A study of the influence of auralization on speech intelligibility and immersion in multi-party teleconferencing systems using binaural audio

Jose J. Lopez
ITEAM Institute, Universidad Politecnica de Valencia, Valencia, Spain.

Emanuel Aguilera
ITEAM Institute, Universidad Politecnica de Valencia, Valencia, Spain.

Pablo Gutierrez
ITEAM Institute, Universidad Politecnica de Valencia, Valencia, Spain.

Summary
The introduction of spatial audio features into conventional teleconferencing systems is a natural way of creating a realistic communication environment and was introduced many years ago. The human auditory system takes advantage of spatial audio cues to locate, separate and understand multiple speakers when they talk simultaneously. As a result, speech intelligibility is significantly improved if the speakers are simulated to be spatially distributed around each participant and also immersion and sense of presence improves. Moreover, it is possible to increase even more this immersion and sense of presence if auralization is introduced in order to simulate the effects of a virtual room where the virtual meeting is taking place. However, if auralization introduces too much reverberation or harmful early echoes, speech intelligibility can be affected. In this paper, we thoughtfully study the influence of auralization on speech intelligibility and sense on immersion in multi-party teleconferencing systems using binaural audio. For carrying out the experiments specific software for multi-party teleconference that employs binaural audio using headphones has been developed. The software runs on smartphones and tablets so touch screens can be used to move and place participants in the virtual room using avatars. Different subjective experiments have been carried out trying to evaluate the influence of the early echoes, the late reverberation, the size of the room and other details related to binaural processing. Results provide interesting guidelines for developing teleconference systems with auralization effects.

PACS no. 43.66.Qp, 43.66.Pn, 43.60.Dh, 43.55.Hy

1. Introduction
Multi-party teleconference services play an important role in social interaction, providing unquestionable advantages today. These are really useful, especially within a business scope, since they help to save a lot of time and travel expenses [1,2]. This is the main reason why, for a long time now, companies have been using these services through the landline telephone network. Likewise, a mobile device can also use these conference call services provided by operators as fixed telephony does [3]. The first and most obvious advantage is that mobile business users acquire mobility: they are not forced to stay in their offices. In this way, with an appropriate access system and rates, the service may be attractive to home (non-business) users that could make audio meetings with friends and family from anywhere. In addition, today mobile devices have many more features than fixed-telephony terminals, which have been relegated almost to pure voice communications.

The idea of exploiting the capabilities of new mobile devices to make more realistic multi-party conference calls is currently the other important advantage that takes a strong commercial interest. One such improvement is related to the incorporation of spatial audio into the conference [3,4].
In previous works it has been demonstrated that spatial audio (stereo or binaural) is generally preferred by subjects over non-spatial monaural audio [4,5], improves speaker discrimination [5,6] and intelligibility of conversations, leveraging the ability of the human auditory system to attend selectively to sounds from a specific direction, i.e., the “cocktail party effect”.

Moreover, the incorporation of room effect (auralization) into the audioconference produces more controversial results according to previous studies [7], because although it increases distance perception, degrades sound source direction and intelligibility.

In this work we propose a new immersive multiparty conference call system for mobile devices (smartphones and tablets) that substantially improves the identification and intelligibility of the participants by means of binaural sound reproduction through headphones to locate the different speakers within a virtual meeting room and, at the same time, the use of a large touch screen to locate the participants taking part in the conference, with the possibility of changing their spatial location in an interactive way by the user. The system, developed by us, comprises the application in the mobile device and the server that manage the communications.

Using this system, we are going to analyze the effects of adding different levels of auralization to the audioconference by means of subjective testing by a jury or participants. The idea is to investigate individually the influence of the early echoes, the late reverberation, the size of the room and other details that usually are employed in classical auralization methods.

The remainder of the paper is organized as follows: Section 2 describes the developed software employed in the experiments and the related technology including the auralization algorithms employed. Section 3 describes the subjective experiment and its objectives. Section 4 exposes the results and its analysis and discussion. Finally, section 6 presents the conclusions of this work.

