Interpolation of Combined Head and Room Impulse Response for Audio Spatialization

Sanjeev Mehrotra, Wei-ge Chen, Zhengyou Zhang

Microsoft Research
Redmond, WA, USA
{sanjeevm, wchen, zhang}@microsoft.com

***Abstract*—Audio spatialization is becoming an important part of creating realistic experiences needed for immersive video conferencing and gaming. Using a combined head and room impulse response (CHRIR) has been recently proposed as an alternative to using separate head related transfer functions (HRTF) and room impulse responses (RIR). Accurate measurements of the CHRIR at various source and listener locations and orientations are needed to perform good quality audio spatialization. However, it is infeasible to accurately measure or model the CHRIR for all possible locations and orientations. Therefore, low-complexity and accurate interpolation techniques are needed to perform audio spatialization in real-time. In this paper, we present a frequency domain interpolation technique which naturally interpolates the interaural level difference (ILD) and interaural time difference (ITD) for each frequency component in the spectrum. The proposed technique allows for an accurate and low-complexity interpolation of the CHRIR as well as allowing for a low-complexity audio spatialization technique which can be used for both headphones as well as loudspeakers.**

I. Introduction

Realistic 3D audio is becoming increasingly important as immersive video conferencing and gaming is becoming popular. Audio spatialization is one of the key technologies in creating realistic 3D audio [1] and has been a subject of much study. Typically, audio spatialization is thought of as consisting of two components: 1) the head related transfer function (HRTF) which specifies the impulse response of sound coming from a particular location to each ear (left and right), and 2) the room impulse response (RIR) which consists of the direct path, early reflections, and reverberation. Each has been studied extensively on its own and there is a rich literature on each, [2], [3], [4], [5], [6], [7], [8], [9], [10] are just a few of the examples. There has also been work to show the usefulness of spatial audio in multi-party conferencing in disambiguating the various speakers [11], [12]. In addition, there has been work in the area of modeling the acoustic wave function at every point in space [13] which can be used to recreate an accurate rendition of the sound field (using some limited number of measurements) and in determining the quantity and placement of loudspeakers to accomplish this.

In audio spatialization, a combination of the listener's position and orientation (head pose) is used to create a realistic environment where the listener feels as if sounds are coming from various locations in a particular room. The spatialization is usually performed using a convolution with an impulse response (FIR filter) which consists of two portions:

- the room impulse response (RIR) which is a function of the room characteristics, the sound source location, and the listener's location and orientation.
- the head related transfer function (HRTF) which is a function the listener and the relative location between the sound source and the listener's location and orientation.

Although previous work has often treated these two separately, recent work has attempted to combine the two into a single combined head and room impulse response (CHRIR) [3], [8].

If the HRTF and RIR and modeled individually, then the HRTF is usually measured using a real head to simulate the response. The measurement takes place in an anechoic chamber, and thus the HRTF is a function of the only the *relative* distance and orientation between the sound source and the listener [4]. In this case, the RIR can either be measured for particular rooms [14] or modeled by direct path plus early reflections plus artificial reverberation [2].

If the CHRIR (combining both of these) is used, then the CHRIR is usually measured [3], and now the measurement is a function of the *actual* location and orientation of the sound source and the listener (not just the relative distance and orientation). Thus if we are only considering a two dimensional plane for the source and listener (i.e. we are not considering sound locations above and below the listener's plane), then the CHRIR is a function of four parameters whereas the HRTF is only a function of two.

Regardless of which approach is taken (HRTF+RIR or CHRIR), we see that measuring and storing the measurements for any decent sized room becomes prohibitively costly. For example, if we sample a 4mx4m room every 10cm, then 1600 measurements need to be taken resulting in 2560000 2-tuples for all combinations of source and listener positions, even if we ignore the listener's head orientation. Although the sampling procedure described in [13] proposes a method to find the minimal number of samples needed to accurately reconstruct the sound field, this number is still prohibitively high.

Therefore, it is necessary to find intelligent methods to interpolate the measured (or computed using some relatively complicated algorithm [13]) responses (HRTF, RIR, or CHRIR), especially when the sampling is significantly less than that proposed in [13]. In addition, in order to operate in real-time applications, they should be of low complexity. Therefore most schemes used in practice are not the perfect sinc interpolator as proposed in [13] which requires a large number of neighboring responses, but rather linear or bilinear interpolation schemes which only use a few samples to interpolate.

