2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Advanced Intelligent Acoustic Interfaces for Multichannel Audio Reproduction

Danilo Comminiello, Stefania Cecchi, Michele Gasparini, Michele Scarpiniti,
Aurelio Uncini and Francesco Piazza

Abstract— **Nowadays, there is a large interest towards multimedia audio systems as a consequence of the development of advanced digital signal processing techniques. In particular, immersive speech communication system has gaining increasing attention since they allow to reproduce realistic acoustic image, and thus achieving good performance in terms of sound quality and accuracy. In this scenario a fundamental role is played by intelligent acoustic interfaces which aim at acquiring audio information, processing it, and returning the processed information to the audio rendering system. In this paper, an effective intelligent acoustic interface composed of a microphone array and a signal processing system capable to enhance the intelligibility of the desired information of the transmitting room is proposed, combined with an advanced reproduction system based on an efficient application of a wave field synthesis technique capable to reproduce an immersive scenario in the receiving room. The whole system has been assessed within a speech communication application involving a moving desired source in a real scenario: objective and subjective evaluation has been reported in order to show the overall system performance.**

I. Introduction

IMMERSIVE audio reproduction has gaining increasing attention due to modern audio technologies which offer users a new way to enjoy audio services aimed at preserving the perceived quality of sound. In such scenario, a fundamental role is played by techniques used for the processing of audio information and its reproduction under the fulfillment of quality requirements demanded by users. In this context, one of the main application is the immersive speech communication, in which a near-end user must be "perceptually" steeped in a sound field, possibly together with other sound sources, thus having the perception to be share its sound space with a far-end user.

In order to enhance users' experience of immersive speech communication, *intelligent acoustic interfaces* (IAIs) [1], [2] can be adopted. An IAI is generally composed of a reproduction and/or an acquisition front-end which is coupled with an audio signal processor. An IAI is said to be "intelligent" due to its capability of adapting to user requirements and to environment conditions. Therefore, an IAI can be used both for the extraction of desired information from the acquisition of speech and sounds in a certain environment, and for the processing and enhancement of audio signals before reproducing them by loudspeakers.

In this work, we propose a complete system for immersive speech communication application considering moving sources scenario, based on the introduction of two different IAIs. The first IAI is placed in the transmitting room for the enhanced acquisition of the desired acoustic information. The other IAI is placed in the receiving room and aims at the reproduction of the desired information acquired by the first IAI. The main purpose in the transmitting room is to capture acoustic information in an attempt to provide user with a realistic perception of the sound field. To this end, an IAI for audio acquisition must be composed of a microphone array and a signal processing system aiming at reducing unwanted and interfering sounds which affect the intelligibility of the desired information. In this work, we adopt a commercial interface, the Microsoft Kinect, which is provided with a microphone array and camera sensors. The latter ones are used for the localization and the tracking of the desired source, being more robust than microphone arrays in absence of sound production from the desired source [3], [4]. Spatial information of the desired source are used by the IAI to steer its beam towards the located source. The adaptive beamforming technique adopted is based on a *generalized sidelobe canceller* (GSC) configuration involving a combined adaptive scheme for the cancellation of the noise arriving from the sidelobes [5]. Among the learning techniques that can be used in this process, single-layer neural networks, i.e., adaptive filters, are usually well suited to implement online adaptive beamforming [6], [7], [8].

The second IAI, in the receiving room, is appointed to reproduce the sound source tracked in the transmitting room together with its movement. To this end, many efforts have been made by using an acoustic interface in conjunction with reproduction techniques. One of the most popular techniques is the *wave field synthesis* (WFS), which is based on a large number of loudspeakers in order to achieve an optimal acoustic image in a larger area than in traditional systems [9]. Correct representation of moving sound sources using this technique is of utmost importance in immersive speech communication scenarios. Recently, a WFS-based IAI has been used for sound reproduction exploiting the far-end acquisition made by a Microsoft Kinect interface [10]. In this work, the advanced IAI used in transmission aims at enhancing the acquisition made by the Kinect, thus improving the listening quality in the receiving room.

Finally, the presented system involving the two IAIs is

Danilo Comminiello and Michele Scarpiniti and Aurelio Uncini are with the Department of Information Engineering, Electronics and Telecommunications (DIET), "Sapienza" University of Rome, Via Eudossiana 18, 00184 Rome, Italy (email: danilo.comminiello@uniroma1.it). Stefania Cecchi and Michele Gasparini and Francesco Piazza are with the Department of Information Engineering (DII), Universitá Politecnica delle Marche, Via Brecce Bianche, 60131 Ancona, Italy (email: s.cecchi@univpm.it).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

assessed within a real application involving a moving desired source. The evaluation is performed by using objective and subjective measures, to verify the spatial image perception and the localization of a moving desired source. The paper is organized as follows: in Section II the role of IAIs in immersive audio scenarios is discussed. Section III introduces an advanced IAI for extracting information from an acoustic environment in the presence of a moving source, while the IAI for reproducing the spatial sound is described in Section IV. The evaluation of the proposed system is presented in Section V, reporting objective results and the subjective experience of preliminary listening tests. Finally, the conclusions are reported in Section VI.

II. Intelligent Acoustic Interfaces

An acoustic interface is defined as the front-end of an audio and speech signal processing system aiming at the extraction and reproduction of acoustic information. An acoustic interface is generally composed of a microphone array and/or one or more loudspeakers. An acoustic interface is said to be *intelligent* [1], [2] when an audio signal processor is connected to microphones and loudspeakers and controls the information to be extracted or reproduced. Such information processing makes the interaction of a user with the machine more intuitive and helpful, since it translates acoustic information from user to machine, and *vice versa*, in order to allow a homogeneous interaction between parties. An IAI should be as invisible and intuitive as possible for a user, which should be able to concentrate on the task which he is going to perform. Therefore, "intelligence" in this context does not actually imply cognition, but it wants to point out the use of acoustic information in an appropriate manner [11]. An IAI is capable of adapting to user requirements and/or to environment conditions. In many cases, an IAI must learn from user behaviors, mood and personality, in order to yield a response as compliant as possible to user needs. Moreover, an IAI could also exploit feedback from user to improve its processing [12].

