An Immersive Multi-Party Conferencing System for Mobile Devices Using 3D Binaural Audio

Emanuel Aguilera[1], José Javier López[1], Máximo Cobos[2], Luis Maciá[1], Amparo Martí[1]

[1] Instituto de Telecomunicaciones y Aplicaciones Multimedia (iTEAM)
Universitat Politècnica de València,
8G Building - access D - Camino de Vera s/n - 46022 Valencia (Spain)

[2] Computer Science Department
Universitat de València,
Avda. de la Universitat s/n - 46100 Burjassot, Valencia (Spain)
Corresponding author: emagmar@iteam.upv.es

Abstract

The use of mobile telephony, along with the widespread of smartphones in the consumer market, is gradually displacing traditional telephony. Fixed-line telephone conference calls have been widely employed for carrying out distributed meetings around the world in the last decades. However, the powerful characteristics brought by modern mobile devices and data networks allow for new conferencing schemes based on immersive communication, one the fields having major commercial and technical interest within the telecommunications industry today. In this context, adding spatial audio features into conventional conferencing systems is a natural way of creating a realistic communication environment. In fact, 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. This paper describes the development of a new immersive multi-party conference call service for mobile devices (smartphones and tablets) that substantially improves the identification and intelligibility of the participants. Headphone-based audio reproduction and binaural sound processing algorithms allow the user to locate the different speakers within a virtual meeting room. Moreover, the use of a large touch screen helps the user to identify and remember the participants taking part in the conference, with the possibility of changing their spatial location in an interactive way.

Keywords: Audio conferencing, multi-party conference, mobile device, smartphone, tablet, spatial audio, touch use interface.

1. Introduction

Mobile telephony has become nowadays the most important way of remote communication, displacing fixed telephony and other media. In addition, the new generation of smart mobile devices is bringing new communication schemes to their users, with new applications in videoconferencing, instant messaging, social networks, etc.

Multi-party conference services play also 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 do [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 [3,4,5].

The users will listen to the other parties as part of a sound scene where each participant is located at a different spatial position.

This paper describes the development of a new multiparty conference call system with spatial audio and a touch GUI for mobile devices (iPhone, iPad and Android phones). As discussed below, the most interesting part of the project is the use of spatial sound (by means of binaural processing) in the conference. In addition, a novel graphical and touch interface that enhances the usability of the system has been designed, developed and tested.

The system allows a group of users to establish a conference simply by using their mobile stereo headset (although the sound goes beyond the traditional stereo panning). The users will listen to the other parties as part of a sound scene where each participant is located at a different spatial position. The position of the participants can be chosen by means of the mobile touch screen.

Several advantages are obtained by using the above features:

- The user immediately associates a spatial position with a person. This helps the user to identify who is speaking, even if the voices of the other participants are not previously.
- It increases the sense of presence. The conversation is more realistic and provides a feeling of being in a face to face meeting.
- It improves intelligibility when several people are speaking at once. It is much easier to understand the partners in the moments they interrupt each other or speak simultaneously because their voices come from well-defined spatial directions.

The remainder of the paper is organized as follows: Section 2 exposes the intelligibility and identification problems in conferences and the existing multi-party conference solutions. Section 3 outlines the requirements and technical limitations we face in this work. Section 4 describes the signal preprocessing and voice compresión applied in the system. Section 5 exposes the binaural techniques applied for the audio spatialization. Section 6 explains the transmission protocols designed for this service. Section 7 describes the implementation details both for the server and the mobile devices. Finally, Section 8 presents the conclusions of this work.

2. Background

Multiple studies have analyzed the most common difficulties when users take part in multi-party conference services [6]. The most frequent ones are the inability to understand what the other participants are saying (usually because of the poor audio quality, the background noise and/or several people talking at the same time) and the difficulty to identify who is speaking (sometimes even not knowing who is in the meeting).

2.1. Intelligibility and Cocktail Party Effect

The problems derived from traditional conference systems are usually caused by the monophonic nature of these services. In other words, current systems mix all the voices into a single audio signal that is reproduced by a unique loudspeaker in each terminal. This not only makes the conference less realistic but often hinders the intelligibility during the conversation when more than one person is speaking at a time because the voices are simply overlapped. In addition, there is the problem to distinguish who is actually speaking when several participants join the meeting, especially when their voices are not known beforehand (a very common situation in business meetings).

Moreover, there is evidence that the human auditory system, which is binaural by nature, has a special ability to locate sources in the sound scene by combining the signals from both ears. In addition, it also has the ability to separate and understand each person when they speak at the same time or they interrupt each other. This capability is called the Cocktail Party Effect and helps to improve aspects as intelligibility, immunity to interference and speech understanding [7]. A classic example of this ability happens when we are in a noisy room full of people, we can still hear and understand the person we are talking to, while ignoring other conversations and background noise (Fig. 1).