2. Implementation

For the development of the test, a previous developed multiconference application for iPad tablets, developed in the research group, has been employed. The application allows the participants to easily control their execution by means of a graphic and tactile interface. This application employs binaural audio by means of a simple model of the head-related transfer function (HRTF). The users can virtually move and place other interlocutors around their own position in the middle of the bottom edge, see fig. 8. A complete description of the application can be found in [8].

2.1. Binaural sound

In practice, HRIR functions have a length between 128 and 512 samples. Convolving each mono signal with the two HRIR implies a significant computational cost. Despite these convolutions can be performed in the frequency domain by using FFT multiplication (using overlap-add or overlap-save algorithms) in an efficient way, this process still has a significant cost. Although today’s smartphones have enough power to carry out these filtering algorithms in real time, a high percentage of the computational resources have to be employed on this processing stage and also reduce the battery life.

Therefore, an efficient implementation of the HRTF has taken into consideration in this work. In order to highly simplify the computational cost, a rough approximation of the HRTF has been implemented though giving up a good enough synthesis.

The implementation of the HRTF has been split into two parts. First, the ILD has been implemented by means of two IIR filters. Second, the ITD is achieved by adding a time delay between the left and right ear signals. The design of the ILD IIR filters has been done by following a procedure similar to that of [9], where the authors obtained a standard HRTF model by averaging the responses from a database of real HRTF responses. Using the averaged response, a 6-th order parametric IIR filter has been adjusted for each azimuth direction. More details of this implementation can be found in [8].

2.2. Auralization

The application simulates that the participants are in a virtual room whose size extends along the screen. The auralization algorithm has to simulate the reflections and other acoustic effects that occur inside the room taking into account the position of the participants and the listener. There are many accurate methods for modeling the room impulse response (RIR) between two different points in a room, many based on physical models of the sound propagation. For this application a very accurate method is not needed, only one that produces a realistic sensation.
The RIR is generally described as composed by direct signal, early echoes and reverberation tail. Our application computes the early echoes using the image method. Four echoes coming for the first order reflections of lateral walls are considered. Additionally the application has the possibility of processing the echoes binaurally taking into account the direction of arrival or in a monaural way. For the reverberation tail a classical Schroeder reverberator [10] is employed.

3. Experiment description

The objective of the experiment is to investigate individually the influence of the parts that conforms the RIR (early echoes, and late reverberation), also the size of the virtual meeting room and additionally the effect of using binaural filtered early echoes or mono early echoes.

The test was carried out in an iPad tablet were Sennheiser HD 439 headphones were connected. These are closed type for good external noise rejection and circumaural that produce less fatigue and are more comfortable to wear. A total of 10 people participated in the test, 5 males and 5 females, all of them with normal audition with ages from 22 to 40, recruited from the student and staff of different research groups at the University. As audio material, a previously recorded conversation between three people was presented to the participants using the iPad application. Then, participants could freely choose the position of the three speakers and were invited to interactively move their avatars using the touch screen to different positions in the room as much as they wanted during the test in order to perceive how spatial perception changes with position and with the different auralization options presented.

Due to the large number of parameters to be tested it was decided to divide the test in two phases. In the first one two types of early echo processing will be analyzed: binaural filtered or monaural. The first one employs slightly more computer power than the second one, being the first one theoretically more realistic, but the difference was so subtle that it was interesting to test if the listeners were able to detect it, and at the same time, to know their preferences. Besides, this previous evaluation allows us to reduce the amount of possibilities presented to the participants in the second phase of the test.

In the first phase the application shows a button on the screen to freely switch in a blind manner between the two different early echoes processing A/B (nominal direct scaling test [11]). The participants had to fill a form answering the question: “Which one of the two cases generates a better feeling of realism and position?” and additionally they were also asked to quantify the amount of difference that they had perceived within a scale from 1 - Almost the same, to 5 - Totally different.

