4.3. Partikkel Hadron Application

In the last example we use the Partikkel Hadron\cite{1} granular synthesizer with one of the provided pre-set.

Through granular synthesis it is possible to obtain large timbre variation due to the nature of the synthesis, but often the control parameter set is large and challenging to design an interface for. This device exposes just 6 parameters for timbre manipulation thanks to the exploitation of the Modulation Matrix \cite{7}. We analyze the generated audio with 25 Bark bands energies over the whole control space given by all possible combinations of the 6 parameters. Sixteen analysis windows per state \( p \) are computed with a hope size of 512 samples, using C2 as fixed note. In this more complex scenario the performance of ISOMAP is sensibly better than PCA. The data presented in Figure 8 shows how ISOMAP, when compared with PCA, allows the reduction of at least one dimension in the control space \( C \) without losses in the overall descriptors space energy. With the ISOMAP adaptation, we obtain a further reduction of the control space. This enables the use of a simple 2D controller, while still permitting the navigation of the majority of the granularity sonic space spanned by the original 6 parameters.

The current MATLAB implementation of the ISOMAP algorithm is computationally expensive in terms of time and memory, thus we had to limit the dimensionality of \( D \) and \( P \) to 4000. This is too small to handle large numbers of parameters sampled with a high resolution. This limitation of the resolution is reflected in the usability experience. An optimization of the algorithm implementation is thus desirable.

In Section 3 we make some assumptions about the control interface output signals. These are generally true for most of the commercial general-purpose interfaces (e.g. sets of sliders, knobs, touch surfaces, touch screen devices). For other interfaces built with large numbers of sensors, or devices capturing human gesture through image or sound, the assumptions may not hold. Through a statistical study of the interface signals it should be possible to apply a pre-processing stage that produces independent components within the desired range.

5. CONCLUSION AND FUTURE WORK

We presented a generic method to adapt general-purpose interfaces to synthesis engines through unsupervised dimensionality reduction techniques and statistical analysis of the perceptually related features computed over the synthetic sound. The application of the prototype demonstrates the benefits introduced by this adaptive technique, including the linearization of the relationship controller-to-sound and the dimensionality reduction of the control space. However some aspects can be further explored for improvements.

The exploitation of dynamic features in synthetic timbres must be explored more extensively. The computation of the dynamic aspect of the timbre has been tested, but embedding static and dynamic features in the same vector \( d \) may not be appropriate for all cases. Storing this information in two separate matrices and running dimensionality reduction separately on each may result in an adaptive mapping that is easier to use.

REFERENCES


A MICROPHONE ARRAY INTERFACE FOR REAL-TIME INTERACTIVE MUSIC PERFORMANCE

Daniele Salvati

AVIRE 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

Gian Luca Foresti

AVIRE lab
Dep. of Mathematics and Computer Science,
University of Udine, Italy
gianluca.foresti@uniud.it

ABSTRACT

This paper presents a novel digital musical interface for real-time interactive music performance, which uses a microphone array to estimate the sound source position in the plane and to allow a performer to use the two x-y coordinates of position to control a audio processing module in real-time through the spatial movement of a sound source. Musical interfaces are often used to allow non-musicians to enhance the expressive control on the sounds generated by their acoustic instruments in a live electronics context. E.g., in the works by Adriano Guarneri – Medco (2002) and Fili bianco-velati (2005) – produced at the “Centro di Sonologia Computazionale” of Padova, the movement of a musician is followed by a motion capture system based on infrared cameras to control a live electronics patch \cite{4}, and using the robust but very expensive, PhaseSpace optical motion capture system. It is composed by led systems, video cameras, and calibration procedure. In general, those kind of systems have considerable complexity and in some situations there could be problems with the low and/or not always controllable lighting of the concert hall, even when using infrared camera. It has been shown in \cite{14} that there is some potentiality in using the source sound localization to directly control the position of a sound played back through a spatialization system by moving the sound produced by its own musical instrument. This work has been improved in \cite{13} introducing an adaptive parameterized Generalized Cross-Correlation (GCC) PHAT filter to localize musical sounds that are mainly harmonics. Both interfaces \cite{14,13} are been tested in a controlled real environment without verifying how the system works with interfering sources from a sound reinforcement system and other instruments. Thus, in this paper a validation in multi-source scenario is presented, introducing the adaptive parameterized SRP-PHAT with a ZCR threshold (Section 3) that has a better performance than the parameterized GCC-PHAT proposed in \cite{13} as shown in Section 4.

