Multi-Speaker Localization, Separation and Resynthesis for Next Generation Videoconferencing

Máximo Cobos, José J. López, Laura Fuster, Emanuel Aguilera
Instituto de Telecomunicaciones y Aplicaciones Multimedia (iTEAM)
Universidad Politécnica de Valencia
Building 8G, access D, Camino de Vera s/n 46022 Valencia (SPAIN)
Corresponding author: jjlopez@dcom.upv.es

Abstract

Videoconference systems have been around the market for a long time. Their aim is to provide a way of carrying out meetings without the need for having physical presence of the participants. However, the sense of realism achieved by these systems is usually far away from the one expected by the people involved in the communication. In this paper, we present several advances in audio signal processing related to the captation, processing and reproduction of participants in a meeting environment. These novel approaches can be integrated into videoconference systems for making the sense of being there as real as possible. This paper is intended to be a brief summary of the work capacities existent in the iTEAM research institute for solving, from both a technical and practical perspective, all the technological challenges that high immersion videoconferencing will bring in the near future.

Keywords: Source Separation, Direction-of-Arrival, Videoconference, Wave-Field Synthesis, Spatial Sound.

1. Introduction

Videoconferencing is one of the most important applications merging audio and video in telecommunications. A videoconfence can be as simple as a conversation between two people in a private office (point-to-point communication) or it can involve several sites (multi-point communication) with several people in large rooms. Besides the audio and visual transmission of meeting activities, videoconferencing can also provide the possibility of sharing documents, computer-displayed information and even whiteboards. In fact, videoconferencing adds a possible alternative to traditional telephone/email communications when:

- a live conversation is needed;
- visual information is an important component of the conversation;
- the parties of the conversation can not physically come to the same location;
- the expense or time of travel is a consideration.

In addition, an important impact in education, medicine and health, business, law, etc. is expected for future videoconference systems. Despite all the advances in multimedia technologies in the last years, the mass adoption and use of videoconferencing is still relatively low. One of the reasons for this slow adoption is in the fact that participants still feel that the immersive sensation is insufficient. In this paper, we overview some advances in audio signal processing related to high realism communications. The goal is to achieve a video screen that appears to be a virtual window to the other side of the conference. A scheme of the proposed system is depicted in Figure 1.

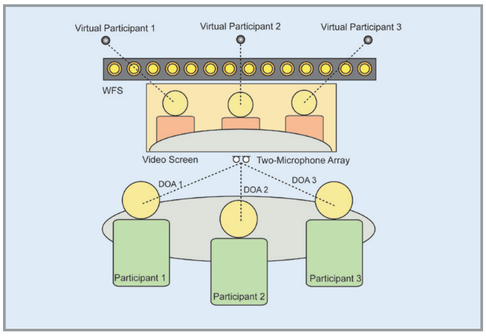

Figure 1. *Two-microphone set-up for DOA estimation.*

Despite all the advances in multimedia technologies, the mass adoption and use of videoconferencing is still relatively low

Firstly, a novel approach for localization of multiple speakers in meetings is used for estimating the azimuth positions (Direction-Of-Arrival or DOA) of the speakers using two close microphones. Then, a real-time source separation technique is applied to the speech mixtures in order to obtain the signal corresponding to each speaker. Finally, the separated speech signals and the positional information are used to set-up the virtual sources at the other side of the communication by means of a Wave-Field Synthesis (WFS) system. Practical issues and emerging loudspeaker technologies (Distributed Mode Loudspeakers) for the combination of WFS and video projection are also discussed.

2. Multi-Speaker Localization

Microphone arrays have been intensively studied in the last years due to their enhanced acoustic properties and their important applications in many speech processing systems, such as hands-free devices or hearing aids. One of the most active research lines in multichannel signal processing is acoustic source localization for videoconferencing. In fact, estimating the direction of arrival of multiple speakers in a real scenario is a very difficult task. Algorithms for acoustic source localization are often classified into direct approaches and indirect approaches [1]. Indirect approaches estimate the time delay of arrival (TDOA) between various microphone pairs and then, based on the array geometry, estimate the source positions by optimization techniques. On the other hand, direct approaches compute a cost function over a set of candidate locations and take the most likely source positions.