Figure 1. *The human auditory system is capable of understanding speech in adverse acoustic situations (like in a Cocktail Party).*

The fact that today most mobile devices have a stereo audio output (with headphones) and processing power to work with digital audio algorithms, gives us the chance to improve multi-party conference services. On one hand it is possible to increase the audio quality using wideband speech coding and denoising algorithms. On the other hand we can incorporate spatial sound to exploit our natural abilities (which are based primarily on a binaural effect) to focus our listening in one direction [8], as will be explained in Section 5.

2.2. Identification and visual cues

While audio cues are the most important ones in order to determine who is talking in an audio conference, visual cues can also be used for identifying and remembering who is participating in a conference [4]. In a face-to-face meeting, we can see who is in the room and notice when they start speaking, but in a traditional multi-party audio conference we don't have any visual information about the rest of participants, making more difficult to have a sense of presence.

Today, smartphones and tablets have large touch screens that can be used to provide the user with visual information about the people involved in a conference. The way we interact with the participants in a conference can also be increased by means of the touch capabilities of the new mobile devices.

2.3. Existing multi-party mobile conference commercial solutions

Currently, there are several commercial solutions to establish a multi-party conference through mobile devices (as well as through fixed telephony). The simplest solutions are usually free, while the rest charge subscribers through premium-rate services that are widely used at enterprise level. These services are offered by telephone operators, websites or mobile applications like Skype.

Technically, some commercial services work with VoIP techniques while others can also connect via GSM or fixed telephony. Regarding sound quality, many of them have poor voice quality, similar to traditional voice services. This means that the voice is sampled at 8 kHz. Only a few commercial solutions based on VoIP improve the quality up to 16 kHz. What all of them have in common is that they do not perform any 3D sound spatialization, i.e. they use mono sound.

Sound spatialization techniques in calls with more than two speakers are also being considered by other companies and research institutes. The mobile operator Docomo has worked in the integration of the sound signals of several speakers into a single stream to transmit it to mobile phones for multi-party conference calls, tele-education and online games [9]. The MPEG SAOC (Spatial Audio Object Coding) working group at Fraunhofer IIS [10] is focused on efficient coding techniques that allow for the manipulation of independent sound objects. However, to the best of the authors' knowledge, there is not currently any commercial product or service using Docomo's developments or MPEG SAOC.

3. Requirements and technical limitations

The evolution of mobile phones in terms of computing power and enhanced features have allowed for the development of advanced signal processing applications oriented to the improvement of the user experience.

The position of the participants can be chosen by means of the mobile touch screen.

Touch screens, cameras and velocity sensors, together with the advances in mobile data networks, provide the user with a powerful device capable of implementing advanced VoIP services within a mobile immersive communication framework.

3.1 User experience

The key of the user experience is intended to be on its interface. Despite being very simple (using the touch capabilities of the screen), it provides a great sense of realism as a result of the improved voice quality (better than traditional telephone calls) and 3D audio features.

Once the user logs into the service with his name and password, the list of contacts and their availability will show up. If they are online and have an available state, the user can ask them to start a conversation in his/her own virtual room or in the other's one. In case they were already active in another conversation, the user can ask to join the virtual room.

During the conversation, the user sees the avatars of the participants in a virtual room. The different speakers can be dragged with the finger to place them in different positions, as seen in Fig. 2. The user will then listen through the headphones the voices from the participants as if they were coming from the location showed in the screen. The avatars will carry their coresponding participant name and will shine whenever they speak. This way, identifying each participant becomes very easy, even if it is the first time they hear their voices.

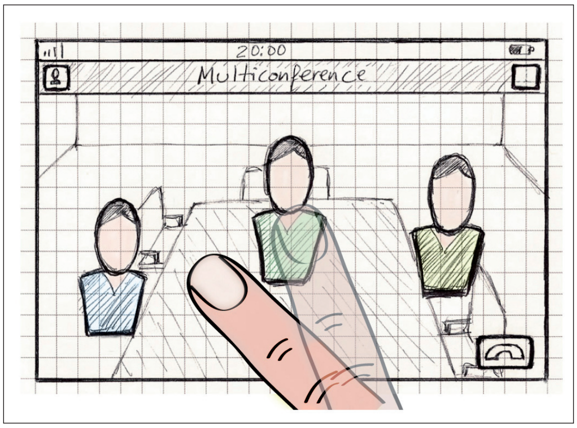


Figure 2. *Sketch of a touch GUI for a multi-party conference.*

3.2. Technical considerations and limitations

Connectivity and bandwidth

Currently, mobile phone calls use the concept of circuit switching. This technology is based on specific standard protocols and mobile phone implementations which are not ready to add more complex voice services like the one presented here.

Given the difficulty of using switching circuit communication for our purposes, it is needed to transmit the voice using the data transmission capabilities of IP networks (known as VoIP), via 3G mobile networks or Wi-Fi access points. Therefore, we are required to minimize the use of bandwidth, not only because of the mobile data connection costs, but also to maintain a high audio quality even in conditions of low network coverage.