1. INTRODUCTION

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 improve the voice recognition using the acoustic source localization and the beam-forming for noise suppression. In the past years, a large number of musical interfaces has been implemented with the goal of providing tools for gestural interaction with digital sounds, using systems played by touching or holding the instrument, interfaces with haptic feedback, systems worn on the body, and interfaces that may be played without any physical contact (electric field sensors \cite{12}, optical sensors \cite{7}, ultrasound systems \cite{10}, and video camera that allows the performer to use their full-body for controlling in real-time the generation of an expressive audio-visual feedback \cite{1}).
vector map to improve the localization performance. Given the Coherence Field (GCF) to construct a spatial analysis between pairs of microphones and using the Global Coherence Function (GCF). It consists of calculating the GCC-PHAT function from the microphone signals. The SRP-PHAT [5] is based on the concept of adding several time delay estimation functions from the microphone pairs. The last component implements the mapping strategy [16] to associate the x-y coordinates with audio processing parameters. The array system is composed of three half-supercardioid polar pattern microphones, which reduce ambient noise and pickup of room reverberation, 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 source sound position in an horizontal plane by providing its Cartesian coordinates. A SRP PHAT method is used to compute the acoustic map analysis. To improve the performance in the case of harmonic sounds, or generally pseudo-periodic sounds, a parameterized SRP-PHAT is proposed. The ZCR function is used to determine if a sound is pseudo-periodic, and to adapt with a threshold value the parametrically control of PHAT filter. A Kalman filter [8] is applied to smooth the time series of observed position to obtain a more robust and accurate x-y values. The last component implements the mapping strategy [16] to associate the x-y coordinates with audio processing parameters.

3. ADAPTIVE PARAMETERIZED SRP-PHAT WITH ZERO-CROSSING RATE THRESHOLD

The SRP-PHAT [5] is based on the concept of adding several time delay estimation functions from the microphone pairs. It consists of calculating the GCC-PHAT function between pairs of microphones and using the Global Coherence Field (GCF) [11] to construct a spatial analysis map to improve the localization performance. Given the vector \( s = [x y]^T \) of space and \( R \) microphone pairs, the SRP-PHAT at time \( t \) can be expressed

\[
\hat{X}(s, f)[\text{GCC}(t)] = \sum_{i=1}^{R} \hat{X}(s, f)[\text{GCC}(t)]
\]

where the \( \text{GCC}(t) \) is the GCC-PHAT of the \( r \)th pair. The position of the source is estimated by picking the maximum peak

\[
\hat{s}_r = \arg\max_x \hat{X}(s, f)[\text{GCC}(t)].
\]

The GCC-PHAT [9] is the classic method to estimate the relative time delay associated with acoustic signals received by a pair of microphones in a moderately reverberant and noisy environment. The GCC-PHAT basically consists of a cross-correlation followed by a filter that aims to reduce the performance degradation caused by additive noise and multi-path channel effects. The GCC-PHAT in the frequency domain is

\[
\hat{X}(s, f) = \frac{1}{L} \sum_{\ell=0}^{L-1} \Psi(f)[\text{GCC}(\hat{S})]
\]

where \( f \) is the Discrete Fourier Transform (DFT) integer frequency index, \( L \) is the number of samples of the observation time, and \( \Psi(f) \) is the frequency domain PHAT weighting function, and the cross-spectrum of the two signals is defined as

\[
\hat{S}_{xixj}(f) = E[X(f)X^T(f)]
\]

where \( X(f) \) and \( X^T(f) \) are the DFT of the signals \( x(t) \) and \( x(t) \) respectively, \( * \) 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 PHAT weighting function places equal importance on each frequency by dividing the spectrum by its magnitude. The PHAT normalizes the amplitude of the spectral density of the two signals and uses only the phase information to compute the GCC

\[
\Psi_{\text{PHAT}}(\beta) = \frac{1}{\hat{S}_{xixj}(f)}.
\]

We note that SRP-PHAT, which uses the sum of the GCCs of the microphone pairs, is equivalent to using a steered response filter and sum beamforming with PHAT weighting. In fact, the SRP of a 2-element array is equivalent to the GCC of those two microphones. The SRP-PHAT algorithm has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments [15]. This algorithm enhances the performance of localization with a network of large arrays. However, the computational cost of the method is very high. To reduce the processing time of search algorithms, improvements have been suggested [2][3].

