A SOUND LOCALIZATION BASED INTERFACE FOR REAL-TIME CONTROL OF AUDIO PROCESSING

Daniele Salvati
AVIRES Lab.
Dep. of Mathematics 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
Sound and Music Computing Group
Dep. of Information Engineering
University of Padova, Italy
roda@dei.unipd.it

ABSTRACT

This paper describes the implementation of an innovative musical interface based on the sound localization capability of a microphone array. Our proposal is to allow a musician to plan and conduct the expressivity of a performance, by controlling in real-time an audio processing module through the spatial movement of a sound source, i.e. voice, traditional musical instruments, sounding mobile devices. The proposed interface is able to locate and track the sound in a two-dimensional space with accuracy, so that the x-y coordinates of the sound source can be used to control the processing parameters. In particular, the paper is focused on the localization and tracking of harmonic sound sources in real moderate reverberant and noisy environment. To this purpose, we designed a system based on adaptive parameterized Generalized Cross-Correlation (GCC) and Phase Transform (PHAT) weighting with Zero-Crossing Rate (ZCR) threshold, a Wiener filter to improve the Signal to Noise Ratio (SNR) and a Kalman filter to make the position estimation more robust and accurate. We developed a Max/MSP external objects to test the system in a real scenario and to validate its usability.

1. INTRODUCTION

Music interaction is an important and new area in the field of audio based Human Computer Interaction (HCI) systems. The development of new interfaces for musical applications has the potential to change and enhance the experience of musical performance and in particular to allow a performer the interaction with a computer for real-time audio processing. The development of digital audio effects has always stimulated the design of interfaces for controlling the processing parameters. A large number of musical interfaces [1] has been implemented and tested with the goal of providing tools for gestural interaction with digital sounds.

In [2], the author divides gestural controllers into four main categories: gestural interfaces played by touching or holding the instrument, interfaces with haptic feedback, interfaces worn on the body and interfaces that may be played without any physical contact. In this last category, the position of the body might be used without the need for the performer to wear or touch any special devices. Examples of such interfaces are: Gesture Wall [3], that uses electric field sensors to measure the position and movement of the player's hands and body in front of a projection screen; Litefoot [4], based on optical sensor; an interface based on video camera that allows the performers to use their full-body for controlling in real-time the generation of an expressive audio-visual feedback [5].

Musical interfaces are often used to allow the performer to enhance the expressive control on the sounds generated by their acoustic instruments in a live electronics context. E.g., in *Medea* by Adriano Guarnieri (2002) the movement of the bell of a trombone is captured by a camera [6] and mapped into parameters for sound spatialization; in *fili bianco-velati* (Guarnieri, 2005), the movement of a violinist is followed by a motion capture system based on infrared cameras.

In general, those kind of systems have considerable complexity and in some situations some problems. In fact, the performer has to wear sensors or devices which can be a hindrance to his/her movements; besides, in the camera-based systems there could be problems with the low and/or not always controllable lighting of the concert hall.

This paper describes the implementation of an innovative musical interface based on the sound localization capability of a microphone array. The interface allows a musician to plan and conduct the expressivity of a performance, by controlling in real-time an audio processing module through the spatial movement of a sound source. In this way a musician, during the performance, is able to interact with the live electronics system through the movement of his/her own musical instrument with an immediate, instinctive and gestural approach. The proposed interface is completely non-invasive (no need for markers, sensors or wires on the musician) and requires no dedicated hardware.

The system uses an algorithm based on an estimate of the Time Difference Of Arrival (TDOA) for sound source localization. Typically, these algorithms tend to reduce their performance in presence of competing sources, high reverberant environment, or low signal to noise ratio. Moreover, in the context of live electronics it is not always possible to have a controlled acoustic scene (there

could be other sources or noise due to the return of audio monitor), thus, we propose an innovative approach that combines an array of supercardioid polar pattern microphones (instead of the classic omnidirectional ones which are usually used in array processing) and a localization process task based on adaptive parameterized GCC-PHAT with ZCR threshold, a Wiener filter to improve the SNR and a Kalman filter to make the position estimation more robust and accurate.

The paper is organized as follow: Section 2 presents the system architecture, both the hardware and software aspects; the algorithms for the sound source localization are detailed in Section 3; finally, Section 4 shows some preliminary results.

2. SYSTEM ARCHITECTURE

The interface consists of three main components: i) a microphone array for signal acquisition; ii) signal processing algorithms for robust sound localization; iii) a mapping strategy to control the audio processing parameters.