IAIs can be widely used in several fields of application such as: speech/audio real-time interaction, automatic speech analysis, automatic music composition and transcription, automatic genre and context recognition in broadcast programs, high-interactivity entertainment, realization of "intelligent rooms" in which both speakers and speech commands must be recognized and the perception of acoustic impulse responses can be controlled. An IAI for immersive audio aims at extracting useful information for computational or human purpose, such as analysis or synthesis of audio signals. At the same time, an IAI must reproduce desired acoustic information taking into account that the listener would hear the sound exactly as in the original sound field.

To this end, in this work we describe an immersive speech communication scenario which involves two IAIs: on one hand an IAI is committed to acquire the information from the desired source while reducing interfering noise; on the other hand, an IAI is used to reproduce the enhanced information recreating the sound space of the transmitting environment. In this context, interfering signals and non-stationarity of signals and sources may degrade quality and intelligibility of the desired information. Therefore, the acquisition of desired signals with high quality is far more difficult and challenging for immersive scenarios than for the classical case when one microphone is close to the user. Moreover, the high-quality reproduction of acquired signals is made more difficult because of the possible movements of the desired source within the transmitting environment. In the next sections we describe the two IAIs used to address these problems, thus providing a high-quality immersive speech communication.

III. Multichannel Acquisition System

In this section we introduce an IAI which is used in the transmitting side of an immersive speech communication to reduce unwanted noise and enhance the desired information. The proposed IAI, depicted in Fig. 1, involves a sensor interface and an adaptive beamforming system, which are detailed below.

A. Adopted sensor interface

As regards the interface, a Microsoft Kinect device is used. The Kinect is one of the most famous among this kind of new devices; although it was originally thought as an innovative video games controller, its functionalities are suitable for many other applications, beyond entertainment. The Kinect is mainly composed of two cameras and a microphone array. The principal features of the device are resumed in Table I.

In the proposed system, the Kinect has been used to get the source position, taking advantage of the Kinect SDK capabilities. The identification of human faces in pictures or videos is a common field of signal processing and many algorithms have been formulated in order to efficiently solve this particular task. The proposed application exploits the face tracking functionalities of the Microsoft Kinect for Windows Developer Toolkit 1.5. The toolkit offers several functions to detect in real time a face in the scene, get its distance from the sensor and the $x - y$ coordinates of some specific points of the face in the 3D space. In order to manage the retrieved data in real time and send them to the WFS algorithm, a NU-Tech plugin has been implemented [13], [10]. At each computation frame, the plugin returns the x and the y coordinates of the lips of the speaker.

Device	**Features**
RGB camera	640×480, 32-bit, 30fps
Depth camera	320×240, 16-bit, 30fps
Audio	4 microphones array, 16kHz, 16-bit

TABLE I
KINECT HARDWARE SPECIFICATIONS.

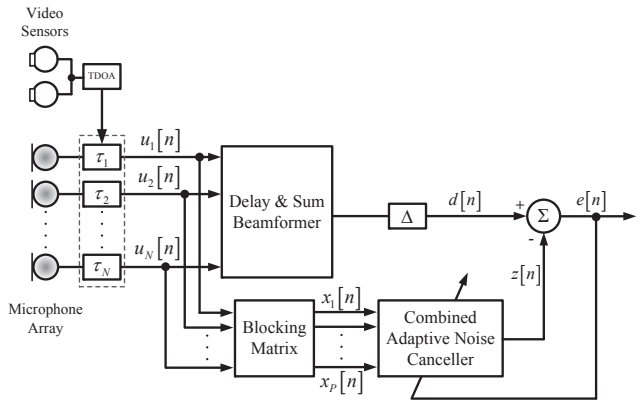

Fig. 1. An intelligent acoustic interface for noise reduction.

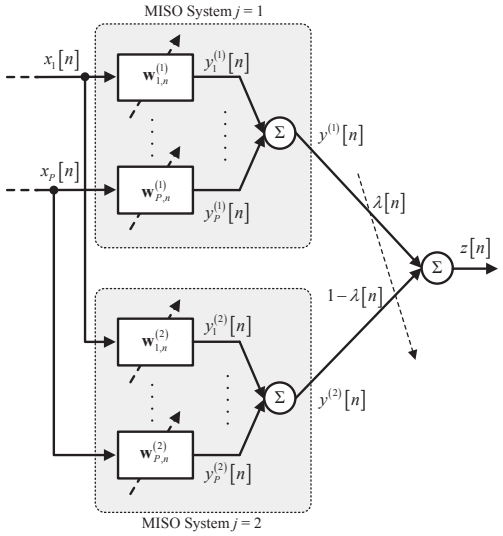

Fig. 2. Combined adaptive noise canceller scheme.

