Virtual Conference Audio Reconstruction Based on Spatial Object

Bo Hang¹², Rui-Min Hu², Ye Ma²

¹. School of Mathematics and Computer Science, Xiangfan University, Xiangfan, China
². National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan, China
bohang@163.com, hrm1964@public.wh.hb.cn, maye08@qq.com

Abstract—This paper proposed a virtual conference audio reconstruction model based on spatial audio object. The aim of the model is to enhance the realistic experience of virtual conference. Firstly, the conference audio synthesis method is given according the principle of the virtual conference. Then the spatial audio parameters interaural level difference (ILD) are used to reconstruct the spatial sound field for each listener based on the theory of spatial audio object coding.

Index Terms— virtual conference, audio reconstruction; spatial audio object; interaural level difference

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 an algorithm 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. And based on the principles the multi-channel audio synthesis method is proposed. Then the new 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 four. Finally, concluding for the algorithm and next research plan are given.

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. As shown in Fig. 1.

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. At this time, the system proposed also needs to make sure that all the participants of the conference can hear the chairman’s speech only. And the chairman can interrupt any other’s speech to terminate the argument when it becomes too intense or goes too far, to make the conference return to normal.

Free discussion mode: In this mode, all of the participants in the conference can discuss freely. Under normal circumstances, the members in the conference should take turns to speak, whereas when the discussion gets heated there may be several members speaking at same time which causes an argument. However, practice shows that the simultaneity of over 3 persons’ speech would cause a handicap for other members in the conference to efficiently access information. So 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. At first all the participants should choose their virtual seat in the virtual conference, and we reconstruct the spatial directions of the speakers’ speech using binaural clues spatial parameters.

Figure 1: Virtual conference working modes
Since each participant’s virtual spatial location is fixed, and the spatial direction’s parameters of the speakers’ speech received by the listeners can be determined by the virtual relative position between the speakers and the listeners, there’s no need to collect the spatial audio signal using microphone array. Of course, for different listeners, the values of the spatial direction parameters are different, as their virtual relative position is different, so is the respective synthesis of spatial effect. As shown in Fig. 2. In 2007, Herre introduced the basic principle of spatial audio object coding (SAOC), according to which the audio was decomposed into object set and parameters were extracted separately for encoding [11]. Each speaker’s speech in the virtual conference naturally becomes an object of audio synthesis, so the SAOC theory is employed for audio synthesis to reconstruct the spatial sound field.

Based on the above principles, we propose virtual meeting’s spatial audio reconstruction model. This model includes conference mode switching, conference speaker selection, speech synthesis for conference listener, spatial audio reconstruction, etc.

III. VIRTUAL CONFERENCE AUDIO SYNTHESIS

A. Conference Mode Switching

The two modes of conference can switch automatically. The default mode in a conference is free discussion mode, and when the conference chairman terminal detects speech, the system would switch to chairman speech mode, in which mode the other participants' speech would be halted. When the chairman’s speech is over, the conference system would switch to free discussion mode automatically. We set the ending flag of the chairman’s speech when the mute duration starts from the speech end exceed $T_1$, under normal circumstances, we set $T_1 = 0.8$ second.

B. Conference Speaker Selection and Audio Synthesis

Assume that there are $n$ participants in the conference, $S_i$ is the chairman’s speech, $S_2, S_3, \cdots S_{n-1}, S_n$ is other participants’ speech, and the input speech vector of the virtual conference is $S_i = \{S_1, S_2, S_3, \cdots S_{n-1}, S_n\}$. We can judge whether the chairman is speaking by detecting $E_i$ (the energy of $S_i$). If $E_i$ is bigger than the threshold $\delta_i$, it’s chairman speech mode, and the output energy value of $S_2, S_3, \cdots S_{n-1}, S_n$ is all set to 0. Then the output speech vector is $S_o = \{S_1, 0, 0, \cdots, 0\}$, and all terminals can only hear the chairman’s speech. If $E_i$ is smaller than the threshold $\delta_i$, it’s free discussion mode, in which the other participants’ input speech energy $E_2, E_3, \cdots E_{n-1}, E_n$ would be detected, and the non-silence speech channels with the 3 biggest energy are selected, besides, if there are less than 3 non-silence input, select all of them. The selected speakers’ speech compose the output speech vector $S_o = \{0, \cdots S_k, \cdots S_{l}, \cdots\}$. Assuming that the number of the output is $N$, the synthesis of the output speech $S_o = \sum_{i=1}^{N} \lambda_i S_i$, and $\sum_{i=1}^{N} \lambda_i = 1$. $\lambda_i$ is the weighted value of $S_i$.