Server and mobile devices capabilities

Unlike traditional one to one phone calls, in a multi-party conference each terminal has to receive the signal from several participants. Multicast transmission is not used because of the problems dealing with complex network configurations. Moreover, in order to use a low bandwidth, peer-to-peer connections between terminals must be discarded if we don't want to saturate the upstream channel (which is often the most limited channel in a mobile connection).

Thus, to make a more efficient use of the available bandwidth, it is necessary to include a server that picks up a single voice signal from each terminal and distributes it to the rest of participants. The server audio streams management must be optimized to ensure a low CPU usage. This is important to enable future server scalability to thousands of users.

Concerning mobile devices, designing signal processing techniques and algorithms that can be executed in real time is not the only big issue. Battery consumption should be minimized where possible, involving several factors. Some of them, like display usage and audio output, are inevitable. However, CPU usage and the amount of transmitted data should be reduced to save battery life.

Speech quality

It has already been highlighted the importance of implementing signal processing algorithms with low computational cost and light data flows. But it should not be forgotten that the original purpose of this service is to improve the speech quality in a conference. Therefore, there are some points to consider:

- Firstly, we cannot use 8 kHz sampled voice signals as in traditional calls. It is essential to use at least 16 kHz sampled signals in order to preserve a natural speech quality.
- Although it is needed to apply strong audio compression before transmitting the data, the encoder must be specifically oriented to speech in order to maintain the highest possible quality.
- When several speakers are talking, the accumulated ambient noise becomes a major problem and it is indispensable to use noise reduction techniques.
- Finally, we want the user to have a real presence feeling in the virtual room, enabling the use of natural directional listening abilities. Thus, we need high performance sound spatialization algorithms.

4. Signal preprocessing and voice compression

The speech processing stage is based on two technologies: high-quality voice codecs and binaural audio spatialization. This section describes the processing methods (voice compression, noise reduction and voice activity detection) used to allow for an optimal transmission of the voice streams over IP.

First, noise reduction is applied to the signal captured by the microphone. This not only increases intelligibility when several speech signals are mixed, but also improves the subsequent voice codec performance. It is not needed to apply echo cancellation, since this service is not going to be used in hands-free mode, only with a stereo headset. Therefore, an expensive part of processing can be saved.

In the next step, a voice activity detector (VAD) is applied. The speech probability in each frame is calculated and a hysteresis thresholding for that value is configured to decide if the frame contains speech or not. Thus, only speech frames will be processed and sent. On the one hand, this reduces the CPU usage of both the transmitter (which will not need to compress the frame) and the receiver (which will not need to decode and spatialize the audio). On the other hand, it reduces the bandwidth required to distribute each signal between all terminals. These benefits become more important when there are a high number of participants, since the speech probability per person decreases when more people are involved.

Voice communication services require audio compression techniques to reduce the bandwidth needs, but a maximum voice quality must be guaranteed. For that reason, one of the most advanced voice codecs (Speex) was selected. This codec has the advantage of being open source, so it is free of patents and allows improving and adapting it to our needs [11].

Speex was created especially for encoding voice, for which there was no suitable open source tool available. Unlike other codecs, Speex is not designed to work with circuit switched protocols (typical on fixed and GSM telephony), but for packet networks and VoIP applications. This means that it is robust against packet loss, but not against corrupted packets, based on the assumption that, with VoIP, packets either arrive unaltered or do not arrive.

The Speex codec, which is based on the CELP algorithm, is designed to be very flexible and allows setting a large range of voice quality and bit-rate. This means that it is not only capable of compressing low bandwidth voice (8 kHz telephone quality) but also large bandwidth voice at 16 kHz and 32 kHz samplings rates. As this codec is targeted at a wide range of devices, it has low complexity and uses very little memory [12].

In this work, the codec has been adapted and optimized for being used in smartphones. A 16 kHz sampling rate has been chosen, since it gives a much better voice quality than traditional 8 kHz with a moderate computational cost. Moreover, it has been achieved a trade-off between the rest of the codec configurable parameters to adjust the voice quality respect to its computational cost on a smartphone. The sound signal is divided into 20 ms frames using 16 bit samples, what produces 640 bytes/frame before compression and becomes into 70 bytes/frame after Speex compression.

5. Sound scene spatialization based on binaural techniques

5.1 Introduction

As commented in Section 2, the human auditory system has the ability to locate sources in a sound scene by combining the signals from both ears (binaural hearing). In addition, it has also the ability to separate and understand different sound sources when they speak at the same time or interrupt each other (Cocktail Party Effect). This section describes the binaural sound synthesis stage that has been implemented in the application in order to improve important aspects such as intelligibility, immunity to interference and speech understanding.

Smartphones and tablets usually feature a stereo audio output. Moreover, they have great processing capabilities to work with digital audio algorithms, so it is possible to incorporate spatial sound to mobile devices with the aim of taking advantage of our natural ability to focus our listening in one spatial direction.

Traditional stereo systems are able to place a sound at a point between left and right loudspeakers, using a simple panning algorithm that applies a different gain to each channel. However this basic spatial sensation is far from a real acoustic experience and does not provide a high degree of immersion.

