ROBUST ACOUSTIC DIRECTION OF ARRIVAL ESTIMATION USING ROOT-SRP-PHAT, A REALTIME IMPLEMENTATION

Anders Johansson and Sven Nordholm

Western Australian Telecommunications Research Institute[†]
39 Fairway, Nedlands, WA 6907 Australia
Email: ajh,sven@watri.org.au

ABSTRACT

Wideband robust direction of arrival estimation using microphone arrays is associated with computationally complex algorithms, unsuitable for implementation on low power devices. This paper improves on the computational complexity by devising an algorithm based on steered response interpolation combined with steered response power estimation employing root solving. The new algorithm is implemented in a PC-based realtime system and has been compared to Root-MUSIC and Far-Field SRP-PHAT. The comparison was performed with regards to computational load and robustness to noise and reverberation in a real room environment. The results show that the new algorithm is the most computationally efficient, while providing superior robustness compared to Root-MUSIC.

1. INTRODUCTION

Finding a robust estimate of the direction of arrival (DOA) to a sound source using a microphone array is a critical feature in a wide array of applications ranging from security applications such as battlefield sensor networks and automatic surveillance systems [1] to office and home applications such as conference telephones and home automation systems [2]. Furthermore the employed DOA estimation algorithm must be computationally efficient in order to reduce power consumption.

The DOA for a sound source is calculated from the time delay of arrival (TDOA) of a sound wavefront across a given microphone pair. Early methods to calculate the TDOA were based on cross-correlation estimation in the time [3] or frequency [4] domain from a single pair of sensors. The time-domain approach was abandoned due to poor resolution, but the frequency domain approach was extended and applied to microphone pairs by Rabinkin [5], and later applied to microphone arrays by Silverman and Brandstein [6]. Their work has underpinned the development of the algorithm featured here.

In parallel, model based narrowband DOA estimation methods such as Root-MUSIC [7] were developed and applied to sonar and radar signals. Advances in coherent wideband processing [8] and shaped response interpolation (SRI) [9] has made it possible to apply narrowband algorithms to wideband data.

The algorithm featured here combines SRI with steered response power (SRP) estimation and root solving to estimate the DOA. The algorithm is denoted Root-SRP-PHAT, and has been implemented and evaluated in realtime on a standard personal computer (PC). The evaluation is performed with regards to robustness to noise and reverberation and to computational load in both a simulated and a real room environment. The results are compared with those of Root-MUSIC [7] and Far-Field SRP-PHAT [10].

[†]WATRI is a joint institute between The University of Western Australia and Curtin University of Technology. The work has been sponsored by ARC under grant no. DP0451111.

2. SIGNAL MODEL

Consider L sound sources, placed at the positions $\mathbf{q}_\ell = [\rho_\ell, \theta_\ell]$; $\ell = 1, 2, \ldots, L$, where ρ and θ denotes radius and DOA to the source respectively. Sound originating in a source, denoted $s_\ell(t)$, impinges on a uniform linear array of I microphone elements, with an inter-element spacing d. Each of the elements are corrupted with noise $\eta_i(t)$; $i = 1, 2, \ldots, I$, which is considered to be spatially and temporally uncorrelated to the sound sources. The impulse response between a sound source at $\mathbf{q}_\ell$ and array element no. i is denoted $h_i(t, \mathbf{q}_\ell)$, and is considered to be linear and time invariant over the measurement period. The microphone signals, $x_i(t)$ are defined as

$$x_i(t) = \sum_{\ell=1}^{L} s_\ell(t) * h_i(t, \mathbf{q}_\ell) + \eta_i(t), \tag{1}$$

for $i = 1, 2, \ldots, I$. In practice these signals are band limited and sampled, and are denoted $x_i[m] = x_i(\frac{m}{F_s})$, where F_s is the sample-frequency. The array output vector is defined by stacking the microphone signals in a vector according to

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) & x_2(t) & \cdots & x_I(t) \end{bmatrix}^T. \tag{2}$$