The array is composed by three microphones arranged in an uniform linear placement. In this way we can localize a sound source in a plane (three microphones are the bare minimum). Signal processing algorithms estimate the the sound source position in a horizontal plane by providing its Cartesian coordinates. The last component implements the mapping strategy [7], so that the x-y coordinates are associated with audio processing parameters. For the purpose of testing, in this paper we have limited ourselves to explore a one-to-one mapping strategy, by using the x-y values to directly control two parameters of an audio effect, e.g. cutoff frequency and resonance of a filter or amount and decay time of a reverb. Of course, this task is closely related to user needs, and in literature there are a lot of works proposing strategies to transform from two-to-many parameters [8] [9] [10] [11] [12]. This paper is mainly focused on the localization task.

Figure 1 summarizes the system architecture. Sound source localization allows to extract information about the location of one or more sources using microphone arrays and signal processing techniques. A widely used approach to estimate the source position consists of 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 position is calculated by means of geometric constraints and using approximation methods such as least-square techniques [13]. The traditional technique to estimate the TDOA between a pair of microphones is the GCC-PHAT [14]. It is highly effective in a moderately reverberant and noisy environment. Unfortunately, concerning musical sounds that are mainly harmonics, the GCC-PHAT approach does not work, because the PHAT filter normalizes the GCC according to the spectrum magnitude. Then, new considerations are required to estimate the TDOA for pseudo-periodic signals. Our proposal is to use a parameterized GCC-PHAT, that weights the contribution of the PHAT filtering, depending on the threshold of the ZCR parameters.

A de-noise algorithm based on Wiener filter is used to improve the SNR of the signals. When the maximum peak detection does not observe any source, it is computed an average estimation of noise (noise print), which will be subtracted in all three signals before the TDOA estimation task.

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

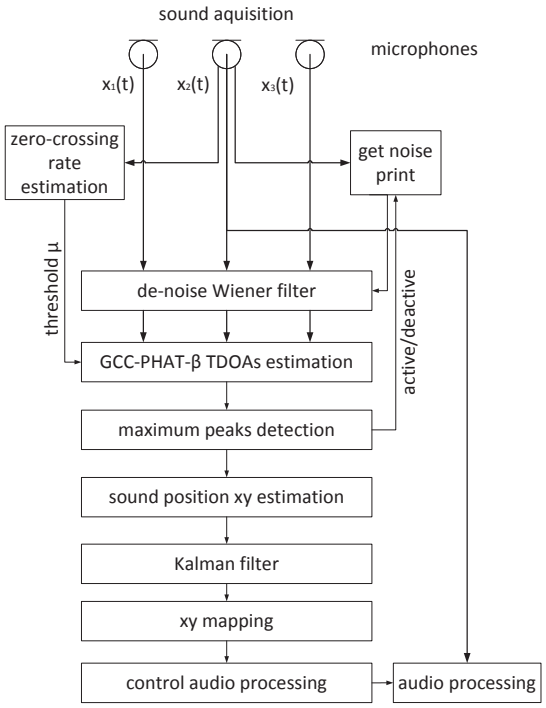

Figure 1: *Block diagram of interface.*

$$\widehat{x} = r\cos(\theta) \tag{1}$$

$$\widehat{y} = r\sin(\theta) \tag{2}$$

where the axis origin is placed in microphone 2, r is the distance between the sound source and the microphone 2, and θ is the angle between r and the x axis

$$\theta = \arccos\Big(\frac{c(\widehat{\tau}_{12}+\widehat{\tau}_{23})(\widehat{\tau}_{12}\widehat{\tau}_{23}c^2-d^2)}{d(2d^2-c^2(\widehat{\tau}_{12}^2+\widehat{\tau}_{23}^2))}\Big) \tag{3}$$

$$r = \frac{\widehat{\tau}_{12}^2c^2-d^2}{2(\widehat{\tau}_{12}c+d\cos\theta)} \tag{4}$$

where c is speed of sound and d is the distance between microphones.

Finally, a second filter provides a more accurate tracking of the source position, by means of the Kalman theory. The Kalman filter is able to provide an estimation of the position of the source, also if the TDOA estimation task misses the target in some frame of analysis.

