A POLYNOMIAL EIGENVALUE DECOMPOSITION MUSIC APPROACH FOR BROADBAND SOUND SOURCE LOCALIZATION

Aidan O. T. Hogg , Vincent W. Neo*, Stephan Weiss†, Christine Evers‡ and Patrick A. Naylor**

*Department of Electrical and Electronic Engineering, Imperial College London, UK
†Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, Scotland
‡Electronics and Computer Science, University of Southampton, UK

ABSTRACT

Direction of arrival (DoA) estimation for sound source localization is increasingly prevalent in modern devices. In this paper, we explore a polynomial extension to the multiple signal classification (MUSIC) algorithm, spatio-spectral polynomial (SSP)-MUSIC, and evaluate its performance when using speech sound sources. In addition, we also propose three essential enhancements for SSP-MUSIC to work with noisy reverberant audio data. This paper includes an analysis of SSP-MUSIC using speech signals in a simulated room for different noise and reverberation conditions and the first task of the LOCATA challenge. We show that SSP-MUSIC is more robust to noise and reverberation compared to independent frequency bin (IFB) approaches and improvements can be seen for single sound source localization at signal-to-noise ratios (SNRs) below 5 dB and reverberation times (T60s) larger than 0.7 s.

***Index Terms*—** Direction of arrival, polynomial eigenvalue decomposition, MUSIC, localization, microphone arrays.

1. INTRODUCTION

Sound source localization is an important task for a multitude of applications, including robot audition [1] and voice-controlled smart devices. Direction of arrival (DoA) estimates are essential in providing angular positional information for localization. In real-world environments, DoA estimation is challenging because of background noise, reverberation, interference and sound source inactivity. DoA estimation approaches include time-delay estimation (TDE)-based, beamformer-based and subspace-based methods [2]. The TDE-based method [3] first computes the time difference of arrival (TDoA) for different microphone pairs and uses *a priori* information about the microphone positions to compute the DoAs. However, TDE approaches such as generalized cross-correlation (GCC)-phase-transform (PHAT) cannot cope with multiple sources in reverberant environments [2]. Beamformer-based methods [4,5] scan the acoustic environment by focusing the microphone array in the directions corresponding to the highest sound intensities. However, beamformer-based methods have been shown to perform poorly in low SNR regimes along with having a higher computational complexity when compared against common subspace-based approaches [6].

The support of the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1) is gratefully acknowledged.

This work was also supported in parts by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/S000631/1 and the MOD University Defence Research Collaboration in Signal Processing.

In a subspace-based approach such as the multiple signal classification (MUSIC) algorithm [7], the correlation matrix is computed from the received signals. An eigenvalue decomposition (EVD) is then used to decompose the correlation matrix into signal and noise subspaces for DoA estimation. The MUSIC algorithm, however, assumes that the source signals are narrowband and uncorrelated. Consequently, its performance is limited in real-world scenarios involving broadband signals such as speech and correlated sources originating from reverberant environments.

A number of broadband extensions have been proposed for MUSIC [8–10]. Most of these extensions rely on transforming the broadband DoA problem into several narrowband problems. This can be achieved by decomposing the broadband signal into several independent frequency bins [11]. The resulting narrowband signals for each frequency bin or filtered output can then be processed independently, or incoherently. This approach, however, is based on a narrowband signal model and ignores phase coherence across different frequency bins [12] which can lead to errors [13].

When broadband signals such as speech signals are involved, time delays cannot be modelled using phase shifts because time delays between different microphones need to be explicitly resolved. Consequently, an EVD cannot completely decorrelate the signals and separate the signal and noise subspaces effectively [2]. Instead, the spatio-spectral polynomial (SSP)-MUSIC approach in [14, 15] is based on a broadband signal model. The approach uses polynomial matrices to model the correlations across different microphones and temporal lags, and a polynomial eigenvalue decomposition (PEVD) to generate the signal and noise subspaces. SSP-MUSIC is shown to be robust and effective for temporally uncorrelated sources in anechoic environments [14, 15].