A. Related Work

There have been several interpolation schemes presented in the literature, some are described in [5], [6], [15], [7], [3], [8]. Although simple interpolation of the time domain response can be used for certain, very low complexity scenarios [8], it is known that straight-forward interpolation of the time domain response can result in destructive interference if the neighboring responses used in the interpolation are out of phase [15]. To solve this, several approaches have been proposed to do interpolation in the time domain. For example, in [7], a technique is proposed which first warps the neighboring responses so that the scale and shift of the responses are aligned. The aligned responses are then interpolated using simple linear interpolation. The vector of parameters used to perform the warping is also interpolated. The interpolated warp vector then re-warps the interpolated response. Other time domain interpolation techniques attempt to find the interaural level difference (ILD) and interaural time difference (ITD) between the left/right channels and makes sure that these two binaural cues are interpolated [3].

Frequency domain techniques have also been developed to perform HRTF interpolation [9], [5] and in [5] it is claimed that frequency domain techniques typically result in better performance then time domain techniques.

We note that previously developed techniques for interpolation have either targeted just interpolation of HRTFs [9], [5], [6], [15] or RIR [7]. Only very simple and low complexity interpolation techniques in the time domain have been studied for CHRIR. In particular, frequency domain interpolation techniques for the combined response (CHRIR) have not been studied which is the focus of this paper.

B. Contributions

In this paper, we present an effective, low-complexity *frequency domain* technique for interpolating the CHRIR. Since this technique interpolates the magnitude and phase of the CHRIR spectrum, it naturally preserves the ITD and ILD interpolation for each frequency component of the CHRIR. The interpolation is done in the polar coordinate system, and ensures that the power of the CHRIR is linearly interpolated in the angular direction and as the inverse of the square of the distance in the radial direction.

We also show how we can use this interpolation technique to perform audio spatialization using both headphones as well as loudspeakers. In particular, the CHRIR for both the *virtual* environment as well as the *real* environment is sampled and measured. Both of the CHRIRs for the real and virtual environment are interpolated using the same low-complexity interpolation technique. The interpolated CHRIRs for these two environments are combined and audio spatialization is performed directly in the frequency domain using the overlap-add method for convolution [16]. This allows for an effective audio spatialization using a limited amount of memory and computational resources.

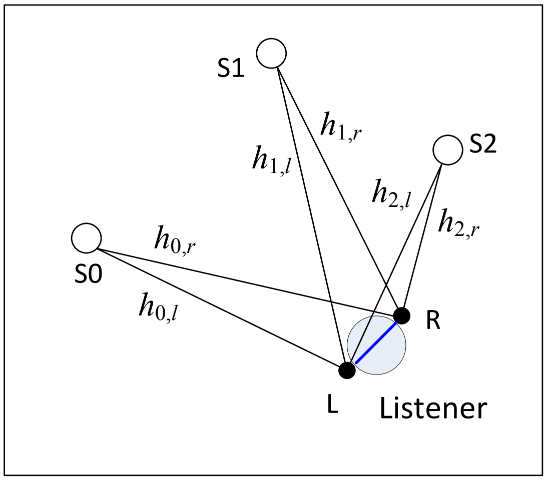


Fig. 1. Setup. There are three sound sources, labeled S_0, S_1, and S_2. The impulse response from each sound source is denoted by $h_{i,l}$ and $h_{i,r}$.

Although we show our results here for a two dimensional interpolation, i.e. the sound source and listener are located in a plane, they can easily be extended to the third dimension where the source and listener are at differing heights.

The remainder of this paper is organized as follows. In Sec. II, we explain the audio spatialization procedure. In Sec. III, the proposed interpolation procedure is presented. In Sec. IV, we show how the CHRIRs for the real and virtual environment are combined to perform audio spatialization using both headphones as well as loudspeakers. In Sec. V, we discuss implementation details and computational costs. In Sec. VI, we show results of the interpolation as well as spatialization, and conclude in Sec. VII.

II. Audio Spatialization

Suppose we wish to simulate a room with N sound source positions and the listener position and orientation as is shown in Fig. 1. Let x_i be the ith sound source ($i = 0, 1, \ldots, N-1$), $h_{i,l}$ be the CHRIR of length L from the location of source i to the left ear of the listener, and $h_{i,r}$ be the CHRIR to the right ear. The CHRIR is the combination of the HRTF and the RIR and is measured from particular sound locations for a given room. Then, the left and right channels of the output signal (denoted as y_l and y_r respectively) is given by,

$$y_l[n] = \sum_{i=0}^{N-1} \sum_{k=0}^{L-1} h_{i,l}[k] x_i[n-k] \quad (1)$$

$$y_r[n] = \sum_{i=0}^{N-1} \sum_{k=0}^{L-1} h_{i,r}[k] x_i[n-k]. \quad (2)$$

Utilizing previous work, we note that we can split the combined impulse response into two portions, one is the direct path and early reflections, which is dependent on direction, and the other is the room reverberation. This is similar to procedures described in [3], [7]. In addition to making it easier to interpolate the response, such a split also results in

