AN ADVANCED SPATIAL SOUND REPRODUCTION SYSTEM WITH LISTENER POSITION TRACKING

Stefania Cecchi[⋆]*, Andrea Primavera*[⋆]*, Marco Virgulti*[⋆]*, Ferruccio Bettarelli*[†]*, Francesco Piazza*[⋆]

[⋆] DII - Università Politecnica delle Marche - Italy
[†] Leaff Engineering - Italy

ABSTRACT

The paper deals with the development of a real time system for the reproduction of an immersive audio field considering the listeners' position. The system is composed of two parts: a sound rendering system based on a crosstalk canceller that is required in order to have a spatialized reproduction and a listener position tracking system in order to model the crosstalk canceller parameters. Therefore, starting from the free-field model, a new model is considered introducing a directivity function for the loudspeakers and considering a three-dimensional environment. A real time application is proposed introducing a Kinect control, capable of accurately tracking the listener position and changing the crosstalk parameters. Several results are presented comparing the proposed approach with the state of the art in order to confirm its validity.

Index Terms— Immersive audio system, Crosstalk cancellation, Head tracking

1. INTRODUCTION

In the field of immersive audio, there are two main approaches for 3D audio rendering: headphones reproduction and loudspeakers reproduction. The first approach attempts to reproduce through headphones, at each eardrum of the listener, the sound pressure of virtual sources using head related transfer functions (HRTFs). In the second case, the binaural signal is delivered to the ears using two or more loudspeakers. In this case, the sound emitted from each loudspeaker is heard by both ears and a network of filters known as crosstalk canceller is usually adopted to avoid this problem (Fig.1) [1, 2]. This is achieved under the assumption that the HRTFs are known and that are dependent on the listener position. This fact implies that the crosstalk canceller is useful for a limited area called "sweet spot". In the last decade, several efforts have been made to develop more effective solutions that are mainly composed of two parts: automatic listener position tracking and sound rendering (or adjustment of the sweet spot) according to the estimated listener position.

A spatial sound reproduction system that considers a vision-based listener-tracking was proposed in [3]. In this system, a HRTFs database is considered and a set of procedures were involved to estimate the listener's head position taking into consideration a motion detection, the segmentation of the moving objects, and the identification of skin-color regions. With this method, an a priori defined HRTFs database is requested and a huge computational load of real time image processing is needed. In [4] a method based on the use of HRTF measurement combined with a head detection algorithm for tracking the location of the listener's ears in real time using a laser scanner is presented. The Karhunen-Loeve expansion is used in order to interpolate the HRTF among listener positions from a small number of HRTF measurements, taking into consideration the listener position data obtained with the accurate tracking system. However, audio systems that used this method are expensive, since this approach requires special equipment (e.g., a 2D laser radar system). A smart virtual sound rendering system consists of a listener position tracking system, which uses infrared and ultrasonic sensors, and an adaptive virtualizer algorithm optimized for the listener position is proposed in [5]. However, the listener's position is estimated along an x-y coordinate, and the use of a remote control is needed. A similar approach was presented in [6], where an optical face tracker which provides information about the listeners x-y position is employed. An interactive audio system that actively combines head tracking and room modeling was also proposed [7]. The listener's head position and orientation are first tracked by a webcam-based head tracker, then the transfer functions are selected from a measured database. All these methods are based on the use of a priori-measured HRTFs and, in addition, require complex system for the listener tracking. In [8], a system based on free-field model capable of approximating the path responses from the loudspeakers to the listening points combined with acoustic-based listener tracking is used. In particular, the exact position is derived using the direction of arrival from the acoustic source generated by the listener, using two horizontally spaced microphones. This method was further improved in [9], where artificial neural network is employed to have a better tracking of the listener position. In this case, a very simple model that gives an estimation of the path along x-y coordinate is used, introducing a low accuracy of the impulse responses.