In this paper, we extend [14, 15] to sound source localization in noisy and reverberant environments. The novel contributions are: (i) proposed enhancements to SSP-MUSIC for sound source localization which include; incorporating a noisy reverberant signal model in the subspace decomposition; modifying SSP-MUSIC to only include the direct-path response in order to reduce the impact of reverberation on localization performance; using SSP-MUSIC to approximate spatial polynomial (SP)-MUSIC for the frequency range of speech; (ii) an analysis on how diffuse noise and reverberation affects the proposed approach; and (iii) a comprehensive evaluation of the proposed method against benchmark algorithms for simulated and real-world recordings.

2. METHOD

In [16], the noisy and reverberant signal, $x_m(n)$, at the m-th microphone for discrete-time sample $n = 0, 1, \ldots, N$, is

$$
\begin{aligned}
x_m(n) &= \mathbf{h}_m^T \mathbf{s}_0(n) + v_m(n) \\
x_m(n) &= \tilde{\mathbf{h}}_{m,\mathrm{dp}}^T \mathbf{s}_0(n) + \tilde{\mathbf{h}}_{m,\mathrm{er}}^T \mathbf{s}_0(n) + \tilde{\mathbf{h}}_{m,\mathrm{lr}}^T \mathbf{s}_0(n) + v_m(n) \\
&= \tilde{s}_m(n) + \tilde{v}_m(n), \quad m = 1, 2, \ldots, M,
\end{aligned} \tag{1}
$$

where $\mathbf{h}_m = [h_{m,0}, h_{m,1}, \ldots, h_{m,J}]^T$ is the m-th acoustic channel, which is modelled as a J-th order finite impulse response filter and decomposed into the direct-path, $\tilde{\mathbf{h}}_{m,\mathrm{dp}}$, early reflections, $\tilde{\mathbf{h}}_{m,\mathrm{er}}$, and the late reflections, $\tilde{\mathbf{h}}_{m,\mathrm{lr}}$ [17], $\mathbf{s}_0(n) = [s_0(n), s_0(n-1), \ldots, s_0(n-J)]^T$ is the anechoic speech signal, $v_m(n)$ is additive noise and $[\cdot]^T$ denotes the transpose operator. The noise signals are assumed to be zero-mean, not perfectly coherent with each other and uncorrelated with the source signals [18]. By exploiting the lack of correlation between the late reflections and anechoic speech signal [19] $\tilde{s}_m(n) = \tilde{\mathbf{h}}_{m,\mathrm{dp}}^T \mathbf{s}_0(n) + \tilde{\mathbf{h}}_{m,\mathrm{er}}^T \mathbf{s}_0(n)$ and $\tilde{v}_m(n) = \tilde{\mathbf{h}}_{m,\mathrm{lr}}^T \mathbf{s}_0(n) + v_m(n)$, can be decomposed into the speech and noise components respectively.

2.1. Review of Polynomial MUSIC

Assuming direct-path-only propagation in the far-field and a noise-free environment, $v_m(n) = 0$, such that (1) simplifies to

$$
x_m(n) = f_{\tau_m}(n) * x_0(n) \tag{2}
$$

where * denotes a linear convolution and $f_{\tau_m}(n)$ is a fractional delay filter [20, 21]. This is required since the m-th relative delay can be fractional, such that

$$
f_{\tau_m}(n) = \frac{\sin(\pi(n - \Delta\tau_m))}{\pi(n - \Delta\tau_m)}. \tag{3}
$$

In the narrowband case, $\Delta\tau_m$ is represented by a simple phase shift and those phase shifts are exploited in the MUSIC algorithm. For broadband sources, however, the delays are frequency-dependent phase shifts corresponding to different time lags.

To capture the temporal correlations of the speech signals at different microphones, the space-time covariance matrix [19] is computed using (1),

$$
\mathbf{R}_{\mathbf{xx}}(\tau) = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}^T(n-\tau)\}, \tag{4}
$$

where the $(p,q)^{\text{th}}$ element, $r_{pq}(\tau) = \mathbb{E}\{x_p(n)x_q(n-\tau)\}$, is the cross-correlation sequence between microphone p and q for discrete-time shift τ. Concatenating the covariance matrix, $\mathbf{R}_{\mathbf{xx}}(\tau)$, for all choices of $\tau \in \{-N, \ldots, N\}$, results in a tensor of dimension $M \times M \times (2N+1)$. The z-transform of (4) is a polynomial matrix, $\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z) = \sum_{\tau=-\infty}^{\infty} \mathbf{R}_{\mathbf{xx}}(\tau) z^{-\tau}$, which can be decomposed by an iterative PEVD algorithm [22–26] to give

