ACOUSTIC DIRECTION OF ARRIVAL ESTIMATION, A COMPARISON BETWEEN ROOT-MUSIC AND SRP-PHAT

Anders Johansson, Greg Cook and Sven Nordholm

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

ABSTRACT

Conference and video telephony, home automation and surveillance appliances use microphone arrays for localisation of sound sources. It is therefore of interest to study different algorithms with respect to implementation and performance in real environments. Three algorithms have been employed in this study: Coherent Wideband Root-MUSIC and near-field and far-field versions of an algorithm based on Steered Response Power - Phase Transform. Results show that Root-MUSIC is the most computationally efficient. However due to simplistic model assumptions it fails in highly reverberant environments. The two algorithms employing Steered Response Power - Phase Transform maintain good performance measures even under reverberant conditions and low signal to noise ratios.

1. INTRODUCTION

Localisation of sound sources is an important feature in several growing areas of technology, such as Conference and video telephony, home automation and surveillance. As such it is used in conference and video telephony to find the active speaker.This information is then used to direct beamformers and steer cameras. Localisation of sound sources is also required in automatic surveillance equipment in order to track the movement of people and vehicles, and furthermore to steer surveillance cameras.

The direction of arrival (DOA) for a sound source is calculated from the time delay of arrival (TDOA) of a sound wavefront across a given microphone pair, using simple trigonometry. Early methods to calculate the TDOA were based on cross-correlation estimation in the time [1] or frequency [2] domains 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 [3], and later applied to microphone arrays by Silverman and Brandstein [4]. Their work has underpinned the development of the two SRP based algorithms considered here.

In parallel, model based narrow-band DOA estimation methods such as Root-MUSIC [5] were developed and applied to sonar and radar signals. Advances in coherent wideband processing [6] and shaped response interpolation (SRI) [7] has made it possible to apply narrow-band algorithms to wideband data. These advances are used as a means to apply Root-MUSIC to microphone array DOA estimation.

In this paper three DOA estimation algorithms have been implemented and evaluated on a standard personal computer (PC) equipped with a multi channel sound card. This means that the evaluation has been done under realistic conditions. The implementation use DFT filterbanks to realise the frequency transformation. The investigated algorithms are:

1. Root-MUSIC using coherent wideband processing via SRI. Root-MUSIC is a far-field, narrow-band DOA estimation algorithm which is computationally very efficient.

2. Steered Response Power - Phase Transform (SRP-PHAT), finds the maximum likelihood estimate of the source position with respect to radius and DOA using a simple free-space model.

3. Far-field SRP-PHAT [8], is a simplification of SRP-PHAT. It operates on the same basic principle as SRP-PHAT but assumes a far-field source model and will thus only estimate the DOA of a sound source.

The evaluation is performed using simulated data from a free-space model and real data from a real room. The accuracy of the algorithms is measured using the root mean square error (RMSE) between DOA estimates and true DOAs.

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 angle respectively. Sound originating in a source, denoted $s_\ell(t)$, impinges on an uniform linear array (ULA) of I microphone elements, placed at the positions $\mathbf{p}_i = [\rho_i, \theta_i];\ i = 1, 2, \ldots, I$, with an inter-element spacing d. Each of the elements are corrupted with noise $\eta_i(t)$, which is considered to be spatially and temporally uncorrelated with 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 stationary over short time periods. The

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

microphone signals, $x_i(t)$ are defined as

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

for $i = 1, 2, \ldots, I$. In practise the signals are band limited and sampled, and are denoted $x_i[m] = x_i(mT)$.

The general assumption is that $h_i(t, \mathbf{q}_\ell)$ 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. In an ideal, free-space scenario $h_i(t, \mathbf{q}_\ell)$ will contain *only* a direct path. In this case, the frequency transfer function is defined for a source at any point $\mathbf{q}$ in space as

$$H_i(f, \mathbf{q}) = \frac{e^{j2\pi f \tau_i(\mathbf{q})}}{4\pi c \tau_i(\mathbf{q})} \quad (2)$$

where c is the speed of sound, and $\tau_i(\mathbf{q})$ is the propagation delay in seconds between point $\mathbf{q}$ and microphone element $\mathbf{p}_i$. The propagation delay is defined according to

