SOUND SPATIALIZATION CONTROL BY MEANS OF ACOUSTIC SOURCE LOCALIZATION SYSTEM

Daniele Salvati
AVIRES Lab.
Dep. of Math. and Computer Science
University of Udine, Italy
daniele.salvati@uniud.it

Sergio Canazza
Sound and Music Computing Group
Dep. of Information Engineering
University of Padova, Italy
canazza@dei.unipd.it

Antonio Rodà
AVIRES Lab.
Dep. of Math. and Computer Science
University of Udine, Italy
antonio.roda@uniud.it

ABSTRACT

This paper presents a system for controlling the sound spatialization of a live performance by means of the acoustic localization of the performer. Our proposal is to allow a performer to directly control the position of a sound played back through a spatialization system, by moving the sound produced by its own musical instrument. The proposed system is able to locate and track the position of a sounding object (e.g., voice, instrument, sounding mobile device) in a two-dimensional space with accuracy, by means of a microphone array. We consider an approach based on Generalized Cross-Correlation (GCC) and Phase Transform (PHAT) weighting for the Time Difference Of Arrival (TDOA) estimation between the microphones. Besides, a Kalman filter is applied to smooth the time series of observed TDOAs, in order to obtain a more robust and accurate estimate of the position. To test the system control in real-world and to validate its usability, we developed a hardware/software prototype, composed by an array of three microphones and a Max/MSP external object for the sound localization task. We have got some preliminary successfully results with a human voice in real moderately reverberant and noisy environment and a binaural spatialization system for headphone listening.

1. INTRODUCTION

The spatialization of sound plays an increasingly important role in electroacoustic music performance from the twentieth century. A first widely studied aspect concerns techniques and algorithms for the placement of sounds in a virtual space. In 1971, John Chowning proposed a pioneering system that simulated the movement of sound sources in the space [1]. Afterwards, Moore [2] developed a general model that drew on basic psychophysics of spatial perception and on work in room acoustics, relying on the precedence effect. To date, many techniques are used for spatialization, such as: holographic approach [3] like 3D panning (Vector Base Amplitude Panning [4]) and Ambisonics [5], Wavefield Synthesis [6], and transaural techniques based on an idea by Schroeder [7]. Besides the methods based on virtual environments using loudspeakers, we mention the theory and practice of 3D sound reproduction using headphones, that requires the filtering of sound streams with Head Related Transfer Functions (HRTFs) [8].

Another important aspect of sound spatialization is related to the control task. Recently, research has begun to investigate control issues, especially related to gesture controlled spatialization of sound in live performance [9]. Most systems of control make use of a separate interface and a specific performer (usually not on stage) to control the movement of sounds. In that sense, the evolution of control systems was mainly related to the design of different equipments, such as multichannel devices with faders, control software with mouse and joystick for two-dimensional movement, sophisticated software with 3D virtual reality display [10], sensors interfaces such as data gloves based system, head trackers and camera-based tracking systems [11].

In [12], the authors propose a system to allow real-time gesture control of spatialization in a live performance setup, by the performers themselves. This gives to the performers the control over the spatialization of the sound produced by their own instrument, during the performance of a musical piece. In the same way, our system provides the capability to control the spatialization of sound by the performer himself, using the potentiality offered by microphone array signal processing. Recently, microphone array signal processing is increasingly being used in human computer interaction systems, for example the new popular interface Microsoft Kinect incorporates a microphone array to conduct acoustic source localization and noise suppression to improve voice recognition. The microphone array approach has the advantage that the performer does not have to wear any sensor or device which can be a hindrance to his/her movements; moreover, it can replace or integrate camera-based tracking systems that can have problems with the low lighting of the concert hall.