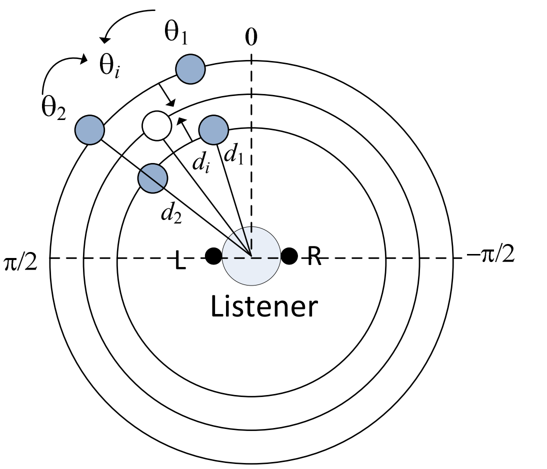


Fig. 2. Interpolation of CHRIR for sound source at a distance d_i and angle θ_i from the listener. The locations of the closest measured (or computed) CHRIRs are shown as solid circles. They are at distances of d_1 and d_2 from the center of the listener, and at angles of θ_1 and θ_2. Interpolation is done in the frequency domain using the polar coordinate system, first interpolating in the angular direction and then in the radial direction. For example, for the left channel, first, $H_{d_1,\theta_1,l}$ and $H_{d_1,\theta_2,l}$ are interpolated in the angular direction to find the response $H_{d_1,\theta_i,l}$. $H_{d_2,\theta_i,l}$ is found similarly. Then interpolation takes place in the radial direction to find $H_{d_i,\theta_i,l}$ from $H_{d_1,\theta_i,l}$ and $H_{d_2,\theta_i,l}$.

a lower computational complexity when performing the audio spatialization, and we can write

$$y_l[n] = \left(\sum_{i=0}^{N-1} \sum_{k=0}^{M-1} h^S_{i,l}[k] x_i[n-k] \right) + \sum_{k=M}^{L-1} h^L[k] \left(\sum_{i=0}^{N-1} x_i[n-k] \right), \tag{3}$$

where the first portion of the convolution (the h^S terms) are dependent on both the sound source location (function of i) and the the channel (left or right); and the second portion of the convolution (the h^L terms) are independent of these two. Here S stands for the "short" portion of the filter, and L represents the "long" portion. The spatialization for the right channel can be written similarly. In [3], there is a more detailed discussion of how this split is done.

III. CHRIR Interpolation

We consider the problem of interpolating the left and right channel CHRIR from locations other than those being measured. In previous work [8], we have attempted a simple linear time-domain interpolation. One of the issues with a simple time-domain interpolation is that the interpolation does not adequately take care of phase differences in the impulse response and can result in destructive interference. As an example, suppose that we are interpolating exactly half-way between two impulse responses (h_0 and h_1) which have a phase difference of 180° (that is $h_1 = -h_0$). Then, the simple time domain interpolation will give an impulse response which is zero, which is obviously incorrect.

As an alternative, we propose to perform the interpolation in the frequency domain, where we can independently interpolate both the magnitude and phase of the frequency response of the CHRIR. In the previous example ($h_1 = -h_0$), the interpolated response would now have the same magnitude as h_0 and h_1, but would simply be shifted by 90° (as opposed to being zero). Temporal domain interpolation techniques may be able to achieve the same effect, but would require complicated techniques to first perform temporal alignment. In the frequency domain, such phase differences are naturally taken care of without additional complexity. Thus the ILD and ITD are naturally interpolated for each frequency component, and the attenuation and phase shift experienced by a pure tone (sinusoid) is interpolated.

Since the room reverberation is assumed to be constant throughout the room and is captured by the h^L portion of the CHRIR, the interpolation is only done for the h^S portion of the CHRIR as described in Eqn. 3. Thus in the discussion in this section, when we refer to the impulse response or the CHRIR, we actually only refer to the h^S terms in Eqn. 3. For simplicity we drop the superscript S from the equations.

To perform the interpolation, we consider the setup as shown in Fig. 2. In this figure, we denote the distance from the sound source to the center of the listener as d, and denote the orientation of the listener relative to the sound source using θ. In this polar coordinate system, $\theta = 0$ is defined to be the direction directly in front of the listener and $d = 0$ is defined to be the center of the listener. Although the actual position and orientation of the listener is also important when finding the CHRIR, we can perform a second stage of interpolation to compensate for the actual position and orientation of the listener (similar to the one proposed here).