The above-mentioned model is utilized to synthesize speech, and the speech signal the speaker receives concludes its own speech. In 2004, Huawei’s invention patent proposed that the terminal should not receive its own speech [12]. But in fact, speakers can not judge whether their own speech is heard by other participants if they can’t hear themselves. So we suggest the speakers’ own speech should be kept in the synthesized audio received by themselves. The related $\lambda_i$ is set to a smaller value for reducing the impact of the echo effect.

Because human speech is not strictly continuous, if we do the speaker selection in each frame, the selected speaker in each frame would be different and the voice of the speaker would be intermittent, always interrupted by others. And this model would cause that every participant will raise their voice to compete for the right to speak, and that is not conducive for the conference to carry on normally. So when the number of the selected speakers reaches 3, the speaker selection will suspend until the duration some selected speaker has kept quiet for is detected beyond $T_1$, which means that the one has dropped out in the speech, then the system will do the speaker selection over again.

But sometimes some speakers are too active and the speech right is unable to be released which causes the other speakers can’t speak, so a mandatory-exit mode should be set. When the number of the speakers reaches 3,
and the time of some speaker’s speech is beyond the threshold $T_2$, the speech right will be released by force, and the system will do the speaker selection again. Furthermore, the conference chairman can stop the others’ speech by force.

C. Spatial Audio Sound Field Reconstruction

In an environment of virtual round table, in order to make the users feel like a conference round a real table, we assume that the users are in a common plane and near to each other. Therefore the height information and distance can be ignored, yet the horizontal direction angle should be considered [1].

The commonly used binaural clues parameters include Interaural level difference(ILD), Interaural time difference(ITD), and Interaural coherence(IC). Because the speaker can be regarded as point sound source, the IC parameters can be ignored. Because ITD dominates judgment for low frequency below 1.6kHz, while the ILD dominates for a larger range of frequency, and the proposed model works in a bandwidth which reaches 8kHz, ITD is not suitable for this model. Above all, this model only uses 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 which describe their relative direction angel. For example, there are participants $a$ and $b$. If a speaks, the binaural interaural level difference of $a$’s speech received by $b$ is $ILD_{ab}$; in the same way, if $b$ speaks, the binaural interaural level difference of $b$’s speech received by $a$ is $ILD_{ba}$. In a round table, we can know that the relative direction angles of $a$ and $b$ is complementary, that is, $ILD_{ab} + ILD_{ba} = 0$. When the number of the participants in a virtual conference and their relative position is fixed, the number of the constructed ILD between each two of the $n$ participants would be $n(n-1)/2$. So we can just keep $n(n-1)/2$ of the ILD datas.

The relative directions of the speakers included in the input speech vector $S_\text{a} = \{0, \cdots, S_j, \cdots, S_k, \cdots, S_l, \cdots\}$ 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 $E_\text{left}$ and $E_\text{right}$, so $E_\text{left}/E_\text{right} = 10^{ILD_{jh}}$. And $E_\text{left} + E_\text{right} = E$, so the energy of left ear and right ear is $E_\text{left}$ and $E_\text{right}$. As shown in (1) and (2).

$$E_\text{left} = (E \cdot 10^{ILD_{jh}})/(10^{ILD_{jh}} + 1)$$  (1)

$$E_\text{right} = E/(10^{ILD_{jh}} + 1)$$  (2)

As shown in Fig. 4, the output synthesized speech received by the left ear and the right ear is $S_{\text{o \_left}}$ and $S_{\text{o \_right}}$.