This paper presents a system for controlling the sound spatialization of a live performance by means of the acoustic localization of the performer. Our proposal is to allow a performer to directly control the position of a sound played back through a spatialization system, by moving the sound produced by its own musical instrument. The proposed system is able to locate and track the position of a sounding object (e.g., voice, instrument, sounding mobile device) in a two-dimensional space with accuracy, by means of a mi-

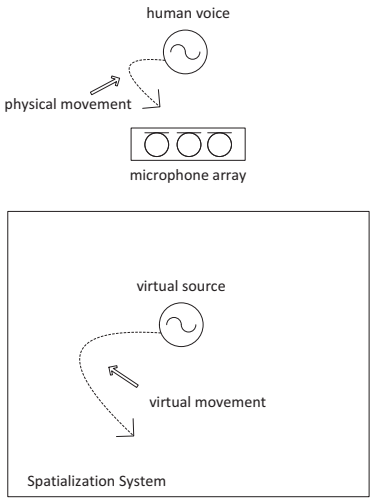

Figure 1. Sound spatialization control setup.

crophone array (see Figure 1).

The paper is organized as follows: after presenting the system architecture in Section 2, we summarize the algorithms for the time delay estimation in Section 3. Section 4 describes the Kalman filter to smooth the observed TDOAs. In Section 5, we illustrate the two-dimensional position estimation. Finally, Section 6 shows the developed prototype and some experimental results with human voice.

2. SYSTEM ARCHITECTURE

The system consists of three main components: i) a microphone array for signal acquisition; ii) signal processing techniques for sound localization; iii) a two-dimensional mapping function for controlling the sound spatialization parameters.

The array is composed by three microphones arranged in an uniform linear placement (in near-field environment, three microphones are the bare minimum to locate source in a plane). Signal processing algorithms estimate the sound source position in a horizontal plane by providing its Cartesian coordinates. Last component regards how to transform the x-y coordinates of the real source into parameters for the virtual source movement, depending on the spatialization setup. To this purpose, we mention the Spatial Sound Description Interchange Format (SpatDIF) [13], a format to describe, store and share spatial audio scenes across 2D/3D audio applications and concert venues. However, this paper is mainly focused on the localization task.

Figure 2 summarizes the block diagram of system. A widely used approach to estimate the source position consists in two steps: in the first step, a set of TDOAs are estimated using measurements across various combinations of microphones; in the second step, knowing the position of sensors and the velocity of sound, the source positions is calculated by means of geometric constraints and using approximation methods such as least-square techniques [14].

The traditional technique to estimate the time delay between a pair of microphones is the GCC-PHAT [15]. Following this approach, the maximum peak detection of the GCC functions provides the estimation of the TDOAs between microphones 1-2 and 2-3. Then, a Kalman filter is applied in order to smooth in time [16] the two estimated TDOAs. The Kalman filter provides a robust and accurate estimation of τ_{12} and τ_{23}, moreover it is able to provide a source position estimation, also if the TDOA estimation task misses the target in some frame of analysis.

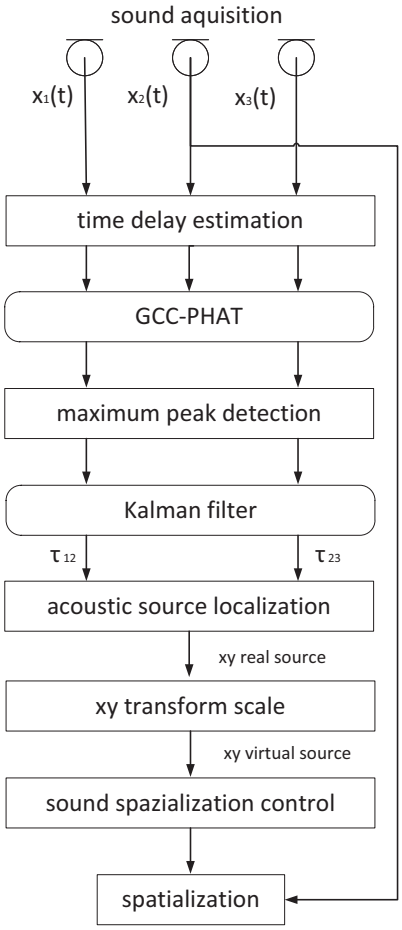

Figure 2. Block diagram of system.

