IMMERSIVE RENDERING OF CODED AUDIO STREAMS USING REDUCED RANK MODELS OF SUBBAND-DOMAIN HEAD-RELATED TRANSFER FUNCTIONS

Pinaki Shankar Chanda[1] *and Sungjin Park*[2]

[1]LG Soft India, Bangalore, India e-mail: pinaki.s@lgsoftindia.com
[2]LG Mobile Handset R&D Center, Seoul, Republic of Korea e-mail: s_park@lge.com

ABSTRACT

We consider immersive rendering of a set of sound sources, apparently positioned at different locations in three-dimensional (3-D) space around a listener. Each sound source is associated with a subband-coded audio stream. A low-complexity rendering approach, based on reduced-rank models of subband-domain Head-related transfer functions (HRTFs), is proposed in this paper. HRTFs are approximated using a set of basis vectors in subband-domain. Immersive rendering of partially decoded audio stream is performed in subband-domain using the basis vectors. Perceptual objective error measurements and subjective test results indicate, despite low computational complexity, this technique is effective in rendering of coded audio streams in 3-D space.

1. INTRODUCTION

Immersive audio rendering can enhance visualization of 3-D multimedia information in applications like home entertainment, gaming, augmented and virtual reality. A prototype system of audio scene rendering for virtual and augmented reality has been reported in [1]. A sound source is simulated at apparent position by filtering the audio stream associated with the sound source with left and right ear HRTFs corresponding to the position of the sound source.

The audio streams associated with the sound objects are often available in coded format. AudioBIFS system, part of the Binary Format for Scene Description (BIFS) tool in MPEG Standard [2], allows construction of sound scenes with coded streaming audio along with 3-D spatialization. The audio streams are mostly transform or subband-coded [3]. Immersive rendering of sound objects, in this framework, consist of two computationally complex operations: decoding of coded audio streams and immersive rendering of the decoded audio streams. Often these operations are performed in the user terminals, where computing resources are premium. This paper proposes a low-complexity approach of rendering multiple subband-coded audio streams using approximated HRTF model.

While rendering subband-coded audio from multiple sound sources two important issues need to be addressed. Separate synthesis of subband-coded audio streams, associated with different sound sources, to time-domain signals needs should be circumvented and rendering of the audio streams should be done in a manner such that the complexity of rendering does not scale up with number of sound sources. Modeling HRTF, as separate transfer function for each direction, is not suitable as computational complexity scales up linearly with the number of sound sources. This presents a heavy computation burden since the orders of the HRTF filters are high. In multi-sound rendering scenario, reduced-rank modeling of HRTF is well suited since a fixed set of basis vectors, irrespective of the number of sound sources, can be used to filter the audio streams [4]-[6]. In this representation each HRTF is approximated as a weighted linear combination of a small set of basis vectors. The scalar weights are functions of the directions of the sound sources. In [7] subband audio signals are manipulated directly using subband-domain basis vectors to generate localized audio. The basis vectors are extracted from full-band HRTF dataset and transported to subband-domain. This approach circumvents the use of separate synthesis banks to convert subband-domain audio streams to time-domain signals. As the audio streams are rendered using a fixed set of basis vectors the complexity does not scale up with number of sound sources. Note that in this approach since the basis vectors are extracted from full-band HRTF dataset before transporting to subband-domain the complexity of rendering in the subbands remains high. This is explained as follows.

In a perfect (or near-perfect) reconstruction filter bank, the synthesis filter bank cancels the aliasing components, introduced by overlapping frequency responses of the analysis filters, to produce perfect (or near-perfect) reconstruction. When a time-domain filter is transported to the subband-domain it is necessary to maintain the relationship between the subband-domain signals so that the synthesis filter bank cancels the aliasing components despite their modification by the subband-domain filter. This ensures that there is transparent equivalence between time-