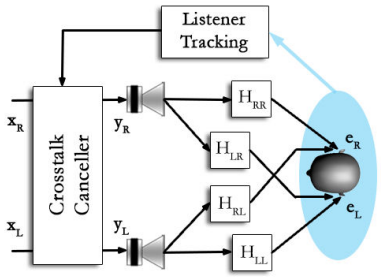


Fig. 1. A block diagram of an audio system that can adjust the sweet spot with relation to the listener position.

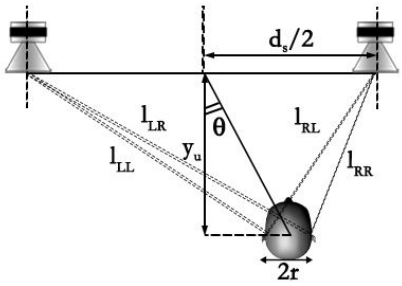


Fig. 2. Geometrical representation of the stereophonic reproduction environment using two loudspeakers.

In this paper, a novel approach for a spatial sound reproduction is proposed. Starting from a new free-field modelling that takes into consideration a three-dimensional environment (e.g., x-y-z coordinate) and the loudspeakers directivity characteristics, a real time application is proposed introducing a Kinect control, capable of accurately tracking the listener's position and changing iteratively the proposed model.

In Section 2 the proposed system is presented, introducing the novelty of the presented approach. Section 3 described the real time implementation of the system focusing on listener's position tracking system. The obtained experimental results are reported in Section 4. Finally, conclusions and future works are described in Section 5.

2. PROPOSED APPROACH

A typical two-loudspeakers listening situation is shown in Fig.1: x_L and x_R are binaural signals sent to the loudspeakers while e_L and e_R are signals perceived at the listener's ears. The system can be described by the following equation

$$\begin{bmatrix} e_L \\ e_R \end{bmatrix} = [H] \cdot [W] \cdot \begin{bmatrix} x_L \\ x_R \end{bmatrix} = \begin{bmatrix} x_L \\ x_R \end{bmatrix} \tag{1}$$

where H represents the paths between the loudspeakers and listener ears and W is the crosstalk canceller matrix. For optimal result, the matrix product $[W]\,[H]$ should be the identity matrix, in order to deliver the binaural signal x_L to the left ear and x_R to the right ear. In this way, unwanted crosstalk terms will be eliminated. According to the free-field model [8, 10], the matrix H can be defined as follows:

$$H = \frac{\rho_0}{4\pi} \begin{bmatrix} \frac{1}{l_{LL}} e^{-jkl_{LL}} & \frac{1}{l_{LR}} e^{-jkl_{LR}} \\ \frac{1}{l_{RL}} e^{-jkl_{RL}} & \frac{1}{l_{RR}} e^{-jkl_{RR}} \end{bmatrix}, \tag{2}$$

where $k = (2\pi f)/c_0$ represents the wave number with c_0 the velocity of sound, $\rho_0 = 1.21 kg/m^3$ the density. l_{LL}, l_{LR}, l_{RL}, l_{RR} represent the distances between the loudspeakers and the listener as reported in Fig. 2. They can be calculated as follows:

$$l_{LL} = \sqrt{\left(\frac{d_s}{2} + y_u tan\theta - r\right)^2 + y_u^2} \tag{3}$$

$$l_{LR} = \sqrt{\left(\frac{d_s}{2} + y_u tan\theta + r\right)^2 + y_u^2} \tag{4}$$

$$l_{RL} = \sqrt{\left(\frac{d_s}{2} - y_u tan\theta + r\right)^2 + y_u^2} \tag{5}$$

$$l_{RR} = \sqrt{\left(\frac{d_s}{2} - y_u tan\theta - r\right)^2 + y_u^2}, \tag{6}$$

where y_u denotes the distance of the listener from the central axis, d_s is the distance between the two loudspeakers, θ is the look direction of the listener, r is the radius of the listener head that can be set equal to $0.10m$. Two are the main limitations of this approach: first of all the system considers the listener's head on the same level of the loudspeakers, then the loudspeaker is represented as an omnidirectional source, neglecting its directivity. In order to overcome the first limitation, a new definition for the distance of Eqs. (3)-(6)) is evaluated taking into consideration the azimuth angle θ and the elevation angle α:

$$l_{LL} = \sqrt{(y_u tan\alpha)^2 + \left(\frac{d_s}{2} + y_u tan\theta - r\right)^2 + y_u^2} \tag{7}$$

$$l_{LR} = \sqrt{(y_u tan\alpha)^2 + \left(\frac{d_s}{2} + y_u tan\theta + r\right)^2 + y_u^2} \tag{8}$$

$$l_{RL} = \sqrt{(y_u tan\alpha)^2 + \left(\frac{d_s}{2} - y_u tan\theta + r\right)^2 + y_u^2} \tag{9}$$

$$l_{RR} = \sqrt{(y_u tan\alpha)^2 + \left(\frac{d_s}{2} - y_u tan\theta - r\right)^2 + y_u^2}. \tag{10}$$

Therefore, in contrast with [8] and Eqs. (3)-(6), this new formulation allows to define the listener position with respect to the rendering system in each position of the environment.

Regarding the source modelling, it is possible to extend the approach considering the sound source approximation as used in [11], that considers the directivity of the source. Assuming that a good approximation is given by the circular piston [12], each element of the H matrix can be modelled as follows:

$$D(r,\alpha) = \frac{2\rho_0}{4\pi}\frac{\mathrm{J}_1(kR\sin\alpha)}{kR\sin\alpha}e^{-jkr}, \qquad (11)$$

where R is the radius of the piston, $J_1(\cdot)$ is the Bessel function of first type with order 1, c is the sound velocity, and ρ is the medium density. Then, the H matrix defined in Eq. (2), can be rewritten taking into consideration Eq. (11) as follows:

$$H = \frac{2\rho_0}{4\pi}\begin{bmatrix} \frac{\mathrm{J}_1(kR\sin\alpha_{LL})}{kR\sin\alpha_{LL}l_{LL}}e^{-jkl_{LL}} & \frac{\mathrm{J}_1(kR\sin\alpha_{LR})}{kR\sin\alpha_{LR}l_{LR}}e^{-jkl_{LR}} \\ \frac{\mathrm{J}_1(kR\sin\alpha_{RL})}{kR\sin\alpha_{RL}l_{RL}}e^{-jkl_{RL}} & \frac{\mathrm{J}_1(kR\sin\alpha_{RR})}{kR\sin\alpha_{RR}l_{RR}}e^{-jkl_{RR}} \end{bmatrix}, \qquad (12)$$

where $\sin\alpha_i = \sqrt{1-(y_u/l_i)^2}$, considering that $\alpha_i \ni \{\alpha_{LL},\ \alpha_{LR},\ \alpha_{RL},\ \alpha_{RR}\}$ are the relative elevation angles computed as a function of loudspeaker and listener's position, and $l_i \ni \{l_{LL},\ l_{LR}, l_{RL},\ l_{RR}\}$ are the relative distances between the listener's ears and each loudspeakers, as reported in Eqs. (7)-(10). At this point, Eq. (12) represents the final H matrix used for the crosstalk canceller, in order to achieve the separation between the right and left channels. Then, the crosstalk cancellation matrix is obtained as follows

$$W = H^{-1} = D^{-1}\begin{bmatrix} H_{22} & -H_{12} \\ -H_{21} & H_{11} \end{bmatrix}, \qquad (13)$$

and the inversion of vector $D = H_{11}H_{22} - H_{12}H_{21}$ is performed using the fast deconvolution algorithm with regularization, as proposed in [13].

3. REAL TIME IMPLEMENTATION

The real time implementation has been developed using the NU-Tech platform [14], a suitable software platform for real time audio processing directly on a PC. An easy plug-In architecture and a free software development kit (SDK) allow the developer write NU-Tech Satellites (NUTSs) in C++ and immediately plug them into the graphical interface design environment. In this context, two plug-In have been realized: the first one is capable of creating the H matrix, to calculate the crosstalk cancellation matrix, and to filter the input signal, while the second one is capable of managing the Kinect control, retrieving the listener's position data and send them to the first one.