$$\tau_i(\mathbf{q}) = \frac{||\mathbf{p}_i - \mathbf{q}||}{c}. \quad (3)$$

Using Eq. 3 we can define the *time difference of arrival* between the two sensors a and b as

$$\tau_{a,b}(\mathbf{q}) = \tau_a(\mathbf{q}) - \tau_b(\mathbf{q}) \quad (4)$$

In the far-field of an array apart from a common bulk delay and attenuation, Eq. 4 becomes

$$\tau_{a,b}(\hat{\mathbf{q}}) = \frac{1}{c}(\mathbf{p}_a - \mathbf{p}_b)^T \hat{\mathbf{q}} = \frac{d\sin(\theta)}{c}(a - b), \quad (5)$$

where $\hat{\mathbf{q}}$ is a unit vector dependent only on θ, and Cartesian coordinates are used.

2.1. Array Covariance

The main parameter of interest in DOA estimation is the cross power spectral density (PSD) between the output signals of any two microphone pairs. For the entire array, this information is contained in the spatial covariance matrix $\mathbf{R}_x$.

Let $\mathbf{h}(f, \mathbf{q})$ denote the *array response vector* describing the transfer function from a point $\mathbf{q}$ to every microphone in the array

$$\mathbf{h}(f, \mathbf{q}) = \begin{bmatrix} H_1(f, \mathbf{q}) & H_2(f, \mathbf{q}) & \cdots & H_I(f, \mathbf{q}) \end{bmatrix}^T \quad (6)$$

and define the array output vector,

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) & x_2(t) & \cdots & x_I(t) \end{bmatrix}^T \quad (7)$$

Since the source and noise processes are uncorrelated, the frequency dependant spatial covariance matrix is given by

$$\begin{aligned} \mathbf{R}_x(f) &= \mathcal{F}\left\{\mathrm{E}\left[\mathbf{x}(t)\mathbf{x}^T(t)\right]\right\} \\ &= \mathbf{H}(f)\mathbf{R}_s(f)\mathbf{H}^H(f) + \mathbf{R}_\eta(f) \end{aligned} \quad (8)$$

where $\mathbf{R}_s(f)$ is an $L \times L$ matrix of source cross-PSDs, $\mathbf{R}_\eta(f)$ is an $I \times I$ matrix of noise cross-PSDs and $\mathbf{H}(f)$ is the transfer function matrix, defined for L signals as

$$\mathbf{H}(f) = \begin{bmatrix} \mathbf{h}(f, \mathbf{q}_1) & \mathbf{h}(f, \mathbf{q}_2) & \cdots & \mathbf{h}(f, \mathbf{q}_L) \end{bmatrix}. \quad (9)$$

In the absence of noise, the cross PSD between microphones a and b for a single, near-field source at $\mathbf{q}_\ell$ with PSD $S_\ell(f)$ is given by

$$\begin{aligned} \mathcal{S}_{a,b}(f, \mathbf{q}_\ell) &= H_a(f, \mathbf{q}_\ell) H_b^*(f, \mathbf{q}_\ell) S_\ell(f) \\ &= \frac{e^{j2\pi f \tau_{a,b}(\mathbf{q}_\ell)}}{(4\pi c)^2 \tau_a(\mathbf{q}_\ell)\tau_b(\mathbf{q}_\ell)} S_\ell(f) \end{aligned} \quad (10)$$

The parameter of interest is the normalised relative phase response between two sensors, given by

$$\psi_{a,b}(f, \mathbf{q}) = e^{j2\pi f \tau_{a,b}(\mathbf{q})}. \quad (11)$$

3. ALGORITHMS

The microphone signals $x_i[m]$ are transformed into frequency domain signals using a DFT filterbank. The output signals from the transformation are denoted $X_i^{[k]}[n]$, where $k = 1, 2, \ldots, K$ is the subband index. These signals can be used to formulate an estimate of the spatial covariance matrix according to

$$\tilde{\mathbf{R}}_{\mathbf{x}}^{[k]}[n] = (1-\alpha)\tilde{\mathbf{R}}_{\mathbf{x}}^{[k]}[n-1] + \alpha \mathbf{X}^{[k]}[n]\left[\mathbf{X}^{[k]}[n]\right]^H \quad (12)$$

where α is the forgetting factor and $\mathbf{X}^{[k]}[n]$ is defined according to

$$\mathbf{X}^{[k]}[n] = \begin{bmatrix} X_1^{[k]}[n] & X_2^{[k]}[n] & \cdots & X_I^{[k]}[n] \end{bmatrix}^T. \quad (13)$$

The estimated spatial covariance matrix is used by the localisation algorithms to estimate $\mathbf{q}_\ell$ (or far-field $\hat{\mathbf{q}}_\ell$). An illustration of the system is presented in Fig. 1.

