesis of binaural signals based on HRTFs, the left and right spectra of the binaural signal are obtained by

$$S_n^L = H_n^L S_n \text{ and } S_n^R = H_n^R S_n. \tag{4}$$

The level differences ΔL_n, time differences τ_n, H_n^L, and H_n^R must be smoothed in time to prevent blocking artifacts in the resulting binaural signal.

3. EXPERIMENTAL RESULTS

To evaluate how useful the proposed method is for a desktop conferencing application (Fig. 6), we gave 12 participants a task which required responding to one of two simultaneous voice messages. This is a variation of the "cocktail party effect" of attending to one voice in the presence of others. The signals were presented to the participants with headphones in an acoustically isolated room. Five different signal kinds were tested for their effect on the ability to respond to one of two simultaneous messages:

1. $diotic$: monophonic signal to both ears
2. ILD_i: ideally synthesized binaural signal with ILDs
3. ITD_i: ideally synthesized binaural signal with ITDs
4. ILD_p: perceptually synthesized binaural signal with ILDs
5. ITD_p: perceptually synthesized binaural signal with ITDs

Each of the participants took all of the tests in randomized order. For the test we used the speech corpus introduced in [7]. Similar tests have been conducted by other authors [7–9]. A typical sentence of the corpus is "READY LAKER, GO TO BLUE FIVE NOW", where LAKER is the call sign and BLUE FIVE is a color-number combination. The possible eight different call signs, four colors, and eight numbers were chosen randomly with the restriction that the call sign assigned to the participant occurred in $50\ \%$ of the cases. In the test the participants were instructed to respond when their call sign was called by indicating the color-number combination by the talker who called their call sign. One out of four female voices was randomly chosen for each of the two talkers in each item of each test. One talker was spatially placed at the right side and the other at the left side for ILD_i and ILD_p (ILD $= \pm 16\ dB$) and ITD_i and ITD_p (ITD $= \pm 500\ \mu s$). Each of the five tests consisted of 10 training items followed by the 20 test items.

Table 1 shows the results for the case when the listener was called ($50\ \%$ of the cases). These results suggest that the percentage of correct identification of the call sign and of the color and number significantly improve for ideally synthesized binaural signals or perceptually synthesized binaural signals over the diotic signal. The perceptually synthesized signals (ILD_p and ITD_p) are almost as good as the ideally synthesized signals (ILD_i and ITD_i). For the case when the listener was not called, the percentages of the listeners responding was below two percent for all tests. The improvement of the binaural cases over the mono case may be explained by informational masking that is reduced by differences in perceived locations of the talkers [5].

	$diotic$	ILD_i	ITD_i	ILD_p	ITD_p
call sign	70 %	78 %	85 %	77 %	78 %
color-number	64 %	98 %	88 %	96 %	91 %

Table 1: Results for the case when the listeners were called by their call sign. The upper row shows the percentage of correct identification of the call sign and the lower row shows the conditional percentage of the correct color-number combination given that the listener's call sign was correctly identified.

4. CONCLUSIONS

In this paper we proposed a new scheme for the simultaneous placement of a number of sources in auditory space. As opposed to previous schemes, the sources are placed taking into account the receiver properties (auditory system), making an assumption about the relevance of localization cues in different critical bands. Given the sum signal of a number of source signals, i.e. a monophonic signal, and a set of parameters, the scheme is capable of individually placing the sources in auditory space by generating a binaural signal which incorporates the cues which are relevant for the perception of the source locations. We refer to a binaural signal generated in this way as a *perceptually synthesized binaural signal* as opposed to an *ideally synthesized binaural signal* which is generated applying spatial cues to each of the separated source signals individually.

Applications for the proposed scheme are desktop conferencing and audio coding. Using the proposed scheme, existing monophonic conferencing systems can be upgraded to stereo conferencing systems in a backwards compatible manner if inclusion of additional side-information is supported. The experimental results suggest that a listener's ability to identify messages in a multitalker environment using the perceptually synthesized binaural signal is much better than for a diotic signal and nearly as good as an ideally synthesized binaural signal.

The proposed scheme is very robust for speech signals. Preliminary experiments demonstrated that even complex signals such as music can be spatialized using the proposed scheme. Future work will focus on binaural analysis for obtaining the spatial parameters, and extension of the proposed scheme for multi-channel speaker playback.

We thank the authors of [7] for making their speech corpus available to us. Many thanks to Jiashu Chen, Oded Ghitza, Joe Hall, Yair Shoham, and Martin Vetterli for the inspiring discussions. Thanks to the participants of the test.

5. REFERENCES

[1] J. Blauert, *The Psychophysics of Human Sound Localization*, MIT Press, 1983.

[2] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*, Academic Press, Cambridge, MA, 1994.

[3] ISO/JTC1 SC29 WG11, *Subpart 5: MPEG-4 Structured Audio*, Oct 1998, Final Commitee Draft FCD 14496-3: Coding of Audiovisual Objects, Part 3: Audio.

[4] J. Benesty, D. R. Morgan, J. L. Hall, and M. M. Sondhi, "Synthesized stereo combined with acoustic echo cancellation for desktop conferencing," *Bell Labs Tech. J.*, vol. 3, pp. 148–158, July-Sept. 1998.

[5] R. L. Freyman, K. S. Helfer, D. D. McCall, and R. K. Clifton, "The role of perceived spatial separation in the unmasking of speech," *J. Acoust. Soc. Am.*, vol. 106, no. 6, pp. 3578–3588, December 1999.

[6] J. B. Allen, "Short-term spectral analysis, synthesis and modification by discrete fourier transform," *IEEE Trans. On Speech and Signal Processing*, vol. ASSP-25, pp. 235–238, June 1977.

[7] R. S. Bolia, W. T. Nelson, M. A. Ericson, and B. D. Simpson, "A speech corpus for multitalker communications research," *J. Acoust. Soc. Am.*, vol. 107, no. 2, pp. 1065–1066, February 2000.

[8] R. S. Bolia, M. A. Ericson, W. T. McKinley, and B. D. Simpson, "A cocktail party effect in the median plane?," *J. Acoust. Soc. Am.*, vol. 105, pp. 1390–1391, 1999.

[9] W. Spieth, J. F. Curtis, and J. C. Webster, "Responding to one of two simultaneous messages," *J. Acoust. Soc. Am.*, vol. 26, no. 3, pp. 391–396, 1954.