B. Combined adaptive beamforming technique

The beamforming technique adopted for the acquisition IAI is based on the classic *generalized sidelobe canceller* (GSC) composed of a microphone array interface, a fixed *delay-and-sum beamformer* (DSB), and an adaptive sidelobe cancelling path, as depicted in Fig. 1. Let us consider a microphone array interface composed of N elements. The signal $u_i[n]$ received by the i-th microphone, with $i = 1, \ldots, N$, is a delayed replica of the target signal $s[n]$ convolved with the acoustic impulse responses. Exploiting the position sources provided by the video sensors of the Kinect, the *time difference of arrival* (TDOA) is computed and then used to align the microphone signals with reference to the desired source direction. The aligned signals are then processed by the DSB which yields the reference signal $d[n]$. In the adaptive path of the beamformer, the *blocking matrix* (BM) generates the noise references $x_p[n]$, with $p = 1, \ldots, P$, being $P = N - 1$. In this paper, the blocking matrix, denoted with $\mathbf{B} \in \mathbb{R}^{P \times N}$, is implemented by pairwise differences between sensor signals [14] (i.e. the sum of the elements of each row is null). The noise reference signals are then processed by means of the *combined adaptive noise canceller* (CANC), whose goal is to remove any residual noise components in $d[n]$, thus minimizing the output power and yielding the beamformer output signal $e[n]$.

The main characteristic of the proposed beamforming approach is represented by the structure of the CANC. Generally, a conventional *adaptive noise canceller* (ANC) is composed of an adaptive *multiple-input single-output* (MISO) filter. Conversely, the adopted architecture results from the adaptive combination of 2 different MISO systems [5], as depicted in Fig. 2, each one bringing different filtering capabilities to the whole beamforming system. The MISO filters receive the same input signals, which are the noise reference signals resulting from the BM, collected into the data vector:

$$\mathbf{x}_n \in \mathbb{R}^{P \cdot M \times 1} = \left[\begin{array}{ccc} \mathbf{x}_{1,n}^T & \ldots & \mathbf{x}_{P,n}^T \end{array} \right]^T \tag{1}$$

which is a concatenation of P input channel signals:

$$\mathbf{x}_{p,n} \in \mathbb{R}^{M \times 1} = \left[\begin{array}{ccc} x_p[n] & \ldots & x_p[n - M + 1] \end{array} \right]^T \tag{2}$$

where $p = 1, \ldots, P$ and M denotes the length of each filter of the two MISO systems. The coefficient vector of the p-th filter belonging to the j-th MISO system, with $j = 1, 2$, is represented at n-th time instant as:

$$\mathbf{w}_{p,n}^{(j)} \in \mathbb{R}^{M \times 1} = \left[\begin{array}{ccc} w_{p,0}^{(j)}[n] & \ldots & w_{p,M-1}^{(j)}[n] \end{array} \right]^T. \tag{3}$$

For each MISO filter, the coefficient vectors (3), for $p = 1, \ldots, P$, are collected in a unique vector:

$$\mathbf{w}_n^{(j)} \in \mathbb{R}^{P \cdot M \times 1} = \left[\begin{array}{ccc} \mathbf{w}_{1,n}^{(j),T} & \ldots & \mathbf{w}_{P,n}^{(j),T} \end{array} \right]^T. \tag{4}$$

The output of each MISO filter is then achieved as:

$$y^{(j)}[n] = \mathbf{x}_n^T \mathbf{w}_{n-1}^{(j)}. \tag{5}$$

Taking into account the reference signal yielded by the DSB, it is possible to obtain the error signal for each MISO filter:

$$e^{(j)}[n] = d[n] - y^{(j)}[n]. \tag{6}$$

Each MISO filter is adapted by using a multichannel *normalized least mean squares* (NLMS) (see for example [15], [8]):

$$\mathbf{w}_n^{(j)} = \mathbf{w}_{n-1}^{(j)} + \mu_j \frac{\mathbf{x}_n e^{(j)}[n]}{\|\mathbf{x}_n\|^2 + \delta} \tag{7}$$

where μ_j is the *step size* for the j-th MISO filter and δ is the *regularization factor*, which is the same for both the MISO filters.

The outputs $y^{(j)}[n]$ of the two MISO filters are convexly combined according to *system-by-system* combination scheme [5], depicted in Fig. 2, since it involves the minimum necessary number of free parameters for single-stage

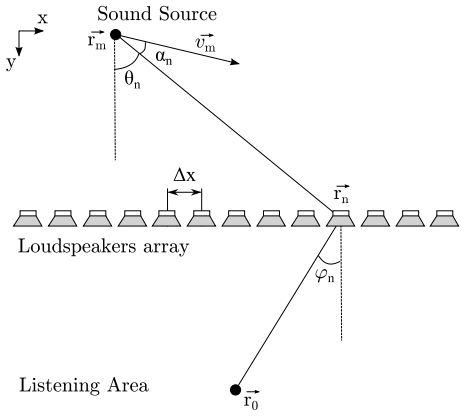

Fig. 3. Geometry used for the driving functions.

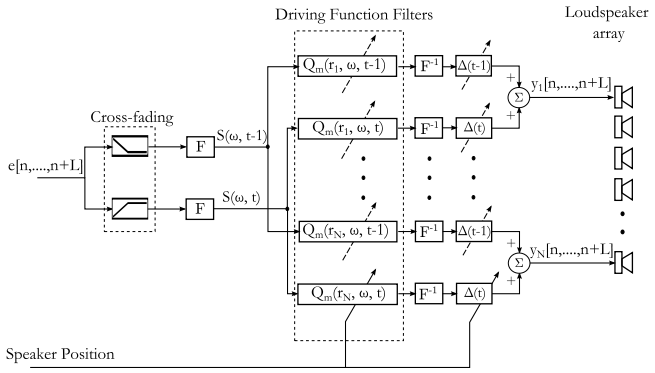

Fig. 4. Multichannel Reproduction System for moving source scenario: the block F represents the transition to the frequency domain, $Q_m(\vec{r_n}, \omega, t)$ are the driving functions, while $\Delta(t)$ indicates the introduced delay.