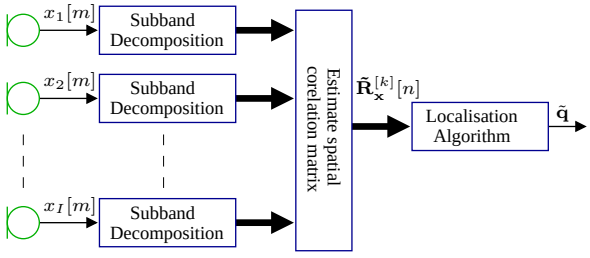

Fig. 1. System block diagram.

3.1. Steered Response Power - Phase Transform

The *relative phase response matrix* $\mathbf{\Psi}^{[k]}(\mathbf{q})$ is defined using the relative phase response for all sensor pairs in the array. For sample frequency F_s,

$$\left[\mathbf{\Psi}^{[k]}(\mathbf{q})\right]_{a,b} = \psi_{a,b}\left(\frac{kF_s}{K}, \mathbf{q}\right). \quad (14)$$

Let $\tilde{\mathbf{\Psi}}^{[k]}[n]$ denote an estimate of the phase response matrix for all sensor pairs and frequency bands. It is

calculated by normalising the elements of the spatial covariance matrix by their magnitude as

$$\tilde{\mathbf{\Psi}}^{[k]}[n] = \tilde{\mathbf{R}}_{\mathbf{x}}^{[k]}[n] \oslash \left|\tilde{\mathbf{R}}_{\mathbf{x}}^{[k]}[n]\right| \tag{15}$$

where $\oslash$ denotes element-wise division, and $|\cdot|$ denotes element-wise absolute value. The normalisation of the array covariance matrix is referred to as phase transform (PHAT) [2].

For a single source environment with a sound source at point $\mathbf{q}$, $\tilde{\mathbf{\Psi}}^{[k]}[n]$ will be an estimate of $\mathbf{\Psi}^{[k]}(\mathbf{q})$, but with amplitude and phase perturbations due to reverberation and background noise.

An estimate of the dominant component of $\tilde{\mathbf{\Psi}}^{[k]}[n]$ for all frequency bands k can be calculated using the relative phase response matrix from Eq. 14 according to

$$\tilde{\mathbf{q}}[n] = \arg\max_{\mathbf{q}} \left(\sum_{k=1}^{K} \mathbf{1} \left(\tilde{\mathbf{\Psi}}^{[k]}[n] \odot \left[\mathbf{\Psi}^{[k]}(\mathbf{q}) \right]^H \right) \mathbf{1}^T \right) \tag{16}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{1}$ is a vector of ones with I elements. This optimisation is a model fit of $\mathbf{\Psi}^{[k]}(\mathbf{q})$ to $\tilde{\mathbf{\Psi}}^{[k]}[n]$, and results in an estimate of the source position $\mathbf{q}$.

Depending on near-field or far-field conditions Eq. 16 is maximised over one or two variables. As such, it is maximised over $[\rho, \theta]$ for SRP-PHAT but only over one variable, θ for Far-Field SRP-PHAT.

3.2. Coherent Wideband Root-MUSIC

If each subband is sufficiently narrowband, $\tilde{\mathbf{R}}_{\mathbf{x}}^{[k]}[n]$ will approximate Eq. 8,

$$\tilde{\mathbf{R}}_{\mathbf{x}}^{[k]}[n] \simeq \mathbf{H}^{[k]} \mathbf{R}_s^{[k]} \left[\mathbf{H}^{[k]} \right]^H + \mathbf{R}_\eta^{[k]} \tag{17}$$

where $\mathbf{H}^{[k]}$, $\mathbf{R}_s^{[k]}$ and $\mathbf{R}_\eta^{[k]}$ are the sampled values of the continuous frequency variables in Eq. 8. The approach used here applies array interpolation at each subband to give

$$\mathbf{g}\left(\hat{\mathbf{q}}\right) \simeq \breve{\mathbf{T}}^{[k]} \mathbf{h}^{[k]}(\hat{\mathbf{q}}) \,;\; k = 1, 2, \ldots, K \tag{18}$$

where $\mathbf{h}^{[k]}(\hat{\mathbf{q}})$ is the sampled array response vector and $\mathbf{g}\left(\hat{\mathbf{q}}\right)$ is the response vector of a *virtual array* which is independent of k. To allow the application of root-MUSIC, the virtual array is by necessity uniform linear.