5.2 Binaural Sound

Binaural sound spatialization systems try to simulate virtually the human hearing function by using the same mechanisms employed by the auditory system. Headphone reproduction helps to recreate a sound pressure in the inner ear similar to that of an equivalent real situation. This makes it possible to discern the direction of sound sources from the relative differences in the sound received by the two ears.

One of the basic binaural processing mechanisms involves the comparison between the time of arrival of the sound to the left and right ears. This difference is commonly known as Interaural Time Difference (ITD). If we assume that the average distance between human ears is about 18 cm, the ITD has a maximum value of about ±0.75 ms.

Another consequence of the presence of the head is that higher frequencies are attenuated or shadowed by the head as they reach the contralateral ear. This attenuation produces an Interaural Level Difference (ILD) which also plays a major role in lateral localization, especially at high frequencies. The ITD and ILD are considered to be the primary cues for the perceived azimuth angle of a sound source, as proposed by Rayleigh in what is known as the "Duplex theory" [13].

The speech processing stage is based on two technologies: high-quality voice codecs and binaural audio spatialization.

These effects and others are considered by the Head-Related Transfer Function (HRTF), which is the transfer function between the sound pressure that is present at the center of the listeners head when the listener is absent and the sound pressure developed at the listener's ear. The HRTF is a function of direction, distance and frequency. The inverse Fourier transform of the HRTF is the Head-Related Impulse Response (HRIR), which is a function of direction, distance, and time.

Each voice coming from each speaker in the conference is a mono sound and it will be considered as a point source. Using the touch interface as explained in Section 3.1, the user can place each speaker at any position in the room. In order to synthesize binaural sound from point sources, HRTF filtering can be employed. Depending on the direction of arrival (i.e. the angle that the point source conforms with the listener) the signal is filtered in time domain with the corresponding values of the HRIR for this direction for the left and right ears as shown in Equation (1).

$$L[n] = HRIR_L\,(\varphi,\theta,n) * s(n)$$
$$R[n] = HRIR_R\,(\varphi,\theta,n) * s(n) \qquad [1]$$

where *L[n]* is the signal for the left ear, *R[n]* for the rigth ear, *s(n)* is the speaker's voice and HRIR is the particularizacion of the HRIR for the arrival direction (φ,θ), as seen in Figure 3.

5.3 Efficient implementation for mobile devices

In practice, HRIR functions have a length between 128 and 512 samples. Convolving each mono signal with the two HRIR implies a significative 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 significative cost.

Although today's smartphones have enough power to carry out these filtering algorithms in real time (we checked this fact during the work development), a high percentage of the computational resources have to be employed on this processing stage. As a result, this issue leaves fewer resources for other tasks in the device. Moreover, the most important factor in mobile application design is battery life: if there is an intensive use of the arithmetic computation resources, the microprocessor will need more electrical power and the battery that supplies that power will get empty much faster.

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. First, we considered the fact that the complex frequency pattern that the HRTF presents at high frequencies (over 5 kHz) can be neglected without significant consequences if the sources are assumed to be located on the horizontal plane. In fact, it is well known that the complex shape of the HRTF over 5 kHz caused by the ear ring plays only a major role in the localization of elevated sources.

Neglecting the effect of the pinna, the ILD is practically reduced to the diffraction effect of the head. This effect is much less complex in frequency and can be modeled by low order IIR filters matching this response. Although head sizes and shapes vary from one person to another, their effect is less important than that of the pinna, allowing for the implementation of a (almost) listener independent spatial audio system.

The implementation of the HRTF has been splitted 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. This model has already been used successfully in other works [14].

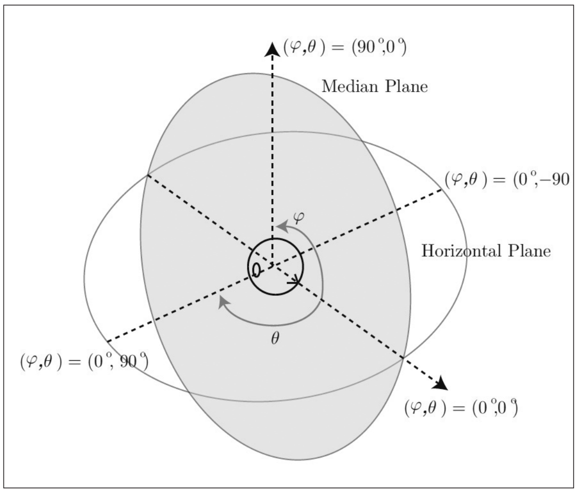
The design of the ILD IIR filters has been done by following a procedure similar to that of [15], where the authors obtained a standard HRTF model by averaging the responses from a database of real HRTF responses. Using the averaged response, a 6-th order parametric IIR filter has been adjusted for each azimuth direction. The parameters of this filter are linked to the direction through a simple polynomial approximation. The details of this implementation are out of the scope of this paper.

