IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 14, 2021, accepted July 3, 2021, date of publication July 7, 2021, date of current version July 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3095345

Spatially Smoothed TF-Root-MUSIC for DOA Estimation of Coherent and Non-Stationary Sources Under Noisy Conditions

RUSLAN ZHAGYPAR, (Student Member, IEEE), KALAMKAS ZHAGYPAROVA, (Student Member, IEEE), AND MUHAMMAD TAHIR AKHTAR, (Senior Member, IEEE)

Department of Electrical and Computer Engineering, School of Engineering and Digital Sciences, Nazarbayev University, Nur-Sultan 010000, Kazakhstan

Corresponding author: Ruslan Zhagypar (ruslan.zhagypar@nu.edu.kz)

This work was supported in part by the Faculty Development Competitive Research Grant Program of Nazarbayev University under Grant 110119FD4525.

ABSTRACT This paper proposes a method for efficient Direction-of-Arrival (DOA) estimation of coherent and non-stationary sources under adverse noise conditions. The method consists of three main parts: 1) derivation of Spatial Time-Frequency Distribution (STFD) matrix; 2) application of the forward-backward spatial smoothing technique; 3) estimating the angles of arrival by solving for the roots of the polynomial. The key significance of the proposed method is that the combination of existing methods and techniques allows an estimation of DOA angles for both coherent and non-stationary source signals under noise. Whereas the individual use of the existing methods does not show adequate performance under these conditions. The experiments allow studying the performance of the proposed method for 1) both coherent and non-coherent clean sinusoidal signals; 2) noisy non-stationary chirp signals; 3) coherent and non-stationary signals under noise. Furthermore, extensive simulations have been carried out to compute the root mean square error (RMSE) performance of the proposed method in comparison with the existing ones. The experiments have been designed for varying number of microphones, level of noise, and value of the time-frequency threshold. As a result of the experiments, we observe the efficacy of the proposed method in comparison with the conventional Root-MUSIC and Time-Frequency MUSIC (TF-MUSIC) methods.

INDEX TERMS Direction-of-arrival estimation, MUSIC, Root-MUSIC, TF-MUSIC, sound source localization.

I. INTRODUCTION

Sound Source Localization (SSL) has been one of the actively studied array signal processing problems. It aims to determine the positions of sound sources via processing the signals, which are recorded by an array of sensors, in terms of two components of a source position: Direction-of-arrival (DOA) estimation and Distance estimation [1]. Fig. 1 illustrates the DOA angle θ as well as the distance d from the sound source to the receiver. Henceforth, the receiver will be represented by a microphone array. DOA estimation finds its applications in many areas, viz. human-robot interaction [1], rescue scenarios with poor visual contact [2], target tracking tasks [3], smart road crossing systems [4], etc. In real-life scenarios, DOA estimation methods should be capable of estimating more than one active sound sources in the environment, which increases the complexity of the problem.

There are many approaches to tackle the DOA estimation problem, which can be divided into four broad categories: time delay-based, beamforming-based, learning-based, and subspace-based approaches. The type of method that provides a high angular resolution relies on the beamforming. Among many beamforming-based methods, researchers are prone to investigate the followings: Bartlett [5], Capon [6], and Minimum Variance Distortionless Response (MVDR) [7]. With the recent advancements of machine learning methods, numerous learning-based methods have started to gain popularity. One of such approaches uses a data-driven

The associate editor coordinating the review of this manuscript and approving it for publication was Hasan S. Mir.

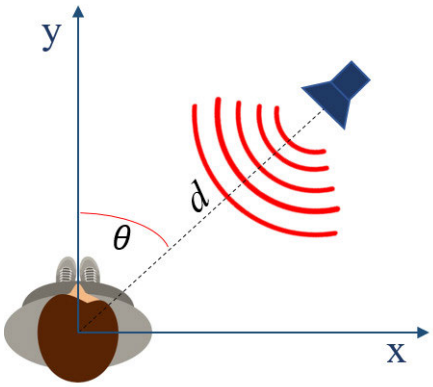


FIGURE 1. A pictorial representation of Sound Source Localization where θ denotes the DOA and d is the distance between source and receiver.

Neural Network (NN), the accuracy of which is solely contingent upon the availability of training data rather than pre-assumptions about array geometries [8]. The methods, falling in the subspace-based category utilize the orthogonality property of source and noise subspaces: Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) [9], minimum variance [10], autoregressive signal model [11], subspace fitting DOA estimation [12], and MUltiple SIgnal Classification (MUSIC) [13]. Particularly, the MUSIC method, which is the classical approach to estimate spatial spectrum of signals, was initially proposed in [13]. The method relies on the orthogonality property of signal and noise subspaces derived as a result of performing the eigenvalue decomposition (EVD) on an input covariance matrix. The scope of the paper is thus concerned with the subspace-based DOA estimation techniques.

A method with less computational complexity compared to the original MUSIC method is known as Root-MUSIC, which estimates the DOA by determining the roots of a polynomial formed from the noise subspace [14]. The aforementioned methods assume narrowband, non-coherent and stationary signal sources with low-level noise for accurate performance, however, these conditions are idealised and rarely met in practical scenarios. The coherent signals are known for having similar frequency components, which can degrade the performances of the frequency estimating methods. Therefore, it is suggested to apply the spatial smoothing technique to circumvent the restriction of non-coherence [16], [17]. Furthermore, the most of the real-world signals, e.g., human speech, are non-stationary, which significantly reduces the effectiveness of the conventional methods. Thus, an advanced approach that leverages the properties of spatial time-frequency (TF) distributions is proposed by [18], [19]. The main advantage of this method is the effect of denoising, which separates components of recorded signals from those of additive noise.