The SRI [7] optimum interpolation matrix $\breve{\mathbf{T}}^{[k]}$ is designed using the least-squares problem formulation

$$\breve{\mathbf{T}}^{[k]} = \arg\min_{\mathbf{T}^{[k]}} \int_{\{\hat{\mathbf{q}}\}} \left\| \mathbf{T}^{[k]} \mathbf{h}^{[k]}(\hat{\mathbf{q}}) - \mathbf{g}(\hat{\mathbf{q}}) \right\|^2 d\hat{\mathbf{q}}. \tag{19}$$

Applying interpolation to the spatial covariance matrices (Eq. 17), and summing over all frequency bands yields

$$\begin{aligned} \mathbf{U}[n] &= \sum_{k=1}^{K} w_k \breve{\mathbf{T}}^{[k]} \tilde{\mathbf{R}}_{\mathbf{x}}^{[k]}[n] \left[\breve{\mathbf{T}}^{[k]} \right]^H \\ &= \sum_{k=1}^{K} w_k \breve{\mathbf{T}}^{[k]} \mathbf{H}^{[k]} \mathbf{R}_s^{[k]} \left[\breve{\mathbf{T}}^{[k]} \mathbf{H}^{[k]} \right]^H + \breve{\mathbf{T}}^{[k]} \mathbf{R}_\eta^{[k]} \left[\breve{\mathbf{T}}^{[k]} \right]^H \\ &\simeq \mathbf{G} \mathbf{R}_s \mathbf{G}^H + \mathbf{N}, \end{aligned} \tag{20}$$

where $\mathbf{G} = \left[\mathbf{g}(\hat{\mathbf{q}}_1) \quad \mathbf{g}(\hat{\mathbf{q}}_2) \quad \cdots \quad \mathbf{g}(\hat{\mathbf{q}}_L) \right]$, w_k is a scalar weight applied to the k^{th} subband data, $\mathbf{R}_s$ is the combined signal covariance matrix and $\mathbf{N}$ is the combined noise covariance matrix.

The Root-MUSIC [5] algorithm can now be applied to the matrix $\mathbf{U}[n]$. Let $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_P\}$ denote the eigenvectors of $\mathbf{U}[n]$, ordered with regards to their corresponding eigenvalue magnitude. The eigenvectors $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_L\}$, span what is commonly termed 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 pre-whitening is applied to $\mathbf{U}[n]$, the L response vectors in $\mathbf{G}$ will be orthogonal to the noise subspace. Thus if $\mathbf{P}$ is a projection matrix onto the noise subspace of $\mathbf{U}[n]$,

$$\|\mathbf{P}\mathbf{g}(\hat{\mathbf{q}}_\ell)\|^2 = 0 \,;\;\; \ell = 1, 2, \ldots, L \tag{21}$$

The zeros of Eq. 21, are found by solving for the roots of a $2(P-1)$ order polynomial.

4. PERFORMANCE EVALUATIONS

The three localisation algorithms described above are implemented in software executed in realtime on a PC. The microphone array consists of eight elements which are mounted on a metal fixture with an inter-element distance of 40mm. The microphones are connected to the PC via a pre-amplifier. The microphone elements, model 2541/PRM902, and the pre-amplifier, model 2210, are from Larson Davis.

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

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} (\tilde{\theta}[n] - \theta)^2}, \tag{22}$$

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

4.1. 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 both environments the stimuli is female speech from a position $\mathbf{q} = [1.5\mathrm{m}, 2.0\mathrm{rad}]$.

The room has the dimensions 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 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 RMSE versus SNR for the two scenarios is shown in Figs. 2 and 3. The evaluation is performed with a forgetting factor $\alpha = 0.1$. The SNR is calculated by averaging the power at all microphone elements, over the frequency range 300 to 3400Hz. Furthermore, the figures show that the SRP-based algorithms are more robust than Root-MUSIC to noise and reverberation. At the given distance to the sound source the far field SRP-PHAT algorithm have an error of less than 300mm at 0dB SNR in the real room environment. The minimum RMSE is limited by the diameter of the loudspeaker in the real room environment.