The general assumption is that the impulse response $h_i(t, \mathbf{q}_\ell)$ will consist of a series of delayed impulses, decreasing in amplitude over time. The algorithms presented in this paper operate under the assumption that the dominant component of $h_i(t, \mathbf{q}_\ell)$ is the direct path between the sound source and the microphone array, and that the sound sources are in the far field of the array. In an ideal, free-space scenario $h_i(t, \mathbf{q}_\ell)$ will contain *only* a direct path, and the sound will impinge on the array as a planar wave. In this case, the array response vector can be approximated by

$$\mathbf{a}(\Omega, \theta) \simeq \frac{e^{j\Omega\Delta}}{4\pi\rho} \begin{bmatrix} 1 & e^{-j\Omega\tau} & \cdots & e^{-j(I-1)\Omega\tau} \end{bmatrix}^T \tag{3}$$

where $\Omega = 2\pi F$ and TDOA $\tau = \frac{d\cos\theta}{c}$ in which F and c denotes frequency and speed of sound respectively. The term Δ is bulk delay and the attenuation $\frac{1}{4\pi\rho}$ is caused by dispersion.

2.1. Array Covariance

The algorithms presented in this paper estimate the DOA from the cross power spectral densities (PSDs) between all microphone pairs. For the entire array, this information is contained in the spatial covariance matrix $\mathbf{R}_{\mathbf{x}}(\Omega)$. Assuming source and noise processes to be uncorrelated, the spatial covariance matrix is given by

$$\begin{aligned}\mathbf{R}_{\mathbf{x}}(\Omega) &= \mathcal{F}\left\{\mathrm{E}\left[\mathbf{x}(t)\mathbf{x}^T(t)\right]\right\} \\ &= \mathbf{A}(\Omega)\mathbf{R}_s(\Omega)\mathbf{A}^H(\Omega) + \mathbf{R}_\eta(\Omega)\end{aligned} \tag{4}$$

where $\mathcal{F}\{\cdot\}$ denotes the Fourier transform and $\mathrm{E}[\cdot]$ denotes the expected value. The matrix $\mathbf{R}_s(\Omega)$ is an $L \times L$ matrix of source cross-PSDs, $\mathbf{R}_\eta(\Omega)$ is an $I \times I$ matrix of noise cross-PSDs and $\mathbf{A}(\Omega)$ is the transfer function matrix, defined for L signals as

$$\mathbf{A}(\Omega) = \begin{bmatrix}\mathbf{a}(\Omega,\theta_1) & \mathbf{a}(\Omega,\theta_2) & \cdots & \mathbf{a}(\Omega,\theta_L)\end{bmatrix}. \tag{5}$$

In the absence of noise, the cross PSD between two microphones indexed a and b for a single source ℓ with PSD $S_\ell(\Omega)$ is given by

$$\mathcal{S}_{a,b}(\Omega,\theta_\ell) = S_\ell(\Omega)\frac{e^{-j(a-b)\Omega\tau_\ell}}{(4\pi\rho)^2}. \tag{6}$$

Thus the parameter of interest τ is contained in the normalised relative phase response between the two microphones

$$\psi_{a,b}(\Omega,\tau) = e^{-j(a-b)\Omega\tau}. \tag{7}$$

3. LOCALISATION ALGORITHMS

The microphone signals $\mathbf{x}[m]$ are transformed into frequency domain signals using a DFT filterbank. The output signals from the transformation are the the discrete array output vectors $\mathbf{X}[k,n]$, where $k = 1, 2, \ldots, K$ is the subband index, and n is the subband time index. These signals are used to formulate an estimate of the spatial covariance matrix according to

$$\hat{\mathbf{R}}_{\mathbf{x}}[k] = \sum_{n=1}^{N}\mathbf{X}[k,n]\left\{\mathbf{X}[k,n]\right\}^H. \tag{8}$$

An estimate of the relative phase response $\hat{\psi}_{a,b}(\omega_k)$, for the frequency band ω_k and microphones a and b, is obtained from $\hat{\mathbf{R}}_{\mathbf{x}}[k]$ as

$$\hat{\psi}_{a,b}(\omega_k) = \frac{\left[\hat{\mathbf{R}}_{\mathbf{x}}[k]\right]_{a,b}}{\left|\left[\hat{\mathbf{R}}_{\mathbf{x}}[k]\right]_{a,b}\right|}. \tag{9}$$

This operation is referred to as the phase transform (PHAT).