3. SOUND SOURCE LOCALIZATION

3.1. TDOA estimation using GCC-PHAT

GCC [14] is the classic method to estimate the relative time delay associated with the acoustic signals received by a pair of microphones in a moderate reverberant and noisy environment. It basically consists of a cross-correlation followed by a filter that aims at reducing the performance degradation due to additive noise and

multi-path channel effects. The signals received at the two microphones $x_1(t)$ and $x_2(t)$ may be modeled as

$$\begin{aligned} x_1(t) &= h_1(t) * s(t) + n_1(t) \\ x_2(t) &= h_2(t) * s(t-\tau) + n_2(t) \end{aligned} \tag{5}$$

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

$$R_{x_1x_2}(t) = \sum_{w=0}^{L-1} \Psi(w) S_{x_1x_2}(w) e^{\frac{jwt}{L}} \tag{6}$$

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

$$S_{x_1x_2}(w) = E\{X_1(w)X_2^*(w)\} \tag{7}$$

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 multipath 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

$$\widehat{\tau} = \underset{t}{\operatorname{argmin}} R_{x_1x_2}(t). \tag{8}$$

PHAT [14] 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 nonreverberant noise conditions [15]. 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

$$\Psi_{\mathrm{PHAT}}(w) = \frac{1}{|S_{x_1x_2}(w)|}. \tag{9}$$

It is widely acknowledged that GCC performance is dramatically reduced in case of harmonic sound, or generally pseudo-periodic sounds. In fact, the GCC have less capability to reduce the deleterious effects of noise and reverberation, when it is applied to pseudo-periodic sound.

3.2. Adaptive parameterized GCC-PHAT with zero-crossing rate threshold

The PHAT weighting can be generalized to parametrically control the level of influence from the magnitude spectrum [16]. This transform will be referred to as the PHAT-β and defined as

$$\Psi_{\mathrm{PHAT}-\beta}(w) = \frac{1}{|S_{x_1x_2}(w)|^\beta} \tag{10}$$

where β varies between 0 and 1. When $\beta = 1$, equation (10) becomes the conventional PHAT and the modulus of the Fourier transform becomes 1 for all frequencies, when $\beta = 0$ the PHAT has no effect on the original signal, and we have the cross-correlation function.

Therefore, in case of harmonic sounds we can use an intermediate value of β so that we can detect the peak to estimate the time delay between signals, and to have a system, at least in part, which exploits the benefits of PHAT filtering to improve performance in a moderately reverberant and noisy environments. To adapt the value of β we use the ZCR to check if sound source is periodic or not. ZCR is a very useful audio feature, and it is defined as the number of times that the audio waveform crosses the zero axis

$$ZCR(t) = \frac{1}{2N} \sum_{i=1}^{N} |sgn(x(t+i)) - sgn(x(t+i-1))|. \tag{11}$$

where sgn(x) is the sign function.

Then, we can express the adaptive parametrized GCC-PHAT, identifying by experimental tests a suitable threshold μ such as

$$\begin{cases} \beta = 1, \; if \; ZCR \geq \mu \\ \beta < 1, \; if \; ZCR < \mu \end{cases} \tag{12}$$

3.3. De-noise Wiener filter

Frequency domain methods, which are based on the Short Time Spectral Attenuation (STSA) [17], require a little information to carry out the filtering (*a priori* information): only an estimate of the noise present is necessary (noise print), since it is assumed to be stationary along the entire signal. Any further information needed (*a posteriori* information) is automatically calculated by the algorithm through the analysis of the characteristics of the signal. Since this method is easy to use and is generally applied to different typologies of audio signals, they are employed in commercial hardware and software systems.

These de-noise systems consist of two important components: a noise estimation method and a suppression rule. These techniques employ a signal analysis through the Short-Time Fourier Transform (STFT) (which is calculated on windowed section of the signal as it changes over time) and can be considered as a non-stationary adaptation of the Wiener filter [18] in the frequency domain. In particular, Short Time Spectral Attenuation (STSA) consists in applying the short-time spectrum of the noise to a time-varying suppression and does not require the definition of a model for the audio signal. Suppose considering the useful signal $s(t)$ as a stationary aleatory process to which some noise $n(t)$ is added (uncorrelated with $x(t)$) to produce the degraded signal $x(t)$. The relation that connects the respective power spectral densities is therefore

$$P_x(w) = P_s(w) + P_n(w). \tag{13}$$

If we hypothesize to succeed in retrieving an adequate estimate of $P_n(w)$, during the silence intervals of the signal $x(t)$, and in the musical portions of $P_x(w)$, we can expect to obtain an estimate of the spectrum of $s(t)$ by subtracting $P_n(w)$ from $P_x(w)$; the initial assumption of stationariness can be considered locally satisfied since short temporal windows are employed. Note that the use of a short-time signal analysis is equivalent to the use of a filter bank. First each channel (that is, the output of each filter) is appropriately attenuated and then it is possible to proceed with the synthesis of the restored signal. The timevarying attenuation applied to each channel is calculated through a determined suppression rule, which has the purpose to produce an estimate (for each channel) of the noise power. Each particular STSA technique is characterized by the implementation of the filter bank and of the suppression rule.

