INTERPOLATION METHODS FOR THE SRP-PHAT ALGORITHM

Sakari Tervo and Tapio Lokki

Helsinki University of Technology
Department of Media Technology
P.O.Box 5400, FI-02015 TKK, Finland

ABSTRACT

The steered response power-phase transform (SRP-PHAT) algorithm is known to be one of the state-of-the-art methods in acoustical source localization (ASL) and it has been shown to outperform other traditional ASL methods such as the steered beamformer-based (SBF) method. The accuracy of the SRP-PHAT algorithm is limited by the time resolution of the PHAT weighted cross correlation functions — the basic building blocks of the SRP-PHAT algorithm. In this article, three methods for interpolating the cross correlation functions of the SRP-PHAT are compared with real concert hall data. Two of the methods are build from previously introduced assumptions on the shape of the global maximum of the cross correlation function. In the experiments it is shown that certain limitations of the SRP-PHAT algorithm can be compensated with interpolation.

Index Terms— Time delay estimation, interpolation, acoustical source localization, and microphone arrays.

1. INTRODUCTION

In acoustical source localization (ASL) and in direction of arrival (DOA) estimation, the steered response power-phase transform (SRP-PHAT) has been of keen interest in recent research. The SRP-PHAT algorithm has been studied extensively [1, 2, 3, 4, 5, 6, 7] and it has been followed by various modifications and optimizations [3, 4, 5, 7].

The SRP-PHAT algorithm can be seen as a two-step localization method. Firstly, the time delay estimation (TDE) functions are calculated from the received microphone signals, where the *time delay* is the difference of the arrival times of a wavefront in two receivers. Many methods have been proposed and studied for the TDE problem over decades. Among the most popular TDE methods are the cross correlation-based approaches [8]. Also the SRP-PHAT algorithm uses the cross-correlation, with phase transform (PHAT)-weighting. Secondly, in the SRP-PHAT algorithm, the steered response power is evaluated by summing the steered TDE functions for each location candidate. The location candidate that gets the highest value is then the location estimate.

The accuracy of the SRP-PHAT algorithm is limited by the time resolution of the discrete TDE functions. Due to limited time resolution the discrete SRP-PHAT function is spatially quantized. In addition, the spatial quantization is also effected by the locations of the microphones. These factors can lead to biased and erroneous estimation with the SRP-PHAT algorithm. To overcome the unwanted effects, interpolation is needed.

In traditional TDE, the interpolation is done usually by fitting a parabola [9] or an exponential function [10] to the maximum peak of the TDE function. TDE and its interpolation leads to a single time delay estimate. In the SRP-PHAT function, the spatial response is build on several values of the TDE functions. This results in the fact that the function fitting TDE interpolation methods can not be used directly for interpolating the TDE functions of the SRP-PHAT function. Therefore, an algorithm for using the function fitting approaches in the SRP-PHAT is developed. In earlier work Do and Silverman applied the interpolation of the SRP-PHAT function in spectral domain with cubic interpolation in [7]. In this article the cubic interpolation is not considered. Traditional interpolation methods such as Fourier-interpolation can be used to increase the time resolution of the signals or the TDE function as in [9].

The performance of the proposed interpolation methods and Fourier-interpolation is examined in a *real* concert hall environment. Although, such an environment is quite demanding for the SRP-PHAT algorithm, it offers a real world data to study interpolation methods in a function of temporal (sampling frequency) and spatial resolutions. In a concert hall, reflections from surfaces are copies of the direct sound and they can be considered as separate sound sources from the directions of which could also be estimated with the SRP-PHAT algorithm. However, in this article the DOA estimation of the direct sound is only considered.

The paper is organized as follows. In the second section, the signal model for reverberant environments, time delay estimation, and the SRP-PHAT algorithm are introduced. In addition, the proposed method for the interpolation is given. In the third section the performance of the methods is examined and the results are presented and discussed. The final conclusions are given in Section 4.