combined architectures. Therefore, the overall output of the CANC $z\,[n]$ can be achieved as:

$$z\,[n] = \lambda\,[n]\, y^{(1)}\,[n] + (1 - \lambda\,[n])\, y^{(2)}\,[n] \tag{8}$$

where $\lambda\,[n]$ is the adaptive *shrinkage parameter*. In order to satisfy the convex constraints, $0 \leq \lambda\,[n] \leq 1$, a variable change is performed involving a sigmoid function:

$$\lambda\,[n] = \beta \left(\frac{1}{1 + e^{-a[n-1]}} - \alpha \right), \tag{9}$$

where

$$\alpha = \frac{1}{(1 + e^4)}, \qquad \beta = \frac{1}{1 - 2\alpha}. \tag{10}$$

The auxiliary parameter $a\,[n]$ is adapted by using a gradient descent rule [16], [5]:

$$\begin{aligned} a\,[n] = a\,[n-1] + \frac{\mu_c}{\beta r\,[n]} e\,[n] \left(y^{(1)}\,[n] - y^{(2)}\,[n] \right) \\ \cdot \left(\lambda\,[n] + \alpha\beta \right) \left(\beta - \alpha\beta - \lambda\,[n] \right) \end{aligned} \tag{11}$$

where $\mu_c / r\,[n]$ represents a normalized step size, $r\,[n] = \gamma r\,[n-1] + (1-\gamma) \left(y^{(1)}\,[n] - y^{(2)}\,[n] \right)^2$ is the estimated power of $\left(y^{(1)}\,[n] - y^{(2)}\,[n] \right)$, and γ is a smoothing factor. The error signal $e\,[n]$ in (11), which represents the overall beamformer output, is derived as:

$$e\,[n] = d\,[n] - z\,[n]\,. \tag{12}$$

IV. Multichannel Reproduction System

The goal of every spatial multichannel audio technique is the exact reproduction of the acoustic image. Over the years Wave Field Synthesis (WFS) has been resulting in one of the most effective method on reaching high spatial precision, especially in video conferencing and entertainment systems. However, WFS high computational load, due to the great number of loudspeakers and processing channels that have to be used in order to obtain a sufficient quality for the reproduced field, constitutes a limit for its applicability. Real time algorithms capable of reducing the computational complexity are hence essentials; to this aim, efficient algorithms have been presented over the years for the synthesis of static sources fields [17]. On the contrary, the reproduction of moving sound sources is still problematic, even if it is of primary importance for a true spatial reconstruction of the sound scene. The movement entails the need of making use of time varying filters with a significant increase of computational load [18]. In this context, the work presented in [19], [20], [10] deals with an efficient real time WFS algorithm capable of reproducing moving sound sources. This is possible taking into consideration that each source trajectory can be approximated with two virtual sources at the same time, considering also phase approximation and fractional delays [17].

In this paper, the proposed idea consists in using a WFS reproduction system for moving sources taking advantage from the proposed IAI used in the transmitting room. The source tracking capability is fundamental every time a moving source, whose trajectory is not known *a priori*, has to be represented. Typical applications of this reproduction systems are remote conference scenarios or theater performances.

Traditional formulation of the WFS is no more correct when source movement is supposed, and some modifications have to be made in the driving functions equations. In particular, in the traditional theory, source velocity is not considered and the source position is assumed to be constant against time. Common conventional approaches sample the source trajectory reducing the calculation to a stationary problem, without explicitly considering the features of the sound field. In the presence of movement, the static driving function Q_m then becomes:

$$\begin{aligned} Q_m(\vec{r_n}, \omega, t) = \frac{S(\omega)}{D_n(\varphi_n, \omega)} \sqrt{\frac{jk}{2\pi}} \times \\ \times \sqrt{\frac{|y_0 - y_n|}{|y_0 - y_m(t)|}} \cos\theta_n(t) \frac{e^{-jk|\vec{r}_m(t) - \vec{r}_n|}}{\sqrt{|\vec{r}_m(t) - \vec{r}_n|}}, \end{aligned} \tag{13}$$

where $S(\omega)$ is the Fourier transform of the source signal, φ_n is the angle between the vector normal to the loudspeakers array $\vec{n}$ and the vector $|\vec{r}_0 - \vec{r}_n|$, θ_n is the angle between the vector normal to the loudspeakers array $\vec{n}$ and the vector $|\vec{r}_n - \vec{r}_m|$ and $D_n(\varphi_n, \omega)$ is the directivity function for the

n-th loudspeaker.

Resulting filters are time varying and are not suitable for a real time implementation. In order to achieve computational efficiency, source signal is divided into frames of a fixed samples number and the source trajectory is sampled considering a different position for each frame. In this way, driving function of eq.(13) can easily be evaluated, since the position is considered fixed in each time frame. Nevertheless, this approach leads to heavy phase discontinuities between frames and can not correctly represent particular features like Doppler effect. An accurate analysis of the main artifacts deriving from this formulation is described in [21], [22]. Moreover, in [21] a basic framework for the derivation of the secondary source driving functions from the field of moving sources is presented. Although this approach allows to avoid some specific artefacts related to the movement of the virtual source, it is not suitable for a real time implementation because of the difficulty in finding a closed form solution for the driving function. In [19] a modified version of the traditional driving function is proposed, in order to take into account the features of the movement. The resulting function is

$$Q_m(\vec{r_n},\omega,t) = \frac{S(\omega)}{D_n(\varphi_n,\omega)}\sqrt{\frac{jk}{2\pi}}\times \\ \times\sqrt{\frac{|y_0-y_n|}{|y_0-y_m(t)|}}\frac{\cos\theta_n(t)}{1-M\cos(\alpha_n)}\frac{e^{-jk|\vec{r}_m(t)-\vec{r}_n|}}{\sqrt{|\vec{r}_m(t)-\vec{r}_n|}}, \tag{14}$$