The ITD delay between the two ears depends also on the direction and can be simplified as in Equation (2):

$$ITD(s) = \frac{c}{r} * (\varphi + \sin\varphi) \qquad [2]$$

where φ is the azimut angle of arrival as shown in Figure 3, c is the speed of sound and r is the head radius.

Note that this delay usually implies a fractional delay, which can be modeled efficiently using FIR or IIR filters. For our ITD model, an 8-th order IIR filter provides enough accuracy and an almost flat frequency response. However, going further, we analysed the effect of just having an integer delay versus having a fractional one. Using an integer delay and fade-in/fade-out mixes between time windows when sources are in movement, the difference between fractional and integer delays is almost imperceptible for most people and the spatial sensation is not degraded. Therefore, in the final implementation we decided to use integer delays to make a better use of the computational resources. The details and a rigorous explanation of the subjective testing carried out to arrive to the above conclusions are also beyond the scope of this paper.



$(\varphi,\theta) = (90^{\circ},0^{\circ})$
Median Plane
$(\varphi,\theta) = (0^{\circ},-90$
φ
Horizontal Plane
θ
$(\varphi,\theta) = (0^{\circ}, 90^{\circ})$
$(\varphi,\theta) = (0^{\circ},0^{\circ})$

Figure 3. *Binaural coordinate system.*

6. Transmission protocols design

In this work, proprietary transmission protocols have been designed to suit the needs of this particular conference call service. As explained before, we work over IP Internet protocol because it is not feasible to modify phone call protocols. We must distinguish two kinds of signals: the audio (voice) and the service management messages. For each of them we have developed a specific protocol that minimizes the required bandwidth.

6.1. Voice stream

The audio stream consists of compressed voice packets that must be sent between the terminals and the server. For transmitting these packets the UDP transport layer protocol has been chosen, the most suitable one for streaming real time audio. This protocol does not require establishing a point to point connection, is more agile and has less data overhead than TCP. The UDP protocol does not guarantee that the packets arrive to their destination, but the speed it provides is preferable for real time audio, because if a packet is lost it is not forwarded later. In addition, the Speex codec is ready (during the decoding stage) to estimate the correct signal when a packet is lost.

For this service, a protocol has been designed where each terminal sends over UDP its audio packets to the server, which replicates this stream to each of the other participants in the virtual room. A structure for the packets has been created where just a very small header is added to the audio compressed data, so the overhead is minimum.

This header contains a user identifier for the multi-party system and a sequence number (Fig. 4). In case the activity detector does not find voice in a frame, the terminal will send only the header, with no audio data. As a result, the receivers maintain synchronization with little bandwidth use. Each voice stream produced by a terminal uses a bandwidth between 2.4 kbps and 30.4 kbps, depending on the voice activity.

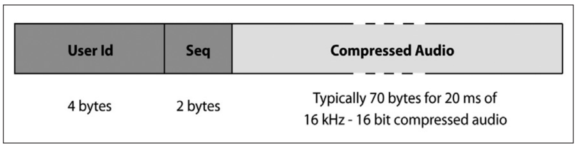

User Id
Seq
Compressed Audio
4 bytes
2 bytes
Typically 70 bytes for 20 ms of
16 kHz - 16 bit compressed audio

Figure 4. *Audio packet structure.*

6.2. Signaling messages

To manage the service, it is needed that terminals and server exchange messages to allow actions such as logging in, sending user's contacts, inviting to a conversation or sending customization pictures. These messages are short and require very little bandwidth, but losing any of these management packets is not acceptable. In addition, there should be a point to point connection between server and terminals to monitor the activity of each terminal.

For these reasons, signaling in this system works over TCP transport layer. A protocol based on the exchange of two kinds of messages, commands and raw data, has been designed. Signaling packets start with a flag indicating the type of message. In the case of containing a command, it is followed by the auxiliary information necessary. In the case of sending a file, it is divided into multiple packets of raw data. Figure 5 shows an example of signaling where the User 1 invites the User 2 to his room and later the User 3 asks permission to User 1 (the room's owner) for joining that room.

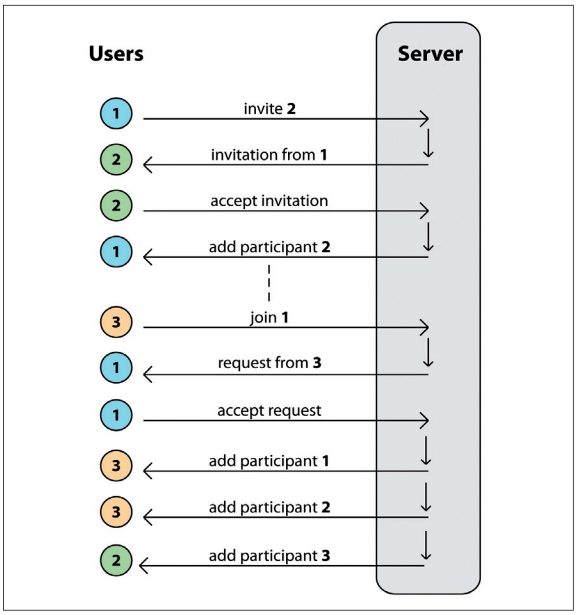


Figure 5. *Example of signaling to start a 3-participants multi-party conference. The number inside each color circle identify each different user.*

Proprietary transmission protocols have been designed to suit the needs of this particular conference call service.