2. SOURCE LOCALIZATION

In a concert hall environment, the source signal $s(t)$ is affected by the channel h from the source to the receiver, and noise $n_{\mathrm{i}}(t)$ [1]: $x_{\mathrm{i}}(t) = h * s(t) + n_{\mathrm{i}}(t)$, where $*$ is convolution, and $x_{\mathrm{i}}(t)$ is the received signal in microphone i. The noise n_i is independent and identically distributed (i.i.d.) for all microphones i.

Time delay is the time difference of arrival between any two received signals x_{i} and x_{j}. A traditional method in obtaining the time delay is to find the maximum argument of the generalized cross correlation (GCC) function [8]:

$$\mathrm{R}_{x_{\mathrm{i}}x_{\mathrm{j}}}(\tau) = \mathcal{F}^{-1}\{\mathcal{W}(f)\mathrm{C}_{x_{\mathrm{i}}x_{\mathrm{j}}}(f)\}, \tag{1}$$

where $\mathcal{W}(f)$, $\mathrm{C}_{x_{\mathrm{i}}x_{\mathrm{j}}}(f)$, and $\mathcal{F}^{-1}$, are the weighting function, cross power spectral density (CPSD) between signals x_{i} and x_{j}, and inverse Fourier transform, respectively. Many weighting functions for the GCC have been listed in [8]. The SRP-PHAT algorithm uses the phase transform (PHAT) weighting [8]:

$$\mathcal{W}_{\mathrm{PHAT}}(f) = |\mathrm{C}_{x_{\mathrm{i}}x_{\mathrm{j}}}(f)|^{-1}. \tag{2}$$

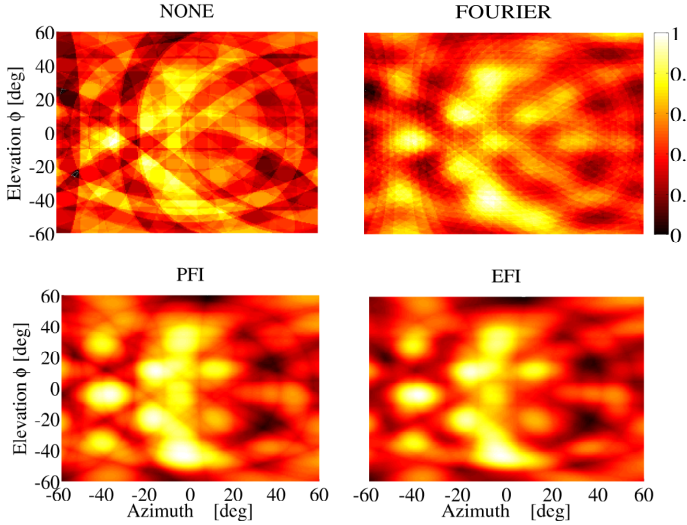

Fig. 1. Normalized SRP-PHAT spatial responses with no interpolation (top left), Fourier-interpolation at 96 kHz (top right), parabolic fitting (bottom left), and exponential fit (bottom right) for signals sampled at 32 kHz.

The SRP-PHAT source localization function is calculated as a sum of the PHAT weighted cross correlation functions [1]:

$$\mathrm{P}(\mathbf{r}) = \sum_{\{\mathrm{i},\mathrm{j}\}=1}^{M} \mathrm{R}_{x_\mathrm{i},x_\mathrm{j}}(\tau(\mathrm{i},\mathrm{j},\mathbf{r})), \tag{3}$$

where $\{\mathrm{i},\mathrm{j}\}$ denotes the microphone pair and M is the number of microphone pairs. The location dependent time delay term is given [1]: $\tau(\mathrm{i},\mathrm{j},\mathbf{r}) = c^{-1}(|\mathbf{r}-\mathbf{r}_\mathrm{j}| - |\mathbf{r}-\mathbf{r}_\mathrm{i}|)$, where $\mathbf{r}_\mathrm{i}$ and $\mathbf{r}_\mathrm{j}$ are the microphone locations, c is speed of sound and $\mathbf{r}$ is a location candidate. Location estimate is the maximum argument of the SRP-PHAT function [1]: $\hat{\mathbf{r}}_s = \arg\max_{\mathbf{r}} \mathrm{P}(\mathbf{r})$.