If we denote the STFT of the $x(t)$ noisy signal with $X(t, w_k)$, where t represents the temporal index and w_k the frequency index (with $k = 1...N$, N represents the number of STFT channels), the result of the suppressing rule application can be interpreted as the application of a $G(t, w_k)$ gain to each value $Y(t, w_k)$ of the STFT of the noisy signal. This gain corresponds to a signal attenuation and is included between 0 and 1. In most of the suppression rules, $G(t, w_k)$ only depends on the noisy signal power level (measured at the same point) and on the estimate of the noisy power at the w_k frequency

$$\widehat{P}_n(w_k) = E\{|N(t, w_k)|^2\} \tag{14}$$

(which does not depend on the temporal index t due to the presumed noise stationariness). At this point a *relative* signal can be defined

$$Q(t, w_k) = \frac{|X(t, w_k)|^2}{\widehat{P}_n(w_k)} \tag{15}$$

which, starting from the hypothesis that the $n(t)$ noise is not correlated to the $x(t)$ signal, we deduce should be greater than 1

$$E\{Q(t, w_k)\} = 1 + \frac{E\{|S(t, w_k)|^2\}}{\widehat{P}_n(w_k)}. \tag{16}$$

A typical suppression rule is based on the Wiener filter [18] and can be formulated as follows

$$G(t, w_k) = \frac{|X(t, w_k)|^2 - \widehat{P}_n(w_k)}{|X(t, w_k)|^2}. \tag{17}$$

3.4. Kalman filter

The Kalman filter [19] is the optimal recursive Bayesian filter for linear systems observed in the presence of Gaussian noise. We consider that the state of the sound localization could be summarized by two position variables, x and y, and two velocities, v_x and v_y. These four variables are the elements of the state vector $\mathbf{x}_t$

$$\mathbf{x_t} = [x, y, v_x, v_y]^T. \tag{18}$$

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

$$\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{w}_{t-1} \tag{19}$$

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

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & dt & 0 \\ 0 & 1 & 0 & dt \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{20}$$

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

$$\mathbf{z}_t = \mathbf{H}\mathbf{x}_t + \mathbf{v}_t \tag{21}$$

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. Hence, we have

$$\mathbf{z}_t = \begin{bmatrix} \widehat{x} \\ \widehat{y} \end{bmatrix}, \tag{22}$$

and then we have

$$\mathbf{H} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{23}$$

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

$$\widehat{\mathbf{x}}_{t|t-1} = \mathbf{F}_t\widehat{\mathbf{x}}_{t-1|t-1}, \tag{24}$$

$$\mathbf{P}_{t|t-1} = \mathbf{F}_t\mathbf{P}_{t-1|t-1}\mathbf{F}^T + \mathbf{Q}_{t-1}, \tag{25}$$

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

$$\widehat{\mathbf{x}}_{t|t} = \widehat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}_t\widehat{\mathbf{x}}_{t|t-1}), \tag{26}$$

$$\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{H})\mathbf{P}_{t|t-1}, \tag{27}$$

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

$$\mathbf{K}_t = \mathbf{P}_{t-1|t-1}\mathbf{H}^T(\mathbf{H}_t\mathbf{P}_{t-1|t-1}\mathbf{H}^T + \mathbf{R}_t)^{-1}. \tag{28}$$

This formulation requires that the dynamic of the system is linear. However our specific problem is non-linear. To accommodate non-linear state transition and observation models, the Extended Kalman Filter (EKF) [20] implements a local linearization of the models. Thus, we need to compute new values for $\mathbf{F}$, at every time step, based on the state $\mathbf{x}$ to approximate the real update.