domain filtered signal and the signal filtered in subband-domain and reconstructed back using the synthesis filter bank. Refer [8] for a generic framework for conversion of time-domain filter to subband-domain. A time-domain filter $h[n]$ ($h(z)$ in z-domain) is transported to subband-domain in M-band critically sampled perfect (or near-perfect) reconstruction filter bank by constructing $(M \times M)$ size matrix $\mathbf{T}(z) = (t_{n,k}(z))$ for filtering the signals in M subbands. The $(n \times k)^{\text{th}}$ element of this matrix $t_{(n,k)}(z)$ is obtained from the $M-1^{\text{th}}$ polyphase component of product filter $H_n(z)h(z)F_k(z)$ where the analysis and synthesis filters are given by $H_n(z)$ and $F_n(z), 0 \le n \le M-1$ respectively. $\mathbf{T}(z)$ is a sparse matrix, only the filters in the main diagonal and adjacent sub-diagonals are non-zero since each subband overlaps only with few adjacent subbands. The subband-domain signals $\mathbf{u}(z) = (u_0(z) \quad \ldots \quad u_{M-1}(z))^T$ is filtered by $\mathbf{T}(z)$ to obtain $\mathbf{v}(z) = (v_0(z) \quad \ldots \quad v_{M-1}(z))^T = \mathbf{T}(z)\mathbf{u}(z)$ where

$$v_k(z) = \sum_{j=-J}^{J} t_{k,(k+j) \bmod M}(z) u_{(k+j) \bmod M}(z), 0 \le k \le M-1 \quad (1)$$

"mod" denotes the modulo operation. In pseudo-QMF filter bank, used in MPEG Layer I and Layer II audio compression [3], $J = 1$ since with each subband k only the adjacent subbands $(k-1) \bmod M$ and $(k+1) \bmod M$ overlap significantly. Hence in subband k the subband-filtered output samples are obtained by filtering the samples of adjacent subbands $(k-1) \bmod M$ and $(k+1) \bmod M$, not only the samples in subband k. Conversions of a time-domain filter $h[n]$ to subband-domain in this pseudo-QMF filter bank results in three subband-domain filters $t_{k,(k+j) \bmod M}(z), j = 0, \pm 1$ in each subband k. Modeling of a time-domain HRTF requires 5-7 basis vectors for reasonable approximation in full-band domain [5], [6]. Conversion of the basis vector set to subband-domain filters hence results in 15-21 filters in each subband. This is computationally complex for real-time implementations.

To obtain a lower complexity multi-sound rendering architecture, we propose a different HRTF modeling technique. Instead of extracting the basis vectors from full-band HRTF dataset and then transporting to subband-domain filters, in this paper the basis vectors are extracted from subband-domain HRTF dataset. The subband audio signals, obtained after partial decoding of coded audio streams, are rendered using the basis vectors in the subbands and finally combined using a synthesis filter bank to generate time-domain 3-D rendered audio. This enables significant reduction in complexity of implementation.

2. HRTF APPROXIMATION TECHNIQUE

The HRTF dataset is diffused-field equalized and critical band smoothed [9]. M-band pseudo-QMF filter bank used in MPEG Layer I and Layer II audio encoding is considered here. Let $h_{\text{L(R)}\,s}(z)$ be the left (or right) ear HRTF corresponding to the position $s, 1 \le s \le S$ in the 3-D space. $h_{\text{L(R)}\,s}(z)$ is converted to the subband-domain filter matrix $\mathbf{T}_{\text{L(R)}\,s}(z), 1 \le s \le S$ using the framework provided in [8]. From k^{th} row of $\mathbf{T}_{\text{L(R)}\,s}(z), s = 1, 2, \ldots, S$ the diagonal and sub-diagonal entries $(k, (k+j) \bmod M), j = 0, \pm 1$ are grouped together to construct a set of filters $T_{\text{L(R)}\,k}(z)$ as follows.

$$T_{\text{L(R)}\,k}(z) = \left\{ t_{\text{L(R)}\,s,(k,(k+j) \bmod M)}(z), j = 0, \pm 1, s = 1, 2, \ldots, S \right\} \quad (2)$$

Constructing the sets for each row $k, 0 \le k \le M-1$, M sets of subband filters $T_{\text{L(R)}\,k}(z), 0 \le k \le M-1$ are obtained. The k^{th} set $T_{\text{L(R)}\,k}(z)$ is subjected to principal component analysis in time-domain to extract C_k most dominant basis vectors $q_{\text{L(R)}\,k,i}, 0 \le i \le C_k - 1$ to model the subband-filters in the set. Refer [5] for principal component analysis of HRTF sequences. In [5] basis vectors are extracted from time-domain full-band HRTFs whereas in this paper we apply principal component analysis on subband-domain HRTF sequences. The subband-domain filters in each set are approximated as follows using these basis vectors.

$$t_{\text{L(R)}\,s,(k,(k+j) \bmod M)}(z) = \sum_{i=0}^{C_k - 1} w_{\text{L(R)}\,s,(k,(k+j) \bmod M),i} q_{\text{L(R)}\,k,i}(z) \quad (3)$$

The weight of the i^{th} basis vector $q_{\text{L(R)}\,k,i}(z)$ corresponding to the filter $t_{\text{L(R)}\,s,(k,(k+j) \bmod M)}(z), j = 0, \pm 1$ is denoted by $w_{\text{L(R)}\,s,(k,(k+j) \bmod M),i}$. The principal basis vectors and weights of the basis vectors corresponding to the HRTFs in the subband-domain are computed offline and they are used in the real-time rendering architecture.