The experiments in the next section are limited to the DOA estimation of the azimuth and elevation angles. Due to DOA estimation, the location candidates are distributed uniformly in a grid that is on the surface of a unit sphere. The region of interest was selected to be from −60 to 60 degrees for both azimuth and elevation angles. An example of the spatial response $\mathrm{P}(\mathbf{r})$ is shown in Fig. 1.

2.1. Proposed Interpolation of the TDE functions

In order to able to extend the function fitting interpolation to the whole TDE function, two known issues of the cross-correlation based TDE functions are considered. Firstly, it is known that in a reverberant environment reflection paths can cause a higher peak in the GCC-PHAT function than the direct path [5]. In other words, n:th highest local maximum in the GCC-PHAT function is caused by the direct path. This also implies that the interesting information of the GCC-PHAT function, and also of the SRP-PHAT function, lies in the local maxima of the GCC-PHAT functions. Secondly, it is known that the global maximum of the cross correlation function has a certain shape [9, 11, 12, 10]. Based on the first introduced assumption, the shape assumption is also valid for some local maxima. In principle any shape can be used for the local maxima. In this article, two previously introduced hypotheses on the shape of global maximum are tested for the local maxima, namely the parabolic shape [11]:

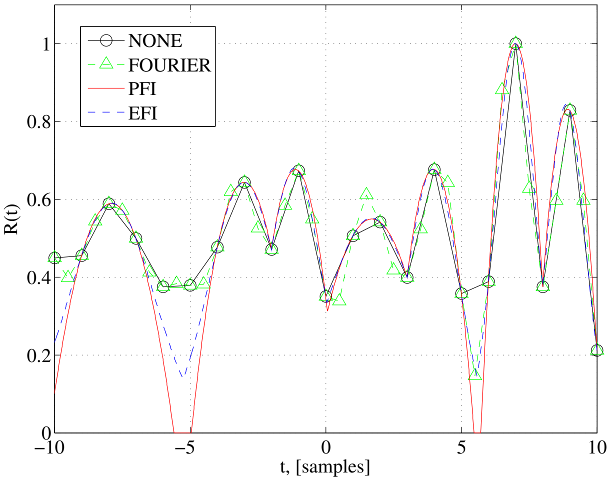

Fig. 2. Normalized discrete phase transform weighted cross correlation at 48 kHz (NONE) with Fourier-interpolation at 96 kHz (FOURIER), parabolic fitting (PFI), and exponential fitting (EFI).

$$f_\mathrm{l}(\tau) = a_l\tau^2 + b_\mathrm{l}\tau + c_\mathrm{l}, \tag{4a}$$

and the exponential shape [10]:

$$f_\mathrm{l}(\tau) = a_\mathrm{l}e^{-b_\mathrm{l}(\tau-c_\mathrm{l})^2}, \tag{4b}$$

where $a_\mathrm{l}, b_\mathrm{l}$, and c_l are the coefficients and f_l is the function for lth local maximum.

Keeping in mind the two introduced assumptions, the interpolation of a TDE function proceeds as follows. Firstly, the TDE function is normalized so that it is positive in the region of interest, which is limited by the distance between the used microphone pair. Secondly, the local maxima are searched from the TDE function in the region of interest. Thirdly, the coefficients in (4a) or in (4b) are solved using the local maximum and two neighboring points on both sides of the maximum. This leads to a function $f_\mathrm{l}(\tau)$ for each local maximum l. Finally, as a result, the interpolated TDE function can be evaluated at any time delay τ:

$$\mathrm{R}_{\mathrm{int}}(\tau) = \max_{\mathrm{l}} f_\mathrm{l}(\tau). \tag{5}$$

In Fig. 2, one example of interpolation with the proposed method is shown for both the parabolic fitting (PFI), the exponential fitting (EFI) and for Fourier-interpolation (FOURIER).