$$
\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z) \approx \boldsymbol{\mathcal{U}}(z)\boldsymbol{\Lambda}(z)\boldsymbol{\mathcal{U}}^P(z), \tag{5}
$$

where the columns of $\boldsymbol{\mathcal{U}}(z)$ are the eigenvectors and the diagonal elements of $\boldsymbol{\Lambda}(z)$ are the eigenvalues. Furthermore, $\boldsymbol{\mathcal{U}}^P(z) = \boldsymbol{\mathcal{U}}^H(1/z^*)$, where $[\cdot]^*$, $[\cdot]^H$ and $[\cdot]^P$ are respectively, the complex-conjugate, Hermitian and para-Hermitian operators.

Thresholding the eigenvalues enables the partitioning of the polynomial matrix into orthogonal signal and noise subspaces, which are associated with $\boldsymbol{\mathcal{U}}_s(z)$ and $\boldsymbol{\mathcal{U}}_v(z)$, respectively. The nullspace of $\boldsymbol{\mathcal{U}}_v(z)$ is probed by the broadband steering vector, which implements fractional delays and is defined as [14]

$$
\mathbf{a}_\theta(z) = \begin{bmatrix} A_0(z) & \cdots & A_{M-1}(z) \end{bmatrix}^T, \tag{6}
$$

where θ is the look direction, $A_\ell(z) = \sum_{n=-\infty}^{\infty} a_\ell(n) z^{-n}$, $a_\ell(n) = \mathrm{sinc}((n - \Delta\tau_\ell)T_s)$ and T_s is the sampling period.

Generalised from MUSIC, the following quantity,

$$
\Gamma_\theta(z) = \mathbf{a}_\theta^P(z)\boldsymbol{\mathcal{U}}_v(z)\boldsymbol{\mathcal{U}}_v^P(z)\mathbf{a}_\theta(z), \tag{7}
$$

is used to compute the pseudo-spectrogram for SSP-MUSIC [14],

$$
\mathbf{P}_{\mathrm{SSP-MU}}(\theta, \Omega) = \frac{1}{\Gamma_\theta(z)}\bigg|_{z=e^{-j\Omega}}, \tag{8}
$$

where frequency Ω is obtained by evaluating z on the unit circle. Therefore, the pseudo-spectrogram can localize sources by exploiting the DoAs in the active range of frequencies.

2.2. Proposed Enhancements for Sound Source Localization

Similar to [19], but unlike [14, 15] which only focuses on non-speech sources in anechoic environments, we incorporate the noisy reverberant signal model in the subspace decomposition. The work in [19] is designed for speech enhancement and incorporates the early reflections that may improve speech intelligibility in some conditions [27, 28] whereas this paper focuses on sound source localization which only requires the direct-path component with the greatest amplitude and shortest time delay. Consequently, the largest delay, W, corresponding to the first maximum peak between every microphone pair is computed and, therefore, the z-transform of (4) is approximated by

$$
\hat{\boldsymbol{\mathcal{R}}}_{\mathbf{xx}}(z) \approx \sum_{\tau=-W}^{W} \boldsymbol{R}_{\mathbf{xx}}(\tau) z^{-\tau}, \tag{9}
$$

using the windowed space-time covariance matrix with dimensions $M \times M \times (2W+1)$. Furthermore, the introduction of W reduces the number of elements used in PEVD and offers computational improvement. While this window choice includes the largest direct-path propagation delay, some reflections are also inevitably captured by microphones that are near the sound sources.

