REALISTIC AUDIO IN IMMERSIVE VIDEO CONFERENCING

Sanjeev Mehrotra, Wei-ge Chen, Zhengyou Zhang, Philip A. Chou

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

ABSTRACT

With increasing computation power, network bandwidth, and improvements in display and capture technologies, fully immersive conferencing and tele-immersion is becoming ever closer to reality. Outside of video, one of the key components needed is high quality spatialized audio. This paper presents an implementation of a relatively low complexity, simple solution which allows realistic audio spatialization of arbitrary positions in a 3D video conference. When combined with pose tracking, it also allows the audio to change relative to which position on the screen the viewer is looking at.

Index Terms— Audio spatialization, Telepresence, Video conferencing

1. INTRODUCTION

Immersive video conferencing and telepresence is taking off with solutions such as Cisco TelePresence and HP Halo finding wide deployment. Realistic audio is a very important component of an immersive video conferencing experience and high-end solutions already have some components of spatialized audio. This allows the user to feel as if the remote participants voices are coming from distinct locations in space and allows for better intelligibility and easier comprehension especially when multiple parties are speaking at once [1, 2].

Although some high-end systems allow the sound of each participant to come from a given fixed location[3], they currently do not create a true 3D sound field since the sound does not adjust to match each viewer's perceived viewpoint and location. To properly create a sound field which represents the correct sound at each location would require a large speaker array or headphones to give individualized renderings.

In addition, due to complexity concerns, most low-end video conferencing systems do not offer any sound spatialization at all. The sound rendered is simply a linear sum of the individual mono captures of the remote participants.

Sound spatialization is sometimes considered difficult even with headphones since many solutions attempt to truly represent the room and listener acoustics making the transfer function used in the spatialization a function of many parameters such as:

- the room,
- the sound source location where the remote participant is to be placed,
- the listener position and viewpoint, and
- the listener's head-related transfer function (HRTF) [4].

This creates a large number of transfer functions which have to be stored as well as makes the true spatialization procedure complex to perform. For example, for a large number of sound source positions and listener positions and viewpoints, we have to compute the room response (direct signal plus early reflections plus reverberation [5]) as well as the HRTFs. Simpler solutions often attempt to parameterize the room response as a function of the source and listener positions combined with an HRTF which often results in a synthetic sound.

In [6], a relatively simple audio spatialization procedure for headphones which directly measures the combined room impulse response and HRTF (using a dummy head) at fixed locations is proposed. The combined head and room impulse responses (CHRIRs) are then used to perform the spatialization. In addition, a simplification which reduces computational complexity by decomposing the CHRIR into two components, a short filter and a long filter, where the long filter is independent of direction is shown. The short filter consists of the directional component and early reflections and the long filter consists mostly of the room reverberation [5]. One of the drawbacks of this approach is that in order to simulate a large number of positions, a large number of measurements need to take place. Although a method of interpolating the CHRIR to simulate virtual sound positions which are close the measured positions is shown, the proposed interpolation is suitable for small deviations from the measured locations.

In this paper, we extend the work from [6] by proposing an even simpler interpolation scheme to simulate various positions for the sound source. We also extend the work so that it can be used in a realistic video conference as opposed to just a multi-party audio conference by dynamically adjusting the spatialization corresponding to the the viewer's viewpoint. We show how the same simple interpolation scheme used to simulate various sound locations can also be used to simulate the effect of the user rotating their head to look at various parties in a multi-party video conference. The pose of the head is obtained by using a pose tracker. In addition we also build an entire audio conferencing system (going over the network) using the proposed audio spatialization and show a demo of