3.1. Steered Response Power Estimation

The normalised steered response power for the TDOA τ and the frequency band ω_k is calculated from the estimated relative phase response according to

$$P_x(\tau,\omega_k) = \sum_{a=1}^{I}\sum_{b=1}^{I}\hat{\psi}_{a,b}(\omega_k)e^{j\omega_k(a-b)\tau}. \tag{10}$$

The peak value with regards to τ of the above function will correspond to the TDOA of the strongest sound source. A robust TDOA estimate $\hat{\tau}$ can be obtained by estimating the global maxima of $P_x(\tau,\omega_k)$ for all frequency bands $k = 1, 2, \ldots, K$ according to

$$\hat{\tau} = \arg\max_{\tau}\sum_{k=1}^{K}P_x(\tau,\omega_k). \tag{11}$$

This algorithm is referred to as Far-Field SRP-PHAT [10]. It is robust but computationally complex, as the results show.

This paper proposes a new approach for calculating the TDOA by only calculating maxima and minima of $P_x(\tau,\omega_k)$ and then comparing their power levels. This can be done by first summing Eq. 10 over the diagonals rather than rows and columns and rearranging it according to

$$\begin{aligned}P_x(\tau,\omega_k) &= \sum_{n=-(I-1)}^{I-1}e^{j\omega_k n\tau}\sum_{m=0}^{I-|n|}\hat{\psi}_{f(n,m),g(n,m)}(\omega_k) \\ &= \sum_{n=-(I-1)}^{I-1}c_n e^{j\omega_k n\tau},\end{aligned} \tag{12}$$

where $f(n,m) = |\min(0,n)| + m$ and $g(n,m) = |\max(0,n)| + m$. Secondly, the maxima and minima are found from the zeros of the derivative of $P_x(\tau,\omega_k)$ from

$$\frac{dP_x(\tau,\omega_k)}{d\tau} = \sum_{n=-(N-1)}^{N-1}j\omega_k n c_n e^{j\omega_k n\tau} = 0. \tag{13}$$

The TDOA $\hat{\tau}$ can be estimated by forming a polynomial from the terms of the above sum and solving for its roots. These roots will identify the maxima and minima of $P_x(\tau,\omega_k)$, where the maximum with the highest power will be an estimate of the TDOA. This algorithm is referred to as Root-SRP-PHAT and is a narrowband algorithm. Here, steered response interpolation is therefore used to transform the problem into a narrowband one.

3.2. Steered Response Interpolation

If each subband is sufficiently narrowband, $\hat{\mathbf{R}}_{\mathbf{x}}[k]$ will approximate Eq. 4,

$$\hat{\mathbf{R}}_{\mathbf{x}}[k] \simeq \mathbf{A}(\omega_k)\mathbf{R}_s(\omega_k)\left\{\mathbf{A}(\omega_k)\right\}^H + \mathbf{R}_\eta(\omega_k) \tag{14}$$

where $\mathbf{A}(\omega_k)$,$\mathbf{R}_s(\omega_k)$ and $\mathbf{R}_\eta(\omega_k)$ are the sampled values of the continuous frequency variables in Eq. 4. The approach used here applies array interpolation at each subband to give

$$\mathbf{g}(\theta) \simeq \breve{\mathbf{T}}(\omega_k)\mathbf{a}(\omega_k,\theta)\,;\ k = 1, 2, \ldots, K \tag{15}$$

where $\mathbf{a}(\omega_k,\theta)$ is the sampled array response vector and $\mathbf{g}(\theta)$ is the response vector of a *virtual array* that is independent of ω_k. To allow the application of Root-SRP-PHAT and Root-MUSIC, the virtual array is by necessity uniform linear.