Consequently, the PEVD of (4) gives [19]

$$
\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z) \approx \begin{bmatrix} \boldsymbol{\mathcal{U}}_{\tilde{s}}(z) \,\big|\, \boldsymbol{\mathcal{U}}_{\tilde{v}}(z) \end{bmatrix} \left[\begin{array}{c|c} \boldsymbol{\Lambda}_{\tilde{s}}(z) & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{\Lambda}_{\tilde{v}}(z) \end{array}\right] \begin{bmatrix} \boldsymbol{\mathcal{U}}_{\tilde{s}}^{\boldsymbol{P}}(z) \\ \hline \boldsymbol{\mathcal{U}}_{\tilde{v}}^{\boldsymbol{P}}(z) \end{bmatrix},
$$

where $\{.\}_{\tilde{s}}$ and $\{.\}_{\tilde{v}}$ represent the orthogonal signal and noise subspace components. The speech subspace comprise predominantly of anechoic speech convolved with the direct-path and some 'leaked' early reflections while the noise subspace contains ambient noise, both early and late reflections associated with the reverberant channel.

To cope with the infinite temporal support of the sinc function in (6), tapered windows have been proposed for truncation [21]. In this paper, the Hamming window defined by

$$
\begin{aligned}
w_{L,\mathrm{Hamm}}(n) &= (0.54 - 0.46\cos\left(\frac{\pi n}{2L}\right)) w_{L,\mathrm{rect}}(n), \\
&\text{where} \quad w_{L,\mathrm{rect}}(n) = \begin{cases} 1, & |n| \le L \\ 0, & |n| > L \end{cases},
\end{aligned} \tag{10}
$$

is used and L is the length of the truncated sinc function.

To compute the DoAs only, the pseudo-spectrogram in (8) is integrated over Ω. For K discrete points evaluated on the unit circle, the spatial-only pseudo-spectrum is approximated by

$$
\hat{\mathbf{P}}_{\mathrm{SSP-MU}}(\theta) = \frac{1}{K} \sum_{k=0}^{K-1} \mathbf{P}_{\mathrm{SSP-MU}}(\theta, \Omega_k), \tag{11}
$$

where $\Omega_k = \frac{2\pi}{K} k$ is the k-th frequency bin. The whole frequency range is considered in SP-MUSIC [14]. However, in this work, only Ω_k in the frequency range of speech (100 Hz to 4000 Hz) [29] are used in (11). A peak detection algorithm [30] is used to estimate the DoAs from (11).

Task		Exp-1				Exp-2			
Algorithm		SSP-MUSIC		IFB-MUSIC		SSP-MUSIC		IFB-MUSIC	
Metric		HR	FAR	HR	FAR	HR	FAR	HR	FAR
SNR [dB]	-15	85.0	15.0	55.0	45.0	56.9	43.1	35.4	64.6
	-10	95.0	5.0	55.0	45.0	60.0	40.0	52.3	46.9
	-5	95.0	5.0	85.0	15.0	64.6	35.4	61.5	38.5
	0	95.0	5.0	85.0	15.0	72.3	27.7	73.8	26.2
	5	100.0	0.0	95.0	5.0	70.8	29.2	84.6	15.4
	10	95.0	5.0	100.0	0.0	73.8	26.2	93.8	6.2
	15	95.0	5.0	95.0	5.0	70.8	29.2	93.8	6.2
	20	95.0	5.0	100.0	0.0	70.8	29.2	93.8	6.2
	25	95.0	5.0	100.0	0.0	76.9	23.1	92.3	7.7
T60 [s]	0.1	100.0	0.0	100.0	0.0	86.2	13.8	96.9	3.1
	0.3	95.0	5.0	95.0	5.0	70.8	29.2	89.2	10.8
	0.5	95.0	5.0	95.0	5.0	67.7	32.3	86.2	13.8
	0.7	95.0	5.0	80.0	20.0	60.0	40.0	73.8	26.2
	0.9	95.0	5.0	85.0	15.0	58.5	41.5	72.3	27.7
	1.1	95.0	5.0	85.0	15.0	60.0	40.0	72.3	27.7
	1.3	95.0	5.0	80.0	20.0	53.8	46.2	70.8	29.2
	1.5	95.0	5.0	75.0	25.0	52.3	47.7	73.8	26.2
	1.7	95.0	5.0	75.0	25.0	56.9	43.1	70.8	29.2

Table 1: Comparison of HR and FAR for Exp-1 and Exp-2.

