SPEAKER LOCALISATION USING THE FAR-FIELD SRP-PHAT IN CONFERENCE TELEPHONY

Anders Johansson[†], *Nedelko Grbić*[‡] *and Sven Nordholm*[†]

atri[†],
Curtin University of Technology
Wark Ave. Bentley, WA Australia
Phone:+61-8-9266-3268,
Email:ajh,sven@atri.curtin.edu.au

Dept. of Electrical & Electronic Engineering[‡],
University of Western Australia
35 Stirling Highway, Crawley, WA 6009
Phone:+61-8-9380-8017
Email:grbic@ee.uwa.edu.au

ABSTRACT

This paper describes a robust algorithm for sound source localization in conference rooms. The method used is a modified steered response power - phase alignment transform algorithm. The results are obtained by processing real data recorded in a typical conference room, and they are compared to data obtained from a simple free-field model. The algorithm demonstrates good accuracy for finding the correct angle of arrival for the dominant speaker in the room and works well for speech sources. The algorithm integrates well with subband decomposition and is suited for real-time applications.

1. INTRODUCTION

Conference telephony and video conferencing are growing areas of communication. It is important to maintain a good sound quality even if the system is operated in hands-free mode. In a hands-free environment, microphones are placed at a remote distance from the speakers causing problems of room reverberation and reduced signal to noise ratio. These problems can be solved by employing spatial selective filtering techniques, [1], [2], [3]. To design such filters a sound source localization algorithm is often required. It is important that this localization algorithm is robust to reverberation and poor signal to noise ratio.

The scenario considered in this study is a typical conference room accommodating 12 people. The persons speaking in the room are simulated with properly placed loudspeakers. The conference telephone with it's microphone array is placed in the center of the conference table. An illustration of the scenario is given in figure 1.

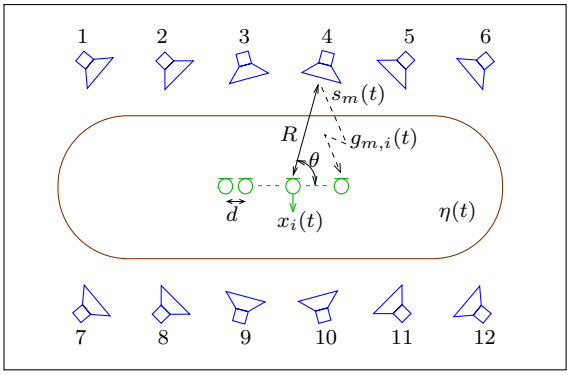

Figure 1: Conference room scenario.

The localization algorithm proposed in this study is a modified steered response power - phase alignment transform (SRP-PHAT) algorithm [4]. It has low computational complexity which makes it suitable for real-time implementation. The algorithm operates in frequency domain, where the transformation is performed using subband decomposition. This implementation strategy allows for a direct link to existing subbands beam-formers [6].

2. SIGNAL MODEL

The source signals originating from the different speakers, are denoted $s_m(t), m = 1, 2, \ldots, M$ and are assumed to be mutually uncorrelated. These signals impinge on an array of I microphone elements, each corrupted with noise $\eta_i(t)$. This noise includes electronic noise and background noise from air-conditioning, etc, and is considered to be uncorrelated with the speech signals. Further, it is considered to be spatially and temporally uncorrelated. The impulse response between speech source no. m and array element no. i is denoted $g_{m,i}(t)$ (see figure 1). The impulse response can be considered to be stationary over short time periods. The microphone signals, $x_i(t)$ are defined as

$$x_i(t) = s_m(t) * g_{m,i}(t) + \eta_i(t) \tag{1}$$

where $*$ denotes convolution. The cross-power density spectrum for the two time-signals $x_l(t)$ and $x_k(t)$ is defined as

$$\Gamma_{lk}(\omega) = \mathcal{F}\Big(E\big[x_l(t)x_k^*(t+\tau)\big]\Big) \tag{2}$$

where $\mathcal{F}(\cdot)$ denotes the Fourier transform, and $E[\cdot]$ is the expectation operator.

3. ALGORITHM

Given the cross-power density spectrum $\Gamma_{lk}(\omega)$, the algorithm estimates the angle of arrival θ of the active speaker.