Small arrays are desirable for practical systems because they are cheaper and can be more easily integrated into practical devices. For this reason, two-microphone arrays have been receiving increasing attention in the last years. When using only two microphones, DOA estimation is usually performed via binaural localization cues. When a source is not located directly in front of the array, sound arrives slightly earlier in time at the microphone that is physically closer to the source, and with somewhat greater energy. This fact produces the interaural time difference (ITD) and the interaural intensity difference (IID) between the two sensors. DOA estimation methods based on binaural models, such as the Jeffress or equalization-cancelation models, have shown to successfully estimate locations of two sources in anechoic environments [3]. The DUET separation technique [2], which is also based on IID and ITD, can be used for estimating with high accuracy the TDOA of several sources in the time-frequency (TF) domain assuming that only one source is active in each TF point. In the next subsections we present a source localization technique developed by the authors based also on the time-frequency analysis of the microphone signals. The signal model and the different steps involved in the system are briefly described.

2.1. Signal Model

We consider a two-sensor array $(M = 2)$ to estimate the location of N sources in the azimuth plane: $\theta \in [0^o, 180^o]$, where θ is measured with respect to the array axis. In a real situation, each sensor captures not only the direct signal arriving from each of the sources, but also multiple reflections due to the effect of multi-path propagation. Therefore, the signal received by each microphone can be modeled as a sum of the original signals convolved with the impulse response corresponding to each source-sensor path. This convolutive mixture can be mathematically expressed as:

$$x_m(t) = \sum_{n=1}^{N} h_{mn}(t) * s_n(t) \quad m = 1, 2, \tag{1}$$

where $s_n(t)$ stands for the different sources, and $h_{mn}(t)$ is the impulse response between source n and sensor m. Considering only the direct path between each source and sensor, the simplified anechoic model is

$$x_m(t) = \sum_{n=1}^{N} s_n(t - \delta_{mn}) \quad m = 1, 2, \tag{2}$$

being δ_{mn} the time delay corresponding to the path between source n and microphone m.

Due to the non-stationarity of speech signals, processing of the microphone signals is usually carried out in the time-frequency domain. The Short-Time Fourier-Transform (STFT) allows to obtain a representation of the spectral content of the signal as it changes over time. The L-point STFT of a time-domain signal $x_m(t)$ sampled at frequency f_s is given by:

$$X_m(k,l) = \sum_{r=-\frac{L}{2}}^{\frac{L}{2}} x_m(l + r)\mathrm{win}(r)e^{(-j\omega_k r)}, \tag{3}$$

where k is a frequency index corresponding to angular frequencies $\omega_k \in \{0, 2\pi \frac{1}{L} f_s, ..., 2\pi \frac{L-1}{L} f_s\}$, win$(r)$ is a window that tapers smoothly to zero at each end, such as a Hann window, $\frac{1}{2}(1 + \cos \frac{2\pi r}{L})$ and l is the new time index. Given the linearity of the STFT, the model of Eq.(2) can be written as:

$$X_m(k,l) = \sum_{n=1}^{N} S_n(k,l)e^{-j\omega_k \delta_{mn}} \quad m = 1, 2, \tag{4}$$

where $S_n(k,l)$ is the STFT of source S_n. The main advantages of working with STFT representations are two: the first one is that convolutive mixtures can be approximated as instantaneous mixtures at each frequency and the second one is that the sparseness is higher. A signal is considered to be sparse if most of their coefficients are zero or close to zero. The sparsity of speech signals in the STFT domain makes able to consider that the sources are *W-disjoint Orthogonal* (WDO). Under this assumption, it is likely that every time-frequency point in the mixture with significant energy is dominated by the contribution of one source [4].

2.2. DOA Estimation

2.2.1. DOA map

Assuming plane wave incidence with angle θ and a inter-microphone distance d, the ITD between the two sources following the model of Eq.(4) is $(\delta_{1n} - \delta_{2n}) = \frac{d}{c}\cos\theta$, where c is the speed of sound (see Fig.2). Therefore, the phase difference observed between the two microphones in a given TF bin will be $\omega_k \frac{d}{c}\cos\theta$. As a result, we can estimate $\cos\theta$ for each TF point using the relation

$$D(k,l) = \frac{c}{\omega_k d} \angle \left(\frac{X_1(k,l)}{X_2(k,l)} \right), \tag{5}$$

where $\angle(\cdot)$ denotes the phase of a complex number. We call $D(k,l)$ the DOA map. Note that phase ambiguity appears for frequencies $\omega_k > \pi \frac{c}{d}$, so d is desirable to be small. Nevertheless, speech signals carry most of their information below 4 kHz and a separation of 5 cm is enough for obtaining good results.