The main contribution of this paper is to propose a method for DOA estimation with good performance for both coherent and non-stationary sources and under noisy conditions. The existing solutions fail to show proper results in such conditions when used separately. Hence, the proposed method incorporates key features of several recent and earlier studies. Instead of the regular covariance matrix, the spatial time-frequency distribution (STFD) matrix is computed. The authors of [20] suggest that the latter one provides improved signal selectivity and noise reduction due to different TF signatures of non-stationary signals corrupted with noise. However, the main contribution of this study is in the application of the forward-backward spatial smoothing technique to the derived STFD matrix, which makes the method immune to coherent as well as to non-stationary signals. Last but not least, the principle of Root-MUSIC is employed which results in an efficient DOA estimation as compared with the spatial spectrum search-based approaches as in [16]- [19]. Hence, the method discussed in the paper expands the scope of aforementioned methods by simultaneously addressing both non-stationary and coherent signals under severe noise conditions.

The remaining of this paper is organized as follows. Sections II and III present the signal model and overview of baseline methods and techniques, respectively. The preliminary TF concepts, key features of the proposed method, and computational complexity analysis are discussed in details in Section IV. Section V reports on the simulation setup and the obtained results. Finally, Section VI concludes the paper and defines the future study direction. Enriched with more simulation results under various conditions, the paper can be viewed as an extension to a short version presented at a conference [21].

II. SIGNAL MODEL

The model comprises of an M-element Uniform Linear Array (ULA) that receives the signals from a P number of sound sources. The study in [22] recommends setting the array spacing d equal to the half of the signal wavelength for higher angular resolution. Considering N time samples (snapshots) of signals, an instantaneous mixing model can be written as

$$\boldsymbol{x}(t) = \boldsymbol{A}(\theta)\,\boldsymbol{s}(t) + \boldsymbol{n}(t), \quad t = 1, \ldots, N, \tag{1}$$

where t represents the discrete-time given in time-samples, $\boldsymbol{s}(t)$ is $\mathbb{C}^{P\times 1}$ vector containing pure source signals, $\boldsymbol{n}(t)$ is $\mathbb{C}^{M\times 1}$ vector for additive white Gaussian noise, and $\boldsymbol{A}(\theta)$ is $\mathbb{C}^{M\times P}$ propagation matrix which contains the delay information of each signal source at every array element. The described basic array signal model will be used in all of the upcoming methodologies with additional assumptions introduced in corresponding sections.

A coherent signal is often formed when a signal is reflected from surfaces before reaching the microphone array. The reflected sound wave has the same properties as the original signal but arrives from a different direction. According to [23], the high resolution subspace-based techniques assume that the covariance matrix of the sources P is non-singular. As this property does not hold when the sources are coherent, the spatial smoothing technique should be applied.

The signals may also display a form that can be described as non-stationary: e.g., a chirp signal, the frequency of which changes linearly with time. The study of [24] represents a collection of TF techniques to process these signals. However, the Wigner-Ville distribution is of major interest for this paper.

III. OVERVIEW OF BASELINE METHODS AND TECHNIQUES

A. TIME DIFFERENCE OF ARRIVAL (TDOA)

The concept of TDOA has been actively used in many advanced DOA estimation methods. It is defined as the delay-time needed for the sound wavefront to propagate through the distance between two receiving microphones [25]. The schematic diagram in Fig. 2 represents the simplest case with one dimensional DOA estimation of a single acoustic source S with a microphone array consisting from only two microphones m_1 and m_2. The assumption of the far-field environment is satisfied by considering the distance from the sound source to the array much larger than the array spacing d. From the geometry of the formed right triangle, it is straightforward to get the expression for TDOA T as follows

$$T = \frac{\Delta l}{c} = \frac{d \sin(\theta)}{c} = T_{\max} \sin(\theta), \tag{2}$$

where Δl difference is the propagation distance, c is the speed of sound, and $T_{\max}$ represents the maximum possible delay when θ equals -90^o. This expression of the delay will be further used to describe the phase information in the subspace-based methods.

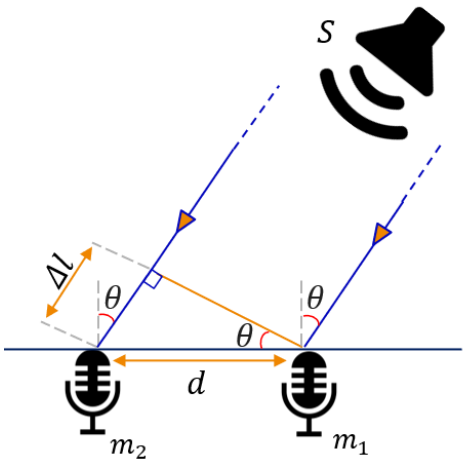


FIGURE 2. A diagram illustrating the principle of TDOA between two microphones.

B. MUltiple SIgnal CLASSIFICATION (MUSIC)

MUSIC is a subspace-based high-resolution method initially developed in [13], which has become the classical approach to spatial spectrum estimation. It requires a covariance matrix $\boldsymbol{R}_{xx}$ of size $\mathbb{C}^{M\times M}$ to be calculated from the instantaneous mixing model in (1) as

$$\begin{aligned}\boldsymbol{R}_{xx} &= \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^H] = \mathbb{E}[(\boldsymbol{A}(\theta)\boldsymbol{s}+\boldsymbol{n})(\boldsymbol{A}(\theta)\boldsymbol{s}+\boldsymbol{n})^H] \\ &= \boldsymbol{A}(\theta)\boldsymbol{R}_{ss}\boldsymbol{A}(\theta) + \sigma_n^2\boldsymbol{I},\end{aligned} \tag{3}$$

where $\mathbb{E}[\cdot]$ indicates the expectation operator, H represents the conjugate transposition, $\boldsymbol{R}_{ss}$ is the $\mathbb{C}^{P\times P}$ covariance matrix of the source signals, σ_n^2 is the additive noise variance, and $\boldsymbol{I}$ is the $\mathbb{R}^{M\times M}$ identity matrix. The MUSIC method leverages the orthogonality property of signal and noise subspaces derived from Hermitian matrix $\boldsymbol{R}_{xx}$. The right-hand-side (RHS) of (3) is known as the form of a matrix after performing EVD. As a result, one obtains M number of eigenvalues given as the entries in the diagonal matrix $\boldsymbol{R}_{ss}$ as well as corresponding eigenvectors stored as the columns of the matrix $\boldsymbol{A}(\theta)$ [26]. Hence, the eigenvectors are functions of DOA angles. When the eigenvalues are put in ascending order, the first P eigenvalues carry sound source information, and the rest belong to noise:

$$\lambda_i = \begin{cases} v_i + \sigma^2 & \text{for } i = 1, 2, \ldots, P \\ \sigma^2 & \text{for } i = P+1, P+2, \ldots, M. \end{cases} \tag{4}$$