where $\vec{r}_m$ points to the position of the virtual source, α_n represents the angle between source velocity vector $\vec{v}_m$ and $\vec{r}_n - \vec{r}_m$, and M is the Mach factor, as reported in Fig. 3. However, in order to reduce the phase discontinuity, each moving virtual source is approximated with two sources as shown in Fig. 4. The former was the virtual source itself and the latter was a copy of the virtual source, named *retarded source*, whose position is delayed by a frame-time. The sources audio signals were then combined using a cross-fading technique based on ramp functions. Maintaining the double source framework just described, the audio processing is based on the implementation of the WFS filtering by shift registers and module weighted fractional delays. Fractional delay filters are computed offline with an upsampling factor of 100; efficient structures for the application of fractional delays in WFS systems can be found in [23]. It is known from the literature that filters orders of 31 or 63 result in a good approximation for the driving functions [17]. In the proposed implementation 100 filters of 64 samples have been employed, resulting in a minimum overlap and save frame of the same length. For each frame position the driving function phase is the following

$$Q_m^{\phi}(\vec{r_n},\omega) = ae^{-jak|\vec{r}_m-\vec{r}_n|}\sqrt{ja\ \mathrm{sgn}(\omega)}, \tag{15}$$

where a is a constant, which is equal to 1 or -1 in case of virtual sources behind or in front of the array, respectively, and it is evaluated and divided into integer and fractional delay part. The former is implemented by mean of a shift register and the latter is employed to select the precalculated fractional delay filter [17] windowed by the module part of the driving function of eq. (14):

$$Q_m^{|\cdot|}(\vec{r_n},\omega) = \frac{|\cos\theta_n|}{\sqrt{|\vec{r}_m-\vec{r}_n|}}\sqrt{\frac{|\omega|}{2\pi c}}\times \\ \times\sqrt{\frac{|y_0-y_n|}{|y_0-y_m|}}\frac{\Delta x}{1-M\cos(\alpha_n)}. \tag{16}$$

Then, the two sources signals obtained in this way are summed together in the output streaming and the whole elaboration has to be repeated for each loudspeaker.

V. Experimental results

In this section we first show the convergence properties of the combined adaptive beamformer used for multichannel audio acquisition, and then the proposed multichannel reproduction system is evaluated in a real scenario.

A. Convergence properties of the combined adaptive beamformer

In the first set of experiments we show the convergence performance of the adaptive beamforming system described in Section III. The simulated scenario is that of a reverberant room with size $6 \times 5 \times 3,3$ m in which a microphone array captures the sound emitted by a desired source which moves in the room. In particular, a *uniform linear array* (ULA) of $P = 4$ microphones is adopted with a spacing of 5 cm. The desired source assumes three positions in the room during the experiments, in all of which it remains for the same time. For each position of the desired source, the 4 acoustic impulse responses (AIRs) are measured by an image source method, using a reflection factor of the walls of $\rho = 0.78$ with a sampling rate of 8 kHz. Each AIR is truncated after $M = 280$ samples. The signal generated by the desired source is a colored signal obtained by using a first-order autoregressive model, whose transfer function is $\sqrt{1-\alpha^2}/\left(1-\alpha z^{-1}\right)$, with $\alpha = 0.8$, fed with an i.i.d. Gaussian random process. The length of the input signal is $L = 12000$ samples. Additive i.i.d. white Gaussian noise signal $v[n]$ is added to each microphone signal in order to provide 20 dB of *signal to noise ratio* (SNR).

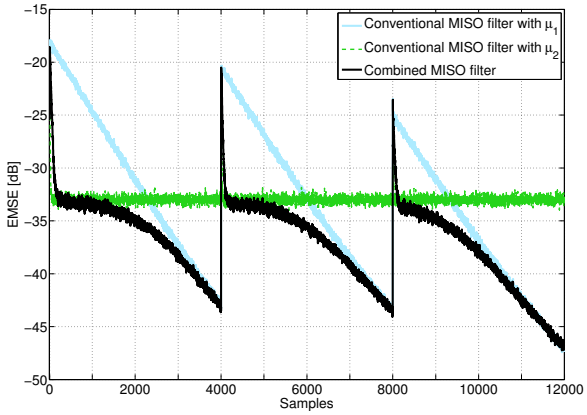

Fig. 5. Convergence performance of the CANC in terms of the ERLE.

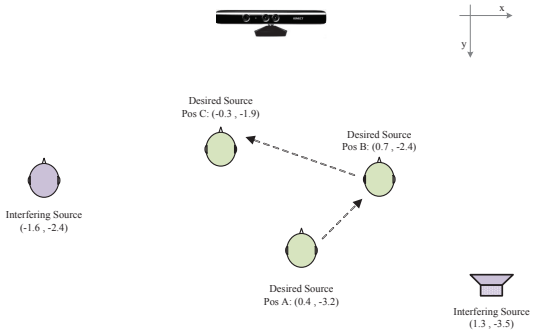

Fig. 6. Experimental setup for the transmitting room.

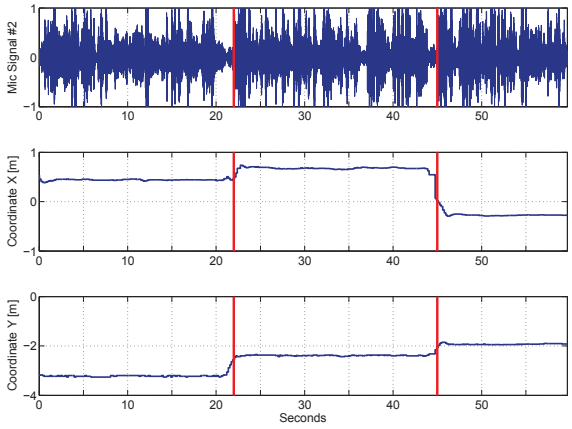

Fig. 7. Tracking of the desired source in the transmitting room.

