Spatial Object Based Virtual Conference Audio Reconstruction
Bo Hang\textsuperscript{1,2}, Rui-Min Hu\textsuperscript{2}, Ye Ma\textsuperscript{2}
\textsuperscript{1}Mathematics and Computer Science College, Xiangfan University, Xiangfan, China
bohang@163.com,
\textsuperscript{2}National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan, China
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\textsuperscript{1-10}, etc., including some research on virtual conference audio synthesis\textsuperscript{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. 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 figure 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 the 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

\textsuperscript{The authors wish to thank the National Natural Science Funds (60832002), under which the present work was possible.}
parameters. 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 figure 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.

![Figure 2: Different Spatial Sound Field for Different Listener](image)

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 members' 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, $T_1 = 0.8$ second.

#### B. Conference speaker selection and audio synthesis

Assume that there are $n$ participants in the conference, $S_1$ is the chairman’s speech, $S_2, S_3, \ldots, S_{n-1}, S_n$ is other participants’ speech, and the input speech vector of the virtual conference is $S_i = \{S_1, S_2, S_3, \ldots, S_{n-1}, S_n\}$. We can judge whether the chairman is speaking by detecting $E_i$ (the energy of $S_1$). If $E_i$ is bigger than the threshold $\delta_1$, it’s chairman speech mode, and the output energy value of $S_2, S_3, \ldots, S_{n-1}, S_n$ is all set to 0. Then the output speech vector is $S_o = \{S_1, 0, 0, \ldots, 0\}$, and all terminals can only hear the chairman’s speech. If $E_i$ is smaller than the threshold $\delta_1$, it’s free discussion mode, in which the other participants’ input speech energy $E_2, E_3, \ldots, E_{n-1}, E_n$ 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 $S_o = \{0, \ldots, S_3, \ldots, S_n, \ldots\}$. Assuming that the number of the output is $N$, the synthesis of the output speech $S_x = \sum_{i=1}^{N} \lambda_i S_i$, 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 receiver 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 $\lambda$ is set to a smaller value for reducing the impact of the echo effect.

Because human speech is not strictly continuous, 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

243
The 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 angle. 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 2 of the $n$ participants would be $(n(n-1))/2$ of the ILD data.

The relative directions of the speakers included in the input speech vector $S_o = \{0, \cdots S_j, \cdots S_k, \cdots S_l, \cdots \}$ are all different. The binaural interchannel level differences received by listener $h$ from speaker $j$, $k$, and $l$ are $ILD_{jh}$, $ILD_{kh}$, and $ILD_{lh}$. The energy of left ear and right ear is $E_{\text{left}}$ and $E_{\text{right}}$, so $E_{\text{left}}/E_{\text{right}} = 10^{10}$. 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).

\[ E_{\text{left}} = (E \cdot 10^{10}) / (10^{10} + 1) \]  
\[ E_{\text{right}} = E / (10^{10} + 1) \]  

As shown in figure 3, the output synthesized speech received by the left ear and the right ear is $S_{o_{\text{left}}}$ and $S_{o_{\text{right}}}$.

\[ S_{o_{\text{left}}} = \lambda_j (E_{j_{\text{left}}} / E_j) S_j + \lambda_k (E_{k_{\text{left}}} / E_k) S_k + \lambda_l (E_{l_{\text{left}}} / E_l) S_l + \lambda_s (E_{s_{\text{left}}} / E_s) S_s \]  
\[ S_{o_{\text{right}}} = \lambda_j (E_{j_{\text{right}}} / E_j) S_j + \lambda_k (E_{k_{\text{right}}} / E_k) S_k + \lambda_l (E_{l_{\text{right}}} / E_l) S_l + \lambda_s (E_{s_{\text{right}}} / E_s) S_s \]

Assume that:

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

So

\[ S_{o_{\text{left}}} = \lambda_j R_{j_{\text{left}}} S_j + \lambda_k R_{k_{\text{left}}} S_k + \lambda_l R_{l_{\text{left}}} S_l + \lambda_s R_{s_{\text{left}}} S_s \]  
\[ S_{o_{\text{right}}} = \lambda_j R_{j_{\text{right}}} S_j + \lambda_k R_{k_{\text{right}}} S_k + \lambda_l R_{l_{\text{right}}} S_l + \lambda_s R_{s_{\text{right}}} S_s \]

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,

\[ R_{h_{\text{left}}} = R_{h_{\text{right}}} = 10^{10} / (10^{10} + 1) = 1/2 \]  

Thus,

\[ S_{o_{\text{left}}} = \lambda_j R_{j_{\text{left}}} S_j + \lambda_k R_{k_{\text{left}}} S_k + \lambda_l R_{l_{\text{left}}} S_l + \lambda_s R_{s_{\text{left}}} S_s / 2 \]  
\[ S_{o_{\text{right}}} = \lambda_j R_{j_{\text{right}}} S_j + \lambda_k R_{k_{\text{right}}} S_k + \lambda_l R_{l_{\text{right}}} S_l + \lambda_s R_{s_{\text{right}}} S_s / 2 \]

This is the signal received by listener $h$.

IV. EXPERIMENTS

In the simulation experiment, 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 4 show the relative positions between the four speakers and the two listeners selected.

As shown in figure 5, speaker 2, 3, and 4 are chosen.
speech signals are constructed for different listener separately.

Figure 6 shows the final synthesized signals received by two listeners.

V. 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. 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.

REFERENCE


[9] Lv Xiaoxing; Sun Lifeng; Li Ling; 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.