Let us denote all the eigenvectors of $\boldsymbol{R}_{xx}$ by $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_M$ each of size $\mathbb{C}^{M\times 1}$. Among these eigenvectors, the ones that correspond to first eigenvalues P eigenvalues in (4) form the source subspace and are denoted as

$$\boldsymbol{e}_j = \boldsymbol{q}_j \quad \text{for } j = 1, 2, \ldots, P. \tag{5}$$

The above-mentioned orthogonality property of the Hermitian covariance matrix results in the following expressions:

$$\begin{aligned}\boldsymbol{e}_j^H \boldsymbol{q}_i = 0 \quad &\text{for } i = P+1, P+2, \ldots, M \\ &j = 1, 2, \ldots, P.\end{aligned} \tag{6}$$

The expression in (6) can be regarded as the Discrete Time Fourier Transform (DTFT) performed on the noise eigenvector:

$$\begin{aligned}\text{DTFT}\{\boldsymbol{q}_i\} = \sum_{k=0}^{M-1} \boldsymbol{q}_i(k) e^{j\omega_j k} = \boldsymbol{e}_j^H \boldsymbol{q}_i \\ \text{for } i = P+1, P+2, \ldots, M \\ j = 1, 2, \ldots, P.\end{aligned} \tag{7}$$

The usefulness of this property becomes clear when (7) is applied to spatial spectrum power function as

$$\begin{aligned}P_{\text{MUSIC}}(\theta) &= \frac{1}{\sum_{i=P+1}^{M} \left|\boldsymbol{e}^H \boldsymbol{q}_i\right|^2} \\ &= \frac{1}{\sum_{i=P+1}^{M} \boldsymbol{e}^H \boldsymbol{q}_i \boldsymbol{q}_i^H \boldsymbol{e}} = \frac{1}{\boldsymbol{e}^H \boldsymbol{Q}\boldsymbol{Q}^H \boldsymbol{e}},\end{aligned} \tag{8}$$

where $\boldsymbol{Q}$ denotes the $\mathbb{C}^{M\times M-P}$ noise subspace matrix [15], given by:

$$\boldsymbol{Q} = \sum_{i=P+1}^{M} \boldsymbol{q}_i. \tag{9}$$

The function in (8) is evaluated for each DOA angle of θ. Whenever the correct source angle is found, the denominator of the function drops to zero, giving a rise to a peak. Although it results in a high angular resolution in a spectrum, the process of checking each angle is time-consuming and can be ineffective in real-time applications.

IV. PROPOSED METHOD

A. SPATIAL SMOOTHING TECHNIQUE

Coherent signals are important to consider as they represent the effect of reverberation [16]. However, the classical MUSIC method requires the covariance matrix to be full rank to keep the noise subspace orthogonal to the signal subspace. The full-rank assumption becomes no longer valid when two or more signals are coherent [17]. Thus, there is a need to apply spatial smoothing technique, which is dedicated to diagonalize a matrix and decorrelate the signal sources [27].

The technique divides the ULA into L overlapping subarrays of size $K \in [P+1; M]$. The subarrays can be created in two ways depending on whether the microphones are grouped in a forward or a backward direction. In the forward direction, the received signal vector similar to (1) is formed at l-th subarray as

$$\boldsymbol{x}_l(t) = [x_l(t), x_{l+1}(t), \ldots, x_{l+K-1}(t)]^T = \boldsymbol{U}_l \boldsymbol{x}(t), \quad (10)$$

where $\boldsymbol{U}_l = [\boldsymbol{0}_{K\times(l-1)}\ \boldsymbol{I}_K\ \boldsymbol{0}_{K\times(M-K-l+1)}]$ is the selection matrix which is used to select the sensors from l to $l+K-1$ when multiplied to complete array model [16], [28]. Separate covariance matrices can be found for each of these subarrays. Hence, the total covariance matrix of the forward smoothing, $\boldsymbol{R}^f$, is derived by averaging all covariance matrices of L subarrays:

$$\boldsymbol{R}^f = \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{R}_l^f = \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{U}_l \boldsymbol{R}_{xx} \boldsymbol{U}_l^H, \quad (11)$$

where $\boldsymbol{R}_l^f$ is the covariance matrix of lth subarray in forward smoothing.

In contrast, the microphones are selected in reverse order in backward smoothing. Therefore, the l-th subarray includes microphones numbered from $l+K-1$ to l. The output vector of the subarray, $\boldsymbol{y}_l(t)$, is found as

$$\boldsymbol{y}_l(t) = \boldsymbol{U}_l \boldsymbol{J} \boldsymbol{x}^*(t), \quad (12)$$

where $\boldsymbol{J}$ is an $\mathbb{C}^{M\times M}$ exchange matrix which has ones as anti-diagonal entries and zeros as others, and $*$ depicts the conjugation. Thus, the counterpart of (11) for backward smoothing is expressed as

$$\boldsymbol{R}^b = \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{R}_l^b = \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{U}_l \boldsymbol{J} \boldsymbol{R}_{xx}^H \boldsymbol{J} \boldsymbol{U}_l^H, \quad (13)$$

where $\boldsymbol{R}_l^b$ is the covariance matrix of lth subarray in backward smoothing, and $\boldsymbol{R}_{xx}^H$ is the conjugate transpose of the non-smoothed covariance matrix in (3).

In the end, the spatially smoothed covariance matrix is found by averaging (11) and (13):

$$\boldsymbol{R}^{fb} = \frac{\boldsymbol{R}^f + \boldsymbol{R}^b}{2}. \quad (14)$$

The smoothed covariance matrix $\boldsymbol{R}^{fb}$ has the full rank when the number of subarrays L is equal to the half of the number of coherent sources. For example, when there are two coherent sources, the value of L becomes equal to unity. At this point, all the eigenvalues become non-zero, which confirms the decorrelation ability of the spatial smoothing technique [17].