$$S_{\text{o \_left}} = \lambda_j(E_{\text{j \_left}}/E_j)S_j + \lambda_k(E_{\text{k \_left}}/E_k)S_k + \lambda_i(E_{\text{i \_left}}/E_i)S_i$$  (3)

$$S_{\text{o \_right}} = \lambda_j(E_{\text{j \_right}}/E_j)S_j + \lambda_k(E_{\text{k \_right}}/E_k)S_k + \lambda_i(E_{\text{i \_right}}/E_i)S_i$$  (4)

Assume that:

$$R_\text{left} = E_\text{left} = E = 10^{ILD_{jh}}/(10^{ILD_{jh}} + 1)$$  (5)

$$R_\text{right} = E_\text{right} = 1/(10^{ILD_{jh}} + 1)$$  (6)

so

$$S_{\text{o \_left}} = \lambda_j R_{\text{j \_left}} S_j + \lambda_k R_{\text{k \_left}} S_k + \lambda_i R_{\text{i \_left}} S_i$$  (7)

$$S_{\text{o \_right}} = \lambda_j R_{\text{j \_right}} S_j + \lambda_k R_{\text{k \_right}} S_k + \lambda_i R_{\text{i \_right}} S_i$$  (8)

The above is the situation when the listeners are silent. However, if the listener is one of the speakers, the ILD of the speech’s relative spatial position to themselves is 0. Generally, assume that listener $h$ is also speaker $l$, so

$$R_{\text{lh \_left}} = R_{\text{lh \_right}} = 10^{ILD_{jl}}/(10^{ILD_{jl}} + 1) = 1/2$$  (9)

Thus,

$$S_{\text{o \_left}} = \lambda_j R_{\text{j \_left}} S_j + \lambda_k R_{\text{k \_left}} S_k + \lambda_i S_i/2$$  (10)

$$S_{\text{o \_right}} = \lambda_j R_{\text{j \_right}} S_j + \lambda_k R_{\text{k \_right}} S_k + \lambda_i S_i/2$$  (11)

This is the signal finally received by listener $h$’s ears.

IV. SYSTEM DESIGN AND IMPLEMENT

A. Audio Display System

In a virtual conference auditory display, the audio output is conveyed to listener either through loudspeakers or through headphones worn by the listener. Both loudspeakers and headphones have their advantages as well as shortcomings.
Loudspeakers cause some problems. The signals from the speakers interfere with each other. It is possible to create the signals for each speaker in such a way that the resulting signal at each ear is still correct, but it has a quite high computational complexity. And, the listener has to be sitting in the right spot, or the sound effect will be different.

If headphones are used, it is not difficult to generate a specific signal for each ear, since there is no interference of the two signals. And the computational complexity is also lower than loudspeakers. So we use headphones in this virtual conference audio reconstruction system.

B. Audio Coding Algorithm

If the audio signal in virtual conference is transmitted without compression, a lot of network bandwidth will be used and the audio signal delay will increase, that will impact the communication effect in virtual conference. So we need to compress the audio signal to audio stream in the proposed system with some encoding algorithm.

As different participants in virtual conference would be in different environment and their network condition may be different, too, the network bandwidth of the proposed system should be restricted. Therefore, we use mid range or low bit-rate wideband audio encoding algorithm to get better sound quality in low bit-rate. The currently used main wideband audio encoding algorithm for mid range or low bit-rate includes HE-AACv2 standardized by MPEG [13], G.729.1 standardized by ITU [14], and AMR-WB+ standardized by 3GPP [15]. The comparison of the three audio codec is shown in Table I.

China AVS (Audio Video coding Standard) organization has drafted AVS-P10 (Part 10), with China’s own independent intellectual property [16]. We were actively involved in the work for AVS audio coding standards, and made important contributions [17-24]. Compared with AMR-WB+, AVS-P10 achieves almost equivalent sound quality at same bitrate, so we use the audio compression algorithm in AVS-P10 to compress the virtual conference audio in the proposed system.

C. Client/Server System Design

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 Fig. 5. The server and the clients access to the same network or Internet. The difference between the two systems is where the spatial audio signals are synthesized? The server or clients.