7. Server and mobile devices implementation

The system has a client-server architecture where clients are the mobile devices (smartphones and tablets) and a PC has been used as a server (Fig. 6).

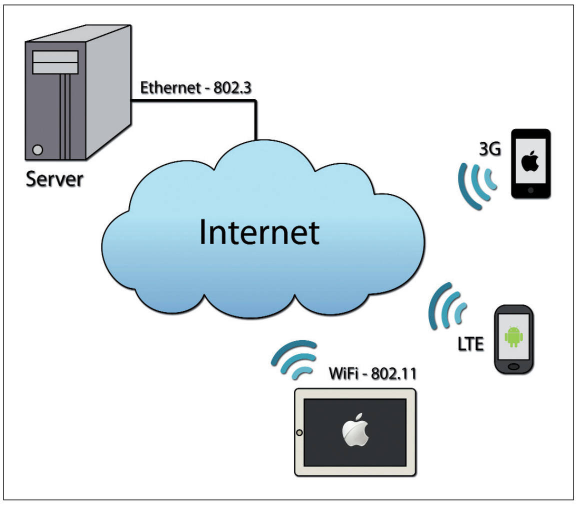


Figure 6. *Client-server architecture.*

The server has been programmed to perform the following functions at the lowest possible computational cost, thinking of future scalability issues:

- To manage the user's connection to the server, providing the information and status of their contacts.
- To manage conference invitations and participants in each room.
- To distribute the voice signal from each participant to rest in a room.
- To collect statistical data to analyze the strengths and weaknesses of the service.

The most important functions performed in the mobile devices are:

- Allow the user to log in, customize his room and ask for a conference to his contacts in a fast and intuitive way, so no previous training is needed.
- To provide a visual and tactile interface for the virtual room that allows each user to place the rest of participants at any position of the scene in comfortable way.
- To encode the voice signal captured by the microphone and send it to the server at the same time that decodes the signals of the rest of speakers received from the server.
- To synthesize the spatial audio scene from the audio data received and set the position of each participant.

According to several market studies, such as Gartner's report [16], currently, the most interesting platforms to implement and distribute applications are Android and iOS. Thus, to capitalize on the efforts of this work, we decided to implement this new service in these two platforms. In the case of smartphones, the application has been targeted to both iPhone and Android phones. In the case of tablets, we have chosen the iPad since it is the dominant tablet in the market by now.

The iOS operative system was created by Apple for the iPhone in 2007. In 2008 it launched the App Store, where any developer or company can sell their applications for iOS (after the approval of Apple). Applications are developed by means of their own SDK, being the programming language Objective-C [17]. This operative system is not open, in the sense that only Apple devices can use it, but has the advantage of being also used in other Apple devices as the iPod Touch or the iPad tablet.

Android is the biggest rival of iOS. It was released in 2008 by Google as an open operative system that can be used by different smartphone manufacturers. Nowadays Android is the platform with greater worldwide presence in smartphones. Google owns the Google Play store where third-party applications can be published, developing them through the provided SDK with Java as the main programming language [18]. The advantage of being used by very different devices sometimes becomes a disadvantage since it is difficult for the applications to fit perfectly on all models.

As for the graphical user interface, we aimed at achieving a trade-off in both platforms, sharing the same use philosophy but also respecting the differences to which users are used to. For this part of the application, high level libraries from each platform SDK have been used. Figures 7, 8 and 9 show the most important screens developed within the application. When programming all the audio processing, native C has been used for maximum performance in real time without involving the upper SDK layers that could slow down the processing.