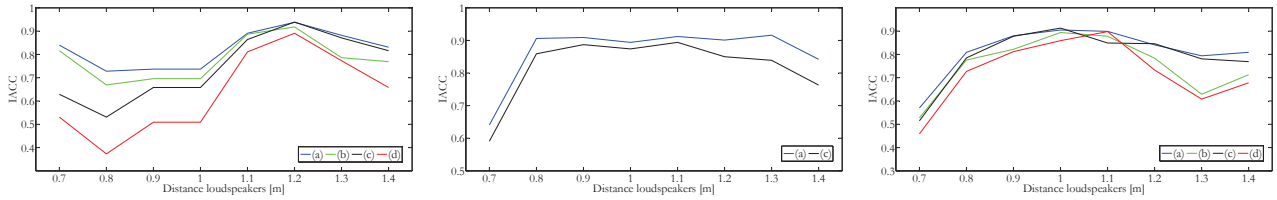
3.1. Listener tracking using Kinect control

The Microsoft Kinect® is one of the most popular among this class of new devices; although it was originally thought as an innovative video games controller, its functionalities are suitable for many other applications [15]. The Kinect is mainly composed of two cameras and a microphone array. In the proposed system, the Kinect has been used to obtain the listener's position, taking advantage of the Kinect SDK capabilities. 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 coordinates of some specific points of the face in the 3D space. While NU-Tech manages the main processing thread that performs the crosstalk cancellation algorithm, data from the Kinect sensor are continuously retrieved and analyzed in a separate process. In particular, the device is accessible by the creation of a INuiSensor interface, that allows to read color and depth data from the cameras as soon as they are ready, minimizing the streaming latency. Coordinates are derived using the IFTFace-Tracker interface, that is capable of detecting faces in a video frame and return the position in pixels of some reference face points. Then, the selected spot is associated with its depth value, allowing to retrieve the real 3D space coordinates of the target. The device is recognized by the PC as a normal capture device using Windows DirectSound drivers. NU-Tech is capable of using the DirectSound drivers to capture the audio stream directly, so that no other elaboration is needed. The validity of the proposed approach in detecting the listener's position has been reported in [15], where an application for the Wave Field Synthesis was proposed. In this case, using this control, starting from the monitored listener position, the parameters α_i, l_i, and y_u of Eq. (12) are derived in real time and supplied to the proposed algorithm.

Fig. 3. Experimental session within the semi-anechoic chamber.

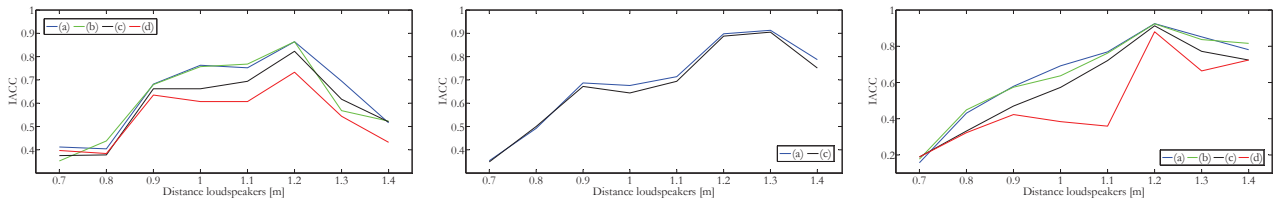
4. EXPERIMENTAL RESULTS

In order to validate the proposed method, experimental results concerning audio signals are shown in this section, from the point of view of the stereophonic perception enhancement, considering objective and subjective measurements. In particular, the interaural cross-correlation (IACC) is used since