3. RENDERING OF CODED AUDIO STREAMS

We consider S sound sources indexed by their positions $s, 1 \le s \le S$ in different positions of audio scene. $\mathbf{u}_s(z) = (u_{s0}(z) \ldots . u_{s\,M-1}(z))^T$ denotes M-band partially

decoded audio signal obtained from sound source s, $1 \le s \le S$. By partial decoding it is meant that the coded audio signal is dequantized and rescaled but not synthesized into time-domain signal using the synthesis filter bank. To simulate a sound source in position $s, 1 \le s \le S$ using the subband-domain filter matrix the M-band audio input $\mathbf{u}_s(z)$ is filtered with subband-domain HRTF matrix $\mathbf{T}_{\mathrm{L(R)}\,s}(z)$ to generate left (right) channel audio $\mathbf{v}_{\mathrm{L(R)}\,s}(z) = \left(v_{\mathrm{L(R)}\,s,0}(z) \quad \ldots \quad v_{\mathrm{L(R)}\,s,M-1}(z)\right)^T$. The k^{th} element of the vector $\mathbf{v}_{\mathrm{L(R)}\,s}(z)$ i.e. $v_{\mathrm{L(R)}\,s,k}(z)$ is the left (or right) channel 3-D rendered audio in subband k from sound source s.

$$v_{\mathrm{L(R)}\,s,k}(z) = \sum_{j=-1}^{1} t_{\mathrm{L(R)}\,s,(k,(k+j) \bmod M)}(z)\, u_{s(k+j) \bmod M}(z) \quad (4)$$

Combining the 3-D rendered subbands from all sound sources $1 \le s \le S$ we obtain (5).

$$v_{\mathrm{L(R)}\,k}(z) = \sum_{s=1}^{S} \sum_{j=-1}^{1} t_{\mathrm{L(R)}\,s,(k,(k+j) \bmod M)}(z)\, u_{s(k+j) \bmod M}(z) \quad (5)$$

The reduced-rank models of $t_{\mathrm{L(R)}\,s\,(k,(k+j) \bmod M)}(z), j = 0, \pm 1$ given in (3) are substituted in (5) to generate $v_{\mathrm{L(R)}\,k}(z)$ for $0 \le k \le M-1$. It can be realized in a manner such that the complexity of implementation does not increase with increase in the number of sound sources. The steps are described as follows.

For each subband $k, 0 \le k \le M-1$ the audio streams $x_{\mathrm{L(R)}\,(k,(k+j) \bmod M),i}(z), j = 0, \pm 1, 0 \le i \le C_k - 1$ are constructed by multiplying the subband audio from k, $k+1$, and $k-1$ subbands with the weights of the basis vector $q_{\mathrm{L(R)}\,k,i}(z), 0 \le i \le C_k - 1$ corresponding to the $\left(k,(k+j) \bmod M\right)^{\text{th}}$, $j = 0, \pm 1$ entries of subband-domain filter matrix $\mathbf{T}_{\mathrm{L(R)}\,s}(z), 1 \le s \le S$.

$$x_{\mathrm{L(R)}\,(k,(k+j) \bmod M),i}(z) = \sum_{s=1}^{S} w_{\mathrm{L(R)}\,s\,,(k,(k+j) \bmod M),i}\, u_{s(k+j) \bmod M}(z) \quad (6)$$

The weighted audio streams in overlapping subbands $(k+j) \bmod M, j = 0, \pm 1$ are added up as follows.

$$x_{\mathrm{L(R)}\,k,i}(z) = \sum_{j=-1}^{1} x_{\mathrm{L(R)}\,(k,(k+j) \bmod M),i}(z), 0 \le k \le M-1 \quad (7)$$

$x_{\mathrm{L(R)}\,k,i}(z), 0 \le i \le C_k - 1$ are filtered with the basis vectors to generate $v_{\mathrm{L(R)}\,k}(z)$ in each subband $k, 0 \le k \le M-1$.