The proposed interpolation method is summarized in Algorithm 1. One can see that the interpolation method is suitable for other TDE functions than GCC-PHAT function as well, and that the shape assumption is not limited to the ones presented in this article. Also, if the number of the local maxima is reduced similarly as in [5] the method will be more efficient. In addition, an advantage of the proposed algorithm over e.g. the Fourier-interpolation is that the TDE function is always presented with a limited number of coefficients, when in the Fourier-interpolation the number of samples increases with the sampling frequency.

Algorithm 1 Proposed interpolation method

```
Estimate TDE function $\mathrm{R}(\tau)$ for example with (1)
Normalize the TDE function
Find the local maxima $\mathrm{l}$ of the TDE function
Solve the coefficients of (4a) or (4b) for each $\mathrm{l}$ and $f_{\mathrm{l}}$
Evaluate the interpolated function with (5)
```

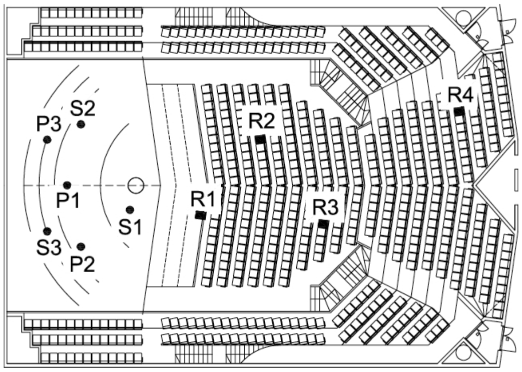

Fig. 3. Source (Sn) and receiver (Rn) positions in the concert hall.

3. CONCERT HALL EXPERIMENTS

The performance of the interpolation methods was examined with impulse responses measured in a 700-seat concert hall located in Pori, Finland. The wall constructions of the hall diffuse the sound field severely and the reverberation time at mid frequencies is 2.2 s. The floor plan of the hall and the source and receiver positions are seen in Fig. 3. The sound source was an omnidirectional loudspeaker of 26 cm diameter, consisting of 12 driver elements. It covers the frequency range of about 100 Hz to 10 kHz. The applied custom made microphone array, depicted in Fig. 4, consisted of 12 miniature electret microphones arranged in six pairs. Three pairs were set with a spacing of 10 mm between capsules and another three pairs with 100 mm spacing. The detailed description of the measurement equipment can be found in [13].

Impulse responses measured at the receiver and the source positions 1-3 (S1, S2 ,S3, R1, R2, and R3 in Fig. 3) were used. Although the sources were stationary, it was not used as an assumption in the experiments. The impulse responses were convolved with 5 different 1 s dry source signals, namely the contrabass, the violin, the flute, the french horn, and a female soprano singer, all recorded in an anechoic chamber. Then, the convolved signals were divided in to 50 % overlapping frames of length 512 samples at 48 kHz. This led to some 8400 directions estimated for each method in each test condition. Moreover, the six outer microphones of the TKK 3-D array were used, leading to 15 TDE functions and microphone pairs. The TDE functions were calculated as in (1) and (2), using the fast Fourier transform (FFT). Moreover, the TDE functions were interpolated with the EFI and PFI as in Section 2.1 and with Fourier-interpolation. In the Fourier-interpolation, the TDE functions were upsampled to 96 kHz with Matlab's `resample`-routine. From the 15 interpolated TDE functions the SRP-PHAT was formed as in (3) and by steering the azimuth and elevation from -60 to 60 degrees as in Fig. 1. Then, the direction that got the highest amount of evidence was the direction estimate for the current frame.

The performance of the methods was first experimented by varying the sampling frequency of the convolved signals from 8 kHz to 48 kHz (with Matlab's `resample`-routine), keeping the spatial resolution in 1 degrees. Secondly, the performance was tested by changing the spatial resolution from .5 to 5 degrees while the sampling frequency was fixed to 48 kHz.

Fig. 4. The six outer microphones of the TKK 3-D array were used.