<table>
<thead>
<tr>
<th>Codec</th>
<th>Bit Rate [kbit/s]</th>
<th>Bandwidth [Hz]</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE-AAC V2</td>
<td>16–128</td>
<td>0–16900</td>
</tr>
<tr>
<td>G.729.1</td>
<td>8–32</td>
<td>50–7000</td>
</tr>
<tr>
<td>AMR-WB+</td>
<td>6–48</td>
<td>50–19200</td>
</tr>
</tbody>
</table>
There are two functional parts in client in centralized computing system. The first is sending audio signal from clients to the server. The processing flow is shown in Fig. 7. The second is receiving the synthesized audio signal from the server, as shown in Fig. 8.

In the centralized computing system, most computational work is completed by the server. What the clients need to do are only encoding and decoding audio signal.

Since each client's sound heard by the listener is different from others, server need to compute and synthesize the audio signal for each client respectively, and send to clients by the way of unicasting. The signal frames from the server to clients include only the synthesized audio signals, as shown in Fig. 9.

(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. One of the clients receives the audio signals and speakers ID to synthesize the spatial audio signal, according to the virtual relative location of the speakers and the listener. The processing flow of server is shown in Fig. 10.

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 Fig. 11.

Clients send mono audio signals of speakers to the server. And the server sends the synthesized stereo audio signals to each client respectively.
Figure 11: Client processing flow of receiving frames in distributed computing system

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 Fig. 12.

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.

In the centralized computing system, as a result of the work of audio synthesis completed on the server, the server firstly needs to receive all the audio signal and then decoding, analyzing, synthesizing, and finally re-encoding, sending out. In the centralized computing system, spatial audio synthesis is done in client, and then forwarding out. The server does not need to decode audio signals. The codec processing of the two systems are shown in Fig. 13. Therefore, the distributed computing system can effectively reduce the computing complexity of the whole system.

Client to Server

Server to Client

Figure 12: Signal frame structure between Server and Clients in distributed computing system

Figure 13: system codec processing

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, 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.

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

To sum up, distributed computing system is better than centralized computing system, based on the analysis of the computational complexity and network payload. We therefore adopted the distributed computing approach to build the virtual conference audio reconstruction 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. Fig. 14 show the relative positions between the four speakers and the two listeners selected. As shown in Fig. 15, speaker 2, 3, 4 are chosen, since the energy of speaker 1 is lowest.

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. As shown in Fig. 16, the three stereo speech signals are constructed for different listener separately.

VI. CONCLUSIONS

This paper proposes a 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. But by using only ILD in the synthesis of spatial audio signal, it is not possible to make a distinction between front and back or above and below. And the sound seems to be coming from inside the head without the distance cues. To solve these problems, more spatial audio cues calculated from Head-Related Transfer Functions (HRTFs) will to be used in future research. As in a real conference, the sound field received by a speaker will change when a participant’s head turns to another side with the changing attention direction. So the next step is to study the audio object attention model, and the way of analyzing the change of participants’ attention, adjusting the spatial direction parameters, and dynamically reconstructing the virtual conference’s spatial sound field.

ACKNOWLEDGMENT

The authors wish to thank the National Natural Science Funds (60832002) and Wuhan University Ph.D candidate independent research project (2008206210002), under which the present work was possible.

REFERENCE


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


Bo Hang received the M.S. degree in computer science from Harbin Engineering University, Harbin, P.R.China, in 2003. He is currently a Ph.D candidate in communication and information system at Wuhan University, Wuhan, P.R.China.

He came into National Engineering Research Center for Multimedia Software, Wuhan University, in 2007, and has been a student member of IEEE since 2008. His current research interests include audio signal compression and processing for mobile communication and surveillance.

Rui-Min Hu received the Ph.D degree in communication and electronic system at Huazhong University of Science and Technology, Wuhan, P.R.China in 1994.

Prof.Hu is the dean of the National Engineering Research Center for Multimedia Software, Wuhan University, and the vice-dean of Computer School of Wuhan University. His current research interests include multimedia signal compression and processing, multimedia communication

Ye Ma received the B.S. degree in computer science from Wuhan University, Wuhan, P.R.China. She is currently pursuing the M.S. degree in National Engineering Research Center for Multimedia Software, Wuhan University.