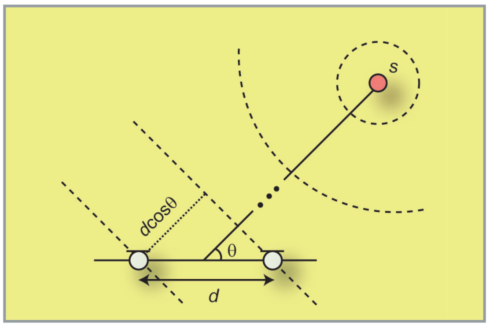

Figure 2. *Two-microphone set-up for DOA estimation.*

2.2.2. Coherence-Based Pre-Selection

The robustness of the source direction estimates to reverberation is improved by discarding TF bins where reverberation is dominant. These bins can be selected using the short-time coherence function [5], defined as:

$$\Phi(k,l) = \frac{|\Phi_{12}(k,l)|}{\sqrt{\Phi_{11}(k,l)\Phi_{22}(k,l)}}. \tag{6}$$

The statistics $\Phi_{ij}(k,l)$ are a practical way of computing the inter-channel correlation $E\{X_i(k,l)\, X^*_j(k,l)\}$, given by:

$$\Phi_{ij}(k,l) = (1-\lambda)\Phi_{ij}(k,l-1) + \lambda X_i(k,l)X_j^*(k,l), \tag{7}$$

where * denotes complex conjugation. Due to the non-stationarity of speech, the forgetting factor λ is introduced to compute the cross-correlation between the observation channels on a block of time frames. The coherence function $\Phi(k,l)$, has values close to one in TF regions where a source is present and it is usually smaller when sounds from different directions overlap. Our experiments suggest that taking values with $\Phi(k,l)>0.9$ gives good results. The effect of applying a coherence-based selection is a sharper histogram in which the sparse nature of speech is highly emphasized. Next, we describe how to obtain the DOAs of the sources using a fitted Laplacian Mixture Model (LMM).

2.2.3. Laplacian Mixture Model

Speech signals can be considered as having a sparse distribution in the STFT domain. There are a number of models that can be used to represent sparsity. One common probabilistic model is the Laplacian density function. The Laplacian density function is given by:

$$L=\beta e^{-2\beta|\theta-\gamma|} \tag{8}$$

where γ is the mean of the distribution of the random variable θ and $\beta>0$ controls the "width" or approximate standard deviation. Our purpose is to model the distribution of selected DOA estimates as a mixture of Laplacian distributions. Then, we take the set of selected points as the observed distribution to be fitted:

$$\theta_n = D(k,l)|^{(k,l)\in S}, \tag{9}$$

where $S=\{(k,l) \mid \Phi(k,l)>0.8\}$ is the set of TF bins selected based on their short-time coherence. The Expectation-Maximization (EM) algorithm [6] is employed to train a LMM over a training set

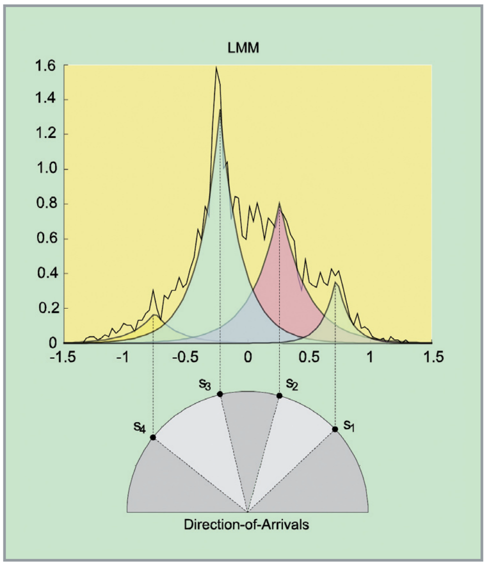

Figure 3. *LMM fitted to the observed distribution and corresponding DOAs.*

Simultaneous conversations among participants usually appear in live debates and business encounters, resulting in a degradation of speech intelligibility.

Wave-Field Synthesis is able to synthesize a desired sound field in a large listening area by means of loudspeaker arrays.

(batch-EM) or even adapt the parameters of the LMM in real time (On-line EM). The set of means obtained after convergence, γ_i will be the final DOAs of the sources. Figure 3 shows the observed distribution of DOA estimates and the fitted LMM. The peaks of the Laplacian functions are directly related to the cosines of the DOAs.