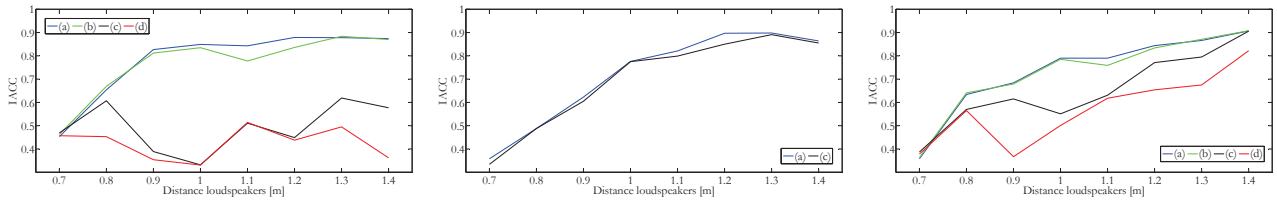
(i) (ii) (iii)

Fig. 4. IACC calculated with an azimuth $\theta = 0°$ and three elevation angles ((i) $\alpha = -10°$,(ii) $\alpha = 0°$,(iii) $\alpha = 10°$), considering (a) free-field model, (b) free-field model with elevation improvement, (c) proposed approach without elevation improvement, (d) proposed approach.



(i) (ii) (iii)

Fig. 5. IACC calculated with an azimuth $\theta = 10°$ and three elevation angles ((i) $\alpha = -10°$,(ii) $\alpha = 0°$,(iii) $\alpha = 10°$), considering (a) free-field model, (b) free-field model with elevation improvement, (c) proposed approach without elevation improvement, (d) proposed approach.



(i) (ii) (iii)

Fig. 6. IACC calculated with an azimuth $\theta = -15°$ and three elevation angles ((i) $\alpha = -10°$,(ii) $\alpha = 0°$,(iii) $\alpha = 10°$), considering (a) free-field model, (b) free-field model with elevation improvement, (c) proposed approach without elevation improvement, (d) proposed approach.

it represents a measure of spatial impression perception comparing the similarity between the signals reaching the left and right ears in a sound field [16]. The IACC is defined as [17]:

$$\mathrm{IACC}(\tau) = \frac{\int_0^{T_0} P_{\mathrm{L}}(t) P_{\mathrm{R}}(t+\tau) dt}{\sqrt{\int_0^{T_0} P_{\mathrm{L}}^2(t) dt \int_0^{T_0} P_{\mathrm{R}}^2(t) dt}}, \qquad (14)$$

where P_{L} and P_{R} represent the left and right channels powers of the binaural recording, argument τ is in the range of ± 1 ms, and T_0 is the integration period, defined as the ratio between the signal frame size and the sampling frequency f_s. As described in [16], less similar the input signals are (i.e., small value of the IACC), the greater the perception of spatial impression is, that it is essential considering a crosstalk algorithm. Tests were carried out using a dummy head, considering the proposed approach and the free-field approach [8], with a separate analysis of the elevation improvement (Eqs. (7)-(10)). The test equipment for simulations includes two loudspeakers (Genelec 6010A), a Bruel&Kjaer Head & Torso Simulator 4128, with right and left Ear Simulators 4158 and 4159, a professional MOTU Traveler sound card, an Intel Centrino-2 laptop PC, and the Kinect module as shown in Fig.3. A stereo decorrelated white noise has been used for all tests and considering Eq. (12), all the parameters of Eqs. (7)-(10) were calculated by the Kinect control module except for $R = 0.04$ m. A mean value of the IACC over the interval τ has been derived. Figs. 4, 5, 6 show the obtained results considering three azimuth and three elevation angles. It is clear that, if the listener is perfectly positioned on the same level of the loudspeakers (i.e., $\alpha = 0$), the proposed approach is slightly better in comparison with the approach of [8] as shown in Figs. 4(ii), 5(ii), 6(ii). But different results

are obtained considering different elevation angles: Figs. 4(i), 5(i), 6(i) show that for $\alpha = -10°$, several improvements are introduced taking into account the proposed approach using both the directivity of the loudspeakers and the elevation angle introduction in the free-field model. The same results are achieved applying a positive elevation angle, as reported in Figs. 4(iii), 5(iii), 6(iii). To assess the spatial perception, we also performed some informal listening tests by playing the original and the modified signals. All these tests show that the spatial impression of the modified signal is enhanced from that of the original one, also in comparison with [8]; this effect is more evident when the listener is positioned at a different level (i.e., different elevation angle) with respect to the loudspeaker placement.