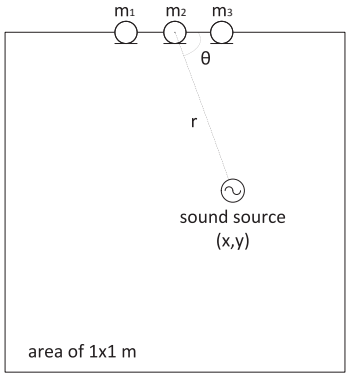

Figure 2: *The map of the considered control area.*

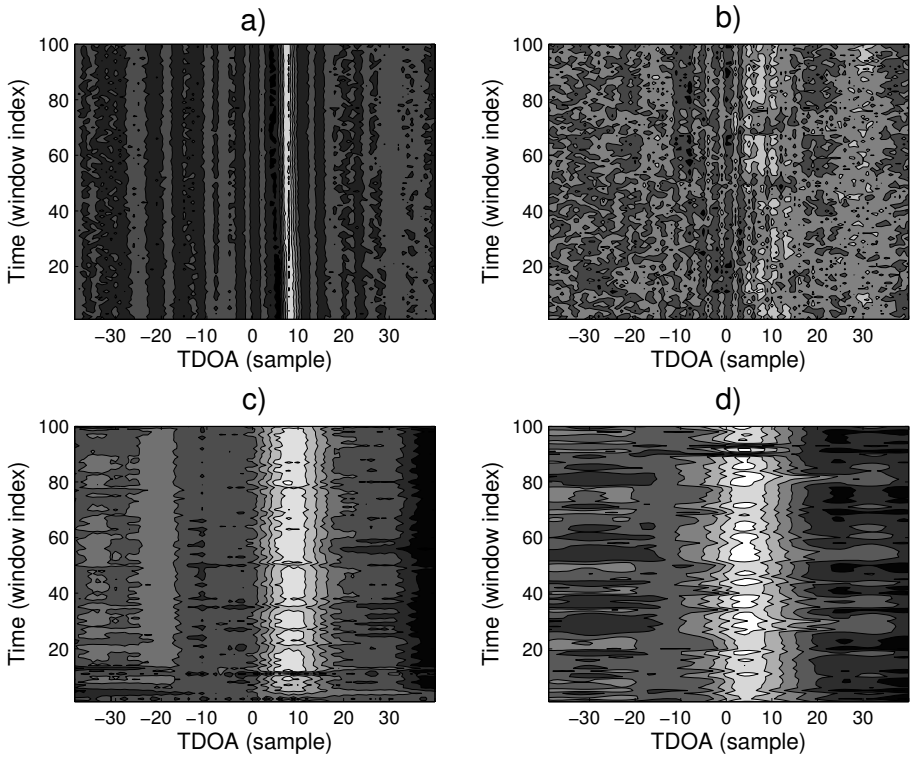

Figure 3: *Comparison of parameterized PHAT-β TDOA estimation performance. All sound sources are approximately located in (0,50) cm. a) White noise played on mobile device, $\beta = 1$. b) Flute, $\beta = 1$. c) Flute, $\beta = 0.65$. d) Flute, $\beta = 0.65$ and de-noise Wiener filter. The variances of TDOA estimation are: a) $\sigma_a^2 = 0.04$; b) $\sigma_b^2 = 80$; c) $\sigma_c^2 = 0.65$; d) $\sigma_d^2 = 0.5$.*

4. RESULTS

Some experimental results related to the localization performance of the interface in a real scenario are presented. To verify and validate our approach to the localization of pseudo-periodic sounds we used and compared three types of sources: white noise played on a mobile device, a flute played by a musician and a human voice. The interface works with sampling rate of 96 kHz, a Hanning analysis window of 42 ms, a time window for the estimation of the average noise (noise print) of 4.2 s. We used three microphones with supercardioid pickup pattern, which are the most frequently used microphones to acquire sound signals in electroacoustic music. It is important to highlight that the classic microphone for array processing is the omindirectional polar pattern, but its use is not appropriate in this context because of possible interference of the loudspeakers during the application in live performance. However, as we shall see, the use of directional microphones allows the localization of an acoustic source in the small area of interest (Figure 2). The distance between microphones is $d = 15$ cm. The working area is included in a square of side 1 meter. The axis origin coincides with microphone 2 (m_2) position, and x axis can vary between -50 cm and 50 cm and y axis between 0 and 100 cm (Figure 2).