Suppose the sound source location and orientation that we wish to spatialize is given by distance d_i and θ_i in this coordinate system. Also, suppose we have measured (or computed using some other relatively complicated model) the CHRIR when the sound source is at locations given by the solid circles in the figure, that is at distances d_1 and d_2 from the center of the listener and at angles θ_1 and θ_2. That is we find measured distances d_1 and d_2 such that $d_1 \leq d_i < d_2$ while minimizing $d_i - d_1$ and $d_2 - d_i$. Similarly, we find θ_1 and θ_2 such that $\theta_1 \leq \theta_i < \theta_2$. If the CHRIR is not uniformly sampled in the coordinate system, then we can alternatively find the four closest points where the CHRIR has been measured. Bilinear interpolation is then performed in the frequency domain using these measured samples.

Define $h_{d_1,\theta_1,l}$, $h_{d_1,\theta_2,l}$, $h_{d_2,\theta_1,l}$, and $h_{d_2,\theta_2,l}$ to be the four CHRIRs from the measured points to the listener's left channel. Similarly, we can define the four CHRIRs to the right channel. We also define $H_{d_1,\theta_1,l}$, $H_{d_1,\theta_2,l}$, $H_{d_2,\theta_1,l}$, and $H_{d_2,\theta_2,l}$ to the discrete Fourier transform (DFT) of the impulse responses, that is the frequency response of the CHRIR.

We show the interpolation for the left channel, l, here. The right channel CHRIR can be found using the same method. We propose to do the interpolation in two stages, the first is to find the frequency response at angle θ_i at the distances d_1 and d_2. To do this, we do a linear interpolation of the angle and phase of each component, K, of the frequency response, that is we find $H_{d_1,\theta_i,l}[K]$ using

$$|H_{d_1,\theta_i,l}| = \alpha \left(|H_{d_1,\theta_1,l}| \frac{\theta_2 - \theta_i}{\theta_2 - \theta_1} + |H_{d_1,\theta_2,l}| \frac{\theta_i - \theta_1}{\theta_2 - \theta_1} \right), \tag{4}$$

$$\angle H_{d_1,\theta_i,l} = \angle H_{d_1,\theta_1,l} \frac{\theta_2 - \theta_i}{\theta_2 - \theta_1} + \angle H_{d_1,\theta_2,l} \frac{\theta_i - \theta_1}{\theta_2 - \theta_1}, \tag{5}$$

where α is chosen to maintain a linear interpolation of the overall power level. The subscript K has been dropped from $H_{d_1,\theta_i,l}[K]$. α is chosen so that

$$\|H_{d_1,\theta_i,l}\|^2 = \|H_{d_1,\theta_1,l}\|^2 \frac{\theta_2 - \theta_i}{\theta_2 - \theta_1} + \|H_{d_1,\theta_2,l}\|^2 \frac{\theta_i - \theta_1}{\theta_2 - \theta_1}. \tag{6}$$

In practice, $\alpha \approx 1$ in most cases. $H_{d_2,\theta_i,l}$ can also be found using the same method.

Once we have the frequency response at an angle of θ_i for the two distances d_1 and d_2, the next step is to interpolate between $H_{d_1,\theta_i,l}$ and $H_{d_2,\theta_i,l}$ to find $H_{d_i,\theta_i,l}$. We do this in the same manner as before, that is

$$|H_{d_i,\theta_i,l}| = \beta \left(|H_{d_1,\theta_i,l}| \frac{d_2 - d_i}{d_2 - d_1} + |H_{d_2,\theta_i,l}| \frac{d_i - d_1}{d_2 - d_1} \right), \tag{7}$$

$$\angle H_{d_i,\theta_i,l} = \angle H_{d_1,\theta_i,l} \frac{d_2 - d_i}{d_2 - d_1} + \angle H_{d_2,\theta_i,l} \frac{d_i - d_1}{d_2 - d_1}. \tag{8}$$

Here β is not chosen to simply linearly interpolate the power between d_1 and d_2, but is instead chosen so that the power level changes inversely proportional to the square of the distance, that is β is found so that

$$\|H_{d_i,\theta_i,l}\|^2 = \|H_{d_1,\theta_i,l}\|^2 \frac{d_1^2}{d_i^2} \frac{d_2 - d_i}{d_2 - d_1} + \|H_{d_2,\theta_i,l}\|^2 \frac{d_2^2}{d_i^2} \frac{d_i - d_1}{d_2 - d_1}. \tag{9}$$

To extrapolate distances outside the region of measured CHRIRs (for example if there is no d_2 such that $d_1 \leq d_i < d_2$), we can simply use the frequency response of the closest CHRIR measurement at distance d_1, $H_{d_1,\theta_i,l}$, and then scale the impulse response by $\frac{d_1}{d_i}$. The interpolation along the angular direction, θ_i, can always take place since there will always be two neighboring angles.