5. CONCLUSIONS

In this paper a real time system for the reproduction of an immersive audio system considering the listener's position has been presented. It has been realized using a crosstalk canceller that changes the path between the loudspeakers and the listener ears as a function of the listener position, derived from a Kinect control module. Taking into account the free-field model, a new model was introduced using a directivity function for the loudspeakers and a correction factor for the three-dimensional environment. Several tests were performed in a semi-anechoic chamber and results were presented comparing our approach with the state of the art. The experiments have shown that the proposed approach is capable of achieving better results in term of spatial perception, especially in those situations where the listener is positioned in a different elevation level with respect to the loudspeakers position.

REFERENCES

[1] W. G. Gardner, *3-D Audio using Loudspeakers*. Kluwer Academic Publishers, 1998.

[2] J. S. Kim, S. G. Kim, and C. D. Yoo, "A Novel Adaptive Crosstalk Cancellation using Psychoacoustic Model for 3D Audio," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, Honolulu, HI, USA, Apr. 2007, pp. 185–188.

[3] C. Kyriakakis, T. Holman, J. Lim, H. Hong, and H. Neven, "Signal processing, acoustics, and psychoacoustics for high quality desktop audio," *Journal of Visual Communication and Image Representation*, vol. 9, no. 1, pp. 51 – 61, 1998.

[4] P. G. Georgiou, A. Mouchtaris, S. I. Roumeliotis, and C. Kyriakakis, "Immersive sound rendering using laser-based tracking," in *Proc. 109th Audio Engineering Society Convention*, Sep 2000.

[5] S. Bang, S. Jang, S. Kim, and D. Kong, "Adaptive virtual surround sound rendering method for an arbitrary listening position," in *Proc. 30th Audio Engineering Society Conference*, Mar 2007.

[6] S. Merchel and S. Groth, "Analysis and implementation of a stereophonic play back system for adjusting the sweet spot to the listeners position," in *Proc. 126th Audio Engineering Society Convention*, May 2009.

[7] M.-S. Song, C. Zhang, D. Florencio, and H.-G. Kang, "An interactive 3-d audio system with loudspeakers," *IEEE Multimedia*, vol. 13, no. 5, pp. 844–855, 2011.

[8] K.-S. Lee and S. pil Lee, "A real-time audio system for adjusting the sweet spot to the listener's position," *IEEE Trans. Consum. Electron.*, vol. 56, no. 2, pp. 835–843, 2010.

[9] K.-S. Lee, "Position-dependent crosstalk cancellation using space partitioning," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 21, no. 6, pp. 1228–1239, 2013.

[10] D. Ward and G. Elko, "A new robust system for 3d audio using loudspeakers," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 2, 2000, pp. II781–II784 vol.2.

[11] P. Peretti, S. Cecchi, F. Piazza, M. Secondini, and A. Fusco, "A Mixed Mechanical/Digital Approach for Sound Beam Pointing with Loudspeakers Line Array," in *Proc. 129th Audio Engineering Society Convention*, San Francisco, CA, USA, Oct. 2010.

[12] R. M. Aarts and A. J. E. M. Janssen, "Sound radiation from a resilient spherical cap on a rigid sphere," *J. Acoust. Soc. Am.*, vol. 127, p. 22622273, 2010.

[13] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante, "Fast Deconvolution of Multichannel Systems using Regularization," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 189–194, Mar. 1998.

[14] 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.

[15] 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.

[16] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Springer-Verlag, 1990.

[17] S. George, S. Zielinski, and F. Rumsey, "Feature Extraction for the Prediction of Multichannel Spatial Audio Fidelity," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 14, pp. 1994–2005, Nov. 2006.