3. EXPERIMENTAL SETUP

We used sequential matrix diagonalisation (SMD) [31] to perform an iterative PEVD as it has been shown to give a higher resolution for SSP-MUSIC than sequential best rotation algorithm (SBR2) [15]. The proposed approach is benchmarked against independent frequency bin (IFB)-MUSIC [30], with MUSIC applied to each frequency bin independently to estimate the DoAs. A frame size of 100 ms was used for all the experiments so that SSP-MUSIC could take advantage of temporal correlations. This allows for the exploitation of the strong decorrelation of the PEVD approach, i.e. the extracted subspaces are not just decorrelated in isolated IFBs but coherently separated across all lag values τ.

3.1. Evaluation Metrics

The performance of SSP-MUSIC and IFB-MUSIC is evaluated using the following metrics. A 'HIT' is when a sound source (speaker) has been detected once within a $\pm\xi$ collar applied around the ground-truth azimuth. A 'MISS' is when a sound source has not been detected within this collar and a false alarm (FA) is when a detection falls outside of a ground-truth azimuth collar. The HIT rate (HR) and false alarm rate (FAR) are, therefore, defined as,

$$\text{HR} = \frac{\text{HITs}}{\text{HITs} + \text{MISSs}}\%, \quad \text{FAR} = \frac{\text{FAs}}{\text{HITs} + \text{FAs}}\% \tag{12}$$

respectively wherein this work the collar ξ is set to 15°. To further evaluate the accuracy of the DoA estimates, the absolute errors for all the HITs are shown in the form of boxplots.

3.2. Simulated Data Generation

In this work, the performance of SSP-MUSIC is first evaluated on data generated using a simulated room of dimensions $3 \times 3 \times 2$ m generated using [30]. A uniform circular array of 8 microphones, with a diameter of 4.2 cm, is positioned at the centre of the room. Since the microphone array geometry is known, the largest direct-path delay can inform the choice of the temporal lag support W (see Section 2.2). This largest delay is 5.87 samples in our array configuration and, therefore, we have set $W = 10$. Two experiments are run using this scenario. Exp-1 evaluates the performance when a single active speaker is placed at a distance of 1.5 m from the centre of the array at an angle of 50° where the anechoic speech used was a 3 s recording taken from the LOCATA

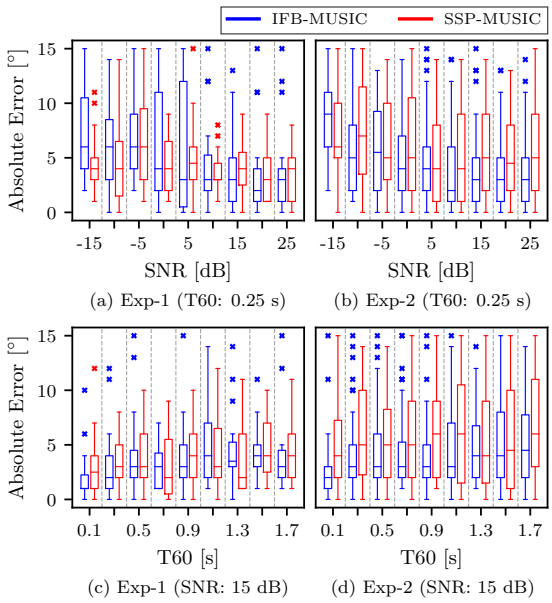

Figure 1: Comparison of absolute errors of all HITs. (a) Exp-1 performance when the T60 is 0.25 s and the SNR is varied. (b) same as (a) for Exp-2. (c) Exp-1 performance when the SNR is 15 dB and the T60 is varied. (d) same as (c) for Exp-2.

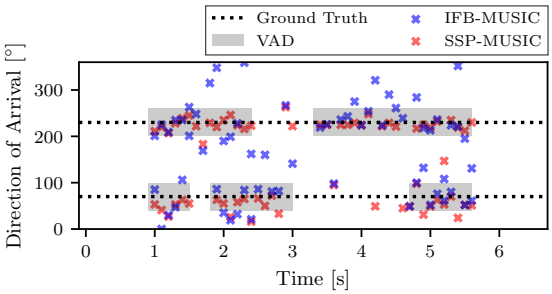

Figure 2: Illustrative example of 2 active sources in a simulated room where the SNR is -15 dB and the T60 is 0.25 s.

corpus [2] (Task 1, Recording 1). Exp-2 evaluates the performance for two speakers where the anechoic speech is taken from LOCATA (Task 2, Recording 1) and the sources are placed at the same distance as Exp-1 but at angles of 70° and 230°.