3. TIME DELAY ESTIMATION

GCC [15] is the classic method to estimate the relative time delay associated with the acoustic signals received by a pair of microphones in a moderately reverberant and noisy environment [17, 18]. It basically consist in a cross-correlation followed by a filter that aims at reducing the performance degradation due to additive noise and multipath channel effects. The signals received at the two microphones $x_1(t)$ and $x_2(t)$ may be modeled as

$$\begin{aligned} x_1(t) &= h_1(t) * s(t) + n_1(t) \\ x_2(t) &= h_2(t) * s(t-\tau) + n_2(t) \end{aligned} \tag{1}$$

where τ is the relative signal delay of interest, $h_1(t)$ and $h_2(t)$ represent the impulse responses of the reverberant

channels, $s(t)$ is the sound signal, $n_1(t)$ and $n_2(t)$ correspond to uncorrelated noise, and * denotes linear convolution. The GCC in the frequency domain is

$$R_{x_1x_2}(t) = \sum_{w=0}^{L-1} \Psi(w) S_{x_1x_2}(w) e^{\frac{jwt}{L}} \tag{2}$$

where w is the frequency index, L is the number of samples of the observation time, $\Psi(w)$ is the frequency domain weighting function, and the cross-spectrum of the two signals is defined as

$$S_{x_1x_2}(w) = E\{X_1(w)X_2^*(w)\} \tag{3}$$

where $X_1(w)$ and $X_2(w)$ are the Discrete Fourier Transform (DFT) of the signals and * denotes the complex conjugate. GCC is used for minimizing the influence of moderate uncorrelated noise and moderate multi-path interference, maximizing the peak in correspondence of the time delay.

The relative time delay τ is obtained by an estimation of the maximum peak detection in the filter cross-correlation function

$$\widehat{\tau} = \underset{t}{\operatorname{argmax}} R_{x_1x_2}(t). \tag{4}$$

PHAT [15] weighting is the traditional and most used function. It places equal importance on each frequency by dividing the spectrum by its magnitude. It was later shown that it is more robust and reliable in realistic reverberant conditions than other weighting functions designed to be statistically optimal under specific non-reverberant noise conditions [19]. The PHAT weighting function normalizes the amplitude of the spectral density of the two signals and uses only the phase information to compute the GCC

$$\Psi_{\mathrm{PHAT}}(w) = \frac{1}{|S_{x_1x_2}(w)|}. \tag{5}$$

GCC works very well with human voice, and it is traditional used with human speech. Instead, it is widely acknowledged that GCC performance is dramatically reduced in case of harmonic sound, or generally pseudo-periodic sounds. In fact, segments of pseudo-periodic sound, when filtered by GCC, have less influence on the deleterious effects of noise and reverberation. Thus, sound objects in which the harmonic component greatly prevails on the noisy part (for example musical instruments like flute and clarinet) require new considerations for the localization task that have to be investigated.

4. TIME DELAY FILTERING USING KALMAN THEORY

The Kalman filter [20] is the optimal recursive Bayesian filter for linear systems observed in the presence of Gaussian noise. We consider that the state of the TDOA estimation could be summarized by two variables: the position τ and velocity v_τ. These two variables are the elements of the state vector $\mathbf{x}_t$

$$\mathbf{x_t} = [\tau, v_\tau]^T. \tag{6}$$

The process model relates the state at a previous time $t-1$ with the current state at time t, so we can write

$$\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{w}_{t-1} \tag{7}$$

where $\mathbf{F}$ is the transfer matrix and $\mathbf{w}_{t-1}$ is the process noise associated with random events or forces that directly affect the actual state of the system. We assume that the components of $\mathbf{w}_{t-1}$ have Gaussian distribution with zero mean normal distribution with covariance matrix $\mathbf{Q}_t$, $\mathbf{w}_{t-1} \sim N(0, \mathbf{Q}_t)$. Considering the dynamical motion, if we measured the system to be at position x with some velocity v at time t, then at time $t + dt$ we would expect the system to be located at position $x + v \cdot dt$, thus this suggests that the correct form for $\mathbf{F}$ is

$$\mathbf{F} = \begin{bmatrix} 1 & dt \\ 0 & 1 \end{bmatrix}. \tag{8}$$

At time t an observation $\mathbf{z}_t$ of the true state $\mathbf{x}_t$ is made according to the measurement model