Since we are only interpolating the short portion of the filter, the $H_{d_i,\theta_i,l}$ found here is actually $H^S_{d_i,\theta_i,l}$. We can then find the frequency response of the overall CHRIR as $H_{d_i,\theta_i,l}[K] = H^S_{d_i,\theta_i,l}[K] + H^L[K]$, where $H^L[K]$ is the frequency response for the reverberation portion (the long filter). $H_{d_i,\theta_i,r}$ can be similarly found. The frequency response of the CHRIR is *directly* used in the frequency domain to perform the spatialization.

IV. Combining CHRIR for Real and Virtual environments

The audio spatialization shown above is valid for when the spatialized signals y_l and y_r are directly going to the left and right ears of the listener respectively. This is the case when the listener is using headphones. However, if the listener is playing the audio through loudspeakers as shown in the setup in Fig. 3, then the actual signal that needs to played through the loudspeakers is not y_l and y_r since the room response of the actual environment needs to be taken into account.

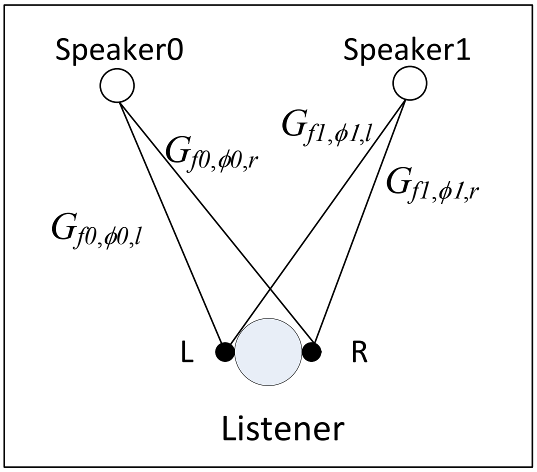


Fig. 3. Setup with loudspeakers. Each loudspeaker i is at a distance f_i and at an angle ϕ_i from the listener. The frequency response of the CHRIR is given by $G_{f_i,\phi_i,l}$ and $G_{f_i,\phi_i,r}$.

Let Y_l and Y_r be the transform domain representation of the signals that we want the listener to hear and let X_i be the transform domain of the sound sources, with $H_{d_i,\theta_i,l}$ being the frequency response of the desired CHRIR for the ith sound source. Then, we can write

$$Y_l = \sum_{i=0}^{N-1} H_{d_i,\theta_i,l} X_i, \tag{10}$$

$$Y_r = \sum_{i-0}^{N-1} H_{d_i,\theta_i,r} X_i. \tag{11}$$

Let $G_{f_0,\phi_0,l}$ and $G_{f_1,\phi_1,l}$ be the CHRIR of the *actual* room response and HRTF from each of two loudspeakers (0 and 1) being used to render the audio to the listener's left ear as shown in Fig. 3. Loudspeaker $i = 0, 1$ is at a distance of f_i and at an angle ϕ_i from the listener. This information can be obtained using head-tracking as in [17]. Similarly, Let $G_{f_0,\phi_0,r}$ and $G_{f_1,\phi_1,r}$ be the CHRIRs from loudspeakers to the the right ear. We can measure and interpolate these CHRIRs using exactly the same mechanism used to compute the CHRIR of the desired room environment. Let Z_0 and Z_1 be the transform domain of the signals that need to be rendered by the loudspeakers. Using this, we can write

$$Y_l = G_{f_0,\phi_0,l} Z_0 + G_{f_1,\phi_1,l} Z_1, \tag{12}$$

$$Y_r = G_{f_0,\phi_0,r} Z_0 + G_{f_1,\phi_1,r} Z_1. \tag{13}$$

Using Eqn. 11 and 13, we can solve to get

$$Z_0 = \sum_{i=0}^{N-1} \frac{G_{f_1,\phi_1,r} H_{f_i,\theta_i,l} - G_{f_1,\phi_1,l} H_{f_i,\theta_i,r}}{G_{f_1,\phi_1,r} G_{f_0,\phi_0,l} - G_{f_1,\phi_1,l} G_{f_0,\phi_0,r}} X_i \tag{14}$$

$$Z_1 = \sum_{i=0}^{N-1} \frac{G_{f_0,\phi_0,l} H_{f_i,\theta_i,r} - G_{f_0,\phi_0,r} H_{f_i,\theta_i,l}}{G_{f_1,\phi_1,r} G_{f_0,\phi_0,l} - G_{f_1,\phi_1,l} G_{f_0,\phi_0,r}} X_i. \tag{15}$$