The DOA of each speaker is the spatial information needed in the other side of the communication to resynthesize the acoustic scene, as will be later described in Section 4.

3. Source Separation

In many meeting situations more than one speaker may be speaking at the same time. Simultaneous conversations among participants usually appear in live debates and business encounters, resulting in a degradation of speech intelligibility. The problem can be even more important if the meeting is being registered by an automatic speech recognition system. In order to deal with this situation, the signal for each speaker is separated from the mixture by means of a source separation technique. This enables to obtain a signal from each speaker without the need for having individual microphones.

The source separation problem can be stated as follows: given M linear mixtures of N sources mixed via an unknown MxN mixing matrix $\mathbf{A}$, estimate the underlying sources from the mixtures. When $M=N$, this can be achieved by estimating an unmixing matrix $\mathbf{W}$, which allows to estimate the original sources up to a permutation and a scale factor. Independent Component Analysis (ICA) algorithms [7] perform the separation assuming that the sources are non-Gaussian and statistically independent. If the mixture is underdetermined, $M<N$, the estimation of the sources becomes more difficult and sparse methods are used [8]. As described in the previous section, source localization and separation when there are more sources than mixtures is easier under sparse representations. Here, we also take profit of the sparseness given by the STFT in order to perform separation by means of a powerful technique: time-frequency masking [3].

Time-frequency masking attempts to construct a set of masks that can be applied to the mixtures in order to obtain the estimates of the sources:

$$Y_{mn}(k,l) = M_n(k,l)X_m(k.l), \tag{10}$$

being $Y_{mn}(k,l)$ the STFT of the image of s_n in sensor m and $M_n(k,l)$ is the separation mask. The estimates of the sources in the time domain are obtained applying the inverse STFT operator.

3.1. Separation based on DOA Segmentation

In this subsection, we describe a source separation algorithm based on TF masking [9]. Inspired by image segmentation techniques [10], separation is achieved by using a maximum interclass variance criterion between the angular distribution of the sources. With this criterion, it is possible to obtain a set of thresholds that divide the azimuth plane into angular sections corresponding to different speakers. We call this algorithm Convolutive Multi-Level Thresholding Separation (CMuLeTS). Multilevel thresholding can be exploited to achieve fast separation in reverberant scenarios by identifying different angular areas wherein the speakers are located with a strong likelihood. The method is based on a framework similar to the one described in Section 2. The idea is to treat the DOA map as a gray-level image that contains several objects. The objects in the image are extracted using the multi-level extension of the Fast Otsu Algorithm. Thus, DOA estimates are considered as being different gray-levels and segmentation is carried out by analyzing the distribution of a weighted histogram. The output of the algorithm is a set of thresholds that define different angular regions in the histogram. TF points lying between each pair of thresholds define the non-zero points of the binary masks used for separation. Figure 4 shows the spectrogram of the right input mixture and the binary masks obtained after segmentation for a mixture of four speakers. The estimated sources can be post-processed for reducing inter-source residuals and improve their isolation [11].

4. Wave-Field Synthesis for Videoconferencing

The simplest and best known method for providing spatial sound is stereo, which is able to position a source in the space using a pair of loudspeakers and amplitude panning. However, only a listener located in a middle position between the loudspeakers is able to localize correctly the source, otherwise, the localization accuracy is severely degraded. On the other hand, multichannel surround sound systems (5.1, 6.1, 7.1), well established in the cinema industry, are not suitable for videoconferencing. This is due to the fact that their main objective is the reproduction of special effects in movies, and the rear loudspeakers would not add any significant contribution in a meeting situation. In [12], Wave-Field Synthesis was proposed as a spatial sound system for videoconference, showing that the sweet spot extension offered by WFS is completely suitable for live-size videoconference systems with multiple participants. The spatial quality of resynthesized WFS scenes using source separation algorithms has been recently studied by the authors in [13].

4.1. Practical Constraints

Wave-Field Synthesis is able to synthesize a desired sound field in a large listening area by means of loudspeaker arrays. This makes the reproduced sound scene independent from the listening position, and therefore, the relative acoustic perspective perceived by a listener changes as he moves (Figure 5). The main idea of WFS was developed in the late 1980s. The Delft University of Technology worked on the idea of WFS, leading to the firsts prototypes [14][15]. WFS is also capable of synthesizing virtual sources both in front of and behind the array, and with a certain directivity characteristic. All of these properties make WFS the most powerful spatial sound reproduction system. However, creating a copy of a sound field is not completely possible due to some practical constraints:

- The discretization of an ideal continuous secondary source distribution to a loudspeaker array leads to spatial aliasing, resulting in both spatial and spectral errors in the synthesized sound field at high frequencies.