$$v_{\mathrm{L(R)}\,k}(z) = \sum_{i=0}^{C_k - 1} x_{\mathrm{L(R)}\,k\,i}(z) q_{\mathrm{L(R)}\,k\,i}(z), 0 \le k \le M-1 \quad (8)$$

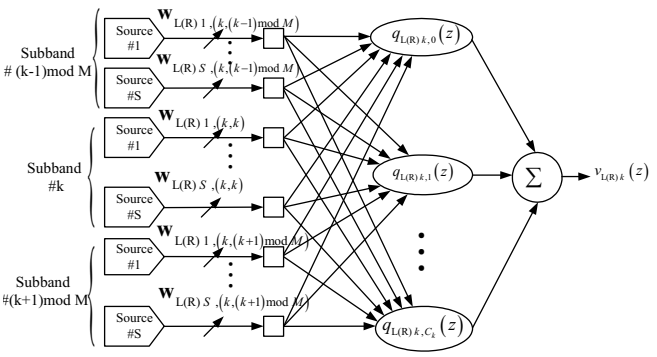

Figure 1. Rendering in subband $k, 0 \le k \le M-1$

Rendering of compressed audio streams in subband $k, 0 \le k \le M-1$ is illustrated in Figure 1. The advantage of the proposed architecture is that audio from subbands $k-1$, k, and $k+1$ are filtered with common set of basis vectors to produce output audio in subband k. Implementations of separate sets of basis vectors to render the audio from $k-1$, k, and $k+1$ subbands are necessary if the basis vectors are extracted from the time-domain HRTFs before coneversion to the subband-domain. $v_{\mathrm{L(R)}\,k}(z)$, $0 \le k \le M-1$ are combined using synthesis filter bank to generate 3-D rendered audio $v_{\mathrm{L(R)}}(z)$ for left (or right) channel. Note that, inter-aural time delay (ITD) is not extracted from the HRTFs while modeling. This circumvents the implementation of fractional delay filters in the subbands.

4. OBJECTIVE AND SUBJECTIVE EVALUATION

The HRTFs used in this study are obtained from KEMAR database [10] sampled at 44.100 KHz. The audio streams are assumed to be coded using 32-band pseudo-QMF filter bank of MPEG Layer I and Layer II endoding. In each subband 5 principal basis vectors of length 35 taps are used to model the subband-domain HRTF filters. Figure 2 shows the approximation of a pair of HRTFs using the proposed model. We also plot the HRTFs modeled using the algorithm in [5]. In [5] 7 basis vectors extracted from full-band time-domain minimum-phase HRTF sequences are used to approximate the HRTF sequences. Note that the approximation technique, proposed in this paper, is able to model the HRTFs with better precision and with lower computational complexity.

Perceptual RMS error metric [11] is obtained comparing the KEMAR HRTF spectra and the approximated HRTF spectra, both critical-band-smoothed and Bark scale warped. The mean RMS error is 0.2 dB considering left ear and right ear HRTFs. The RMS error in modeling the entire KEMAR HRTF set is less than 0.6 dB

suggesting a good fit of the proposed model with the KEMAR HRTF database. The distribution of the RMS error is shown in Figure 3(a).

Subjective tests, using double-blind triple-stimulus with hidden reference technique [11], are conducted using headphone listening experiments. A set of pink noise samples, of duration 1 second with 50 millisecond onset and offset time, localized at five different apparent source positions, are used as a test stimulus. Along with this, a set of sound clips consisting of sounds of musical instruments and speech signal, rendered together at different apparent source locations, is used. Subjects are asked to rate the localization similarity of each test stimulus rendered using proposed approximated HRTF models compared to the test stimulus rendered using original HRTFs on a continuous five point impairment scale given in Table I. Figure 3(b) shows the result of the subjective tests. The subjective tests produced a mean localization similarity grade of 4.64 across the set of listeners with a standard deviation of .40.

TABLE I. LOCALIZATION SIMILARITY SCALE

Grade	Localization Similarity
5	No difference
4	Very similar
3	Slightly similar
2	Slightly different
1	Very different

5. CONCLUDING REMARKS

A low-complexity approach of modeling HRTF dataset, using reduced rank approximations in subband-domain, has been presented in this paper. Based on this HRTF model, a multi-sound rendering architecture is proposed for coded audio streams. Total cost of implementation of the subband-based rendering is of 6 FIR filters of 35 coefficients at 44.100 KHz sampling rate. From simulations and subjective test results it is evident that the proposed model is very effective in design of audio displays containing a large number of sound sources.