$$\mathbf{z}_t = \mathbf{H}\mathbf{x}_t + \mathbf{v}_t \tag{9}$$

where $\mathbf{H}$ is the observation model which maps the true state space into the observed space and $\mathbf{v}_t$ is the observation noise which is assumed to be zero mean Gaussian white noise with covariance $\mathbf{R}_t$,$\mathbf{v}_t \sim N(0, \mathbf{R}_t)$. We only measure the position variables, i.e. the maximum peak detection of GCC-PHAT. Hence, we have

$$\mathbf{z}_t = \widehat{\tau} \tag{10}$$

and then we have

$$\mathbf{H} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{11}$$

The filter equations can be divided into a prediction and a correction step. The prediction step projects forward the current state and covariance to obtain an a priori estimate. After that the correction step uses a new measurement to get an improved a posteriori estimate. In predication step the time update equations are

$$\widehat{\mathbf{x}}_{t|t-1} = \mathbf{F}_t\widehat{\mathbf{x}}_{t-1|t-1}, \tag{12}$$

$$\mathbf{P}_{t|t-1} = \mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}^T + \mathbf{Q}_{t-1}, \tag{13}$$

where $\mathbf{P}_t$ denotes the error covariance matrix. In the correction step the measurement update equations are

$$\widehat{\mathbf{x}}_{t|t} = \widehat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}_t\widehat{\mathbf{x}}_{t|t-1}), \tag{14}$$

$$\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{H})\mathbf{P}_{t|t-1}, \tag{15}$$

where $\mathbf{I}$ is the identity matrix and so-called Kalman gain matrix is

$$\mathbf{K}_t = \mathbf{P}_{t-1|t-1}\mathbf{H}^T(\mathbf{H}_t\mathbf{P}_{t-1|t-1}\mathbf{H}^T + \mathbf{R}_t)^{-1}. \tag{16}$$

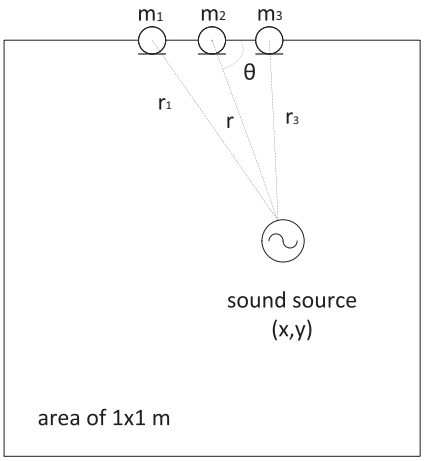

Figure 3. The map of the considered control area.

5. ACOUSTIC SOURCE LOCALIZATION

Starting from the estimated TDOAs between microphones $\widehat{\tau}_{12}$ and $\widehat{\tau}_{23}$, it is possible to calculate the coordinates of the sound source by means of geometric constraints. In near-field environment we have

$$\widehat{x} = r\cos(\theta) \tag{17}$$

$$\widehat{y} = r\sin(\theta) \tag{18}$$

where the axis origin is placed in microphone 2, r is the distance from source and microphone 2, and θ is the angle between r and x axis. Then, we have

$$r_1 = r + \tau_{12}c \tag{19}$$

$$r_3 = r + \tau_{23}c \tag{20}$$

and we obtain

$$\theta = \arccos\Big(\frac{c(\tau_{12}+\tau_{23})(\tau_{12}\tau_{23}c^2 - d^2)}{d(2d^2 - c^2(\tau_{12}^2+\tau_{23}^2))}\Big) \tag{21}$$

$$r = \frac{\tau_{12}^2c^2 - d^2}{2(\tau_{12}c + d\cos\theta)} \tag{22}$$

where c is speed of sound and d is the distance between microphones. Figure 3 show the map of considered area.