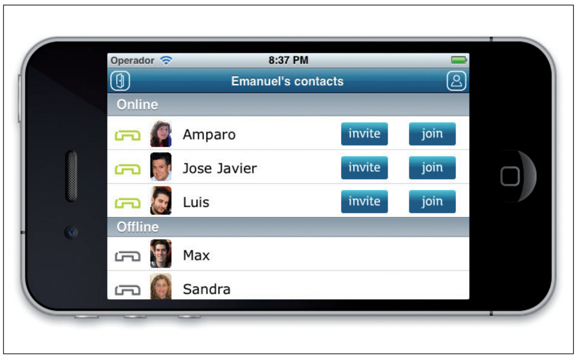


Figure 7. *User's contacts screen.*

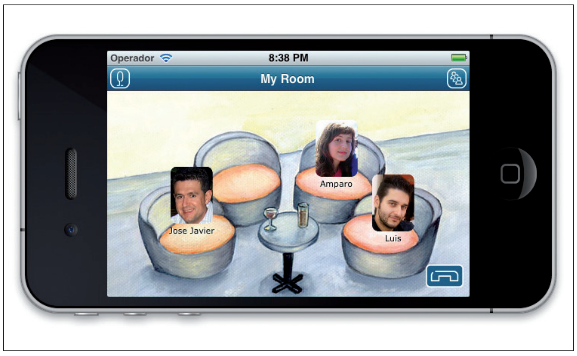


Figure 8. *Virtual room screen.*

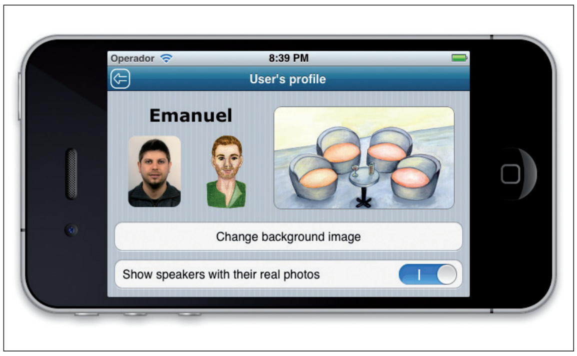


Figure 9. *User's profile and customization screen.*

8. Conclusions

In this paper, the development of an innovative multiparty conferencing system with 3D audio for mobile devices (smartphones and tablets) has been described. It can be very useful not only for business meetings but also for home and casual users. The system is based on a novel union of four existing technologies:

- Voice over IP on mobile devices.
- Multi-party conference services.
- Realistic 3D audio techniques based on binaural sound.
- Graphical interfaces over touch screen in mobile devices.

The combination of the above technologies substantially improves the intelligibility and identification of the speakers taking part in a multi-party mobile conference. In this context, an intensive research and development work has been carried out to optimize and adapt these techniques to run efficiently over a wide range of devices with limited resources and battery life.

The developed system has been extensively tested to verify that it meets the initial objectives in terms of robustness, usability and listening experience. The preliminary results show that the implemented protocols, despite being very simple, are very efficient and provide the application with a reliable data flow that rarely blocks the communication.

The user of this new service has the sense of being surrounded by other participants within a virtual room and experiences a more natural and realistic communication environment than that offered by other conventional services.

Future work will be aimed at scaling this experimental service to support a large number of users, adding more customization options and incorporating other spatial audio features, such as the use of a low cost head tracking [19] to adapt the perceived location of the participants dynamically.

Acknowledgments

This work has been partially supported by the government of Spain grant TEC-2009-14414-C03-01 and by the new technologies department of Telefónica.

References

[1] G. M. Olson and J. S. Olson, "Distance matters", Human-Computer Interaction, Volume 15, pp. 139-178, Lawrence Erlbaym Associates Inc, 2000.

[2] J. Sussman, J. E. Christensen, S. Levy, W. E. Bennett, T. V. Wolf, T. Erickson and W. A. Kellogg, "Rendezvous: Designing a VoIP conference call system", UIST, 2006.

[3] S. Deo, M. Billinghurst., N. Adams and J. Lehikioinen, "Experiments in spatial mobile audio-conferencing". Proceedings of the 4th international conference on mobile technology applications and systems and the 1st international symposium on Computer human interaction in mobile technology, Volume 7, ACM Press, pp. 447-451, 2007.

[4] S. Goose, J. Riedlinger and S. Kodlahalli, "Conferencing3: 3D audio conferencing and archiving services for handheld wireless devices", Wireless and Mobile Computing, Volume 1, 1, 2005.

[5] G. N. Marentakis and S. A. Brewster, "Effects of reproduction equipment on interaction with a spatial audio interface". Conference on Human Factors in Computing Systems CHI '05, pp. 1625-1628, Portland, USA, 2005.

[6] N. Yankelovich, W. Walker, P. Roberts, M. Wessler, J. Kaplan and J. Provino, "Meeting Central: Making distributed meetings more effective", in Proceedings of Computer Supported Cooperative Work (CSCW 2004), ACM Press, NY, pp. 419-428, 2004.

[7] A. Bronkhorst. "The Cocktail Party Phenomenon: A review on speech intelligibility in multiple-talker conditions". Acta Acustica united with Acustica, 86, pp. 117–128, 2000.

[8] J. J. Baldis, "Effect of Spatial Audio on Memory, Comprehension, and Preference during Desktop Conferences". Proceedings of the ACM Computer Humarn Interaction, Human Factors in Computing Systems Conference, pp. 166-173, Washington, USA, 2001.

[9] NTT Docomo Inc, "DOCOMO develops spatial audio transmission technology for mobile phones", Press Release (http://www.nttdocomo.com/pr/2009/001438.html), Tokio, Japan, 2009.

[10] J. Engdegard et al., "Spatial Audio Object Coding (SAOC) – The upcoming MPEG standard on parametric object based audio coding", AES Convention Paper, 2008.

[11] J. M. Valin, "Speex: A Free Codec For Free Speech", linux.conf.au, 2006.