We note that in case the user is using headphones, we can write $G_{f_0,\phi_0,l} = 1$, $G_{f_0,\phi_0,r} = 0$, $G_{f_1,\phi_1,l} = 0$, and $G_{f_1,\phi_1,r} = 1$, and thus $Z_0 = Y_l$, and $Z_1 = Y_r$. If the filters in Eqn. 15

are non-minimum phase, then we can use various methods to obtain the inverse, for example [18].

If there are multiple listeners in the actual room, additional loudspeakers can be used to create the realistic virtual environment at multiple locations. In particular for L listeners, we would need $2L$ loudspeakers to create a realistic virtual environment and solve the mapping between the two environments. We note that additional, more sophisticated cross-talk cancellation methods can be used as in [19]. Regardless, interpolation of the responses can still be used.

V. Implementation

The given interpolation and spatialization can be performed with relatively little computational and memory requirements. The DFT of the measured filters can be directly stored in memory. Suppose we store $H^S_{d_v,\theta_v,l}$ and $H^S_{d_v,\theta_v,r}$ for V virtual positions in the virtual environment at a distance d_v and orientation θ_v from the listener. We also store H^L. Similarly, we store the CHRIR for the actual room environment, $G^S_{f_a,\phi_a,l}$ and $G^S_{d_a,\phi_a,r}$ for A actual positions along with G^L. If the short filter is of length S, the long filter is of length L, and the total length $T = S + L$, then we need $(A^2 + V^2)S + 2L \approx (A^2 + V^2)S$ amount of memory for the storage. For example, if we quantize the room into 10 possible distance values and 10 possible angle values, and the length of the short filter is 100, then we need approximately 4MB of memory to store the filters (assuming 4 bytes per coefficient).

In order to perform the spatialization, we interpolate the filters directly in the frequency domain, and perform the spatialization via the overlap-add [16] method for performing convolution. An additional step is also needed to interpolate the frequency spectrum of the short filters from length S to length T. This additional interpolation step plus the interpolation from Eqn. 5-9 and the combining of the real and virtual CHRIRs from Eqn. 15 can be seen to be $O(T)$ every time we need to interpolate a new filter. If we perform convolution using frames of length M, then the majority of computation from the overlap add is in the forward and inverse FFT which requires $O(Mlog_2(M))$ complexity. Suppose that we wish to spatialize 16kHz audio using this method. Then, even if we update the filters every 2048 samples (125ms) and if the total filter is also of length 2048, and additionally we perform the overlap-add method on frames of length 2048, then the number of operations per sample is on the order of $c_1 log_2(2048) + c_2$, for some constant c_1, c_2, which is very manageable.

VI. Results

We show two sets of results. One is to show the accuracy of the interpolation method proposed. The other is to show the performance of audio spatialization when using the measured CHRIRs for a particular room to perform audio spatialization of a moving sound source.

We compare the results of this interpolation method with the simple time domain interpolation technique presented in [8]. Although previous interpolation schemes have mostly focused on interpolation of HRTFs, we realize these techniques could also potentially be extended to perform interpolation of CHRIRs. We leave comparison with these techniques as future work.