Experiments have been done in a rectangular room of 3.5×4.5 m, with a moderately reverberant and noisy environment. Figure 3 shows a comparison of parameterized PHAT-β TDOA estimation performance. We made four tests with different parameters of interface configuration. We consider the TDOA estimation between microphone 2 and 3. All sound sources are approximately located in the center of interested map, (a) (5,52) cm, (b) (4,51) cm, (c) (5,53) cm, (d) (3,51) cm. In the first test (a), we played a continuous white noise signal by a mobile device with $\beta = 1$ interface configuration. In this way we checked the whole efficiency offered by the PHAT filter to optimize the TDOA estimation, reducing the degradation effects due to noise and reverberation. We can see in Figure 3 how the maximum peak detection is clearly visible (white line). We can also see the effects of multipath reverberation represented by the other parallel gray lines. The value of TDOA estimation is $\widehat{\tau}_{23} = 7$ (sample). The variance of TDOA maximum peak during the whole reproduction of sound is $\sigma_a^2 = 0.04$. The TDOA estimation is extremely accurate. In test (b), is considered a flute again with $\beta = 1$ parameter. As expected, the source is not detected ($\sigma_b^2 = 80$). Subsequently, in test (c) we examined a flute with $\beta = 0.65$ setting. The source is detected as shown in Figure 3. The mean value of TDOA estimation is $\widehat{\tau}_{23} = 7$ (sample) and it corresponds to the correct position of the source, the variance results $\sigma_c^2 = 0.65$. In the last test (d), we considered once more a flute with $\beta = 0.65$ and the de-noise Wiener filter task. The mean value of TDOA estimation is $\widehat{\tau}_{23} = 5$ (sample), the variance results $\sigma_c^2 = 0.5$. Hence, in this case, a lower value of variance indicates a less swinging of the TDOA than the average value, which is the correct location of the source.

Therefore, the parameterized PHAT-β allows the TDOA estimation of harmonic sounds, and de-noise component can improve the accuracy. However, the comparison with test (a), whose robust and well-defined result we aim to obtain, does not give yet satisfac-

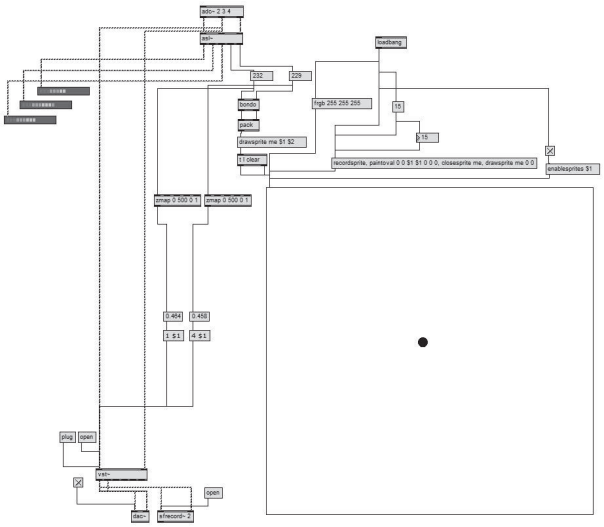

Figure 6: *The Max/MSP interface with the external object asl∼.*

tory results. A parameterization of PHAT with value of $\beta = 0.65$, in according to [16], is a good compromise between filtering and detection.

Before considering the experiments with the Kalman filter, we show the results of a test with a human voice, ranging from harmonic to noise sounds, in order to verify the threshold value of zero-crossing rate for the activation of PHAT-β. The human voice source is located in (-20,70) cm. Figure 4 shows the results of the ZCR and adaptive parameterized GCC-PHAT-β with threshold value of $\mu = 0.03$, $\beta = 0.65$ when ZCR < μ and de-noise Wiener filter. We believe that this value μ is enough to achieve an adequate adaptation of GCC. Still in Figure 4, we can note that when the sound becomes harmonic, and then we have partially filtered GCC with PHAT, the TDOA peak tends to widen, reducing its robustness, but still allowing the estimation of source position.

Finally, the last test on the localization performance shows the effectiveness of the Kalman filter to make the xy coordinates more accurate and usable in the interface. Once again we used a flute moving within the mapped area. The threshold value of ZCR is $\mu = 0.03$, $\beta = 0.65$, and de-noise task is active. As you can see from Figure 5, the black lines, which represents the data after the Kalman filtering, are reported in order to have less stability problems due to reverberation. In fact, the estimated raw data (gray lines) present very high swinging values, which would make the interface inappropriate to control the processing parameters.

In conclusion, we implemented the system by developing a Max/MSP external object, named asl∼, in order to validate the interface in real-world music application. The object receives incoming audio signals acquired by the three microphones and, as output, it gives the position of the sound source. The object performs all the signal processing techniques described in the previous sections. Moreover, a simple Max/MSP patch (see Figure 6) has been developed in order to control in real-time an audio processor. As mentioned, xy values have been used to directly control the parameters of an audio effect. We made use of different VST plug-ins, such as reverb and delay effects, with encouraging results.