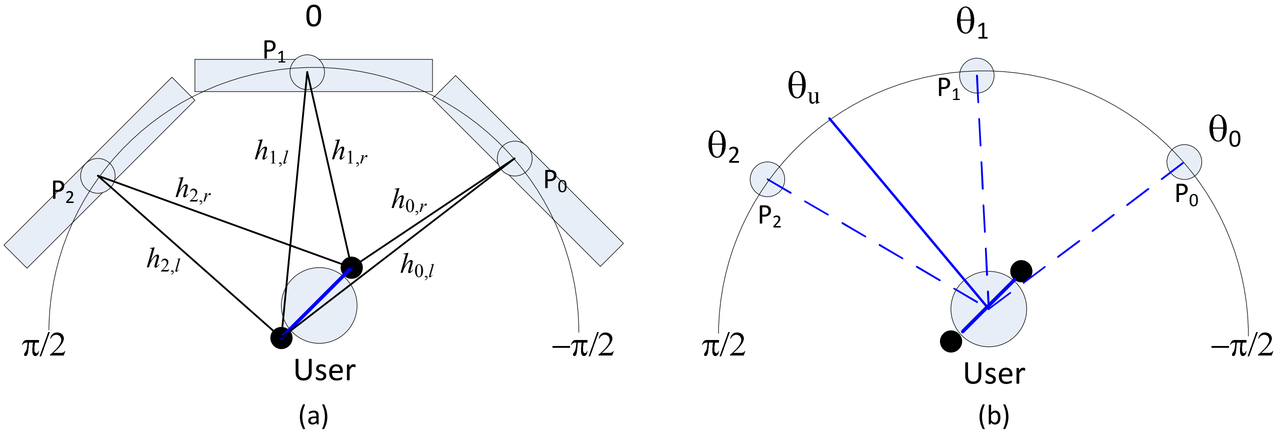


Fig. 1. System Setup. (a) shows a four party video conference where three remote participants are labeled P_0, P_1, and P_2. The participants here are placed on three different monitors. The local user is facing towards somewhere between P_1 and P_2. The ears of the local user are shown as the solid circles with the CHRIRs from each remote party i to the local user shown as $h_{i,l}$ and $h_{i,r}$. (b) shows the angles at which the remote participant i is placed, labeled as θ_i, and the user's angle as θ_u.

how it performs.

In Section 2, we describe the proposed multi-party video conferencing system. In Section 3, we describe the procedure to interpolate the measured CHRIRs. In Section 4, we show how the system can be modified to accommodate loudspeakers instead of headphones. In Section 5, we present some results, the demo setup, and audio simulation clips.

2. SYSTEM SETUP

We consider a multi party video conferencing system as shown in Fig. 1(a). The display can consist of multiple monitors (as shown in the figure) or a large curved display surface. The multi-party conference consists of N participants each of which is in a different location (fully distributed). The user is using headphones which are shown as solid circles in the Fig. 1(a). From the viewpoint of one participant, there are $(N-1)$ remote participants which are assumed to be placed in a circular fashion around the one local participant as shown in Fig. 1(a).

For this discussion, we consider a coordinate system where angle 0 is to the front of the user, angle $\pi/2$ is to the left side, and angle $-\pi/2$ is to the right side as shown in the figure. Let's assume there are $(N-1)$ remote parties labeled P_i, $i=0,\ldots,N-2$. For example there are three remote parties P_0, P_1, and P_2 in the four party conference shown in Fig. 1(a). The local user is labeled U. Assume that remote party P_i is placed at a certain location as shown in the figure. The CHRIR from the remote party i to the user's left ear is given by $h_{i,l}$ and to the right ear by $h_{i,r}$. Note that the user can in general be in any position and need not be facing the front of the display.

If the mono (single-channel) audio signal coming from the remote party i is given by x_i and if the CHRIR is of length L, then we can write the output audio signal in the left and right channels respectively as

$$y_l[n] = \sum_{i=0}^{N-2}\sum_{k=0}^{L-1} h_{i,l}[k]x_i[n-k] \qquad (1)$$

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

This is done by simply summing up the convolution of the audio signal with the CHRIR over the $(N-1)$ remote parties. For the low complexity solution proposed in [6], we can write a simplified version as

$$y_l[n] = \left(\sum_{i=0}^{N-2}\sum_{k=0}^{M-1} h_{i,l}^{S}[k]x_i[n-k]\right) + \sum_{k=M}^{L-1} h^{L}[k]\left(\sum_{i=0}^{N-2} x_i[n-k]\right), \qquad (3)$$

where a portion of the convolution (the h^L terms of length $L-M$) is independent of location. The output for the right channel can be similarly written as in Eqn. 3

The one thing we note about this approach is that in order to fully recreate the spatialization we need to measure the CHRIR for many given source locations and head positions which becomes infeasible. In order to generalize the setup, we consider the case as shown in Figure 1(b). Here we only show the angle between the remote party and the viewpoint of the user instead of showing all paths to both the listener's ears. We say that remote party i is at an angle of θ_i, and that the user is looking towards angle θ_u as determined by the pose tracker. The goal is to then spatialize the audio of remote party P_i at an angle of $\rho_i = P_i - P_u$.