The SRI [9] optimum interpolation matrix $\breve{\mathbf{T}}$ is designed using the least-squares problem formulation

$$\breve{\mathbf{T}}(\omega_k) = \arg\min_{\mathbf{T}(\omega_k)}\int_{-\pi}^{\pi}\left\|\mathbf{T}(\omega_k)\mathbf{a}(\omega_k,\theta) - \mathbf{g}(\theta)\right\|^2 d\theta. \tag{16}$$

Applying interpolation to the spatial covariance matrices, and summing over all frequency bands yields

$$\begin{aligned}\mathbf{U}_{\mathbf{x}} &= \sum_{k=1}^{K} \breve{\mathbf{T}}(\omega_k)\hat{\mathbf{R}}_{\mathbf{x}}[k]\left[\breve{\mathbf{T}}(\omega_k)\right]^H \\ &= \sum_{k=1}^{K} \breve{\mathbf{T}}(\omega_k)\mathbf{A}(\omega_k)\mathbf{R}_s(\omega_k)\left\{\breve{\mathbf{T}}(\omega_k)\mathbf{A}(\omega_k)\right\}^H \\ &\quad + \breve{\mathbf{T}}(\omega_k)\mathbf{R}_\eta^{[k]}\left\{\breve{\mathbf{T}}(\omega_k)\right\}^H \\ &\simeq \mathbf{G}\mathbf{R}_s\mathbf{G}^H + \mathbf{N},\end{aligned} \tag{17}$$

where $\mathbf{G} = \begin{bmatrix}\mathbf{g}(\theta_1) & \mathbf{g}(\theta_2) & \cdots & \mathbf{g}(\theta_L)\end{bmatrix}$, $\mathbf{R}_s$ is the combined signal covariance matrix and $\mathbf{N}$ is the combined noise covariance matrix. Thus $\mathbf{U}_{\mathbf{x}}$ contains the DOA information for all subbands in a single matrix. Both Root-MUSIC and Root-SRP-PHAT operates on $\mathbf{U}_{\mathbf{x}}$.

3.3. Root-MUSIC

Let $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_P\}$ denote the eigenvectors of $\mathbf{U}_{\mathbf{x}}$, ordered with respect to their corresponding eigenvalue magnitude. The eigenvectors $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_L\}$ span the *signal subspace* and the remaining eigenvectors $\{\mathbf{e}_{L+1}, \mathbf{e}_{L+2}, \ldots, \mathbf{e}_P\}$ span the *noise subspace*.

Assuming the signals are not highly correlated and noise prewhitening is applied to $\mathbf{U}_{\mathbf{x}}$, the L response vectors in $\mathbf{G}$ will be orthogonal to the noise subspace $\mathbf{N}$. Thus if $\mathbf{P}$ is a projection matrix onto the noise subspace of $\mathbf{U}_{\mathbf{x}}$,

$$\|\mathbf{P}\mathbf{g}(\theta_\ell)\|^2 = 0\,;\ \ell = 1, 2, \ldots, L \tag{18}$$

The zeros of Eq. 18, are found by solving for the roots of a polynomial of order $2(P-1)$. The DOA with the highest steered response power will identify the strongest sound source.

4. IMPLEMENTATION

The three localisation algorithms described above are implemented in software executed on a standard PC. The microphone array is connected to the PC using a multi-channel analog input/output (I/O) card.

The microphone array consists of eight elements which are mounted on a metal fixture with an inter-element spacing of 40mm. The microphone outputs are connected to a preamplifier which in turn is connected to the I/O card. The microphone elements, model 2541/PRM902, and the preamplifier, model 2210, are from Larson Davis. The I/O card has 24-bit analog to digital converters with built in anti-aliasing filters and is operated at a sample frequency of 8kHz. The card is an M-Audio Delta-1010LT.

All algorithms operate over a frequency range of 800Hz to 3200Hz. The frequency range and sample frequency are chosen to give the algorithms optimum performance for speech for the given background noise and room reverberation. The optimisation algorithm used for Far-Field SRP-PHAT is an iterative one-dimensional search. The array interpolation projects onto a virtual array with the same array geometry as the physical array and with a centre frequency of 2500Hz for Root-MUSIC, and 800Hz for Root-SRP-PHAT. The spatial covariance matrix is estimated using exponential averaging with forgetting factor $\alpha = 0.1$.