B. ROOT-MUSIC

Similar to conventional MUSIC, Root-MUSIC is devised to work with non-coherent and stationary signals only. It was shown that the classical MUSIC estimates the DOA angles by performing a spatial search through all angles. Root-MUSIC takes a different approach to estimate the DOA angles by determining the roots of a polynomial formed from the noise eigenvectors [14]. In fact, the denominator of the power function in (8) can be regarded as the $\mathcal{Z}$-transform of $\boldsymbol{q}_i$

$$Q_i(z)|_{z=e^{jw_1}} = \sum_{n=0}^{M-1} q_i(n)\, z^{-n} = \boldsymbol{e}_1{}^H \boldsymbol{q}_i = 0. \quad (15)$$

Thus, it is evident that Root-MUSIC considers the denominator of (8) as a polynomial in $\mathcal{Z}$-domain

$$\begin{aligned} F(z) &= \boldsymbol{e}^H(z) \boldsymbol{Q}\boldsymbol{Q}^H \boldsymbol{e}(z) \\ &= \boldsymbol{e}^H(z) \boldsymbol{C} \boldsymbol{e}(z), \end{aligned} \quad (16)$$

where $\boldsymbol{e}^H(z) = [1 \quad z^{-1} \quad z^{-2} \quad \ldots \quad z^{-(M-1)}]$, and $\boldsymbol{C}$ is the $\mathbb{C}^{(M-1)\times(M-1)}$ matrix containing the information about the coefficients of the polynomial.

Ideally, after solving the polynomial, there will be P number of roots that would lie on the unit circle and represent the DOA angles for the incoming signals [29]. However, the presence of noise might deviate the roots from the unit circle. The actual angles of the sources are determined for each root as

$$\theta_k = \arcsin\left[\frac{\lambda}{2\pi d}\arg(z_k)\right], \quad k = 1, \ldots, P, \quad (17)$$

where $\arg(z_k)$ is the argument of the kth root on the unit circle, λ is the signal wavelength, and θ_k is the corresponding DOA angle for kth source. The method provides faster solution to the DOA estimation problem, eliminating the burden of spectral search present in the classical MUSIC. Moreover, Root-MUSIC is often preferable as it gives the DOA angles in scalar numbers in contrast to the spectral peaks in the classical method.

C. TF CONCEPTS

Time-frequency (TF) distribution of a signal can be derived using different approaches, which include Short Time Fourier Transform (STFT), Spectrogram and Gabor Transform,

Rihaczek Distribution and Wavelet Transform [30]. However, the smoothed version of the Wigner-Ville Distribution (WVD) is used in this paper. The main reason for using the WVD is that it represents a basic TF distribution, which can be transformed into other distributions by applying a certain kernel. We denote an analytic associate of the source signal $s(t)$ received by a microphone as $z(t)$. Assuming that $z_1(t)$ and $z_2(t)$ are two microphone signals, their distributions can be defined by the smoothed cross-WVD [30]:

$$\begin{aligned} D_{z_1\, z_2}(t,f) &= \mathcal{F}_{\tau\to f}\{G(t,\tau) \circledast (z_1(t+\frac{\tau}{2})\, z_2^*(t-\frac{\tau}{2}))\} \\ &= \mathcal{F}_{\tau\to f}\{G(t,\tau) \circledast K_{z_1 z_2}\}, \end{aligned} \tag{18}$$

where f is the discrete frequency defined with frequency bins, τ is the time lag, $\mathcal{F}_{\tau\to f}$ stands for the Discrete Fourier Transform (DFT) operator, and $\circledast$ is the convolution operation. $K_{zz}(t,\tau)$ is known as an instantaneous autocorrelation function (IAF). The time-lag kernel, $G(t,\tau)$, is applied to reduce the unnecessary artifacts of cross-terms in the TFD [31]. In fact, the time-lag kernel acts as a filter implemented by different window functions such as Hann and Hamming, Rectangular, Flat-top and other windows.

D. TF-MUSIC

TF-MUSIC method proposed in [18], [20] is the variation of the classical MUSIC method. The method incorporates the knowledge in the TF domain to preprocess the signals such that it has better noise performance for non-stationary sources. The TF preprocessing stage implies the operations performed on the conventional covariance matrix. As a result, the covariance matrix is replaced with a spatial time-frequency distribution (STFD) matrix. The STFD matrix contains auto- and cross-TFDs of signals from all array elements. As the matrix accounts for the number of microphones, the third spatial dimension is introduced. If the stationarity assumption is removed [31], the covariance matrix in (3) becomes time-dependent:

$$\boldsymbol{R}_{xx}(t,\tau) = \boldsymbol{A}(\theta)\boldsymbol{R}_{ss}(t,\tau)\boldsymbol{A}(\theta)^H + \sigma_n^2\boldsymbol{I}, \tag{19}$$

where $\boldsymbol{R}_{xx}(t,\tau)$ is the $\mathbb{C}^{M\times M}$ covariance matrix of the recorded signals at specific time instant and corresponding time-lag, $\boldsymbol{R}_{ss}(t,\tau)$ is the $\mathbb{R}^{P\times P}$ covariance matrix of the source signals, σ_n^2 is the additive noise variance, and $\boldsymbol{I}$ is the $\mathbb{R}^{M\times M}$ identity matrix.