We compare the performance of two adaptive beamformers employing a conventional ANC, one with a small step size, $\mu_1 = 0.001$, and the other one with a larger step size, $mu_2 = 0.01$, with that of the proposed adaptive beamforming involving the combination of the two individual ANCs. The ANCs (and consequently the CANC) are adapted by using a multichannel NLMS algorithm, as described in Section III, with a regularization factor $\delta = 0.001$. The combined MISO filter of the CANC uses a step-size value $\mu_c = 0.5$ and the following initial setting for the adaptation of the auxiliary parameter: $a[0] = 0$, $r[0] = 1$, and a smoothing factor $\beta = 0.9$.

Convergence is evaluated in terms of the *excess mean square error* (EMSE), which is defined (in dB) as: $\mathrm{EMSE}[n] = \mathrm{E}\left\{(e[n] - v[n])^2\right\}$, where the operator $\mathrm{E}\{\cdot\}$ denotes the mathematical expectation. The EMSE is evaluated over 1000 independent runs.

Results are depicted in Fig. 5, where it is worth noting that the conventional ANC with μ_1 shows a slow convergence rate when the desired source assumes a new position but a good precision at steady state when the source keeps its position. Conversely, the ANC with μ_2 provides faster convergence rate but lower precision. The combined ANC is capable of exploiting the advantages of both the MISO filters, thus showing fast convergence rate and good precision at steady state.

B. Evaluation of the multichannel reproduction system

1) Experimental setup: In order to evaluate the effectiveness of the proposed approach a real conference scenario has been set up. The scenario involves a transmitting and a receiving room, which are both office rooms with typical furniture (e.g. desks, cabinets, etc.).

In the transmitting room, the IAI described in Sec. III acquires the desired speech, together with interferences and background noise, and send it enhanced to the receiving room. The experimental scenario is depicted in Fig. 6 and involves three sources, a desired speech and two interfering sources. The desired source is a male speaker which moves head on the IAI within the aperture area of the microphone array comprised between -45 and 45 degrees. The interfering source on the left of the desired source is a male speaker, while on the right of the desired source a loudspeaker reproducing music is located.

The length of the experiment is about 60 seconds, during which the desired source assumes three different positions, while the interfering sources keep their positions all along the experiment. In Fig. 6 the movement of the desired source is depicted, starting from an initial position A, which is the farthest from the IAI, to positions B and C. The processing has been carried out at a sampling frequency of 48 kHz. The parameter setting of the CANC of the beamformer is the same of the previous experiment.

The IAI tracks the movements of the desired source by using the camera sensors of the Kinect interface, and the source coordinates are used by the IAI to steer the proposed adaptive beamforming system. The source tracking with respect to the acquired signals is represented in Fig. 7, in which it is evident the coordinate change when the desired source approaches positions B and C.

Once enhanced, the signals are sent to the transmitting room for the reproduction. In the receiving room, a linear array of 8 loudspeakers has been used for the synthesis. The array was driven by a workstation running NU-Tech with the WFS plugin [10], [13].The setup for the receiving room is

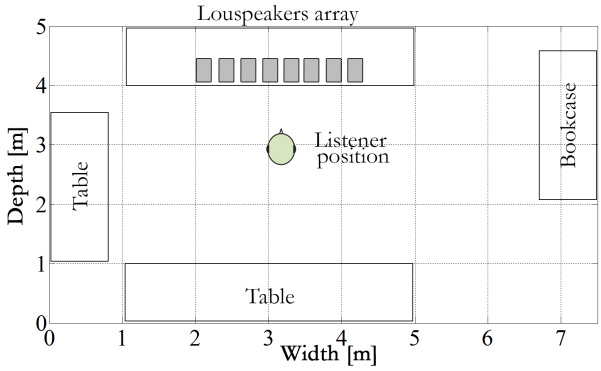

Fig. 8. Setup of the receiving room.

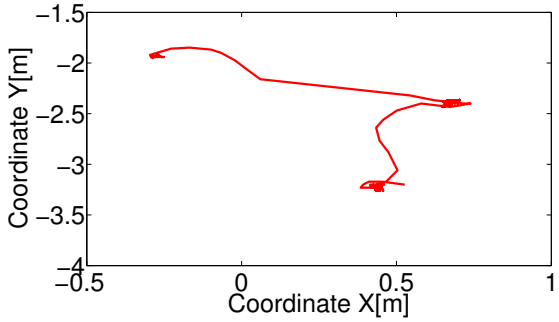

Fig. 9. Plot of the coordinates identified by the kinect control using the NU-Tech platform.