5. PERFORMANCE EVALUATION

In the following evaluation, the estimation error is presented as RMSE in radians, calculated according to

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}(\hat{\theta}[n] - \theta)^2}, \tag{19}$$

where the values of $\hat{\theta}[n]$ are the estimated DOAs produced by the algorithm and θ is the actual DOA. The evaluations are performed using $N > 1000$ estimates.

5.1. Robustness to Reverberation and Noise

The algorithms are evaluated in terms of robustness to reverberation and background noise power. The evaluation is performed in a real room and using synthesised input. In each environment the stimulus are female speech from a position $\mathbf{q} = [1.5\mathrm{m}, 2.0\mathrm{rad}]$. The signal to noise ratio (SNR) is calculated by averaging the power at all microphone elements over the frequency range 300 to 3400Hz.

The synthesised input is generated using a free space model, with spatially white noise added to the sensor elements. The noise is Gaussian with the same spectral contents as the background noise in the real room.

The dimensions of the real room are 3.1m×3.5m×2.2m and the walls are partly covered with acoustic foam panels to reduce reflections coming from behind the microphone array. A semi-diffuse noise field was created in the room by playing white noise through two loudspeakers placed behind baffles in the corners of the room, facing away from the array.

The RMSE versus SNR for the free space model is shown in Fig. 1. It shows that Far-Field SRP-PHAT reaches a minimum error. This error is a bias error and is caused by the frequency sampling. It is further investigated in [10]. The RMSE versus SNR for the real room environment is shown in Fig. 2. It shows that the SRP-based algorithms are more robust than Root-MUSIC to noise and reverberation for poor SNRs. At the given distance to the sound source the Root-SRP-PHAT algorithm has an error of 150mm at 10dB SNR. The minimum RMSE is limited here by the diameter of the loudspeaker.

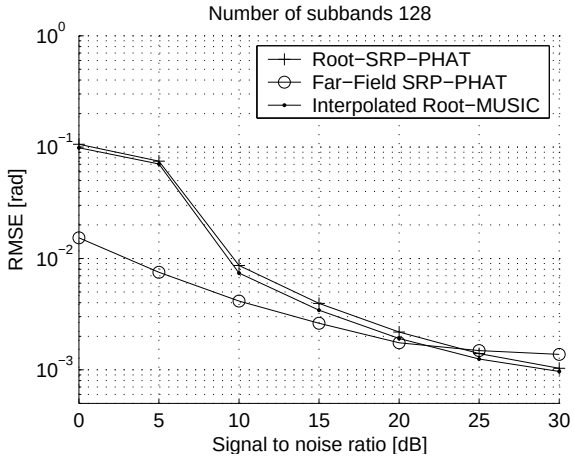

Fig. 1. RMSE versus SNR for female speech in free space model.

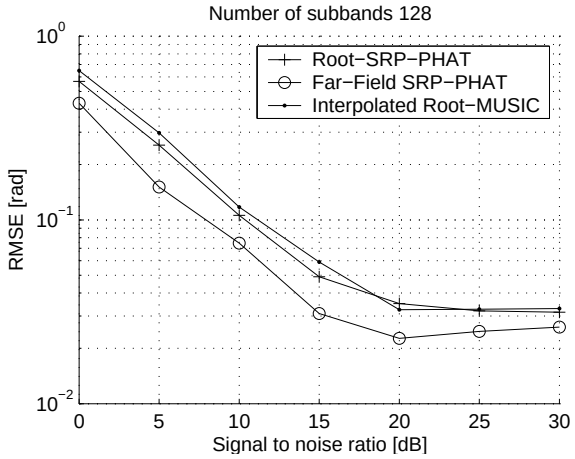

Fig. 2. RMSE versus SNR for female speech in real room environment.

5.2. Computational Complexity Evaluation

The computational load is evaluated by measuring the average number of clock-cycles taken for an algorithm to complete one position estimate. The measurement has been made for different numbers of subbands, and is normalised with respect to the most computationally efficient algorithm. From Fig. 3 it can be seen that Root-SRP-PHAT has the lowest computational load followed by Root-MUSIC, which has doubble the computational load compared to Root-SRP-PHAT. Far-Field SRP-PHAT has a computational load between 3-11 times that of Root-SRP-PHAT using the given implementation.