It should be noted that, in this work, we only evaluate the DoA estimates for frames that are known to contain speech activity. This information is provided by an oracle voice activity detection (VAD) given in LOCATA. The oracle VAD is also used to determine the number of active speakers for both IFB-MUSIC and SSP-MUSIC.

4. RESULTS

To evaluate the performance of SSP-MUSIC, it is compared against the `pyroomacoustics` implementation of IFB-MUSIC [30].

4.1. Exp-1: One Static Speaker

In this experiment, we carried out a comparison for when only 1 static talker is active. Table 1 shows the impact of diffuse white Gaussian noise on the accuracy of the DoA estimates. While a rank-1 decomposition is expected for a single speaker, the PEVD

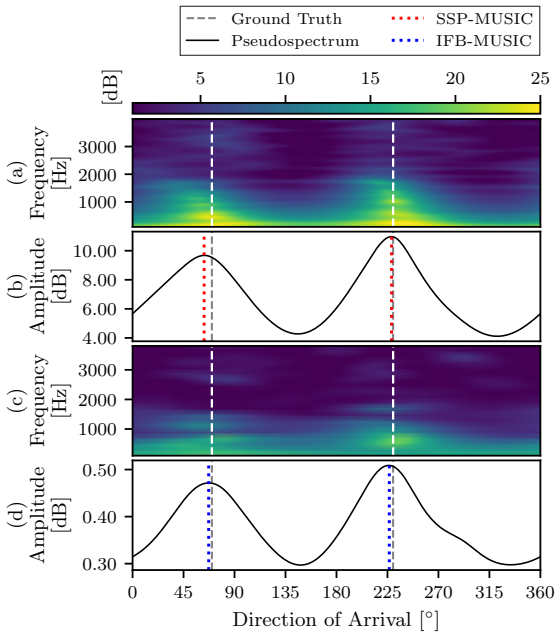

Figure 3: Illustrative example of a 100 ms frame from Exp-2 (SNR: -10 dB, T60: 0.25 s). (a) pseudo-spectrogram of SSP-MUSIC, (b) pseudo-spectrum of SSP-MUSIC, (c) pseudo-spectrogram of IFB-MUSIC, (d) pseudo-spectrum of IFB-MUSIC.

instead produced a rank-2 matrix decomposition, also observed in [19]. The first and second principal eigenvectors capture roughly the direct-path and the early and late reflections, respectively, for reasons given in Section 2.2. In terms of performance, it can clearly be seen that at signal-to-noise ratio (SNR) conditions lower than 5 dB, the performance of IFB-MUSIC has lower HR and a higher FAR when compared to SSP-MUSIC. This is expected as it is well known that subspace methods, such as MUSIC, suffer from the so-called 'threshold effect', which results in a degradation both in terms of resolution and precision at low SNR values [32]. Table 1 goes on to show the degradation that occurs when the reverberation time (T60) is increased. It can be seen that SSP-MUSIC performs better than IFB-MUSIC when the T60 value is larger than 0.5 s. This robustness to reverberation is likely a result of (9) which is one of the proposed enhancements and forces SSP-MUSIC to mainly consider the direct-path component and only allowing a few reflections to be additionally captured. Fig. 1(a) and (c) show the estimates' accuracy by highlighting the absolute errors of all the estimates considered to be HITs.

4.2. Exp-2: Two Static Speakers

In this second experiment, we carried out a comparison of 2 active talkers. In a similar manner to Exp-1, Table 1 shows that as the diffuse white Gaussian noise increases so do the errors in the DoA estimates where IFB-MUSIC is more adversely affected. An illustrative example, when the SNR is -15 dB and the T60 is 0.25 s, is also given in Fig. 2 to highlight the performance improvement of SSP-MUSIC at low SNR values.