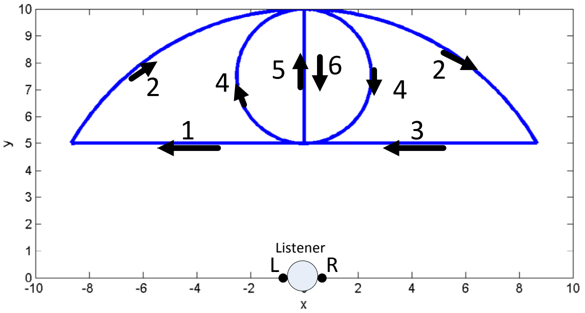


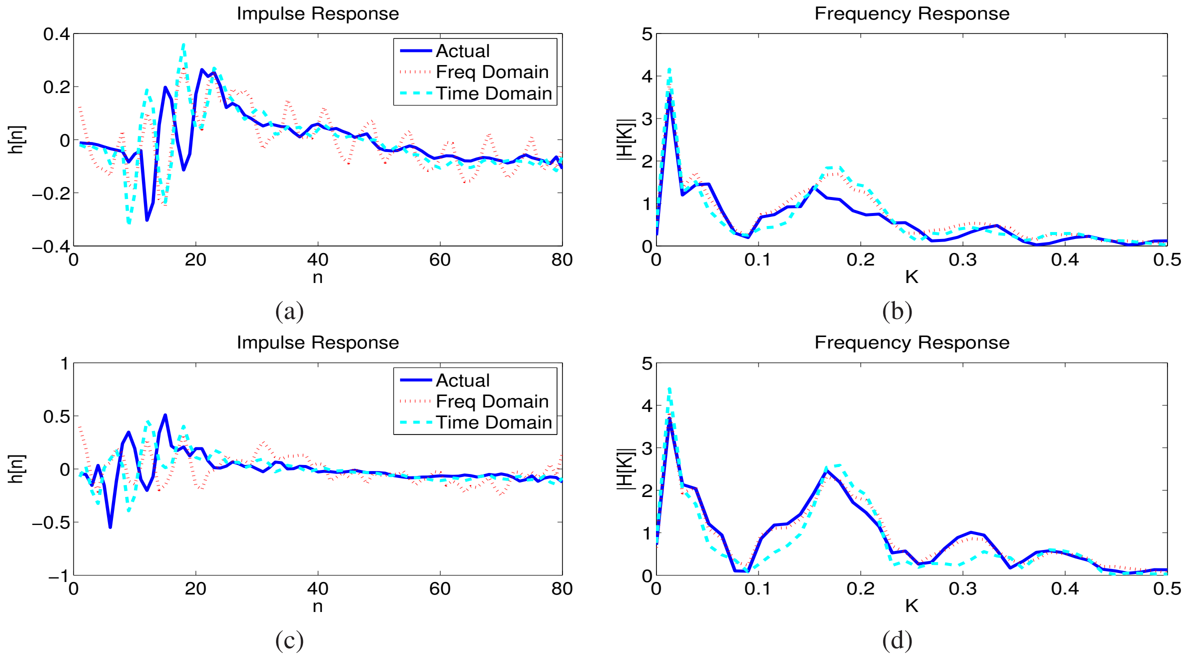
Fig. 4. Pattern used in creating audio clip. The listener is located at $(x, y) = (0, 0)$ and the audio source moves as shown.

For the first part, we compare an interpolated CHRIR with a measured CHRIR to show the accuracy of the interpolation scheme. Using the coordinate system from Fig. 2, we measure the CHRIR at an angle of 0, $\frac{-\pi}{4}$, and $\frac{-\pi}{2}$. In Fig. 5, we compare the measured CHRIR at an angle of $\frac{-\pi}{4}$ with interpolated versions. We compare the simple temporal domain interpolation scheme with the one presented here. We see that although the time domain response of both interpolation schemes is reasonably close to the measured one, the frequency domain response is much better using the scheme presented here. In fact the signal-to-noise (SNR) ratio when comparing the magnitude of the frequency response between the measured and interpolated versions, we see an improvement of over 9x (SNR of 90 vs. 11). The SNR is defined as $\sum_k \frac{\|H_a[k]-H_i[k]\|^2}{\|H_a[k]\|^2}$, where $H_a[k]$ and $H_i[k]$ are the kth frequency coefficient of the measured and interpolated CHRIRs respectively.

For the second part of the comparison, we generate an audio clip showing the spatialization when an audio source is moving in a pattern as shown in Fig. 4. The listener is located at $(x, y) = (0, 0)$ as shown in the figure and the source is initially at $(x, y) = (0, 5)$. The units for all the motion is in feet. Then the audio source moves according to the following at a rate of 1 feet/sec. The audio clip being spatialized consists of 8 clips of distinct 8 second segments for a total of 64 seconds.

1) The source moves left until it is at a distance of 10 from the listener $(x, y) = (-8.66, 5)$ for $\sim$ 8.66 seconds.
2) The source then moves in a circle around the listener at a distance of 10 until $(x, y) = (8.66, 5)$ for $\sim$ 21 seconds.
3) The source moves left again until $(x, y) = (0, 5)$ for $\sim$ 8.66 seconds.
4) The source moves around a circle with diameter 5 centered at $(0, 7.5)$ for 15.7 seconds.
5) The source then moves away from the listener from $(0, 5)$ to $(0, 10)$ for 5 seconds.
6) The source moves back towards the listener from $(0, 10)$ to $(0, 5)$ for 5 seconds.

The spatialized audio clips using both interpolation techniques can be heard at http://research.microsoft.com/en-us/um/people/sanjeevm/audiospatialization.