- The finiteness of the array leads to truncation effects, resulting in diffraction waves that cause after-echoes and pre-echoes.

- The restriction to a line loudspeaker array in the horizontal plane instead of a planar array leads to amplitude errors and restricts the localization to the horizontal plane.

Several methods for dealing with the problems mentioned are found in the literature [16]. Despite these technological issues, there is an inherent problem regarding the combination of WFS with video projection. Conventional loudspeakers have an important visual impact that can degrade the sense of immersion. In addition, there is usually the need for having two line arrays (one above and one below the screen) for giving the sensation that the sound comes from the screen itself. In order to deal with this problem, emerging loudspeaker technologies are being integrated into these systems. Distributed Mode Loudspeakers (DMLs) are a promising solution to this problem. In the next subsections we describe this new technology and how it can be used for WFS by means of Multiactuator Panels (MAPs).

4.2. Distributed Mode Loudspeakers

The DML essentially consists of a thin, stiff panel that vibrates in a complex pattern over its entire surface by means of a electro-mechanic transducer called exciter. The exciter is normally a moving coil device, which is carefully positioned and designed to excite the natural resonant modal structure of the panel optimally. In Figure 6, a graphical representation of a DML is presented, which shows panel, exciter and housing. DMLs are panels of finite extent deploying bending waves. The DML relies on the optimization of its eigenmodes to produce a modal density that is sufficiently high to give the impression of a continuous spectrum [17]. The excitation of bending waves on panels results in sound radiation with distinct qualities with regard to the pistonic motion of typical dynamic loudspeakers. A traditional loudspeaker acts for the most part of its radiation as a phase coherent radiator, and thus, it has a correlated output. However, the uncorrelated output of a DML produces an omnidirectional directivity response over the major part of the audio frequency band [18]. In addition to this, DML sources produce reflections that are less correlated to the direct sound than those radiated from piston sources and thus, constructive and destructive interference of sound is minimized.

One of the practical advantages of DMLs is their ease to mount directly on the wall surface. Besides, they are light-weight loudspeakers with a small back housing that can get unnoticed as part of the decoration. Since the panel surface can be large and the vibration is low enough to be imperceptible to the human eye, they can be integrated into a room interior and simultaneously used as projection screens [12]. This way, image and sound are fully integrated for multimedia applications. Furthermore, the cost of DMLs is generally lower than that of dynamic loudspeakers on baffles. These features make DMLs very suitable for WFS reproduction, which will be introduced in the next subsection. Encouraged by the positive results on sound localization, the applicability of single-exciter DMLs for WFS reproduction was tested for the first time in [19], reporting that individual panels reconstructed the wave field correctly. However, the secondary sources spacing required by the WFS algorithm to acquire a reasonable useful bandwidth, forced the size of panels to be very small. This conferred DMLs weak bass response due to the lack of excited modes in the low frequency region. In [20], Boone proposed to extend the DML technology to a panel with multiple exciters, each driven with a different signal. Such a configuration would act as a WFS array if every exciter on the panel would excite only a small part around the exciter position. Since exciters in a DML operate by converting electrical signals into mechanical movement which is applied to the panel, these panels are also known in the technical literature as Multiac-

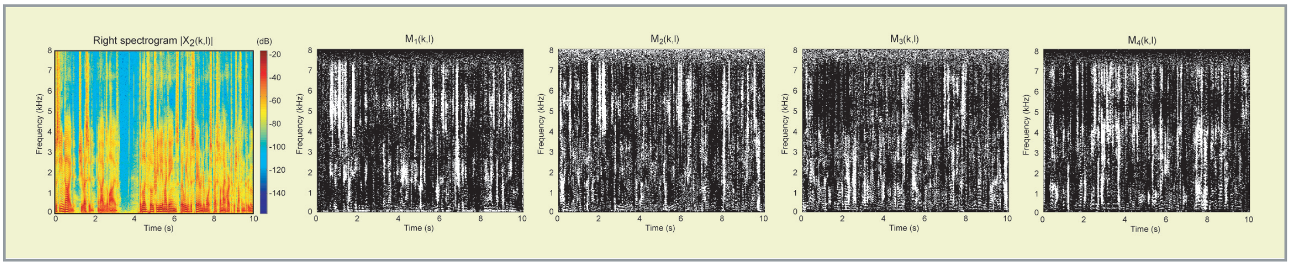

Figure 4. *Spectrogram and binary masks obtained after segmentation for a mixture of four male speakers.*