Fig. 3(a) and (c) compare the two pseudo-spectrograms of both SSP-MUSIC and IFB-MUSIC for a single 100 ms frame. The final DoA estimates have an average absolute error of 4.5° and 3° for SSP-MUSIC and IFB-MUSIC respectively, but the SSP-MUSIC spectrogram has far more accurate components across

Method		SSP-MUSIC		IFB-MUSIC	
Metric		HR	FAR	HR	FAR
Recording	1	90.0	10.0	95.0	5.0
	2	64.7	35.3	67.6	32.4
	3	95.1	4.9	75.6	24.4

Table 2: Comparison of HR and FAR for both SSP-MUSIC against IFB-MUSIC on the first 3 LOCATA recordings for Task 1.

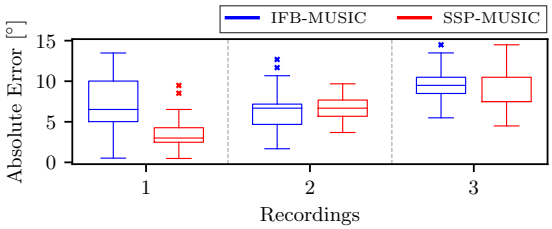

Figure 4: Performance comparison of IFB-MUSIC against SSP-MUSIC across Task 1 LOCATA recordings.

frequencies when compared with IFB-MUSIC. The dynamic range of IFB-MUSIC is also much smaller than SSP-MUSIC which could lead to numerical issues with the peak detection algorithm when applied to the pseudo-spectra of IFB-MUSIC. It can also be observed in Fig. 3(b) and (d) that the resolution of the pseudo-spectrum for SSP-MUSIC was not as sharp as IFB-MUSIC. This is likely due to the fact that fractional delay filters implementing the time delays associated with different angles are not accurate across the entire frequency range [20].

4.3. Exp-3: LOCATA Task 1

To validate SSP-MUSIC as a good alternative to IFB-MUSIC, we compared both algorithms for 3 recordings from Task 1 of the LOCATA challenge. In LOCATA, the ground truth DoAs are given by an optical tracking system, *OptiTrac*. In this experiment, real-world audio signals that were captured from an 8 microphone non-uniform circular array, selected from an Eigenmike, were used for the evaluation. The largest path delay is 5.87 samples and we have chosen $W = 10$. The recordings, measured in a real room with T60 ≈ 0.5 s, also contained low-level background noise.

Table 2 shows that on average a 3.9% better HR and a 3.9% lower FAR can be achieved by SSP-MUSIC when compared against IFB-MUSIC on the 3 recordings studied. Fig. 4 shows that the accuracy of the DoA estimates given by SSP-MUSIC is better or the same in terms of mean absolute error when compared against IFB-MUSIC. It should be noted that, as the SNR values are high and T60 values are low (0.5 s) across all these recordings, we do not expect to see great improvements in the performance. This result, however, still illustrates the benefits of incorporating a broadband signal model for reverberant speech in the subspace decomposition.

5. CONCLUSION

In this paper, we have developed and explored the potential of SSP-MUSIC which is a polynomial extension of MUSIC. In addition, we have proposed some enhancements for sound source localization. This paper has highlighted the benefits of using SSP-MUSIC for localization of a single sound source at SNR values lower than 5 dB or T60 values larger than 0.7 s as it is more robust to noise and reverberation. An evaluation was also carried out on real data, taken from the LOCATA corpus, which has shown that SSP-MUSIC can outperform IFB-MUSIC on real-world signals.

6. REFERENCES

[1] C. Evers and P. A. Naylor, "Optimized self-localization for SLAM in dynamic scenes using probability hypothesis density filters," *IEEE Trans. Signal Process.*, vol. 66, no. 4, pp. 863–878, Feb. 2018.

[2] C. Evers, H. W. Löllmann, H. Mellmann, *et al.*, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1620–1643, Apr. 2020.

[3] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[4] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 5, pp. 1210–1217, Oct. 1983.

[5] W. Bangs and P. Schultheis, "Space-time processing for optimal parameter estimation," *Signal Process.*, pp. 577–590, 1973.

[6] N. A. Baig and M. B. Malik, "Comparison of direction of arrival (DOA) estimation techniques for closely spaced targets," *Int. J. Future Comput. Commun.*, vol. 2, no. 6, pp. 654–659, Dec. 2013.