6. EXPERIMENTAL RESULTS

A hardware/software prototype was developed in order to test the proposed system in a real environment. It is composed by a linear array of three microphones and a Max/MSP external object, which implements all the signal processing tasks needed for the sound localization. The object receives the audio signals captured by the three microphones and gives as output the x-y coordinates of the sound source. We also developed a Max/MSP patch (see Figure 4) the control and the real-time interaction with a sound spatialization tool. A human voice sound has been used to validate the interface. The audio signals, sampled at a rate of 96 kHz, are processed with a Hanning analysis window of

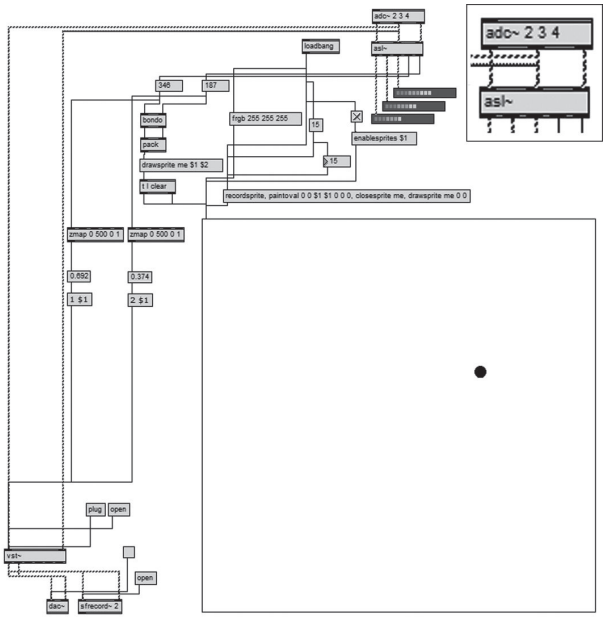

Figure 4. The Max/MSP interface with the external object asl∼.

42 ms. We used microphones with supercardioid pickup pattern, which are the most frequently used microphones for capturing sound signals in electroacoustic music.

It is important to highlight that microphones with omnidirectional polar pattern are commonly used for array processing, but their use is not appropriate in this context, because of possible interferences of the loudspeakers during a live performance. However, as we shall see, the use of supercardioid microphones allows as well the localization of an sound source in a small active area (see Figure 3). With a distance between the microphones of $d = 15$ cm, the useful area for the sound localization is about a square of 1 meter per side. The origin of the reference system coincides with the position of the microphone 2 (m2). Then, the active area is included between -50 cm and 50 cm along x-axis and between 0 and 100 cm along y-axis (Figure 3). Experiments have been done in a room of 3.5×4.5 m with a moderately reverberant and noisy environment.

The first experiment is related to the TDOAs estimation. Figure 5 shows the TDOAs of a human voice moving along the y-axis approaching to microphone 2, with $x = 0$. It can be seen how the values of TDOAs, when the sound source approaches the microphone 2, tend to be swinging due to the supercardioid polar pattern of the microphones, and this happens when the angle of sound incidence increases over the microphone vertical. The comparison between the raw data (gray lines) and the data processed by the Kalman filter (black lines) shows that the filtering allows to obtain more accurate and stable values.

Figure 6 shows the results of the second experiment, related to the two-dimensional movement of the sound source. The test is composed by eight parts. In each part the sound source, still a human voice, is moved from the center of the active area along a different direction each time.

The positions represented by dots are the raw data estimated directly by the GCC-PHAT and the continuous lines