It is important to note that the PHAT performance is dramatically reduced in the case of harmonic sounds, or generally pseudo-periodic sounds. In fact, the PHAT has less capability to reduce the deleterious effects of noise and reverberation when it is applied to a pseudo-periodic sound. An accurate analysis of the PHAT performance for a broadband and narrowband signal can be found in [6]. The results of this work highlight the ability of the PHAT to enhance the detection performance for single or multiple targets in noisy and reverberant environments, when the signal covers most of the spectral range.

Thus, to work with pseudo-periodic sounds the proposal is to use a parameterized SRP-PHAT that weighs the contribution of the PHAT filtering, depending on the threshold of the ZCR parameters. The PHAT weighting can be generalized to parametrically control the level of influence from the magnitude spectrum [6]. This transformation will be referred to as the PHAT-\( \beta \) and defined as

\[
\Psi_{\text{PHAT}}(\beta,f) = \frac{1}{\hat{S}_{xixj}(f)^\beta}
\]

where \( \beta \) varies between 0 and 1. When \( \beta = 1 \), equation (6) 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 the case of harmonic sounds, we can use an intermediate value of \( \beta \) so that we can detect the peak to estimate the time delay between signals, and can have a system, at least in part, that exploits the benefits of PHAT filtering to improve performance in moderately reverberant and noisy environments. The results of localization improvement in case of pseudo-periodic sounds using the PHAT-\( \beta \) are reported in [13]. To adapt the value of \( \beta \), we use the ZCR to determine if the sound source is periodic. ZCR is a very useful audio feature and is defined as the number of times that the audio waveform crosses the zero axis

\[
ZCR(t) = \sum_{i=1}^{L} \left| sgn(x(t+i)) \right| \cdot sgn(x(t+i-1))
\]

where \( sgn(x) \) is the sign function.

![Figure 1. Block diagram of the interface.](image1)

![Figure 2. The sample space position of a square with 100 cm sides considering three microphones with distance of 25 cm and TDOAs between microphones s and m are: a) 44110 Hz; b) 96000 Hz.](image2)

![Figure 3. The analysis map area with microphones and interference sources (pink noise) position (C1 and C2). In the experiments the flute sound source moves from point S to point E.](image3)
2. SYSTEM ARCHITECTURE

The proposed interface has the advantage of being completely non-invasive (no need for markers, sensors or wires on the performance), and requires no dedicated hardware. The architecture consists of a microphone array and digital signal processing algorithms for robust sound localization. Figure 1 summarizes the system architecture of the interface. The array system is composed of three half-supercardioid polar pattern microphones, which reduce ambient noise and pickup of room reverberation, 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 sound source position in an horizontal plane by providing its Cartesian coordinates. A SRP-PHAT method is used to compute the acoustic map analysis. To improve the performance in the case of harmonic sounds, or generally pseudo-periodic sounds, a parameterized SRP-PHAT is proposed. The ZCR function is used to determine if the signal is pseudo-periodic, and to adapt with a threshold of the ZCR parameters. The PHAT weighting function is used to estimate the time delay between signals, and can have a multiple targets in noisy and reverberant environments, when the signal covers most of the spectral range.

3. ADAPTIVE PARAMETERIZED SRP-PHAT WITH ZERO-CROSSING RATE THRESHOLD

The SRP-PHAT [5] is based on the concept of adding several time delay estimation functions from the microphone pairs. It consists of calculating the GCC-PHAT function between pairs of microphones and using the Global Coherence Factor (GCF) [11] to construct a spatial analysis map to improve the localization performance. Given the vector $x = [x_1] \rightarrow$ of space and $K$ microphone pairs, the SRP-PHAT at time $t$ can be expressed

$$\gamma(x, f) = \sum_{i=1}^{K} \hat{\gamma}(x, f) = \sum_{i=1}^{K} f \hat{\gamma}(x, f)$$