3.1. Results and discussion

The results of the experiments are shown in Fig. 5. In the experiments the errors that were greater than 3 degrees were considered as anomalous estimates and omitted from the results. For all the experiments the results are presented as the mean squared error (MSE) of the non-anomalous estimates. The percentage of anomalous estimates is plotted above the MSE for each condition.

For the first experiment (Fig. 5, top and bottom left), the results against sampling frequency indicate that EFI and PFI interpolation methods have better performance than the original SRP-PHAT when the sampling frequency is higher than 24 kHz. Respectively, Fourier-interpolation has better performance than the original SRP-PHAT when the sampling frequency is higher than 40 kHz. In the second experiment (Fig. 5 top and bottom right) the performance of the original SRP-PHAT does not improve when the spatial resolution is smaller than 3 degrees. The limitations in the accuracy of the original SRP-PHAT after 3 degrees are due to sampling frequency and the locations of the microphones i.e. time resolution is too low.

The main observation from the experiments is that the exponential fitting produces the most accurate estimates in total. The slight superiority of EFI over PFI is addressed also in [10] for the direct cross correlation. In overall, the experiments show that the interpolation, not only provides more accurate estimates, but also decreases the number of anomalous estimates in total. This behavior may be explained by the fact that the spatial resolution of the original SRP-PHAT is, in many cases, so low that many direction candidates have the same amount of evidence, as can be seen in Fig. 1. In other words, multiple global maxima exists and direction estimation becomes ambiguous or biased. When interpolation is used, the direction estimation might be erroneous, but the ambiguousness is not present. The expected observation from the experiments is that if the interpolation is used the sampling frequency can be lowered, while the performance stays at the same level, and smaller spatial resolution can be used to achieve more accurate estimates.

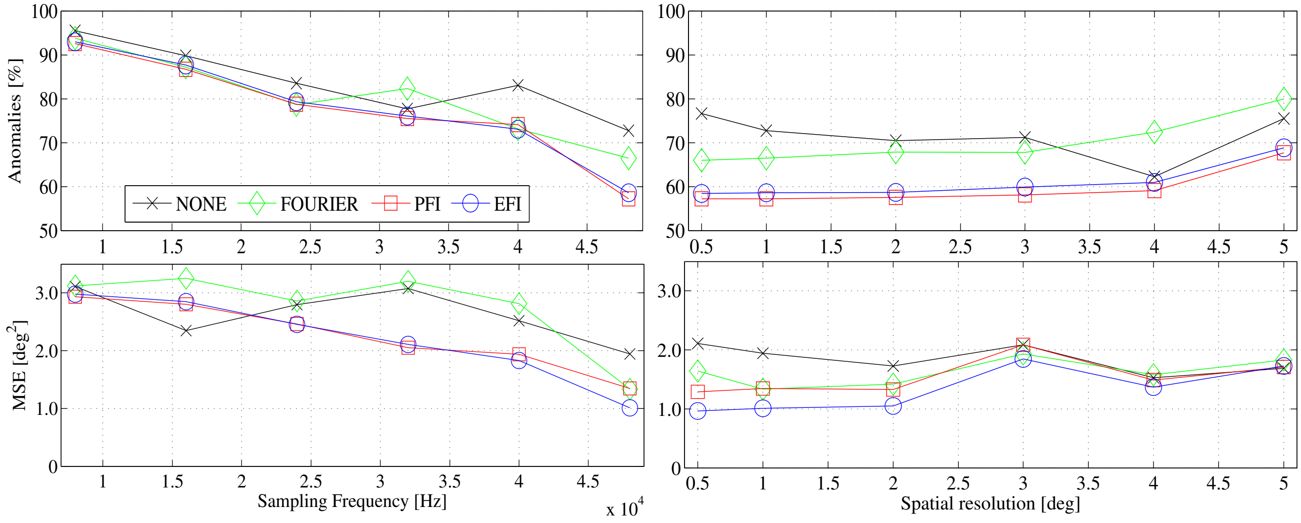

Fig. 5. MSE error and the percentage of anomalies against sampling frequency and spatial resolution. The spatial resolution was 1 degree when the sampling frequency was altered (left-side plots) and when the spatial resolution was varied the sampling frequency was set to 48 kHz (right-side plots).