The covariance function can also be described as

$$\begin{aligned} \boldsymbol{R}_{xx}(t,\tau) &= \mathbb{E}\{\boldsymbol{K}_{xx}(t,\tau)\} \\ &= \mathbb{E}\left\{\boldsymbol{x}\left(t+\frac{\tau}{2}\right)\boldsymbol{x}^H\left(t+\frac{\tau}{2}\right)\right\}, \end{aligned} \tag{20}$$

where $\boldsymbol{K}_{xx}$ is the matrix with the IAF of the corresponding signals as each entry. Rewriting the expression in (19) by using (20) and (18), the STFD matrix of the recorded signals can be derived as

$$\boldsymbol{D}_{xx}(t,f) = \boldsymbol{A}(\theta)\boldsymbol{D}_{ss}(t,f)\boldsymbol{A}(\theta)^H + \sigma_n^2\boldsymbol{I}. \tag{21}$$

where $\boldsymbol{D}_{ss}(t,f)$ represents the STFD matrix of the source signals, where the diagonal and off-diagonal entries correspond to auto-TFDs and cross-TFDs, accordingly.

At the moment, the structure of $\boldsymbol{D}_{xx}(t,f)$ is different from conventional covariance matrix in a way that each entry of the STFD matrix is a TF distribution. Therefore, there is a need to reduce the distributions to a scalar number, which is done by following four important steps. The first step is to find the averaged distribution of diagonal entries of (21). The cross-terms in the spectrum are reduced thanks to this spatial averaging, which is given by

$$\boldsymbol{D}_{\text{avg}}(t,f) = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{D}_{mm}(t,f), \tag{22}$$

where $\boldsymbol{D}_{\text{avg}}$ represents the averaged TFD, and $\boldsymbol{D}_{mm}(t,f)$ is the TFD of the mth sensor signals. This effect is explained by averaging the auto-terms located on the same spots on the TF planes while the cross-terms are allocated in different spots. The second step is to reduce the noise in the TF distribution by applying a threshold to the energy of each TF point in $\boldsymbol{D}_{\text{avg}}(t,f)$. The idea is to select those points with energies higher than a user-defined threshold, Φ, and reject the ones with low energies that may come from noise [31]. The threshold is given by

$$\Phi = \phi\ \max(\boldsymbol{D}_{\text{avg}}(t,f)), \tag{23}$$

where ϕ denotes the threshold in percents. The third step is to multiply the averaged TFD to all entries of $\boldsymbol{D}_{xx}(t,f)$ as if it were a TF kernel. This allows reducing the noise in all TFDs. The final step for obtaining the STFD matrix, which would be suitable for the DOA estimation, is determining the averages of the chosen TF points after setting the threshold:

$$\boldsymbol{D}_{xx} = \frac{1}{V}\sum_{i=1}^{V}\boldsymbol{D}_{xx}(t_i,f_i), \tag{24}$$

where V represents the total number of TF points. After these steps, the derived $\boldsymbol{D}_{xx}$ matrix has scalar values as entries similar to the conventional covariance matrix. Therefore, the same subspace methods can be applied to the STFD matrix. According to [20], the STFD matrix guarantees better signal selectivity thanks to different TF signatures of source signals. Moreover, it is advantageous in terms of noise suppression as the power of noise is spread over the entire TF plane.

E. SPATIALLY SMOOTHED TF-ROOT-MUSIC

Several variations of the conventional MUSIC method were discussed in the previous section. As they involve certain signal processing occurring at different stages of the method, their combination produces an method that promises to simultaneously handle both coherent and non-stationary and to have a better noise performance. The distinctive feature of the method is related to the spatially smoothed STFD matrix.

TABLE 1. The computational complexity analysis of the proposed method in comparison with the methods discussed in this paper.

	Number of Multiplications per iteration		
	Analytical Expression	Example 1	Example 2
MUSIC	$M^3(B+1)+M^2(2B-PB+N)+MB$	98790	772040
Root-MUSIC	$4M^3+M^2(N-P-2)+2P$	19156	67208
TF-MUSIC	$BM^3+M^2(N^3+N^2+N+(5N+2)log(N)+BP)+MB+N^2$	4.8422×10^9	6.8791×10^{10}
Proposed method	$5M^3+M^2(N^3+N^2+N+(5N+2)log(N)-P)+M+N^2+2P$	4.8421×10^9	6.8790×10^{10}

Example 1: $M=6; P=2; B=361; N=512$
Example 2: $M=8; P=4; B=1801; N=1024$

After the TF preprocessing of the input audio signals, the STFD matrix $\boldsymbol{D}_{xx}$ is derived as in (24). The spatial smoothing technique is further applied to the matrix in contrast to the immediate spectral estimation step in TF-MUSIC. Hence, the total STFD matrices of the forward and backward smoothing, $\boldsymbol{D}^f$ and $\boldsymbol{D}^b$, which represent the counterparts of (11) and (13), are derived by averaging all STFD matrices of L subarrays:

$$\boldsymbol{D}^f = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{D}_l^f = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{U}_l\boldsymbol{D}_{xx}\boldsymbol{U}_l^H, \tag{25}$$

$$\boldsymbol{D}^b = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{D}_l^b = \frac{1}{L}\sum_{l=1}^{L}\boldsymbol{U}_l\boldsymbol{J}\boldsymbol{D}_{xx}^H\boldsymbol{J}\boldsymbol{U}_l^H, \tag{26}$$

where $\boldsymbol{D}_l^f$ and $\boldsymbol{D}_l^b$ are the STFD matrices of lth subarray in forward and backward smoothing, respectively, and $\boldsymbol{D}_{xx}^H$ is the conjugate of the non-smoothed STFD matrix in (24). The following step involves averaging (25) and (26) to find the spatially smoothed STFD matrix as

$$\boldsymbol{D}^{fb} = \frac{\boldsymbol{D}^f+\boldsymbol{D}^b}{2}. \tag{27}$$

The aforementioned properties inherent to the smoothed covariance matrix in (14) also apply to the smoothed STFD matrix. However, the main distinction is that the latter allows to estimate the DOA angles not only of coherent signals but also of non-stationary signals. After having the matrix formed, the actual DOA angles are estimated using the rules of Root-MUSIC method. The pseudo code for the proposed method (spatially smoothed TF-Root-MUSIC) is given in Algorithm 1.