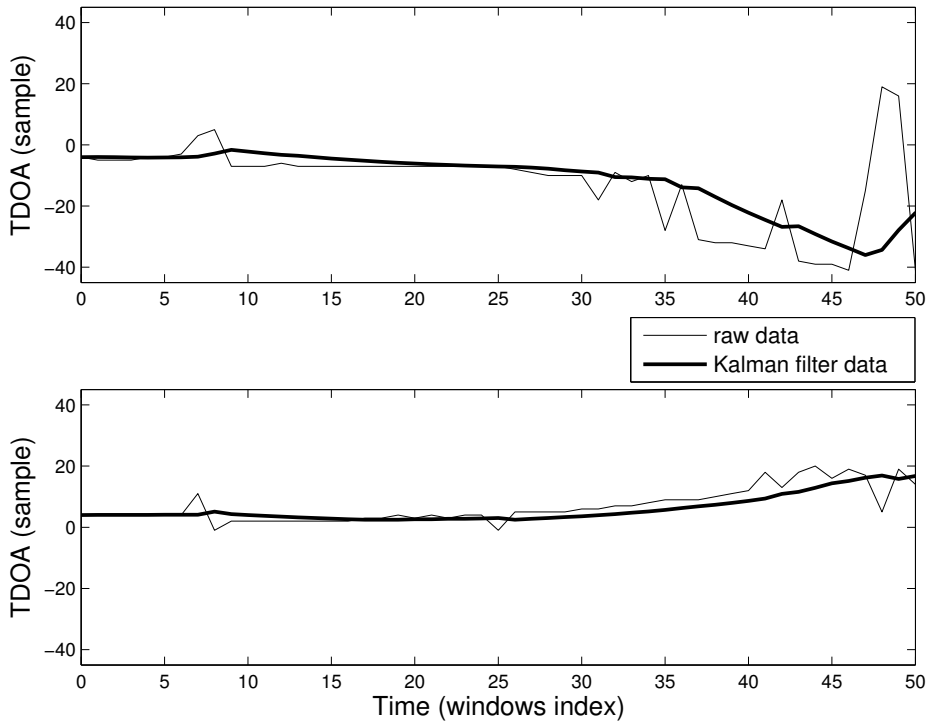

Figure 5. Comparison of TDOA estimation of human voice (on the top between microphone 1 and 2, below between 2 and 3). The Kalman filter data is represented by black lines and raw data by gray lines.

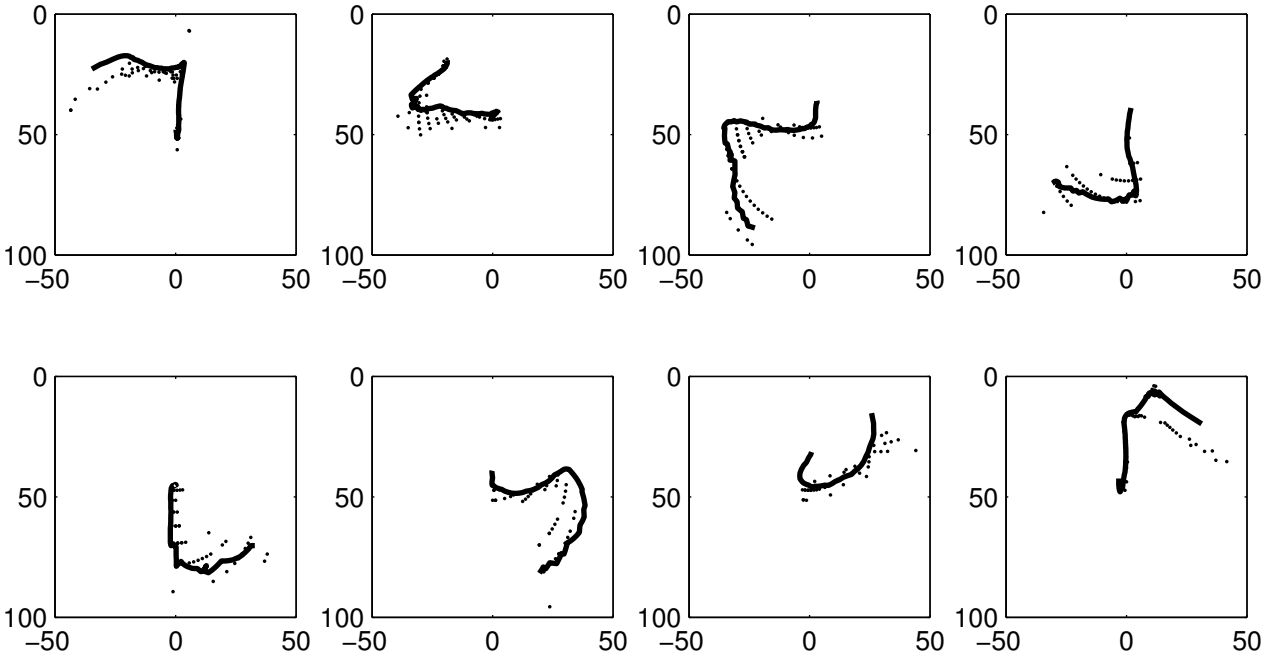

Figure 6. Acoustic source localization performance. Human voice moves in different directions (dots are the raw data), xy axis are in cm.

represent the filtered Kalman data.