[7] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[8] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1327–1339, 2007.

[9] H. L. Van Trees, *Optimal Array Processing. Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 2002.

[10] B. Friedlander, "The root-MUSIC algorithm for direction finding with interpolated arrays," *Signal Process.*, vol. 30, no. 1, pp. 15–29, Jan. 1993.

[11] M. Alrmah, S. Weiss, S. Redif, *et al.*, "Angle of arrival estimation for broadband signals: A comparison," in *Intell. Signal Process. Conf. (ISP)*, Jan. 2013, pp. 1–6.

[12] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 240–254, May 2000.

[13] S. Weiss and I. Proudler, "Comparing efficient broadband beamforming architectures and their performance trade-offs," in *Proc. IEEE Int. Conf. Digital Signal Process. (DSP)*, A. N. Skodras and A. G. Constantinides, Eds., July 2002, pp. 417–424.

[14] M. A. Alrmah, S. Weiss, and S. Lambotharan, "An extension of the MUSIC algorithm to broadband scenarios using a polynomial eigenvalue decomposition," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2011, pp. 629–633.

[15] M. Alrmah, "Broadband angle of arrival estimation using polynomial matrix decompositions," Ph.D. dissertation, University of Strathclyde, Scotland, Oct. 2015.

[16] V. W. Neo, C. Evers, and P. A. Naylor, "PEVD-based speech enhancement in reverberant environments," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2020, pp. 186–190.

[17] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Springer-Verlag, 2010.

[18] Y. Huang, J. Benesty, and J. Chen, "Dereverberation," in *Springer Handbook of Speech Processing*. Springer-Verlag, 2008, pp. 929–943.

[19] V. W. Neo, C. Evers, and P. A. Naylor, "Speech dereverberation performance of a polynomial-EVD subspace approach," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2020, pp. 221–225.

[20] T. I. Laakso, V. Valimaki, M. Karjalainen, *et al.*, "Splitting the unit delay [FIR/all pass filters design]," *IEEE Signal Process. Mag.*, vol. 13, no. 1, pp. 30–60, Jan. 1996.

[21] J. Selva, "An efficient structure for the design of variable fractional delay filters based on the windowing method," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3770–3775, Aug. 2008.

[22] J. G. McWhirter, P. D. Baxter, T. Cooper, *et al.*, "An EVD algorithm for parahermitian polynomial matrices," *IEEE Trans. Signal Process.*, vol. 55, no. 5, pp. 2158–2169, May 2007.

[23] S. Weiss, J. Pestana, and I. K. Proudler, "On the existence and uniqueness of the eigenvalue decomposition of a parahermitian matrix," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2659–2672, May 2018.

[24] S. Weiss, I. K. Proudler, and F. K. Coutts, "Eigenvalue decomposition of a parahermitian matrix: Extraction of analytic eigenvalues," *IEEE Trans. Signal Process.*, vol. 69, pp. 722–737, 2021.

[25] V. W. Neo and P. A. Naylor, "Second order sequential best rotation algorithm with Householder transformation for polynomial matrix eigenvalue decomposition," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2019, pp. 8043–8047.

[26] S. Redif, S. Weiss, and J. G. McWhirter, "An approximate polynomial matrix eigenvalue decomposition algorithm for parahermitian matrices," in *Proc. Int. Symp. on Signal Process. and Inform. Technol. (ISSPIT)*, 2011, pp. 421–425.

[27] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am.*, vol. 113, no. 6, pp. 3233–3244, June 2003.

[28] H. Kuttruff, *Room Acoustics*, 4th ed. London: Taylor & Francis Ltd., 2000.

[29] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, New Jersey, USA: Prentice Hall, 1978.

[30] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2018, pp. 351–355.

[31] S. Redif, S. Weiss, and J. G. McWhirter, "Sequential matrix diagonalisation algorithms for polynomial EVD of parahermitian matrices," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 81–89, Jan. 2015.

[32] J. K. Thomas, L. L. Scharf, and D. W. Tufts, "The probability of a subspace swap in the SVD," *IEEE Trans. Signal Process.*, vol. 43, no. 3, pp. 730–736, 1995.