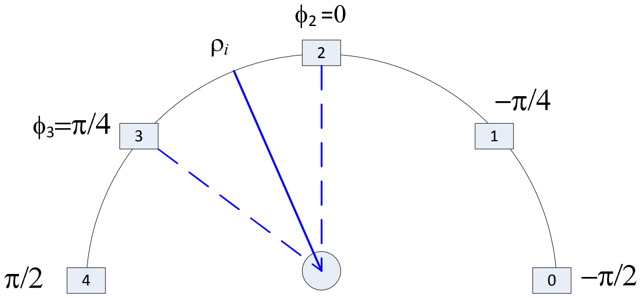


Fig. 2. Interpolation of CHRIR using nearest measured CHRIRs. For example in this figure, ρ_i is interpolated using the measured CHRIRs at positions ϕ_2 and ϕ_3. The five measured locations for CHRIRs are shown and the difference between the remote party's angle and the local user's angle. In this figure, it is assumed the local user is looking straight ahead when the CHRIRs are measured.

3. SYNTHESIZING CHRIR AT DIFFERENT POSITIONS

Suppose we have measured the CHRIR for various positions in the room at various angles ϕ_j, $j = 0, \ldots, K-1$ as shown in Fig. 2. ϕ_j is defined to be the difference in angle between the remote party and the local user's angle. For example, in Fig. 2, we show five measured CHRIRs at $\phi_0 = -\pi/2$ to $\phi_4 = \pi/2$. The CHRIR for these measured locations for the short filter is given by $h^S_{\phi_j,l}$ and $h^S_{\phi_j,r}$ for the left and right channels respectively.

Now suppose we wish to find the CHRIRs for a different angle ρ which has not been measured. We define $\hat{h}^S_{\rho_i,l}$ and $\hat{h}^S_{\rho_i,r}$ to be the interpolated versions of the left and right CHRIRs at this new location. Using Eqn. 3 with the interpolated CHRIRs we can perform the spatialization.

The method we use to do the interpolation at different angles is one inspired from remapping of channels when playing back multi-channel audio through loudspeakers. We can think of each remote participant as a channel location in a multi-channel audio recording. Suppose we wish to play this captured audio channel using loudspeakers placed at other locations which correspond to the locations of the measured CHRIRs. One way to do this is to distribute each audio channel (corresponding to each remote party) to the two nearest speaker locations (corresponding to each location where the CHRIR is measured).

Let ϕ be ordered such that $\phi_j < \phi_{j+1}, \forall j$. Then, we find j' such that $\phi_{j'} \le \rho < \phi_{j'+1}$. The audio signal at angle ρ is now distributed to locations j' and $j'+1$ using weights w_0 and w_1 respectively, where

$$w_0 = \cos\left(\frac{\rho - \phi_{j'}}{\phi_{j'+1} - \phi_{j'}}\frac{\pi}{2}\right) \quad (4)$$

$$w_1 = \sin\left(\frac{\rho - \phi_{j'}}{\phi_{j'+1} - \phi_{j'}}\frac{\pi}{2}\right). \quad (5)$$

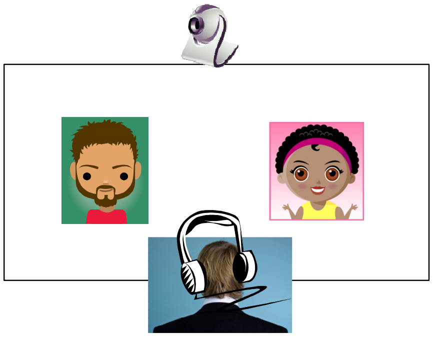


Fig. 3. Demo setup. The demo will show a live three-party video conference. The video will be setup on a single screen. The audio will be spatialized to correspond with the positions of the windows. As the listener turns their head, the audio spatialization will change accordingly. For example, if the user turns their head from the middle towards the left speaker, the left speaker will come into center and the right speaker will move further right.

Note that $w_0^2 + w_1^2 = 1$, and that $w_0 = 1$ if $\rho = \phi_{j'}$ and $w_1 = 1$ if $\rho = \phi_{j'+1}$. Now the synthesized CHRIR at location ρ is simply given by

$$\hat{h}^S_{\rho,l} = w_0 h^S_{\phi_{j'},l} + w_1 h^S_{\phi_{j'+1},l} \quad (6)$$

$$\hat{h}^S_{\rho,r} = w_0 h^S_{\phi_{j'},r} + w_1 h^S_{\phi_{j'+1},r}. \quad (7)$$