According to the results of this first test part (analyzed in next section), binaural processing of early echoes was selected for the second part. In the second part the participants were able to switch between six different auralization cases combining early echoes, late reverberation and size of the room. The combinations are:

1. Direct signal (without early echoes or reverberation). \{DS\}
2. Direct signal + early echoes in a small room (5x3.75m). \{EE(S)\}
3. Direct signal + early echoes in a large room (10x7.5m). \{EE(L)\}
4. Direct signal + early echoes in a small room (5x3.75m) + reverberation \{EE(S)+R\}
5. Direct signal + early echoes in a large room (10x3.75m) + reverberation \{EE(L)+R\}
6. Direct signal + reverberation \{R\}

Using the same conversation employed in test 1, the participants were able to freely move the speakers and freely switch among the different cases in a blind manner, figure 8. The application changes the signal processing algorithm for auralization in real time depending on the button selected by the user. The buttons were randomly assigned to the different auralization cases and for each of the participants.

Then, participants were asked to evaluate 4 different features related with the auralization.

1. Intelligibility: Do you understand the conversation easily? How much?
2. Distance: Do you have a sense of distance when you move the speakers? How much?
3. Immersion: Do you feel immersed in a real environment or room? How much?
4. Overall preference: How much do you like each option? Order them according to your personal preference.

For each feature, the participants have to order the 6 different auralization cases, from 0-low, to 5-high (ordinal direct scaling test [11]). In total, they have to complete a table of 4x6 cells.
4. Results and Discussion

4.1. Part 1 Results

The results of first part of the experiment are surprising. Almost the same number of participants found that monaural early echoes (40%) produced better feeling of realism and positioning than binaural processed echoes (60%) that in theory are more realistic. Anyway, they found very difficult to decide between those cases because they were very similar (a mean of around 2 in a scale of similarity from 1 to 5), table I.

Table I. Results of the first test part.

<table>
<thead>
<tr>
<th>Early Echoes</th>
<th>Participants Choice</th>
<th>Perceived Difference (1-5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monaural</td>
<td>40 %</td>
<td>1.75</td>
</tr>
<tr>
<td>Binaural</td>
<td>60 %</td>
<td>2</td>
</tr>
</tbody>
</table>

Taking into account these results we continued in the next part of the test applying binaural processed early echoes, so different combination of auralization characteristics were reduced to a small amount of cases of study and avoid very similar cases.

4.2. Part 2 Results

Four features (intelligibility, sense of distance, immersion feeling and overall preference) for six auralization cases were evaluated in the second part of the test. Each of these features was sorted by participants in a score from 0 (the worst case) to 5 (the best case). Mean score and 95% confidence intervals (CI) have been calculated for each case and feature.

According to intelligibility, as figure 1 shows, the better performance was obtained for the just direct signal case (DS) and participants found the most difficult to understand the combination of early echoes and reverberation in the large room (EE(L)+R).

The evaluation of the sense of distance (fig. 2) indicates that the preferred case is the direct signal with only reverberation (R), being the just direct signal case (DS) the second one. The auralization cases that include early echoes do not achieve much sense of distance and their confidence interval indicates that perception of distance is very similar in these cases.

When asked for immersion (fig. 3), participants prefer a plain direct signal (DS) and followed by early echoes cases. Reverberation is not well valued for the immersion feeling.

Surprisingly, in the overall preference (fig. 4) the direct signal (DS) is clearly the preferred method for a multi-party teleconferencing system and the sum of reverberation with the early echoes in a large room (EE(L)+R) is the worst for this use.

4.3. Discussion

4.3.1. Features

From previous figures we can see that intelligibility, immersion feeling and overall preference have similar results for all the different auralization cases, having the sense of distance an opposite tendency (with the direct signal case exception).

Intelligibility casts the expected results. A cleaner signal is preferred, if we consider that adding early echoes to a direct signal makes it dirtier and even more if we add reverberation.