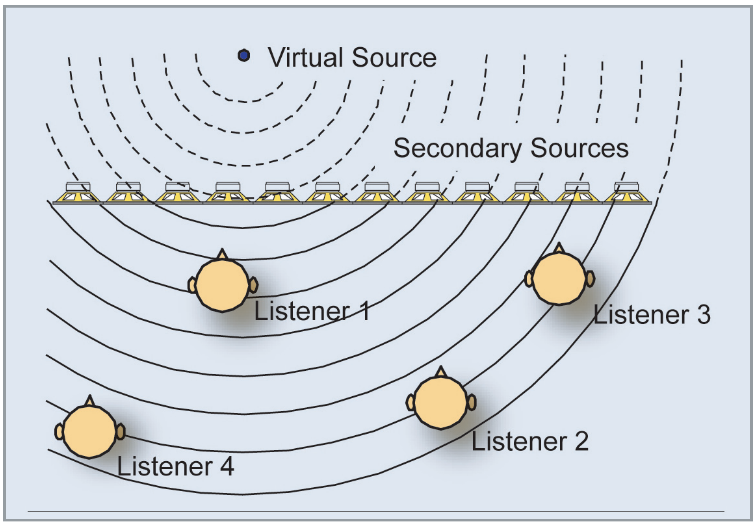

Figure 5. *Several listeners perceive correctly the location of a virtual source in a WFS system.*

tuator Panels (MAP). There are some benefits for MAPs to be used in WFS reproduction. They can be easily integrated into a living room because of its low visual profile. Furthermore, the vibration of the surface is almost negligible so that it can be used as projection screens.

4.3. Large MAPs for WFS and Video Projection

In this subsection we describe a prototype for video projection using MAPs. The well-known 3D displays that require the viewer to wear special glasses present two different images in the same display plane. The glasses select which of the two images is visible to each of the viewer's eyes. Technologies for this include polarization, shuttering or anaglyph. In this prototype we selected the shuttering technology were a double framerate was employed (left and right eye emitted alternatively) in combination with shutter glasses that blocked the opposite image. The projector employed was an InFocus DepthQ working at 120 Hz with DLP technology. For the projection screen a large MAP was especially designed and built (Fig. 7), to meet the demands of immersive audio applications. For that purpose, it included a horizontal line of exciters composed of 13 exciters with 18 cm spacing, presenting an aliasing frequency of approximately 1 kHz.

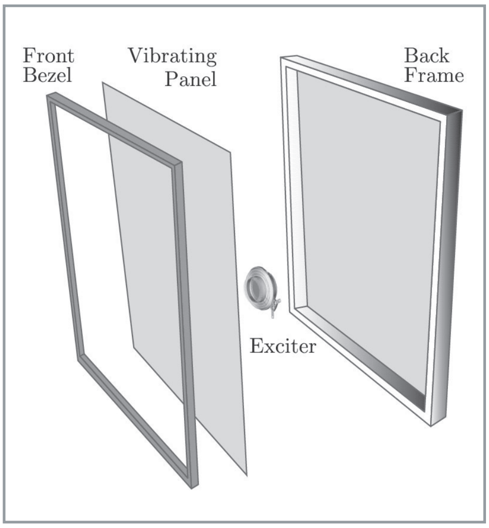

Figure 6. *Block diagram of a Distributed Mode Loudspeakers with only one exciter (wiring is omitted).*

