Distributed Virtual Conference Stereo Audio Reconstruction
Bo Hang, Lei Yuan, and Zhao Wu
Mathematics and Computer Science College, Xiangfan University, Xiangfan, China
bohang@163.com

Abstract—To enhance the realistic experience of virtual conference, this paper proposed a distributed virtual conference stereo audio reconstruction model. The spatial audio parameters inter-aural level difference (ILD) is used to reconstruct the spatial sound field for each listener. The distributed synthesis system is designed to get a lower network payload.

Index Terms—virtual conference; audio reconstruction; inter-aural level difference; distributed synthesis

I. INTRODUCTION
Virtual conference system makes it able for the remote participants to communicate during a traditional-like conference in a common virtual environment. Virtual conference breaks up the limits of the space, and realizes the cross-boundary information interaction. At the same time it saves the participants’ time and cost to attend a conference. Speech and audio information is one of the main communication information in virtual conference system, therefore, high-quality audio synthesis and reconstruction is one of the key technologies in virtual conference. In recent years, the relative technology in virtual conference has also become a research focus, in which area some in-depth research has been made by the National University of Defense Technology, Tsinghua University, and Zhejiang University [1-10], etc., including some research on virtual conference audio synthesis [1-3].

This paper proposes a system of virtual conference audio synthesis and reconstruction. The rest of paper is organized as follows. At first, the principles of the virtual conference audio synthesis and reconstruction are introduced in section 2. Then the algorithm of spatial audio object reconstruction is proposed based on the spatial audio object coding theory. And system design and implement are introduced in section 4.

II. THE PRINCIPLES OF VIRTUAL CONFERENCE AUDIO SYNTHESIS AND RECONSTRUCTION
Conferences may fall into disorder and confusion without any control method when argument occurs between several persons. Therefore, according to real conference experience, two virtual conference working modes are provided to keep the conference in order: chairman speech mode and free discussion mode.

Chairman speech mode: There must be a chairman in each conference who plays the role of the organizer to control the conference topics and process. The other members should keep quiet during the chairman’s speech. Free discussion mode: In this mode, all of the participants in the conference can discuss freely. we require that at most 3 persons can speak at same time. Otherwise the system will choose the 3 speaks based on a competition principle to let other members to hear their speech only.

In order to increase the realistic experience of the virtual conference, we need to reconstruct the audio’s spatial sound field for conference.

III. VIRTUAL CONFERENCE AUDIO SYNTHESIS
This model uses Inter-aural Time Difference (ILD) as the parameter to describe the virtual conference audio’s spatial direction.

The number of the participants in the virtual conference is known as \( n \), and between each two users there’s a group of ILD parameters. The relative directions of the speakers included in the input speech vector \( S_o = \{o_1, o_2, \ldots, o_j, \ldots, o_l\} \) are all different. The binaural interchannel level differences received by listener \( h \) from speaker \( j, k, \) and \( l \) are \( ILD_{jh}, ILD_{kh}, ILD_{lh} \). The energy of left ear and right ear is \( E_{left} \) and \( E_{right} \), so \( E_{left} / E_{right} = 10^{ILD} \). And \( E_{left} + E_{right} = E \).

As shown in figure 1, the output synthesized speech received by the left ear and the right ear is \( S_{o \_left} \) and \( S_{o \_right} \). Assume that:

\[
R_{left} = E_{left} / E = 10^{ILD} / (10^{ILD} + 1) \quad (1)
\]

\[
R_{right} = E_{right} / E = 1 / (10^{ILD} + 1) \quad (2)
\]
So

\[ S_{o\_left} = \lambda_j R_{i\_left} S_j + \lambda_k R_{k\_left} S_k + \lambda_l R_{l\_left} S_l \]  

(3)

\[ S_{o\_right} = \lambda_j R_{i\_right} S_j + \lambda_k R_{k\_right} S_k + \lambda_l R_{l\_right} S_l \]  

(4)

IV. SYSTEM DESIGN AND IMPLEMENT

In this virtual conference audio reconstruction system, a conference server, which is used to receive the audio from speakers and send out the synthesized audio to listener, is needed. Each virtual conference participant has a client terminal. The client terminal is used to send the audio signal if the participant is a speaker, and receive the audio from server and synthesize the audio to be displayed to listener.