(a) (b) (c) (d)

Fig. 5. Impulse response and frequency response of CHRIR. In the figures, we show the actual response at an angle of $-\pi/4$, the interpolated response using simple temporal domain linear interpolation [8], and the interpolated response using the interpolation scheme presented here. (a) Impulse response of left channel (b) Frequency response of left channel (c) Impulse response of right channel (d) Frequency response of right channel.

VII. Conclusion

In this paper, we have shown a frequency domain interpolation scheme which can be used to effectively interpolate the CHRIR at source and listener location and orientations which have not been measured. The interpolation is a two step interpolation which interpolates both the magnitude and phase of the CHRIR in both the angular and radial directions. This interpolation is done for both the virtual environment and the real environment and can be used to perform audio spatialization using either headphones or loudspeakers. It is shown that such an interpolation and spatialization can give good results using reasonably low memory and computational resources making it suitable for immersive games and conferencing applications.

References

[1] J. R. Stuart, "The psychoacoustics of multichannel audio," in *Audio Engineering Society Conference: UK 11th Conference: Audio for New Media (ANM)*, 3 1996.

[2] M. R. Schroeder, "Natural-sounding artificial reverberation," *Journal Audio Engineering Society*, vol. 10, no. 3, pp. 219–233, 1962.

[3] W. ge Chen and Z. Zhang, "Highly realistic audio spatialization for multiparty conferencing using headphones," in *Proc. Workshop on Multimedia Signal Processing*. IEEE, Oct. 2009, pp. 1–6.

[4] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio*, Oct. 2001, pp. 99–102.

[5] K. Hartung, J. Braasch, and S. J. Sterbing, "Comparison of different methods for the interpolation of head-related transfer functions," in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*, Mar. 1999.

[6] F. Freeland, L. W. Biscainho, and P. S. R. Diniz, "Interpolation of head-related transfer functions (hrtfs): A multi-source approach," in *Proc. 2004 EUSIPCO- European Signal Processing Conference*, Sep. 2004, pp. 1761–1764.

[7] C. Masterson, G. Kearney, and F. Boland, "Acoustic impulse response interpolation for multichannel systems using dynamic time warping," in *Audio Engineering Society Conference: 35th International Conference: Audio for Games*, 2 2009.

[8] S. Mehrotra, W. ge Chen, Z. Zhang, and P. A. Chou, "Realistic audio in immersive video conferencing," in *Proc. Int'l Conf. Multimedia and Expo*, Jul. 2011.

[9] B. Carty and V. Lazzarini, "Frequency-domain interpolation of empirical hrtf data," in *Audio Engineering Society Convention 126*, 5 2009.

[10] J.-M. Jot, A. Philp, and M. Walsh, "Binaural simulation of complex acoustic scenes for interactive audio," in *Audio Engineering Society Convention 121*, 10 2006.

[11] J. Baldis, "Effects of spatial audio on memory, comprehension, and preference during desktop conferences," in *Proc. of SIGCHI Conference on Human Factors in Computing Systems*, vol. 3. ACM, Apr. 2001, pp. 166–173.

[12] M. Chignell and R. Kilgore, "Listening to unfamiliar voices in spatial audio: Does visualization of spatial position enhance voice identification?" in *Proc. of the International Symposia on Human Factors in Telecommunication*, Mar. 2006.

[13] T. Ajdler, L. Sbaiz, and M. Vetterli, "The plenacoustic function and its sampling," *IEEE Trans. Signal Processing*, vol. 54, no. 10, pp. 3790–3804, Oct. 2006.

[14] A. Marquez-Borbon. Room impulse response measurement and analysis. [Online]. Available: \url{https://ccrma.stanford.edu/∼adnanm/SCI220/Music318ir.pdf}

[15] F. Freeland, L. W. Biscainho, and P. S. R. Diniz, "Interpolation of head-related transfer functions (hrtfs): A multi-source approach," in *Proc. 2004 EUSIPCO- European Signal Processing Conference*, Sep. 2004, pp. 1761–1764.

[16] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1989.

[17] M.-S. Song, C. Zhang, D. Florencio, and H.-G. Kang, "Personal 3d audio system with loudspeakers," in *Proc. Int'l Conf. Multimedia and Expo*, Jul. 2010, pp. 1600 –1605.

[18] A. Mouchtaris, P. Reveliotis, and C. Kyriakakis, "Non-minimum phase inverse filter methods for immersive audio rendering," in *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, vol. 6, Mar. 1999, pp. 3077–3080.

[19] ——, "Inverse filter design for immersive audio rendering over loudspeakers," *Multimedia, IEEE Transactions on*, vol. 2, no. 2, pp. 77 –87, Jun. 2000.