F. COMPUTATIONAL COMPLEXITY ANALYSIS

In practice, the computational complexity of the methods is as important as their DOA estimation performance. This section is thus dedicated to analyse and compare the computational complexity of the proposed and baseline methods. The computational complexity analysis is carried out on the basis of the number of multiplications required per iteration of the respective algorithms, and is summarized in Table 1 where B, M, P, and N denote the sampling grid of potential DOA angles, the number of microphones, the number of sources, and the snapshots, respectively. Here the last two columns represent numerical examples for the simulation situations considered later. It can be observed that TF-MUSIC and the proposed method are computationally more complex than MUSIC and Root-MUSIC as they include TF preprocessing. When comparing MUSIC and Root-MUSIC, the cost of the former one is larger due to the spectral search, i.e., the evaluation of (8) on the sampling grid of $B > M$ spatial frequency. It is noticed that the Root-MUSIC offers the lowest computational complexity, followed by that of the basic MUSIC method. TF-MUSIC and the proposed methods offer a comparable computational complexity, which is in fact, higher as compared with the computational complexity of the basic MUSIC and Root-MUSIC methods. The reason for this high computational complexity lies in the fact that these methods require TF preprocessing and the spatial search.

As demonstrated by extensive computer simulations (described later in Section V), the basic MUSIC and Root-MUSIC may fail in many practical scenarios for DOA estimation. Furthermore, the proposed (spatially smoothed TF-Root-MUSIC) method outperforms the TF-MUSIC method for coherent and both stationary and non-stationary signals with low SNR. It is also more precise compared to MUSIC and TF-MUSIC as it does not depend on the resolution of the spatial grid. Therefore, the increased computational complexity may be considered as the price paid for an improved performance not achievable with the basic methods.

V. SIMULATIONS

To test the performances of the discussed and the proposed methods, the simulation setup in Fig. 3 is used as a baseline. Depending on the type of a case study, additional assumptions and settings will be applied.

In the ULA, there are M omnidirectional microphones. The number of sound sources is equal to P. The relevant assumptions about the types of arriving signals from these sources will be discussed separately in the upcoming simulations. The reference direction is chosen from the geometrical middle of the microphone array and is labelled as $\theta_{\text{ref}} = 0^o$. The field-of-view is considered to be 180^o from -90^o to 90^0. To simulate the noisy environment, the white Gaussian noise will be added to the source signals. The assumptions of

Algorithm 1 Pseudo Code of the Proposed Method for DOA Estimation Using Spatially Smoothed TF-Root-MUSIC

Input:

$\boldsymbol{S} \leftarrow$ the matrix containing all microphone signals

Output:

DOA $\leftarrow$ a vector containing angular directions of all sources

Parameters:

$N \leftarrow$ the length of signals
$M \leftarrow$ the number of microphones
$P \leftarrow$ the number of sources
$d \leftarrow$ the array elements spacing
$w \leftarrow$ time-lag window
$\phi \leftarrow$ the percent value of the threshold
$\boldsymbol{D}_{\mathrm{TL}} \leftarrow$ time-lag matrix

for $i = 1$ to M **do**
  for $j = 1$ to M **do**
    $s_1 = \boldsymbol{S}(i, :)$
    $s_2 = \boldsymbol{S}(j, :)$
    $K = 2\,(2^{\mathrm{nextpow2}(N)})$ ▷ new signal length for FFT calculations
    $S_1 = \mathrm{fft}(s_1, K)$
    $S_2 = \mathrm{fft}(s_2, K)$
    if $f \leq K/2$ **then**
      double $S_1(f)$ and $S_2(f)$
    else
      set $S_1(f)$ and $S_2(f)$ to zero
    end if
    $z_1 = \mathrm{ifft}(S_1)$ ▷ find the analytic associate signal
    $z_2 = \mathrm{ifft}(S_2)$
    $\boldsymbol{D}_{\mathrm{TL}}(\tau, t) = w(\tau) \circledast z_1(t + \tau/2) \circledast (z_2(t + \tau/2)'$
    $\boldsymbol{D}(i, j) = \mathrm{fft}(\boldsymbol{D}_{\mathrm{TL}}, K/2)$ ▷ time-frequency representation of s_1 and s_2
  end for
end for
$\boldsymbol{D}_{\mathrm{avg}} \leftarrow$ average of diagonal entries of $\boldsymbol{D}$
$\Phi = \phi \max(\boldsymbol{D}_{\mathrm{avg}})$ ▷ user-defined threshold
if $\mathrm{abs}(\boldsymbol{D}_{\mathrm{avg}}(t, f)) \geq \epsilon$ **then**
  $\boldsymbol{D}_{\mathrm{avg}}(t, f) = 1$
else
  $\boldsymbol{D}_{\mathrm{avg}}(t, f) = 0$
end if
for $m_1 = 1$ to M **do**
  for $m_2 = 1$ to M **do**
    $\boldsymbol{D}(m_1, m_2) = (1/n_{\mathrm{points}})\mathrm{sum}(\boldsymbol{D}(m_1, m_2)\boldsymbol{D}_{\mathrm{avg}})$
  end for
end for
$\boldsymbol{J} = \mathrm{flip}(\mathrm{eye}(M))$ ▷ exchange matrix
$\boldsymbol{D}_y = \boldsymbol{J}(\boldsymbol{D})'\boldsymbol{J}$
$\boldsymbol{D}^{fb} = \boldsymbol{D} + \boldsymbol{D}_y$ ▷ forward-backward spatially smoothed STFD matrix
$\boldsymbol{q} \leftarrow$ eigenvectors of $\boldsymbol{D}^{fb}$
$\boldsymbol{Q} = \boldsymbol{q}(:, P : M)$ ▷ noise subspace
$\boldsymbol{C} = \boldsymbol{Q}\boldsymbol{Q}'$ ▷ matrix with diagonal entries representing the polynomial coefficients
$\boldsymbol{B} \leftarrow$ store the diagonal elements of $\boldsymbol{C}$
$\boldsymbol{r} \leftarrow$ solve for roots of $\boldsymbol{B}$ and store P number of roots close to unit circle
for $p = 1$ to P **do**
  $\mathrm{DOA}(n) = \arcsin\left(\mathrm{phase}(\boldsymbol{r}(p))\lambda\right)/(2\pi d)\ 180/\pi$
end for