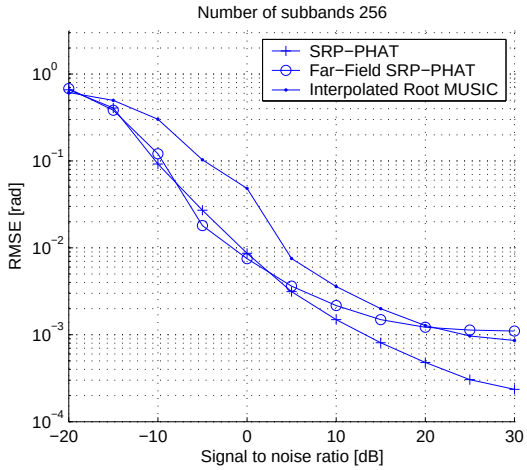

Fig. 2. RMSE versus SNR for female speech in free space model.

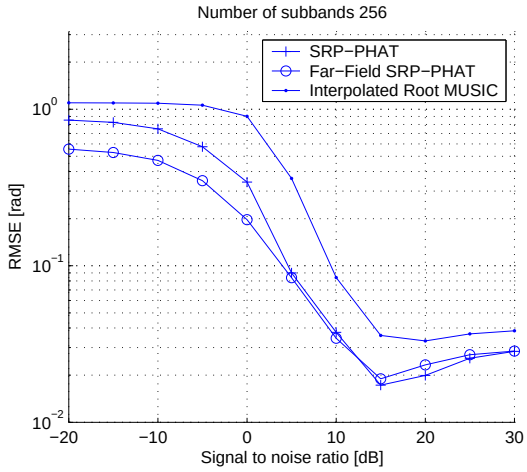

Fig. 3. RMSE versus SNR for female speech in real room environment.

4.2. Computational Complexity Evaluation

The computational load was evaluated by measuring the average number of clock-cycles taken for an algorithm to complete one position estimate for different number of subbands. The results showed that Root-MUSIC has the lowest computational load followed by Far-Field SRP-PHAT, which has 3-15 times higher computational load than Root-MUSIC. SRP-PHAT has a computational load nearly 20-100 times higher than Root-MUSIC.

5. CONCLUSIONS AND FUTURE WORK

Three localisation algorithms have been implemented in a real-time system and evaluated using a microphone array. The algorithms were evaluated with regards to robustness and computational load.

The results show that the two versions of SRP-PHAT have nearly identical performance with regards to robustness. However, the computational complexity of SRP-PHAT is more than 10 times that of Far-Field SRP-PHAT.

Root-MUSIC has the lowest computational complexity, but demonstrates comparatively poor robustness to reverberation and low SNR. As such the Root-MUSIC is extremely sensitive to parameter settings and the results presented above were obtained after considerable experimentation. In contrast the SRP-PHAT algorithms are comparatively insensitive to parameter adjustments.

6. REFERENCES

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

[2] C.H. Knapp and G.C. Carter, “The generalized correlation method for estimation of time delay”, IEEE Transactions on ASSP, 24(4):320-327,1976.

[3] D. Rabinkin, R. Renomeron, J. French and J. Flanagan. “Estimation of Wavefront Arrival Delay Using the Cross-Power Spectrum Phase Technique”, 132nd Meeting of the Acoustical Society of America, Honolulu, HI, USA, Dec. 1996.

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

[5] A. J. Barabell, “Improving the resolution performance of eigenstructure-based direction-finding algorithms,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Boston, MA, 1983, pp. 336–339.

[6] M. A. Doron, E. Doron and A. J. Weiss, “Coherent wide-band processing for arbitrary array geometry,” IEEE Trans. Signal Processing, vol. 31, pp. 414–417, Jan 1993.

[7] G. J. Cook, B. K. Lau and Y. H. Leung, “An alternative approach to interpolated array processing for uniform circular arrays,” in Proc. IEEE Asia-Pacific Conference on Circuits and Systems, Bali, Indonesia (relocated to Singapore), Oct. 2002, vol. 1, pp. 411–414.

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