The sense of distance is better perceived if reverberation is present. Considering the results, early echoes do not help to improve the sense of distance, probably because they add more energy to the signal, so the decreasing level when the speaker is supposed to be farther is not well perceived. When reverberation is added to any of the first 3 cases, the sense of distance is increased, as we can see in figure 7.

The immersion feeling results were not so obvious. We could expect that adding room auralization to the signal would increase the sense of realism but participants said that a cleaner signal makes them feel more immersive. Many of them reported that they expected to be speaking in a well-conditioned non-reverberant room, so any kind of echoes or reverberation made the experience less realistic. There was not a correspondence between what they hear and the place they think they are placed for those cases.

According to the overall preference, participants prefer cleaner signals giving preference to intelligibility and immersion and do not care about the sense of distance. This means that in this kind of application, the realistic sense of distance is not an important feature.

4.3.2. Cases

Just direct signal (DS): in a multi-party teleconferencing application, this is the best option for intelligibility, immersion and it is the preferred case. Even for the sense of distance is a good option, because the decreasing amplitude of the signal is well appreciated when moving the avatars farther away.
Figure 1. Mean intelligibility score with 95% CI

Figure 2. Mean sense of distance score with 95% CI

Figure 3. Mean immersion feeling score with 95% CI

Figure 4. Mean overall preference score with 95% CI

Figure 5. Room size comparison when using early echoes.

Figure 6. Room size comparison when using early echoes with reverberation.

Figure 7. Mean sense of distance score, with and without reverberation.

Figure 8. Main screen of the tablet application
Early echoes (EE(S) and EE(L)): as the first part of the test showed, binaural early echoes are slightly better than monaural according to realism and position. Considering that the CPU usage is almost the same for both techniques, this would be the suitable option. With the second part of the test we found that early echoes are quite good for intelligibility, immersion and a good final choice, being the small room better than the large room (fig. 5). If we consider only the sense of distance, use early echoes are the worst option, being the small room (EE(S)) worse than the large room (EE(L)). When adding reverberation to early echoes, the difference between the small (EE(S)+R) and large (EE(L)+R) room behavior increases, as figure 6 shows.

Reverberation (R): adding just reverb to the direct signal has similar intelligibility, immersion and preference results that using just early echoes (EE) for this kind of application. But reverberation is the best auralization processing technique to increase the sense of distance. If we add the reverb to a signal (just direct signal or with early echoes) the distance feeling is better, as figure 7 shows.

Early echoes and reverberation (EE(S)+R and EE(L)+R): it is the worst case (mainly in the large room) in terms of intelligibility, immersion and overall preference and it is not good for the sense of distance, so there are no reasons to use these combinations in this kind of application.

5. Conclusions and Future Work

In this paper different subjective experiments have been carried out trying to evaluate the influence of the early echoes, the late reverberation, the size of the room and other details when auralization is applied to binaural teleconference system. Although in the discussion section the particular conclusions for each case have been reported, some general conclusions can be obtained. Participants prefer cleaner signals giving preference to intelligibility and immersion and do not care about the sense of distance. This means that in this kind of application, the realistic sense of distance is not an important feature. Just the binaural direct signal is the best option for intelligibility, immersion and it is the preferred case. Even for the sense of distance is a good option because the decreasing amplitude of the signal is well appreciated when moving the avatars farther away. Oppositely, a more complete auralization algorithm including early echoes and reverberation report the worst results.

Anyway, in some cases the confidence intervals overlap too much to consider the whole test very confident. As a future work the test should be extended to a greater number of people in order to reduce the C.I. Moreover, the level of the early echoes (i.e. the reflection coefficient of the room walls) should be analyzed by subjective testing, in order to test if lower values produce better results than this test and also better than not using it.

Acknowledgement

The Spanish Ministry of Science and Innovation supported this work under the project TEC2012-37945-C01.

References