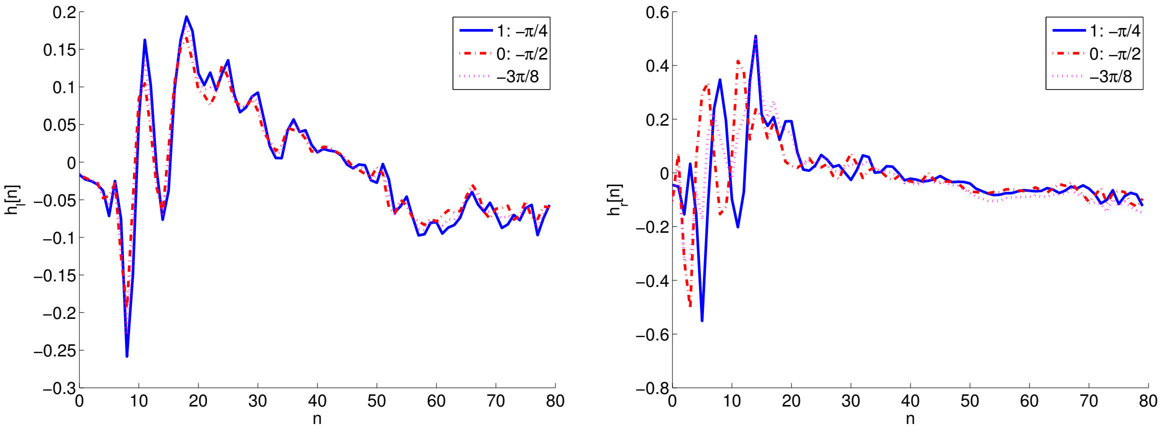
The filters are also scaled by a scale factor to ensure that the $\|\hat{h}^S_{\rho,l}\|^2 = \frac{\rho-\phi_{j'}}{\phi_{j'+1}-\phi_{j'}}\|h^S_{\phi_{j'+1},l}\|^2 + \frac{\phi_{j'+1}-\rho}{\phi_{j'+1}-\phi_{j'}}\|h^S_{\phi_{j'},l}\|^2$, that is the energy of the filter changes linearly when going from one position to another. The audio spatialization is then done using the synthesized CHRIR as in Eqn. 3.

4. USING LOUDSPEAKERS

If the users are not using headphones with close microphones, then additional steps have to take place. For example, on the capture side, one needs to use AEC potentially followed by dereverberation to get the true speaker signal. On the playback side, if the loudspeakers are close to the user's ear, then potentially one can simply use the spatialized audio as is done for the headphones. If the speakers are reasonably far from the listener, then one may have to compensate for speaker cross-talk and the room impulse response from the loudspeakers to the listener. However, these issues are orthogonal to the main spatialization procedure except for the fact that combining the filters prior to convolution will save on complexity.

5. RESULTS AND DEMO

In Fig. 4, we show the CHRIRs for the left and right channel when using measurements at angles $-\pi/4$ and $-\pi/2$. We also show the interpolated position at $-3\pi/8$. We see that



(a)

(b)

Fig. 4. CHRIRs for (a) left and (b) right channel showing measured positions $-\pi/4$ and $-\pi/2$ along with interpolated position $-3\pi/8$.

Table 1. CPU utilization when performing spatialization for both variable and fixed locations.

	$N-1=1$	$N-1=2$	$N-1=3$
Variable	19.46%	20.20%	20.89%
Fixed	19.43%	20.16%	20.84%

the interpolation resembles the general shape of the measured responses and looks like a reasonable interpolation.

In Table 1, we show the CPU complexity for performing spatialization on a 265 second 16kHz audio clip on a 3GHz x64 processor (single core is used). We show the results when changing the spatialization position (angle) every 150ms, for one, two and three inputs. We also show the CPU complexity when the spatialization angles are fixed. We see that thee is very little difference between increasing the number of participants in the conference or when changing between fixed and variable locations. This is because the spatialization filters have been divided into directional short filters (of length 80) and long filter (of length 1930). The interpolation and directional spatialization only take place on the short filter and thus there is little increase in complexity for varying the angle or increasing N. The majority of complexity comes from the single directionally independent long filter.

We will also show a fully working three party video conference as a demo as shown in Fig. 3. The listeners will each use headphones with a close microphone as is commonly done for video web conferences. The web camera at top will track the listener's pose to adjust the spatialization.

We have also created audio clips using the spatialization and submitted as supplemental material (in the final version of the paper, these clips will be given as web links). The clips should be listened to using headphones and show two things, one is the spatialization at fixed positions which are measured as shown in Fig. 2, the second is a sweep of the spatialization in 1° increments from left to right and back. The spatialization angle is changed every 150ms. This shows that the interpolation gives a reasonable feeling of moving one's head.

6. REFERENCES

[1] 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*. ACM, Apr. 2001, vol. 3, pp. 166–173.

[2] 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.

[3] "HP offers industry leading audio with new enhancements for HP Halo telepresence solutions," `http://h71028.www7.hp.com/enterprise/us/en/downloads/hp_haloaudio_enhancements_mediaalert.pdf`.

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

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