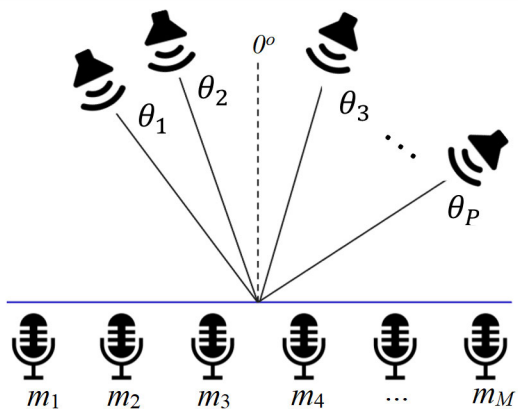


FIGURE 3. Simulation setup diagram with ULA microphones.

free- and far-field propagation space are made in the model. The size of subarrays, K, is equal to the number of microphones in all experiments where the spatial smoothing is applied.

The performances of the discussed methods will be examined throughout several experiments. The Root Mean Square Error (RMSE) is chosen as the performance metric in all simulations, given by

$$\text{RMSE} = \sqrt{\frac{1}{P}\sum_{p=0}^{P}(\hat{\theta}_p - \theta_p)^2}, \tag{28}$$

where $\hat{\theta}_p$ and θ_i represent the predicted and ground truth DOA angles of the source s_p. Being the difference between two angles, RMSE is measured in degrees.

A. CASE 1: STATIONARY SINUSOIDAL SOURCES

In this case study two experiments have been performed. The first experiment is dedicated to study the behaviours of the methods for $P = 2$ non-coherent and stationary sinusoidal source signals arriving from $\theta_1 = -40^o$ and $\theta_2 = 20^o$. The signals are considered to be clean. The array consists of $M = 6$ microphones. The angular frequencies of the signals are equal to $\pi/2$ and $\pi/4$. It should be noted that as the outputs of the classical MUSIC and TF-MUSIC are the spectral plots, their peak values are manually selected. Fig. 4 represents the spectral plots for the first simulation. The numerical results of Root-MUSIC and the proposed methods are derived and presented in Table 2. It is observed that all four methods estimate DOA values accurately primarily thanks to the simplicity of the applied signal conditions.

The second experiment aims to observe the results of the methods when given $P = 3$ sinusoidal sources arriving from $\theta_1 = -40^o$, $\theta_2 = 20^o$, and $\theta_3 = 30^o$, where the first and the last sources are coherent. Hence, the corresponding angular frequencies are equal to $\pi/2$, $\pi/4$, and $\pi/2$. The spectral plots of MUSIC and TF-MUSIC for this scenario are given in Fig. 5. Both of the methods failed to locate the coherent sources at -40^o and 30^o. The results of all methods

TABLE 2. The numerical results of the methods for clean, non-coherent, and stationary sinusoidal sources in Case 1.

Method	$S_1(-40^o)$	$S_2(20^o)$
MUSIC	-40	20
Root-MUSIC	-40	20
TF-MUSIC	-40	20
Proposed Method	-40	20

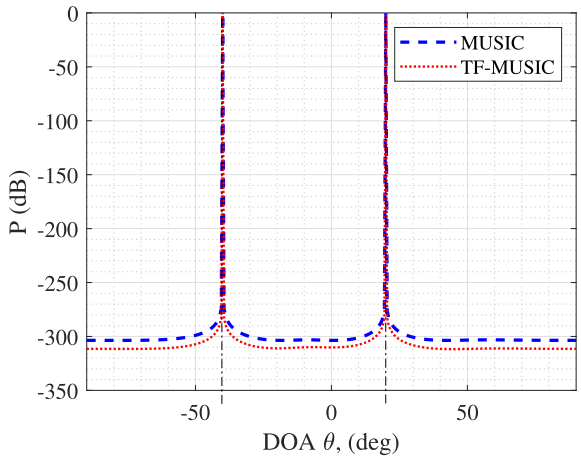


FIGURE 4. Spatial spectrum plots of MUSIC and TF-MUSIC for clean, non-coherent, and stationary sinusoidal sources in Case 1.

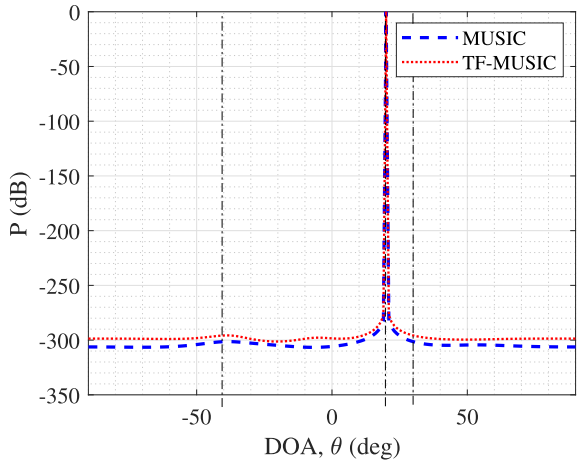


FIGURE 5. Spatial spectrum plots of MUSIC and TF-MUSIC for clean, coherent and stationary sinusoidal sources in Case 1.

are illustrated in Table 3. Although Root-MUSIC determined all three source signals, there are some discrepancies with the first and the third sources. In contrast, the proposed method is the only one which estimated all DOA values accurately.

B. CASE 2: NON-STATIONARY SOURCES

This case study also comprises of two experiments simulated to observe the effects of: 1) non-coherent and non-stationary sources, and 2) coherent and non-stationary sources, both corrupted with noise. The first experiment considers $P = 4$

TABLE 3. The numerical results of the methods for clean, coherent and stationary sinusoidal sources in Case 1.