Finally, the control interface was tested in connection with a sound spatialization system. VST plug-in based on binaural spatialization for headphone listening was used. An informal test of the system showed encouraging results: the performer has in fact been able to control in real time the position of a virtual sound source by small movements (of the order of tens of centimeters) of his/her mouth.

7. CONCLUSIONS

This paper presented a system that exploits microphone array signal processing to allow a performer to use the movement of a sounding object (voice, instrument, sounding mobile device) to control a sound spatialization system. A hardware/software prototype, composed by a linear array of three supercardiod microphones and a Max/MSP external object, was developed. Preliminary results with human voice show that the system can be used in a real scenario. GCC-PHAT and Kalman filter provides an accurate time delay estimation in moderately reverberant and noisy environment. However, new investigation must be done in order to work with harmonic sounds, or generally pseudo-periodic sounds, such as those traditional musical instruments in which the harmonic component greatly prevails on the noisy part. This is the main focus of our future work, which also will regard the use of the interface in a real live performance setup with a loudspeaker based spatialization system.

8. REFERENCES

[1] J. Chowning, "The simulation of moving sound sources," *Journal of the Audio Engineering Society*, vol. 19, no. 1, pp. 2–6, 1971.

[2] F. R. Moore, "A general model for spatial processing of sounds," *Computer Music Journal*, vol. 7, no. 3, pp. 6–15, 1982.

[3] A. J. Berkhout, "A holographic approach to acoustic control," *Journal of the Audio Engineering Society*, vol. 36, no. 12, pp. 977–995, 1988.

[4] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Acoustical Society of America*, vol. 45, no. 6, pp. 456–466, 1997.

[5] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *Journal of the Acoustical Society of America*, vol. 33, pp. 959–871, 1985.

[6] D. de Vries, E. W. Start, and V. G.Valstar, "The wavefield synthesis concept applied to sound reinforcement restriction and solutions," in *Audio Engineering Society Convention*, 2 1994.

[7] M. Schroeder, "Improved quasi-stereophony and colorless artificial reverberation," *Journal of the Acoustical Society of America*, vol. 33, no. 8, pp. 1061–1064, 1961.

[8] F. Wightman and D. Kistler, "Headphone stimulation of free field listening I: stimulus synthesis," *Journal of the Acoustical Society of America*, vol. 85, pp. 858–867, 1989.

[9] M. Marshall, J. Malloch, and M. Wanderley, "Gesture control of sound spatialization for live musical performance," in *Gesture-Based Human-Computer Interaction and Simulation*. Springer Berlin / Heidelberg, 2009, vol. 5085, pp. 227–238.

[10] M. Naef and D. Collicott, "A VR interface for collaborative 3d audio performance," in *Proc. International Conference on New Interfaces for Musical Expression*, 2006, pp. 57–60.

[11] M. Wozniewski, Z. Settel, and J. Cooperstock, "A framework for immersive spatial audio performance," in *Proc. International Conference on New Interfaces for Musical Expression*, 2006, pp. 144–149.

[12] M. Marshall, N. Peters, A. Jensenius, J. Boissinot, M. Wanderley, and J. Braasch, "On the development of a system for gesture control of spatialization," in *Proc. International Computer Music Conference*, 2006.

[13] N. Peters, S. Ferguson, and S. McAdams, "Towards a spatial sound description interchange format (SpatDIF)," *Canadian Acoustics*, vol. 35(3), pp. 64–65, 2007.

[14] R. O. Schmidt, "A new approach to geometry of range difference location," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-8 Issue: 6, pp. 821–835, 1972.

[15] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, May 1976.

[16] U. Klee, T. Gehrig, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–15, 2006.

[17] B. Champagne, S. Berdard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 148–152, 1996.

[18] J. Chen, Y. Huang, and J. Benesty, "A comparative study on time delay estimation in reverberant and noisy environments," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 21–24.

[19] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum based technique," in *Proc. IEEE ICASSP*, vol. 2, 1994, pp. 273–276.

[20] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.