where the $\hat{\gamma}(x, f)$ is the GCC-PHAT of the $kth$ pair. The position of the source is estimated by picking the maximum peak

$$s_n = \arg \max \gamma(x, f)$$

The GCC-PHAT [9] is the classic method to estimate the relative time delay associated with acoustic signals received by a pair of microphones in a moderately reverberant and noisy environment. The GCC-PHAT basically consists of a cross-correlation followed by a filter that aims to reduce the performance degradation caused by additive noise and multi-path channel effects. The GCC-PHAT in the frequency domain is

$$\gamma_{GCC}(f) = \frac{1}{L} \sum_{l=-L}^{L} \gamma(x, f)[e^{i2\pi f t}]$$

where $f$ is the discrete Fourier Transform (DFT) integer frequency index, $L$ is the number of samples of the observation time, and $\gamma(x, f)$ is the frequency domain PHAT weighting function, and the cross-spectrum of the two signals is defined as

$$S_{xixj}(f) = E[\{X(f)Y(f)]$$

where $X(f)$ and $Y(f)$ are the DFT of the signals $x(t)$ and $y(t)$ respectively, and $\ast$ 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 PHAT weighting function places equal importance on each frequency by dividing the spectrum by its magnitude. The PHAT normalizes the amplitude of the spectral density of the two signals and uses only the phase information to compute the GCC

$$\Psi_{PHAT}(\beta, f) = \frac{1}{S_{xixj}(f)}$$

We note that SRP-PHAT, which uses the sum of the GCCs of the microphone pairs, is equivalent to using a steered response filter and sum beamforming with the PHAT weighting. In fact, the SRP of a 2-element array is equivalent to the GCC of those two microphones.

The SRP-PHAT algorithm has been shown to be one of the most robust sound source localization approaches operating in noisy and reverberant environments [15]. This algorithm enhances the performance of localization with a network of large arrays. However, the computational cost of the method is very high. To reduce the processing time of search algorithms, improvements have been suggested [2][3].

It is important to note that the PHAT performance is dramatically reduced in the case of harmonic sounds, or generally pseudo-periodic sounds. In fact, the PHAT has less capability to reduce the deleterious effects of noise and reverberation when it is applied to a pseudo-periodic sound. An accurate analysis of the PHAT performance for a broadband and narrowband signal can be found in [6]. The results of this work highlight the ability of the PHAT to enhance the detection performance for single or multiple targets in noisy and reverberant environments, when the signal covers most of the spectral range.

Thus, to work with pseudo-periodic sounds the proposal is to use a parameterized SRP-PHAT that weights the contribution of the PHAT filtering, depending on the threshold of the ZCR parameters. The PHAT weighting can be generalized to parameterically control the level of influence from the magnitude spectrum [6]. This transformation will be referred to as the PHAT-$\beta$ defined as

$$\Psi_{PHAT}(\beta, f) = \frac{1}{S_{xixj}(f)}$$

where $\beta$ varies between 0 and 1. When $\beta = 0$, equation (6) becomes the conventional PHAT and the modulus of the transfer function becomes 1 for all frequencies; when $\beta = 1$, the PHAT has no effect on the original signal, and we have the cross-correlation function.

Therefore, in the case of harmonic sounds, we can use an intermediate value of $\beta$ so that we can detect the peak to estimate the time delay between signals, and can have a system, at least in part, that exploits the benefits of PHAT filtering to improve performance in moderately reverberant and noisy environments. The results of localization improvement in case of pseudo-periodic sounds using the PHAT-$\beta$ are reported in [13]. To adapt the value of $\beta$, we use the ZCR to determine if the sound source is periodic. ZCR is a very useful audio feature and is defined as the number of times that the audio waveform crosses the zero axis

$$ZCR(t) = \sum_{i=1}^{N} \{sgn[x(i)] - sgn[x(i+1)]\}$$

where $sgn(x)$ is the sign function.
sides; the peak of GCF map outside the area of interest is automatically discarded from the system. A flute sound (continuous G5 note) was used as source of interest, and sides; the peak of GCF map outside the area of interest is automatically discarded from the system. A flute sound (continuous G5 note) was used as source of interest, and the localization source estimation is not possible for both grades the localization performance, in particular for the microphone array, we obtained some experimental step to obtain a more accurate value of position mobile device reference position of 40 cm from the microphone array. We can see a good localization of the source up to SIR=20 dB with SRP-PHAT-β, but at 15 dB of SIR the localization source estimation is not possible for both algorithms by the system.