4. CONCLUSION

The interpolation of a popular source localization algorithm, the steered response power-phase transform algorithm (SRP-PHAT), was considered. The SRP-PHAT was shown to suffer some limitations in time and spatial resolution. These limitations can be overcome by interpolating the SRP-PHAT. In the tested interpolations methods, the interpolation is applied in time-domain to the PHAT weighted cross correlation functions — the basic building blocks of the SRP-PHAT algorithm. The best interpolation method for the SRP-PHAT was found to be exponential fitting.

5. ACKNOWLEDGEMENTS

The work is partly performed while S. Tervo visited the Digital Signal Processing Group of Philips Research, Eindhoven, The Netherlands. S. Tervo thanks Dr. Steven van De Par and Dr. Aki Härmä for the supervision of the work at Philips Research as well as for the discussions and comments regarding this article. M.Sc. Pasi Pertilä from Tampere University of Technology is also thanked for the comments. In addition, the partial funding from Tekniikan edistämissäätiö (TES) and the Academy of Finland (#129092) is thanked for.

6. REFERENCES

[1] J.H. DiBiase, H.F. Silverman, and M.S. Brandstein, *Microphone arrays: signal processing techniques and applications*, chapter 8. Robust localization in reverberant rooms, pp. 157–180, Eds: Michael Brandstein and Darren Ward, Springer-Verlag, 2001.

[2] P. Aarabi, "The Fusion of Distributed Microphone Arrays for Sound Localization," *EURASIP J. Applied Signal Processing*, vol. 2003, no. 4, pp. 338–347, 2003.

[3] D.N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Trans. Speech and Audio Processing*, vol. 12, pp. 499–508, 2004.

[4] B. Mungamuru and P. Aarabi, "Enhanced Sound Localization," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 3, pp. 1526–1540, 2004.

[5] J.M. Peterson and C. Kyriakakis, "Hybrid algorithm for robust, real-time source localization in reverberant environments," in *Proc. ICASSP*, 2005, vol. 4, pp. iv/1053–iv/1056.

[6] A. Johansson, G. Cook, and S. Nordholm, "Acoustic direction of arrival estimation, a comparison between Root-MUSIC and SRP-PHAT," in *Proc. TENCON*, 2004, vol. B, pp. 629–632.

[7] H. Do and H.F. Silverman, "A Fast Microphone Array SRP-PHAT Source Location Implementation Using Coarse to Fine Region Contraction (CFRC)," in *Proc. WASPAA*, 2007, pp. 295–298.

[8] C.H. Knapp and G.C. Carter, "The generalized cross correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech, and Audio Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[9] X. Lai and H. Torp, "Interpolation methods for time delay using cross-correlation for blood velocity measurement," *IEEE Trans. Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 46, no. 2, pp. 277–290, 1999.

[10] L. Zhang and X. Wu, "On cross correlation based discrete time delay estimation," in *Proc. ICASSP*, 2005, vol. IV, pp. 981–984.

[11] G. Jacovitti and G. Scarano, "Discrete time techniques for time-delay estimation," *IEEE Trans. Signal Processing*, vol. 41, no. 2, pp. 525–535, 1993.

[12] J. Chen, Y.A. Huang, and J. Benesty, "Performance of GCC- and AMDF-based Time Delay Estimation in Practical Reverberant Environments," *EURASIP J. Applied Signal Processing*, vol. 1, pp. 25–36, 2005.

[13] T. Peltonen, T. Lokki, B. Gouatarbes, J. Merimaa, and M. Karjalainen, "A system for multi-channel and binaural room response measurements," in *the 110th Audio Engineering Society (AES) Convention*, Amsterdam, the Netherlands, May 12-15 2001, preprint no. 5289.