6. REFERENCES

[1] D.N. Zotkin, R. Duraiswami, and L.S. Davis, "Rendering localized spatial audio in a virtual auditory display," IEEE Trans. Multimedia, vol. 6, no.4, August 2004

[2] Int. Std. (IS) 14496:2000. Information Technology—Coding of Audiovisual Objects (MPEG-4). Second Ed., ISO/IEC 14496, 2000.

[3] M. Bosi and R.E Goldberg, *Introduction to Digital Audio Coding and Standards*, Springer, 2002

[4] J. Chen, Van Veen, and K. E. Hecox., "A spatial feature extraction and regularization model for the head related transfer function", J. Acoustic. Soc. Am., 97. pp. 439-452.

[5] Zhenyang Wu , Francis H. Y. Chan, and F. K. Lam, "A time domain binaural model based on spatial feature extraction for the head related transfer functions", J. Acoust. Soc. Am. 102(4), pp. 2211-2218, October, 1997.

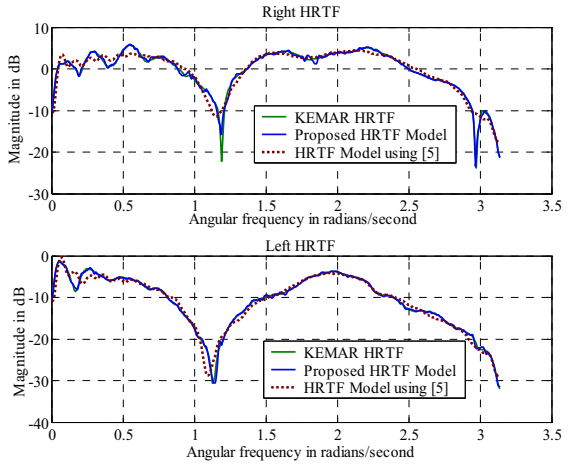

Figure 2. (a) Approximation of KEAMR HRTFs at 30 degree azimuth and 0 degree elevation.

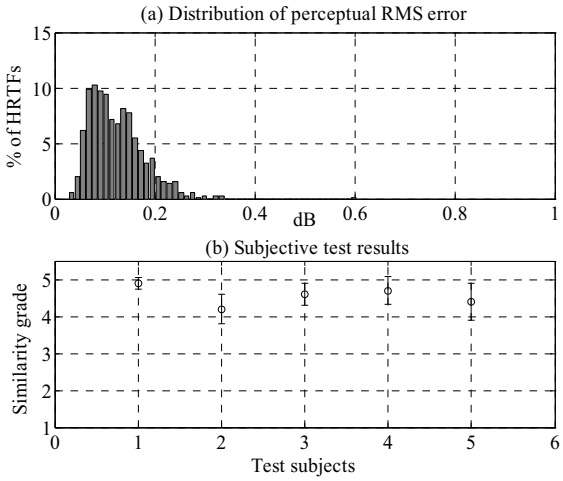

Figure 3. (a) Distribution of perceptual RMS error metric in modeling KEMAR HRTFs (including left and right ear HRTFs) (b) Subjective test results for a set of 5 listeners over the entire test stimuli set. This figure shows (mean grade ± standard deviation) for each listener.

[6] P.S. Chanda and S.J. Park, "Low Order Modeling for Multiple Moving Sound Synthesis using Head-related Transfer Functions' Principal Basis Vectors," Int. Joint Conf. on Neural Networks 2005 (IJCNN05), Montreal, Canada, 2005

[7] A. Benjelloun Touimi, M. Emerit, J.M. Pernaux ,"Efficient method for multiple compressed audio stream spatialization," In Proceedings of the 3rd international conference on Mobile and ubiquitous multimedia MUM '04, October, 2004

[8] A. Benjelloun Touimi, "A generic framework for filtering in subband domain," In Proceedings of IEEE 9th Workshop on Digital Signal Processing, Hunt, Texas, USA, October 2000

[9] J.M. Jot, V. Larcher, and O. Warusfel, "Digital signal processing issues in the context of binaural and transaural stereophony," 98th Audio Eng. Soc. Convention, France, February 1995.

[10] MIT Media Lab Machine Listening Group, "HRTF Measurements of a KEMAR Dummy-Head Microphone", obtained from http://sound.media.mit.edu/KEMAR.html.

[11] J. Huopaniemi, N. Zacharov, and M. Karjalainen, "Objective and Subjective Evaluation of Head-Related-Transfer Function Filter Design," 105th Audio Engineering Society Convention, preprint no. 4805, August 1998