The generalized cross correlation - phase alignment transform (GCC-PHAT) algorithm [5], is defined as

$$\hat{\tau}_{opt} = \arg\max_{\hat{\tau}_{lk}} C_{lk}(\hat{\tau}_{lk}) = \arg\max_{\hat{\tau}_{lk}} \left(\frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{\Gamma_{lk}(\omega)}{|\Gamma_{lk}(\omega)|} e^{j\omega\hat{\tau}_{lk}} d\omega \right) \tag{3}$$

where $\hat{\tau}_{lk}$ is the time delay-difference of arrival (TDOA) for the incoming sound for the microphone pair l and k. The generalization of the above is the SRP-PHAT algorithm

$$\hat{\mathbf{q}}_{opt} = \arg\max_{\hat{\mathbf{q}}} P(\hat{\mathbf{q}}) = \arg\max_{\hat{\mathbf{q}}} \left(\sum_{l=1}^{I} \sum_{k=1}^{I} \int_{-\infty}^{+\infty} \frac{\Gamma_{lk}(\omega)}{|\Gamma_{lk}(\omega)|} e^{j\omega\Delta_{lk}(\hat{\mathbf{q}})} d\omega \right) \tag{4}$$

where $\Delta_{lk}(\hat{\mathbf{q}})$ denotes time-delay difference between spatial location $\hat{\mathbf{q}}_{opt}$ of the dominant source and microphone pair l, k. This is done by optimizing the function with regards to all possible pairs of indexes l and k, hence the optimization spans over a $\binom{I}{2}$ dimensional space, requiring a large number of computations. This makes the algorithm less suitable for real-time applications when the number of microphone elements is large, but does however result in a robust location estimate.

By introducing a far-field assumption, we assume that the source signals will impinge on the array as plane waves. This implies that the TDOA $\hat{\tau}_{lk}$ for the two microphones l and k can be expressed as

$$\hat{\tau}_{lk} = (l-k)\hat{\tau} \tag{5}$$

where $\hat{\tau}$ is the TDOA for two adjacent microphones (see figure 2). This assumption is not restrictive in a conference room scenario, due to the fix physical locations of the conference participants.

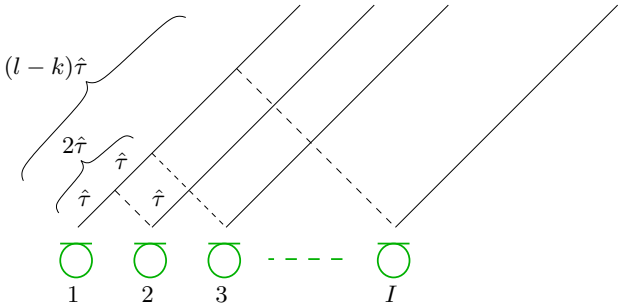

Figure 2: Plane wave arrival delay.

By using the far-field assumption a greatly simplified algorithm may be realized. By inserting (5) into (3) and by using the averaging from equation (4) we get

$$\hat{\tau}_{opt} = \arg\max_{\hat{\tau}} C(\hat{\tau}) = \arg\max_{\hat{\tau}} \left(2\pi \sum_{l=1}^{I} \sum_{k=1}^{I} \int_{-\infty}^{+\infty} \frac{\Gamma_{lk}(\omega)}{|\Gamma_{lk}(\omega)|} e^{j\omega\hat{\tau}(l-k)} d\omega \right). \tag{6}$$

This algorithm is denoted the far-field SRP-PHAT algorithm. The algorithm requires far fewer computations in comparison to the original SRP-PHAT, since it optimizes the function with regards

to only one parameter. The drawback is that it will only find the *angle of arrival* of the sound not the *position*. This is however adequate for our application.

In the subband approach the integration of (6) is approximated using a summation, which introduces a discretisation in frequency domain. This leads to approximation errors which becomes negligible as the number of subbands increases, as will be shown.

For a linear array the angle of arrival θ can easily be calculated from $\hat{\tau}_{opt}$ as

$$\theta = \arccos\Big(\frac{c}{d\,F_s}\hat{\tau}_{opt}\Big) \tag{7}$$

where c is the speed of sound propagation, d is the distance between the microphones and F_s is the sample frequency.

4. EVALUATION RESULTS

The far-field SRP-PHAT algorithm has been evaluated using data from three different scenarios:

1. Simple free-field model without background noise using white noise source.
2. Real room environment using white noise source.
3. Real room environment using speech source.

The real room environment data is recorded in a typical conference room using loudspeakers to simulate people speaking in the room (see figure 1). The recordings were made using a linear 7 microphone element array with an inter-element distance of 4 cm. The average signal to noise ratio (SNR) in the room was 20 dB at the time of recording, and a sample frequency of 8 kHz was used. The loudspeakers are numbered $m = 1, \ldots, 12$. An ideal speech activity detector has been assumed for the recorded speech signals.

The evaluations in section 4.1 and 4.3 are performed using 256 subbands.

4.1 Time evaluation

The algorithm has been evaluated for speaker position $m = 1 \ldots 6$. The angle of arrival estimates are presented in figure 3, where the speakers emits sound in a sequential order starting from speaker nr. 1. The small offset from the true position is mainly caused by real room reverberation.