[12] J. M. Valin, "The Speex Codec Manual Version 1.2 Beta 3", 2007.

[13] J. W. S. Rayleigh, "The Theory of Sound" (2nd Edition), New York, Dover Publications, 1945.

[14] J. Mackenzie, J. Huopaniemi, V. Välimäki and I. Kale, "Low-order modeling of head-related transfer functions using balanced model truncation", IEEE Signal Processing Letters, Volume 4, no. 2, pp. 39-41, Feb. 1997.

[15] J.J. López, M. Cobos, B. Pueo, "Elevation in wavefield Synthesis using HRTF cues", Acta Acustica, 96 (2), pp. 340-350, 2010.

[16] Gartner report about smartphone sales in first quarter 2011, http://www.gartner.com/it/page.jsp?id=1689814

[17] iOS Dev Center website, http://developer.apple.com/devcenter/ios/index.action

[18] Android Developers website, http://developer.android.com/index.html

[19] M. Billinghurst, S. Deo, N. Adams and J. Lehikoinen, "Motion-tracking in spatial mobile audio-conferencing", presented at Workshop on Spatial Audio for Mobile Devices (SAMD'07), Mobile HCI'07, Singapore, 2007.

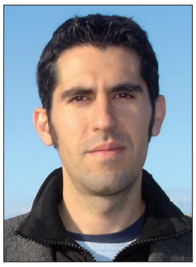
Biographies

Emanuel Aguilera

received a telecommunications engineering degree in 2004 and a M.S. degree in Artificial Intelligence, Pattern Recognition and Digital Image in 2011, both from the Universitat Politècnica de València, Spain. He is a researcher and senior programmer at the Institute of Telecommunications and Multimedia Applications (iTEAM), where he has been working since 2006 on the area of digital signal processing for audio, multimedia, virtual reality and mobile devices applications. He is interested in wave-field synthesis, image processing, real-time multimedia processing for telecommunications and audio applications for mobile platforms.

José Javier López

was born in Valencia, Spain, in 1969. He received the telecommunications engineer degree and the Ph.D. degree, both from the Universitat Politècnica de València, Valencia, Spain, in 1992 and 1999, respectively. Since 1993, he has been involved in education and research at the Communications Department, Universitat Politècnica de València, where he is currently a Full Professor. His research activity is centered on digital audio processing in the areas of spatial audio, wave field synthesis, physical modeling of acoustic spaces, efficient filtering structures for loudspeaker correction, sound source separation, and development of multimedia software in real time. He has published more than 160 papers in international technical journals and at renowned conferences in the fields of audio and acoustics and has led more than 25 research projects. Dr. Lopez was workshop co-chair at the 118th Convention of the Audio Engineering Society in Barcelona and has been serving on the committee of the AES Spanish Section for nine years, currently as secretary of the Section. He is a full ASA member, AES member and IEEE senior member.



Máximo Cobos

was born in Alicante, Spain, in 1982. He received the telecommunications engineer degree, the M.S. degree in telecommunication technologies, and the Ph.D. degree in telecommunications, all of them from the Universitat Politècnica de València, Valencia, Spain, in 2006, 2007, and 2009, respectively. His Ph.D. dissertation was awarded with the Ericsson Best Thesis Award on Multimedia Environments from the Spanish National Telecommunications Engineering Association (COIT). He completed with honors his Ph.D studies under the University Faculty Training program (FPU). In 2010, he was awarded with a "Campus of Excellence" post-doctoral fellowship to work at the Institute of Telecommunications and Multimedia Applications (iTEAM). In 2009 and 2011, he was a visiting researcher at Deutsche Telekom Laboratories, Berlin, Germany, where he worked in the field of audio signal processing for telecommunications. Since 2011, he has been an Assistant Professor in the Computer Science Department, Universitat de València. His work is focused on the area of digital signal processing for audio and multimedia applications, where he has published more than 50 technical papers in international journals and conferences. Dr. Cobos is a member of the Audio Engineering Society (AES) and the Institute of Electrical and Electronics Engineers (IEEE).

Luis Maciá

was born in Agost (Alicante), Spain, in 1983. He received the telecommunications engineering degree at the Universitat Politècnica de València, Spain, in 2007. He is currently studying the M.S. degree in electronic engineering taught at the same university. He performs his work as an electronic designer at the Institute of Telecommunications and Multimedia Applications (iTEAM) since 2008. His areas of work are FPGAs and DSPs programming, electronic circuits prototyping, embedded Linux programming, and software development for mobile devices.

Amparo Martí

was born in Valencia, Spain, in 1983. She received the degree in Electrical Engineering from the Universitat Politècnica de València, Spain, in 2008 and the MSc. Degree in Telecommunication Technologies in 2010. Currently, she is a PhD grant holder from the Spanish Ministry of Science and Innovation under the FPI program and is pursuing her PhD degree in Electrical Engineering at the Institute of Telecommunications and Multimedia Applications (iTEAM) working in the field of multichannel audio signal processing.