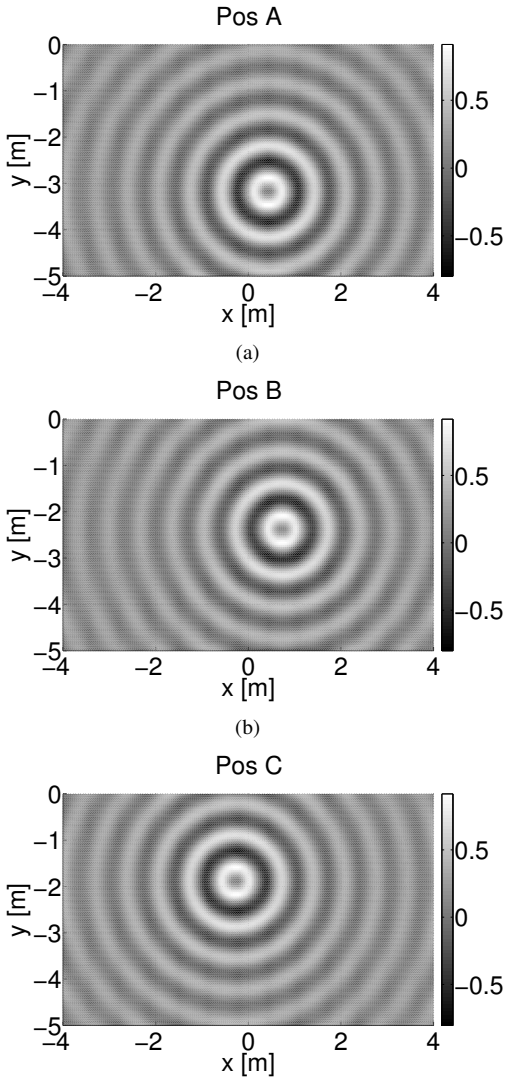

Fig. 10. Sound field emitted by the loudspeakers array in the receiving room: (a) desired source at position A, (b) desired source at position B, (c) desired source at position C.

reported in Fig. 8.

2) System evaluation: In order to evaluate the performance of the proposed advanced multichannel audio reproduction system, two validations has been performed. First of all, an analysis of the generated sound field is presented, then a subjective listening test is performed.

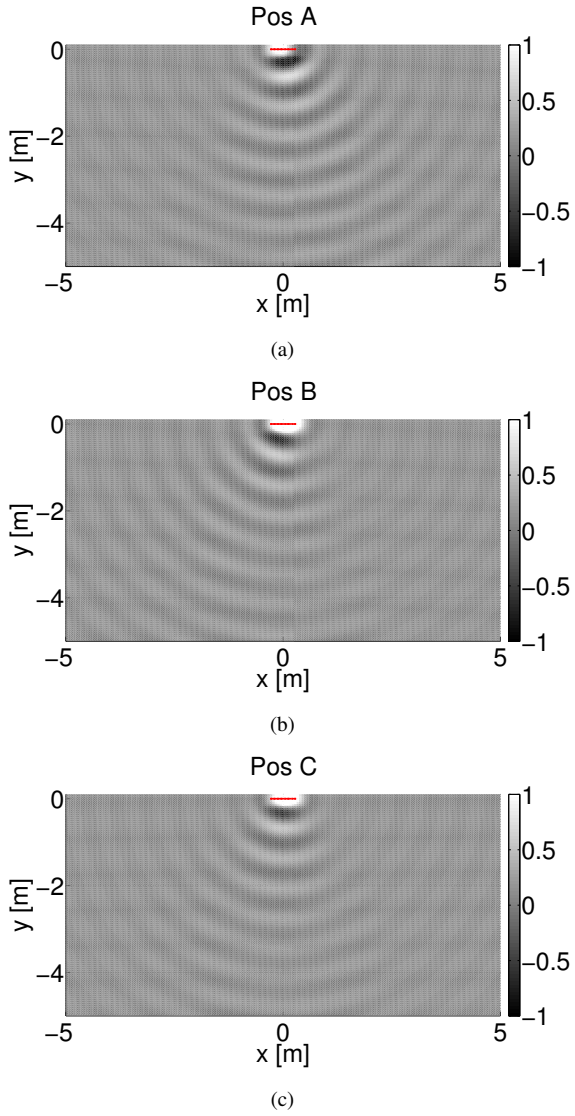

Fig. 11. Sound field emitted by the loudspeakers arrary in the receiving room: (a) desired source at position A, (b) desired source at position B, (c) desired source at position C.

Fig. 9 shows the moving speaker position coordinates identified by the Kinect control and used in the WFS system: it is evident that the speaker position is well identified and followed in comparison with Fig. 6. This is also confirmed by Figs. 10 and 11 that show the sound field emitted by the speaker in the transmitting room scenario. Figs. 10 and 11 show the sound field generated by the loudspeakers array in the receiving room using the proposed approach: each sub-figure represents the three main positions of the moving sound source in the transmitting room.

Starting from this results, subjective listening tests have been carried out in order to assess the system capabilities considering the listener point of view. In particular, taking into account the transmission room of Fig. 6, in the receiving

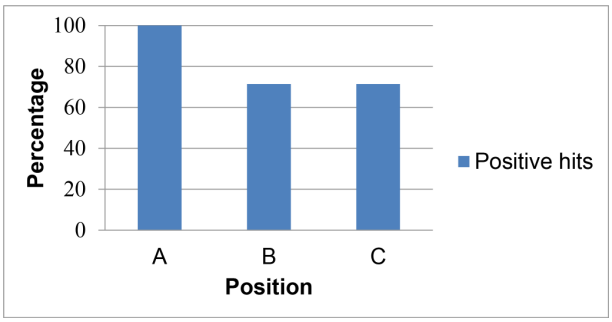

Fig. 12. Results of the subjective listening tests: percentage of positive hits for each source position.

room we asked a listening panel composed of 8 people to identify the direction of the speaker taking into consideration the path from point A to B and the path from B to C as reported in Fig.8. The results in terms of positive hints are reported in Fig. 12 and confirm that most of the listeners were able to identify the correct position of the virtual speaker, following well his movements within the transmitting room. It is worth noting that, in the few cases in which the listener has not been able to identified the exact position, he has identified a position very close to the real one.

Furthermore, the intelligibility of the speech has been tested asking to the listener to grade the quality of the speech considering a scale from 0 to 10. A mean value of 8.4 was obtained, giving also a subjective evaluation of the IAI in the transmitting room, that is capable to reduce the contribution of interfering sources reported in Fig. 6.