We designed two optional systems for the audio reconstruction: centralized computing system, and distributed computing system. The physical structures of the two systems are same, as shown in figure 2. The difference between the two systems is the place where the spatial audio signals are synthesized.

(1) Centralized computing system

In the centralized computing system, the server receives the speakers’ audio signal, computes the synthesized audio signal for each participant respectively, according to the relative positions of the participants. Client terminals receive audio signals from the server, decode and output the signals.

In detail, the processing of the server in centralized is shown in figure 3.

There are two functional parts in client in centralized computing system. The first is sending audio signal from clients to the server. The second is receiving the synthesized audio signal from the server. The processing flow is shown in figure 4.

(2) Distributed computing system

Different from centralized computing system, the server in distributed computing system receives the audio signal from speakers, makes judgment for conference mode, selects the audio signals, and then sends the selected audio signals and the speakers ID to the clients. The processing flow of server is shown in figure 6.
Like the centralized computing system, there are two functional parts in client in distributed computing system. The processing of sending from client to server is same as centralized computing system. But the receiving processing is different. The client needs not only to decode but also to synthesize the audio signal, as shown in figure 7.

In the distributed computing system, the computing is completed by the server and clients respectively. Since the server does not separately compute the synthesis of audio signals for each client, the server will send the selected audio signals and speakers ID all clients by broadcasting. Data frame from server to client include not only the audio signals, but also speakers ID, as shown in figure 8.

Clients send mono audio signals of speakers to the server. And the server sends to all the listeners the mono audio signals selected, which will be synthesized in client and construct the spatial audio signals.

(3) Comparison and Analysis

By comparing processing flow and the signal frame structure of the two systems, we can see that in centralized computing system, the computing workload is mainly put on the server; and in distributed computing system, client take on the work of synthesis of audio signal and effective reduce the server computing load.

Here we analyze the network payload of both systems. Frame lengths from the client to the server are same in both systems. And from the server to the client, the frame length in the centralized computing system is shorter than the frame length in the distributed computing system, because each frame in the distributed computing system consists of three speakers' audio signals. However, the server in centralized computing systems send signal frame to each client respectively by unicasting. Assuming there are \( n \) participants, then the server needs to send \( n \) signal frames. The client terminals of the distributed computing system receive same signals, so broadcasting is used. Assuming there are \( m \) speakers are selected, we can get \( m \leq n \), because the number of speakers will never larger than the number of participants. The server of centralized computing system sends stereo signals to client terminals. The distributed computing system send mono audio signals to client terminals, and the attached speaker ID for each mono audio signal require less bits than the number of bits for side information of stereo audio signal.

Therefore, the length of each audio signal from server to client in the distributed computing system is always less than the centralized computing system, \( L_{\text{Dis}} < L_{\text{Cen}} \).

The payload of centralized system is \( P_{\text{Cen}} = L_{\text{Cen}} \times n \). And the payload of distributed computing system is \( P_{\text{Dis}} = L_{\text{Dis}} \times m \). From the analysis above, we can get \( P_{\text{Dis}} < P_{\text{Cen}} \), which means the network payload of distributed computing system is smaller than centralized computing system.

V. EXPERIMENTS

In the experiments, four in eight virtual participants speak at the same time. Three speakers with larger speech energy are chosen and constructed as spatial audio objects for different listener. Figure 9 show the relative positions between the four speakers and the two listeners selected. As shown in figure 10, speaker 2, 3, 4 are chosen.

Since the relative position between the three speakers and the two listeners are different, the ILD of each speech for different listener are also different. The three stereo speech signals are constructed for different listener separately. Figure 11 shows the final synthesized signals received by two listeners.
VI. CONCLUSIONS

This paper proposes a distributed virtual conference spatial sound field reconstruction model, based on the features of virtual conference and the theory of spatial audio object coding. This model can effectively synthesize the virtual conference audio object and reconstruct the virtual spatial sound field for each listener, enhancing the realistic experience of virtual conference, with prefer low network payload compared with centralized system.

ACKNOWLEDGMENT

The authors wish to thank the National Natural Science Funds (60832002), under which the present work was possible.

REFERENCE


[9] Lv Xiaoxing; Sun Lifeng; Li Jing; Li Fang; “Research of video synthesize and transmission technique in the virtual space conference system.” Journal of Beijing Radio & TV University, pp. 40-44, March 2004.