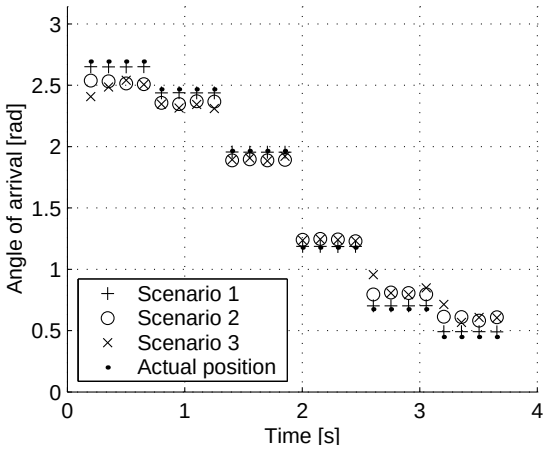

Figure 3: Angle estimates versus time for different positions and scenarios.

4.2 Accuracy

The algorithm is evaluated with regards to bias and standard deviation for different number of subbands using data obtained from the scenarios described in section 4. The results are presented in figure 4 and 5. Both plots shows a more rapid decrease for the free-field scenario. This effect comes mainly from the real room reverberation.

4.3 Background noise

The robustness of the algorithm has been evaluated with regards to the signal to noise ratio for the source signal. The bias and standard deviation for the signals obtained from the scenarios described in section 4 is presented in figure 6 and 7, respectively. The results show that the bias and standard deviation decreases as the SNR increases. The background noise used for the speech is multiple voices, often denoted "babble". The standard deviation at 0 dB for speech is 0.08 radians, this equals an error of less than 15 cm at the physical location of the speaker.

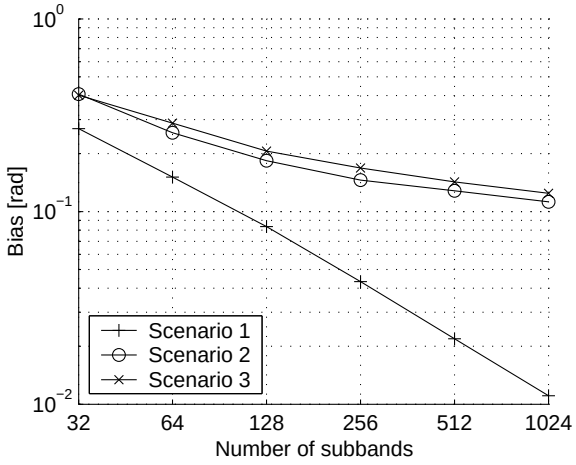

Figure 4: Bias versus the number of subbands for different scenarios.

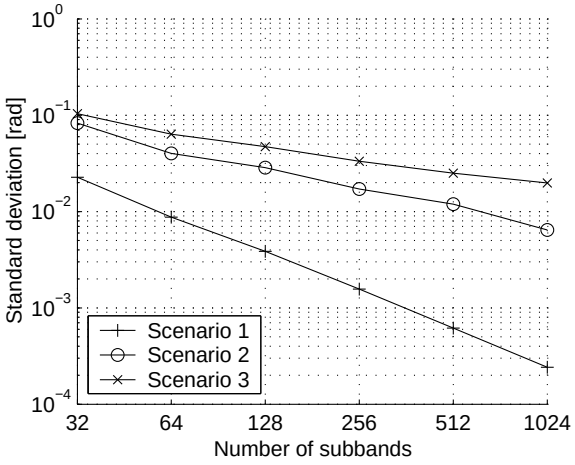

Figure 5: Standard deviation versus the number of subbands for different scenarios.

4.4 Near-field and discretisation effects

The algorithm has also been studied with regards to changes in the bias in relation to the number of subbands and the error caused by near-field effects. The results are presented in figure 8 and 9, respectively. The data used is the white noise from scenario 1. The y-axis in the figures shows the bias for the angle estimate. The first plot shows that the bias caused by the discretisation is reduced as the number of subbands increases. It also shows that the bias is non-linear with respect to the angle of arrival. The second plot shows that the bias caused by near-field effects is only noticeable for the extreme near-field, i.e. far from the actual positions at the conference table.

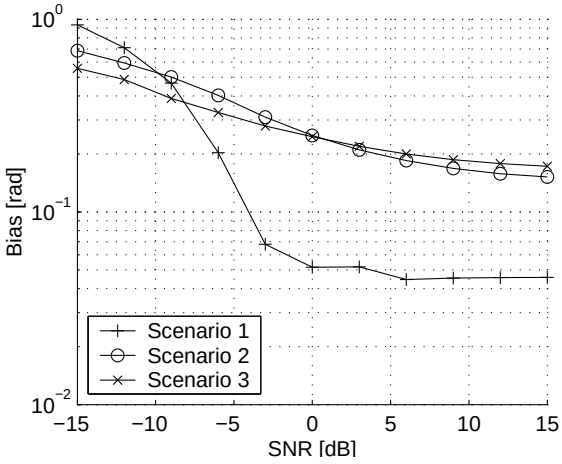

Figure 6: Bias versus SNR for different positions and scenarios.