5. CONCLUSIONS

We described a digital interface that incorporates real-time sound source localization for gestural control without any physical contact, which can be used as audio HCI system to enhance the experience of a musical performance. In order to work with harmonic sounds, we proposed a system consisting of adaptive parameterized GCC-PHAT with zero-crossing rate threshold. We have seen that this solution allows to locate sources such as musical instruments, but it is less robust in moderate reverberation and noisy environments, comparing to the standard GCC-PHAT. For this reason, we included two filters. The first one has been set up using the STSA with Wiener filter, before the TDOA estimation task, in order to improve the SNR of signals. The second one has been formulated using the Kalman filter theory, after the estimation of the source position. In this way, we obtained an accurate localization system. We used a linear array of three supercardioid polar pattern microphones, and we have seen that we are able to locate the sound source inside an area of one square meter. The usability of the interface was validated by developing a Max/MSP external object, so that we can map the xy position of the sound source (i.e. voice, traditional musical instruments and sounding mobile devices) into control parameters.

Future works include the test and use of the sound localization based interface in real application of live performance to verify how the system works with interfering sources from a sound reinforcement system and other instruments. In addition, we plan to test other mapping strategies in order to obtain a more articulate, complex and interesting system.

6. REFERENCES

[1] E. R. Miranda and M. M. Wanderley, *New Digital Musical Instruments: Control and Interaction Beyond the Keyboard*, A-R Editions, 2006.

[2] T. Todoroff, *DAFX: Digital Audio Effects*, chapter Control of Digital Audio Effects, pp. 465–498, Wiley, 2001.

[3] J. Paradiso and N. Gershenfeld, "Musical applications of electric field sensing," *Computer Music Journal*, vol. 21(3), pp. 69–89, 1997.

[4] N. Griffith and M. Fernstrom, "Litefoot - a floor space for recording dance and controlling media.," in *Proc. International Computer Music Conference*, 1998, pp. 475–481.

[5] G. Castellano, R. Bresin, A. Camurri, and G. Volpe, "Expressive control of music and visual media by full-body movement," in *Proc. International Conference on New Interfaces for Musical Expression*, 2007, pp. 390–391.

[6] A. de Götzen, "Enhancing engagement in multimodality environments by sound movements in a virtual space," *IEEE Multimedia*, vol. 11, pp. 4–8, 2006.

[7] V. Verfaille, M. Wanderley, and P. Depalle, "Mapping strategies for gestural and adaptive control of digital audio effects," *Journal of New Music Research*, vol. 35, pp. 71–93, 2006.

[8] J. Allouis and J. Y. Bernier, "The SYTER project: Sound processor design and software overview," in *Proc. International Computer Music Conference*, 1982, pp. 232–240.

[9] M. Spain and R. Polfreman, "Interpolator: a twodimensional graphical interpolation system for the simultaneous control