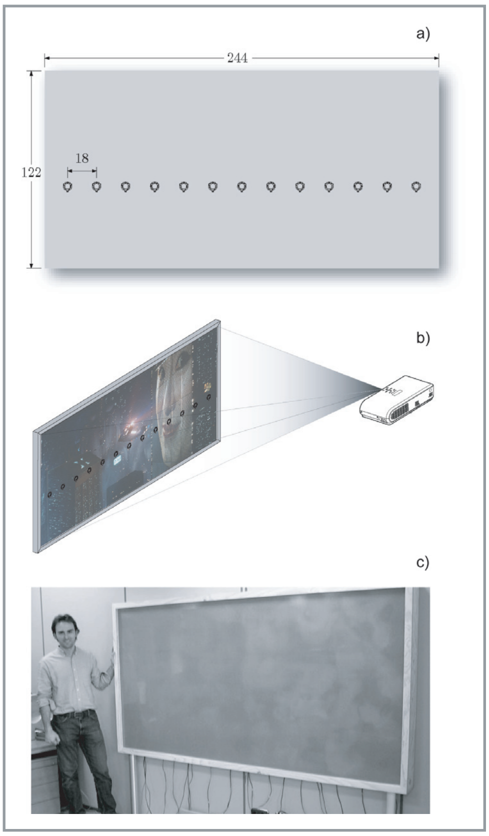

Figure 7. *Large MAP, a) block diagram and measures, b) employment in conjunction with a projector, c) photograph of the resulting prototype panel assembled and ready for use.*

The panel is a sandwich of polyester film bonded to an impregnated paper honeycomb 5 mm thick using a thermoplastic adhesive (cell size = 4.8 mm). Its bending rigidity is 4.23 and 2.63 Nm in the x and y directions respectively and has an areal density of 0.51 kg/m2. Due to its size, frequencies until 100 Hz can be reproduced successfully. More about the acoustic performance and audio quality of this panel was analyzed and previously presented by the authors in [21].

5. Conclusion

Videoconferencing is a complete telecommunication system that combines audio and video technologies in a challenging way. Although practical systems have been around the market for long time, there are still open problems regarding the sense of immersion of the participants. In this paper, several advances in audio signal processing and electroacoustics for future videoconference systems have been presented. These advances are related to the localization and separation of participants and their posterior resynthesis by means of Wave-Field Synthesis. Localization and separation are achieved by using a pair of omnidirectional microphones and applying time-frequency processing techniques to the input mixtures. On the other hand, a multiexcited DML in the form of MAP had also been presented as an alternative technology to conventional cone loudspeakers in WFS videoconferencing. The big size of the screen in conjunction with the realistic sound provided by WFS produced a better sense of immersion. This paper has summarized the capacities of the iTEAM research institute for solving, both from a technical and practical perspective, all the technological challenges that high immersion videoconferencing will bring in the near future.

Acknowledgment

This work has been partially supported by the Spanish Ministry of Education and Science under project TEC2006-13883-C04-01 and by the Spanish Administration Agency CDTI, under project CENIT-VISION 2007-1007.

References

[1] N. Madhu and R. Martin, Advances in Digital Speech Transmission,Wiley-Interscience, 2008.

[2] C. Liu, B. C. Wheeler, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," Journal of the Acoustical Society of America, vol. 108, no. 4, pp. 1888–1905, 2000.

[3] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Transactions on Signal Processing, vol. 52, no. 7, pp. 1830–1847, July 2004.

[4] S. Rickard and O. Yilmaz, "On the w-disjoint orthogonality of speech," in IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 529-532, Orlando, Florida, May 2002.

[5] C. Avendano and J.-M. Jot, "Frequency domain techniques for stereo to multichannel upmix," in Proc. AES 22nd Conf. on Virtual, Synthetic and Entertainment Audio, 2002, pp. 121–130.

[6] A. P. Dempster, N. Laird, and D. Rubin, "Maximum likelihood for incomplete data via the EM algorithm," Journal of the Royal Statistical Society, Ser. B, vol. 39, pp. 1–38, 1977.

[7] J. F. Cardoso, "Blind signal separation: Statistical principles," in Proccedings of the IEEE, vol. 86, no. 10. IEEE Computer Society Press, October 1998, pp. 2009–2025.

[8] S. Pedersen, J. Larsen, U. Kjems, and L. Parra, Springer Handbook of Speech Processing. Springer Press, 2007, ch. A Survey of Convolutive Blind Source Separation Methods.

[9] M. Cobos and J. J. Lopez, "Stereo audio source separation based on time-frequency masking and multilevel thresholding," Digital Signal Processing, vol. 18, no. 6, pp. 960–976, 2008.

[10] N. Otsu, "A threshold selection method from gray-level histogram," IEEE Transactions on System Man Cybernetics, vol. SMC-9, no. 1, pp. 62–66, 1979.

[11] M. Cobos and J. J. Lopez, "Improving isolation of blindly separated sources using time-frequency masking," IEEE Signal Processing Letters, vol. 15, pp. 617–620, 2008.

[12] W. de Brujin and M. Boone, "Application of wave-field synthesis in life-size videoconferencing," in Audio Engineering Society 114th Convention, Amsterdam, Netherlands, March 2003.