5. CONCLUSIONS AND FUTURE WORK

The SRP-PHAT sound source localization algorithm has been modified to fit the conference room environment and has been evaluated using simulated and recorded data. The algorithm has proven to be very robust when used to localize speech in the proposed environment, even down to SNR's as low as 0 dB. The modifications have also led to a substantial reduction in the number of required computations, which makes the algorithm suitable for real-time applications.

Future work includes incorporating the localization into an adaptive beam-forming structure and evaluation in a real-time system.

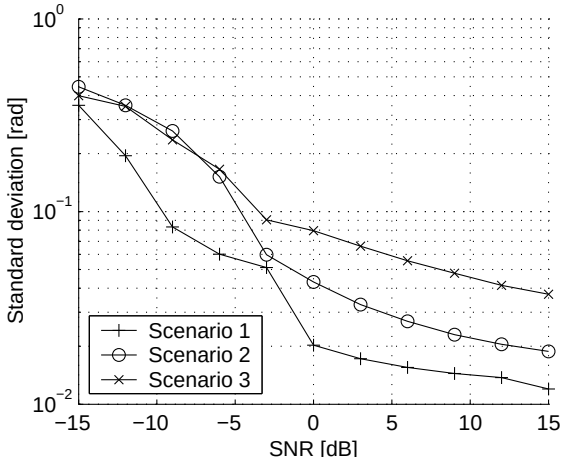

Figure 7: Standard deviation versus SNR for different positions and scenarios.

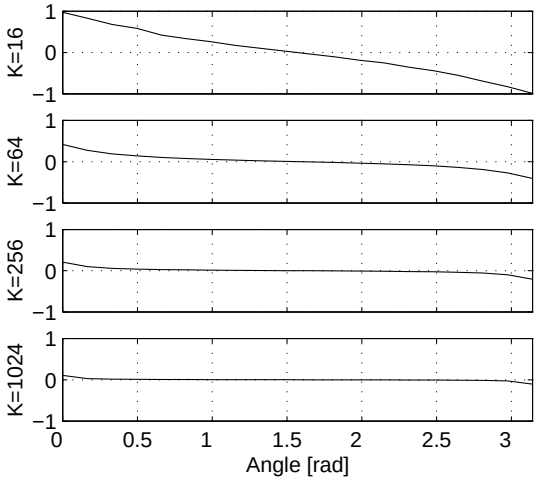

Figure 8: Bias versus angle of arrival for different number of subbands.

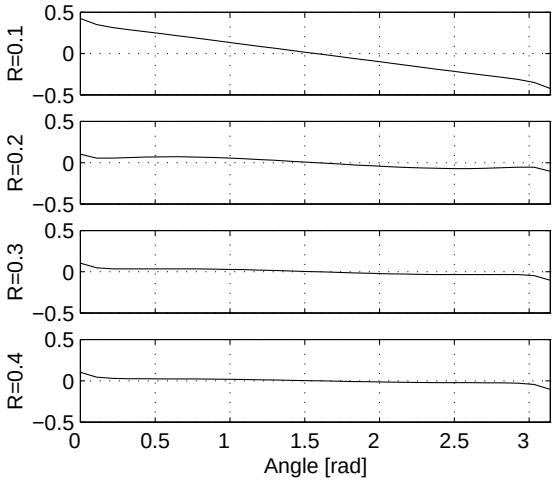

Figure 9: Bias versus angle of arrival for different radii to the source.

References

[1] H.F. Silverman, W.R. Patterson, J.L. Flanagan, D. Rabinkinn, "A digital processing system for source location and sound capture by large microphone arrays", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 251-254 vol.1, 1997.

[2] J. Flanagan, D. Berkeley, G. Elko, J. West, M. Sondhi, "Autodirective Microphone Arrays", Acustica, vol 73, pp. 58-71, 1991.

[3] S. Nordholm, I. Claesson, M. Dahl, "Adaptive Microphone Array Employing Calibration Signals. An Analytical Evaluation.", IEEE Transaction on Speech and Audio Processing, vol. 7, no. 3, pp. 241-252, May 1999.

[4] "Microphone arrays, Techniques and Applications", editors Michael S. Brandstein and Darren B. Ward, by Springer Verlag, Ch. 8, Jun. 2001.

[5] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay", IEEE Trans. Acoustic Speech Signal Processing, vol. ASSP-24,pp. 320-327, August 1976.

[6] N. Grbić and S. Nordholm, "Soft Constrained Subband Beamforming for Hands-Free Speech Enhancement", accepted for presentation at 2002 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Orlando, Florida, May 2002.