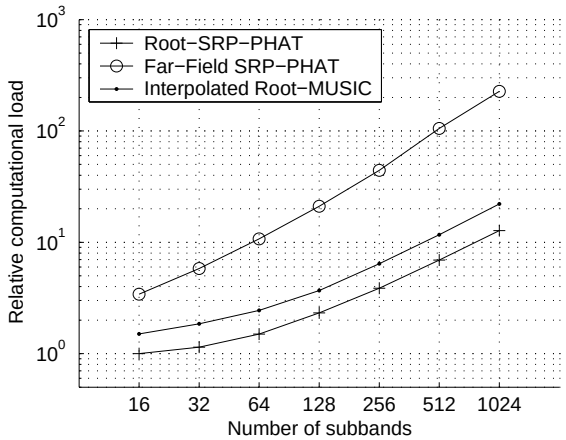

Fig. 3. Normalised computational load versus number of frequency bands.

6. CONCLUSION

A new localisation algorithm, Root-SRP-PHAT, was compared to two common existing algorithms in a realtime implementation. The algorithms were evaluated with regards to computational load and robustness to background noise and reverberation in a real and a simulated environment.

The results show that Root-MUSIC and Root-SRP-PHAT has nearly identical performance with respect to robustness, with Root-SRP-PHAT showing slightly higher robustness to reverberation and low SNRs in a real room environment. Far-Field SRP-PHAT is however capable of sustaining nearly the same RMSE at 5dB lower SNR than the other two algorithms in a real environment. However, the new algorithm outperforms Root-MUSIC by a factor of two, and Far-Field SRP-PHAT by a factor of 3 or more with regards to computational complexity.

7. REFERENCES

[1] J.D. de Jesus, J.J.V. Calvo, and A.I. Fuente, "Surveillance System Based on Data Fusion From Image and Acoustic Array Sensors," *IEEE Aerospace and Electronic Systems Magazine*, vol. 15, no. 2, pp. 9–16, Feb. 2000.

[2] S.G. Goodridge and M.G. Kay, "Multimedia Sensor Fusion for Intelligent Camera Control," in *EEE/SICE/RSJ International Conference on Multisensor Fusion and Integration for Intelligent Systems*, NY, USA, Dec. 1996, pp. 655–662.

[3] F. A. Reed, P. L. Feintuch, and N.J. Bershad, "Time Delay Estimation Using the LMS Adaptive Filter-Static Behavior," *IEEE Transactions Acoustics, Speech and Signal Processing*, June 1981.

[4] C.H. Knapp and G.C. 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, Aug. 1976.

[5] D. Rabinkin, R. Renomeron, J. French, and J. Flanagan, "Estimation of Wavefront Arrival Delay Using the Cross-Power Spectrum Phase Techniqu," in *132nd Meeting of the Acoustical Society of America*, Honolulu, USA, Dec. 1996, vol. 100, p. 2697.

[6] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust Localization in Reverberant Rooms," in *Microphone Arrays, Techniques and Applications*, Michael S. Brandstein and Darren B. Ward, Eds., number ISBN 3-540-41953-5, chapter 8, pp. 157–178. Springer Verlag, June 2001.

[7] A. J. Barabell, "Improving the Resolution Performance of Eigenstructure-Based Direction-Finding Algorithms," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Boston, MA, Apr. 1983, pp. 336–339.

[8] M. Doron, E. Doron, and A. Weiss, "Coherent Wide-Band Processing for Arbitrary Array Geometry," *IEEE Transactions on Signal Processing*, vol. 41, no. 1, pp. 414–417, Jan. 1993.

[9] G. J. Cook, B. K. Lau, and Y. H. Leung, "An Alternative Approach to Interpolated Array Processing for Uniform Circular Arrays," in *IEEE Asia-Pacific Conference on Circuits and Systems, Bali, Indonesia (relocated to Singapore)*, Feb. 2002, vol. 1, pp. 411–414.

[10] A. Johansson, N. Grbic, and S. Nordholm, "Speaker Localisation Using the Far-Field SRP-PHAT in Conference Telephony," in *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, Kaohsiung, Taiwan, Nov. 2002.