VI. Conclusion

Intelligent acoustic interfaces play a fundamental role in capturing acoustic information, enhancing desired audio signals and reproducing them under the quality constraints desired by user. In this work we have proposed advanced IAIs for multichannel sound reproduction in immersive speech communication scenarios. In particular, in the transmitting room an IAI exploiting a Microsoft Kinect sensor interface allows to track a moving desired source to acquire the acoustic information that is enhanced by a combined adaptive beamformer. On the other hand, in the receiving room an IAI involving a loudspeaker array reproduces the sound field by employing wave field synthesis, thus improving the immersive listening experience. Experimental results shown the effectiveness of the proposed IAIs in both enhancing the quality of captured signals from a moving desired source and providing a realistic spatial sound reproduction of the desired source. Future works will investigate the possibility to design full-duplex immersive communication IAIs. Moreover, the successful employment of a Kinect interface gives a boost to a robust involvement of media interaction and fusion.

References

[1] D. Comminiello, "Adaptive Algorithms for Intelligent Acoustic Interfaces," Ph.D. dissertation, 'Sapienza' University of Rome, Dec. 2011.

[2] D. Comminiello, M. Scarpiniti, R. Parisi, and A. Uncini, "Intelligent Acoustic Interfaces for Immersive Audio," in *Proc. 134th Audio Engineering Society Convention*, Rome, Italy, May 2013.

[3] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-Latency Real-Time Meeting Recognition and Understanding Using Distant Talking Microphones and Omni-Directional Camera," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 499–513, Feb. 2012.

[4] J.-S. Lee, G.-K. You, J.-M. Yang, and H.-G. Kang, "Unified Framework for User Tracking and Sound Beamforming with Audio/Depth Sensors in Kinect," in *Work. Kinect in Pervasive Computing*, Newcastle, UK, Jun. 2012.

[5] D. Comminiello, M. Scarpiniti, R. Parisi, and A. Uncini, "Combined Adaptive Beamforming Schemes for Nonstationary Interfering Noise Reduction," *Signal Process.*, vol. 93, no. 12, pp. 3306–3318, Dec. 2013.

[6] N. Cesa-Bianchi, A. Conconi, and C. Gentile, "On the Generalization Ability of On-Line Learning Algorithms," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 2050–2057, Sep. 2004.

[7] S. Haykin, *Neural Networks - A Comprehensive Foundation*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, Inc., 1999.

[8] A. Uncini, *Fundamentals of Adaptive Signal Processing*. Springer, 2014, ISBN 978-3-319-02806-4.

[9] A. J. Berkhout, D. De Vries, and P. Vogel, "Acoustic Control by Wave Field Synthesis," *J. Acoust. Soc. Amer.*, vol. 93, no. 5, pp. 2764–2778, May 1993.

[10] M. Gasparini, S. Cecchi, L. Romoli, A. Primavera, P. Peretti, and F. Piazza, "Kinect Application for a Wave Field Synthesis-Based Reproduction System," in *Proc. 133rd Audio Engineering Society Convention*, San Francisco, CA, USA, Oct. 2012.

[11] W. E. Hefley and D. Murray, "Intelligent user interfaces," in *Proc. 1st ACM Int. Conf. Intelligent User interfaces (IUI '93)*, Orlando, FL, Jan. 1993, pp. 3–10.

[12] D. Comminiello, S. Scardapane, M. Scarpiniti, and A. Uncini, "User-Driven Quality Enhancement for Audio Signal Processing," in *Proc. 134th Audio Engineering Society Convention*, Rome, Italy, May 2013.

[13] A. Lattanzi, F. Bettarelli, and S. Cecchi, "NU-Tech: The Entry Tool of the hArtes Toolchain for Algorithms Design," in *Proc. 124th Audio Engineering Society Convention*, Amsterdam, The Netherlands, May 2008, pp. 1–8.

[14] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. New York, NY: Springer, 2001.

[15] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing*. Berlin: Springer-Verlag, 2006.

[16] J. Arenas-García, A. R. Figueiras-Vidal, and A. H. Sayed, "Mean-Square Performance of a Convex Combination of Two Adaptive Filters," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 1078–1090, Mar. 2006.

[17] P. Peretti, L. Romoli, S. Cecchi, L. Palestini, and F. Piazza, "Phase Approximation of Linear Geometry Driving Functions for Sound Field Synthesis," in *Proc. 18th European Signal Processing Conference*, Aalborg, Denmark, Aug. 2010, pp. 1939–1943.

[18] A. Franck, K. Brandenburg, and U. Richter, "Efficient Delay Interpolation for Wave Field Synthesis," in *Proc. 125th Audio Engineering Society Convention*, New York, NY, USA, Oct. 2008.

[19] M. Gasparini, P. Peretti, S. Cecchi, L. Romoli, and F. Piazza, "Real-time Reproduction of Moving Sound Sources by Wave Field Synthesis: Objective and Subjective Quality Evaluation," in *Proc. 130th Audio Engineering Society Convention*, London,UK, May 2011.

[20] M. Gasparini, P. Peretti, L. Romoli, S. Cecchi, and F. Piazza, "Quality and Performance Assessment of Wave Field Synthesis reproducing Moving Sound Sources," in *Proc. 131st Audio Engineering Society Convention*, New York, USA, Oct. 2011.

[21] J. Ahrens and S. Spors, "Reproduction of Moving Virtual Sound Sources with Special Attention to the Doppler Effect," in *Proc. 124th Audio Engineering Society Convention*, Amsterdam, The Netherlands, May 2008.

[22] A. Franck, A. Gräfe, T. Korn, and M. Strauß, "Reproduction of Moving Sound Sources by Wave Field Synthesis: An Analysis of Artifacts," in *Proc. 32nd Audio Engineering Society Conference*, Hillerød, Denmark, Sept. 2007.

[23] H. Zhao and J. Yu, "A Simple and Efficient Design of Variable Fractional Delay FIR Filters," *IEEE Trans. Circuits Syst. II*, vol. 53, no. 2, pp. 157–160, Feb. 2006.