[13] M. Cobos and J. J. Lopez, "Resynthesis of wave-field synthesis scenes from stereo mixtures using sound source separation algorithms," Journal of the Audio Engineering Society, accepted for publication, 2009.

[14] A. J. Berkhout, "A holographic approach to acoustic control," Journal of the Audio Engineering Society, vol. 36, pp. 977–995, 1988.

[15] M. M. Boone, E. N. G. Verheijen, and P. F. van Tol, "Spatial sound field reproduction by wave field synthesis," Journal of the Audio Engineering Society, vol. 43, no. 12, pp. 1003–1012, 1995.

[16] H. Wittek, "Perceptual differences between wavefield synthesis and stereophony," Ph.D. dissertation, School of Arts, Communication and Humanities, University of Surrey, October 2007.

[17] J. W. Panzer and N. Harris, "Distributed-mode loudspeaker radiation simulation," in Audio Engineering Society 105th Convention, San Francisco, USA, September 1998.

[18] J. A. Angus, "Distributed mode loudspeaker polar patterns," in Audio Engineering Society 107th Convention, New York, USA, September 1999.

[19] M. Boone and W. Brujin, "On the applicability of distributed mode loudspeaker panels for wave field syntehsis based sound reproduction," in Audio Engineering Society 108th Convention, Paris, France, February 2000.

[20] M. Boone, "Multi-actuator panels (MAPs) as loudspeaker arrays for wave field synthesis," Journal of the Audio Engineering Society, vol. 52, no. 7-8, pp. 712–723, 2004.

[21] J. J. Lopez, M. Cobos, and B. Pueo, "Conventional and distributed mode loudspeaker arrays for the application of wave-field synthesis to videoconference," in Proccedings of the 11th International Workshop on Acoustic Echo and Noise Control, Seattle, USA, September 2008.

Biographies

Jose Javier Lopez was born in Valencia, Spain, in 1969. He received a telecommunications engineering degree in 1992 and a Ph.D. degree in 1999, both from the Universidad Politécnica de Valencia, Spain. Since 1993 he has been involved in education and research at the Communications Department of the Universidad Politécnica de Valencia, where at present he is an associate professor. His current 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. Dr. Lopez has published more than 100 papers in international technical journals and at renowned conferences in the fields of audio and acoustics and has lead several research projects. He was workshop cochair at the 118th Convention of the Audio Engineering Society in Barcelona and has been serving on the committee of the AES Spanish Section for 6 years, at present as secretary of the section. He is a member of the AES, full member of the ASA, and member of IEEE.

Maximo Cobos was born in Alicante, Spain, in 1982. He received a telecommunications engineer degree in 2006 and an M.S. degree in telecommunications technologies in 2007, both from the Universidad Politécnica de Valencia, Valencia, Spain. In 2009 he was a guest researcher at the audio group of the Deutsche Telekom Laboratories in Berlin, Germany, where he worked in the field of audio signal processing for telecommunications.

Currently, he is a grant holder from the Spanish Government and he is pursuing a Ph.D. degree in telecommunications engineering at the Universidad Politécnica de Valencia, where he works as part of the research staff of the Institute of Telecommunication and Multimedia Applications. His work is focused on the area of digital signal processing for audio and multimedia applications. He is interested in sound source separation, spatial sound, array signal processing and room acoustics. Mr. Cobos is a student member of the AES and the IEEE.

Laura Fuster was born in Valencia, Spain, in 1977. She received a telecommunications engineering degree from the Universidad Politécnica de Valencia, Spain in 2002. Since 2001, she has been working in the Audio and Communications Signal Processing Group of the Institute of Telecommunications and Multimedia Applications, where at present she is a research senior technician. During 2003, she was a collaborative researcher at the Friedrich Alexander Universität Erlangen-Nürnberg, Germany, where she worked in the development of multichannel equalization algorithms for acoustic panels used in wave-field synthesis rendering systems. Her current research interests include multichannel signal processing for audio, spatial sound reproduction and psychoacoustics.

Emanuel Aguilera was born in Buenos Aires, Argentina, in 1978. In 2004, he received a telecommunications engineering degree from the Universidad Politécnica de Valencia. Currently, he combines his M. S. studies in computer science with his research at the Institute of Telecommunications and Multimedia Applications, where he has been working for 3 years on the area of digital signal processing for audio, multimedia and virtual reality. He is interested in wave-field synthesis, image processing, pattern recognition and real-time multimedia processing for telecommunications.