Finally, an experiment involving a real flute instrument with β=0.85 and SIR=35 dB is shown in Figure 6. The post-processing Kalman filter is a crucial and fundamental step to obtain a more accurate value of position and to track the source movement.

5. CONCLUSIONS

A digital musical interface based on sound localization using a microphone array allows a performer to directly interact with a computer by moving a sounding object, and to plan and conduct the expressivity of a performance by controlling an audio processing module. A novel algorithm based on adaptive SRP-PHAT-β was proposed to solve the problem of acoustic source localization when the sound is pseudo-periodic. By developing a real-time software in Max environment, we obtained some experimental results that show the system is able to locate the source in a more accurate and robust way than the GCC-PHAT-β (Figure 5). Moreover, when the SIR is 25 dB the GCC-PHAT-β has more errors, compared to the SRP-PHAT-β. Table 1 summarizes the values of Root Mean Square Error (RMSE) for the variable y and shows the maximum and minimum estimates of the variable x. The RMSE of the variable y is computed considering the maximum and minimum estimates of the variable x.

Besides, it is shown how the angular resolution degrades the localization performance, in particular for the GCC-PHAT-β, in correspondence with the increase of competitive sources; locate is impossible when the SIR is less than 20 dB. An increase in the number of microphones could improve the system performance, especially with an algorithm based on SRP, which tends to have a more robust performance with a large array.

The interface has considerably less complexity than systems based on electric field, optical and video camera sensors. The latter are widely used, especially for great success of Kinect, but in general it is possible to have problems for the low or variable lighting condition during live electronics. One of the most robust way is to use a motion capture system (e.g., PhaseSpace), which is very complex and expensive. Hence, we believe that our approach based on microphone array is presented as a viable alternative. However, in the future, we plan to use and test the microphone array interface in live electronic performances.

6. REFERENCES

Figure 4. The x-y position estimation by the system of a flute sound with SIR=35 dB. The Kalman filter data is the black line and raw data are the dots. a) GCC-PHAT-β b) SRP-PHAT-β

Table 1. The values of RMS\(E(y)\), maximum and minimum of \(x\) related to the experiments.

<table>
<thead>
<tr>
<th>SIR</th>
<th>RMS(E(y))</th>
<th>Max (x)</th>
<th>Min (x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 dB</td>
<td>-10.5</td>
<td>90</td>
<td>-100</td>
</tr>
<tr>
<td>30 dB</td>
<td>-10.2</td>
<td>80</td>
<td>-100</td>
</tr>
<tr>
<td>35 dB</td>
<td>-10.0</td>
<td>70</td>
<td>-100</td>
</tr>
</tbody>
</table>

Figure 5. The x-y Kalman position estimation by the system with different SIR. a) GCC-PHAT-β b) SRP-PHAT-β

Figure 6. The x-y position estimation of a real flute instrument with SRP-PHAT-β and SIR=35 dB.

5. CONCLUSIONS

A digital musical interface based on sound localization using a microphone array allows a performer to directly interact with a computer by moving a sounding object, and to plan and conduct the expressivity of a performance by controlling an audio processing module. A novel algorithm based on adaptive SRP-PHAT-β was proposed to solve the problem of acoustic source localization when the sound is pseudo-periodic. By developing a real-time software in Max environment, we obtained some experimental results that show the system is able to locate the source in a more accurate and robust way than the GCC-PHAT-β, when SIR ≥ 20 dB in a multi-source moderate reverberant (RT60=0.35 s) and noisy room.

6. REFERENCES


[16] V. Verfaille, M. Wenderly, and P. Depalle, “Map-Mod: multimodal interaction and prediction strategies for gestural and adaptive control of spoken Dialogue systems based on electric field, optical and video camera sensors. The latter are widely used, especially for great success of Kinect, but in general it is possible to have problems for the low or variable lighting condition during live electronics. One of the most robust way is to use a motion capture system (e.g., PhaseSpace), which is very complex and expensive. Hence, we believe that our approach based on microphone array is presented as a viable alternative. However, in the future, we plan to use and test the microphone array interface in live electronic performances.