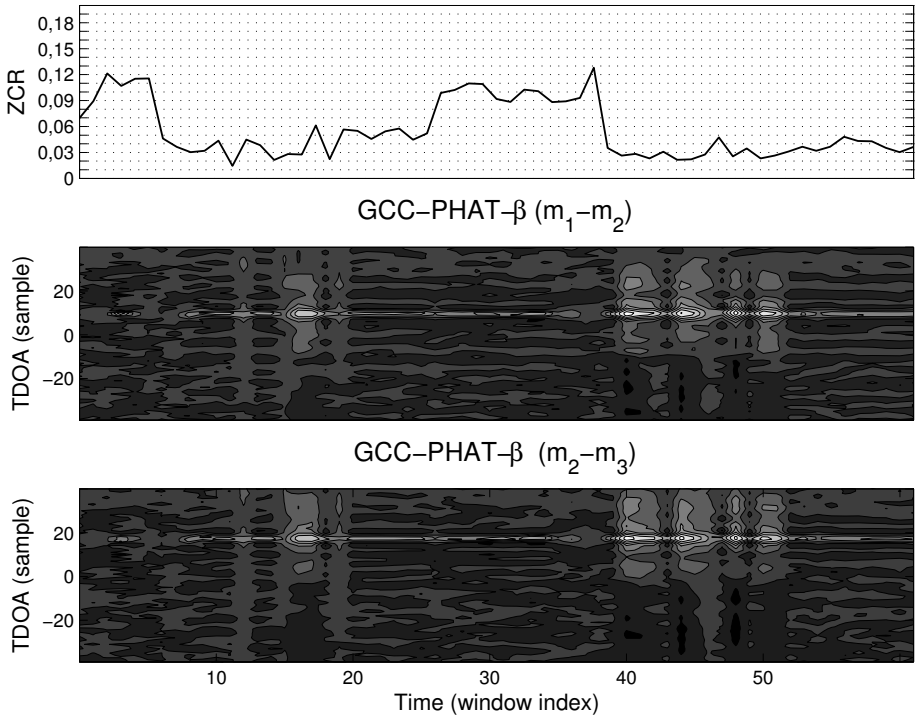

Figure 4: *Human voice located (-20,70) cm. ZCR and parameterized GCC-PHAT-β with threshold value of $\mu = 0.03$; $(m_1 - m_2)$ is referred to TDOA estimation between microphone 1 and 2, whereas $(m_2 - m_3)$ between 2 and 3.*

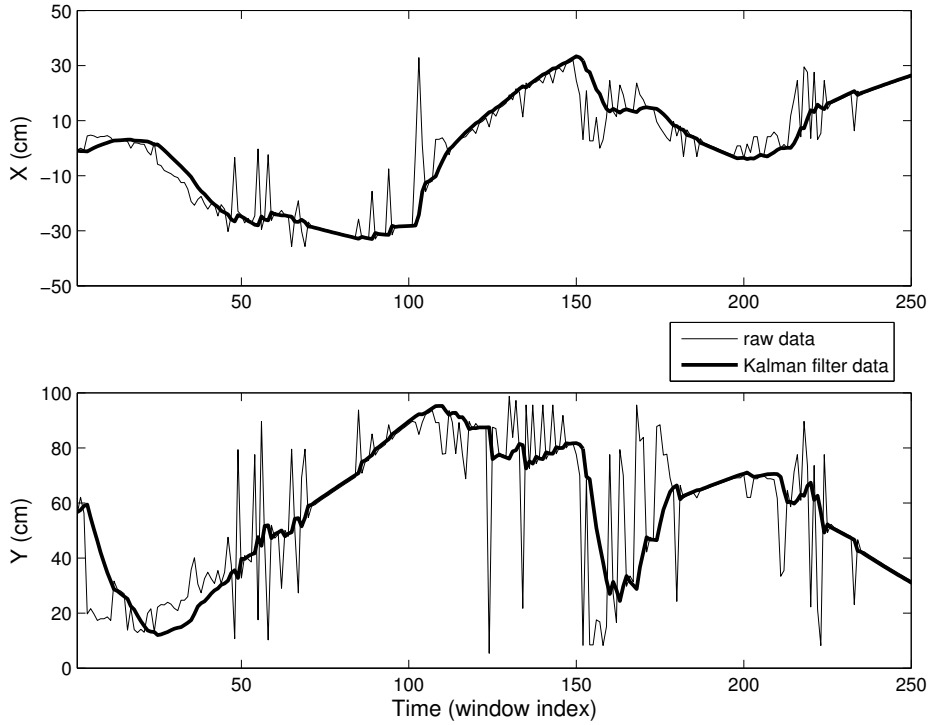

Figure 5: *Flute performance moving within the interested area; X and Y position of Kalman filtering data (black lines) and raw data (gray lines).*

of digital signal processing parameters," *Organised Sound*, vol. 6, no. 2, pp. 147–152, 2001.

[10] A. Momeni and D. Wessel, "Characterizing and controlling musical material intuitively with geometric models," in *Proc. International Conference on New Interfaces for Musical Expression*, 2003, pp. 54–62.

[11] R. Bencina, "The metasurface: applying natural neighbour interpolation to two-to-many mapping," in *Proc. International Conference on New Interfaces for Musical Expression*, 2005, pp. 101–104.

[12] O. Larkin, "INTĹIB - A Graphical Preset Interpolator For Max MSP," in *Proc. International Computer Music Conference*, 2007.

[13] 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.

[14] 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.

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

[16] K. D. Donohue, J. Hannemann, and H. G. Dietz, "Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments," *Signal Processing*, vol. 87, pp. 1677–1691, July 2007.

[17] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSSP-27, no. 2, pp. 113–120, 1979.

[18] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series with engineering applications*, Cambridge, MIT Press, Massachusetts, 1949.

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

[20] S. Schmidt, "Applications of state-space methods to navigation problems," *Advances in Control Systems*, vol. 3, pp. 293–340, 1966.