Method	$S_1(-40^o)$	$S_2(20^o)$	$S_3(30^o)$
MUSIC	-	20	-
Root-MUSIC	-49.5	20	33.5
TF-MUSIC	-	20	-
Proposed Method	-40	20	30

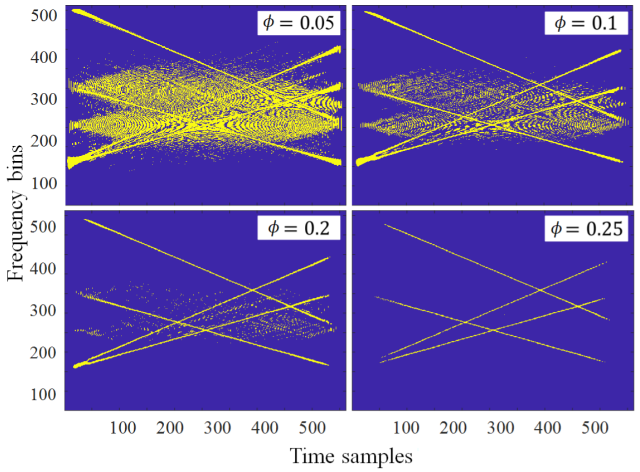


FIGURE 6. TF distribution of the signals with different threshold values ϕ for signals in Case 2.

Linear Frequency Modulated (LFM) source signals arriving from $\theta_1 = -40^o$, $\theta_2 = 20^o$, $\theta_3 = 30^o$, and $\theta_4 = 38^o$ to the array of $M = 6$ with the white Gaussian noise of SNR $= -5$ dB. The angular frequencies of the signals were linearly modulated from $[\frac{\pi}{5}, \pi, \frac{\pi}{5}, \frac{3\pi}{5}]$ to $[\frac{4\pi}{5}, \frac{2\pi}{5}, \frac{3\pi}{5}, \frac{\pi}{5}]$. The effect of noise threshold on the TF distribution of the signals are depicted in Fig. 6. It is noticed that the contribution of noise is removed when the threshold Φ is around 25% of the maximum TF point. By setting the threshold to this value, the corresponding spectral plots of MUSIC and TF-MUSIC are shown in Fig. 7. It can be seen that the conventional MUSIC fails to locate the sources at 20^o and 30^o, whereas TF-MUSIC shows the peaks at correct angles as a result of the applied TF threshold. The overall results in Table 4 suggest that both TF-MUSIC and the proposed TF-Root-MUSIC show accurate performance for non-coherent and non-stationary sources due to the presence of TF preprocessing step.

In the second experiment, $P = 4$ acoustic sources are considered coherent and non-stationary corrupted with the noise of SNR $= -5$ dB. There are $M = 6$ microphones receiving the signals from $\theta_1 = -30^o, \theta_2 = -25^o, \theta_3 = 40^o$, and $\theta_4 = 80^o$. The angular frequencies linearly vary from $[\frac{\pi}{5}, \pi, \frac{\pi}{5}, \frac{3\pi}{5}]$ to $[\frac{3\pi}{5}, \frac{2\pi}{5}, \frac{3\pi}{5}, \frac{\pi}{5}]$. The threshold value for the TF preprocessing is remained as 25% of the maximum TF point's value. The spectral plots of MUSIC and TF-MUSIC are represented in Fig. 8 and the corresponding numerical results for all methods are presented in Table 5. Fig. 8 shows

TABLE 4. The numerical results of the methods for non-coherent and non-stationary sources under noise of SNR $= -5$ dB in Case 2.

Method	$S_1(-40^o)$	$S_2(20^o)$	$S_3(30^o)$	$S_4(40^o)$
MUSIC	-40	25	-	38
Root-MUSIC	-40.7	21.8	-	37.2
TF-MUSIC	-40	19.5	29.5	38
Proposed Method	-40.3	19.3	28.2	40.2

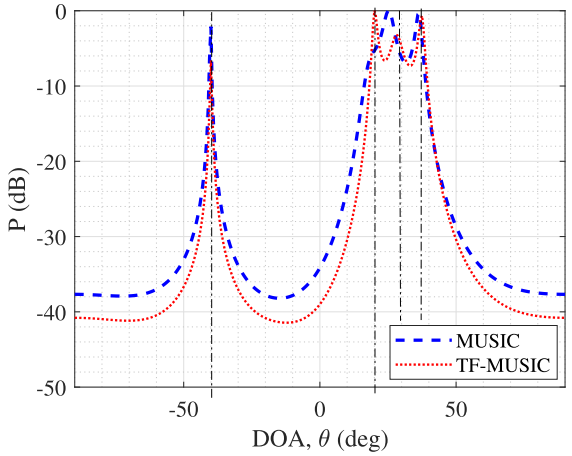


FIGURE 7. Spatial spectrum plots of MUSIC and TF-MUSIC for non-coherent and non-stationary sources under noise of SNR $= -5$ dB in Case 2.

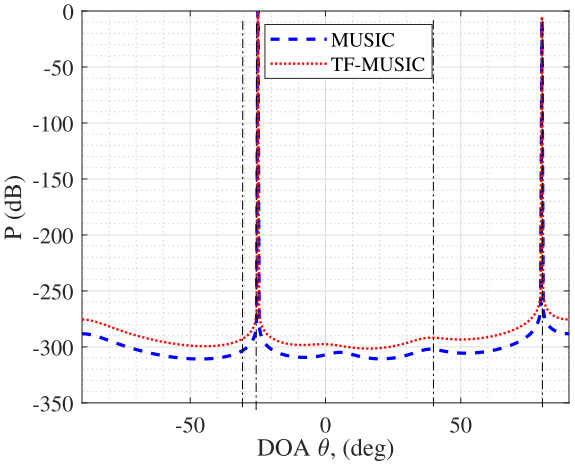


FIGURE 8. Spatial spectrum plots of MUSIC and TF-MUSIC for coherent and non-stationary sources under noise of SNR $= -5$ dB in Case 2.

that both MUSIC and TF-MUSIC could not locate the sources at -30^o and 40^o. The numerical results in Table 5 demonstrate that only the proposed method accurately estimates DOA values of all coherent and non-stationary sources under noisy conditions.

C. CASE 3: EFFECT OF NUMBERS OF MICROPHONES

One of the several parameters that influence the performance of subspace-based methods is the number of microphones M.

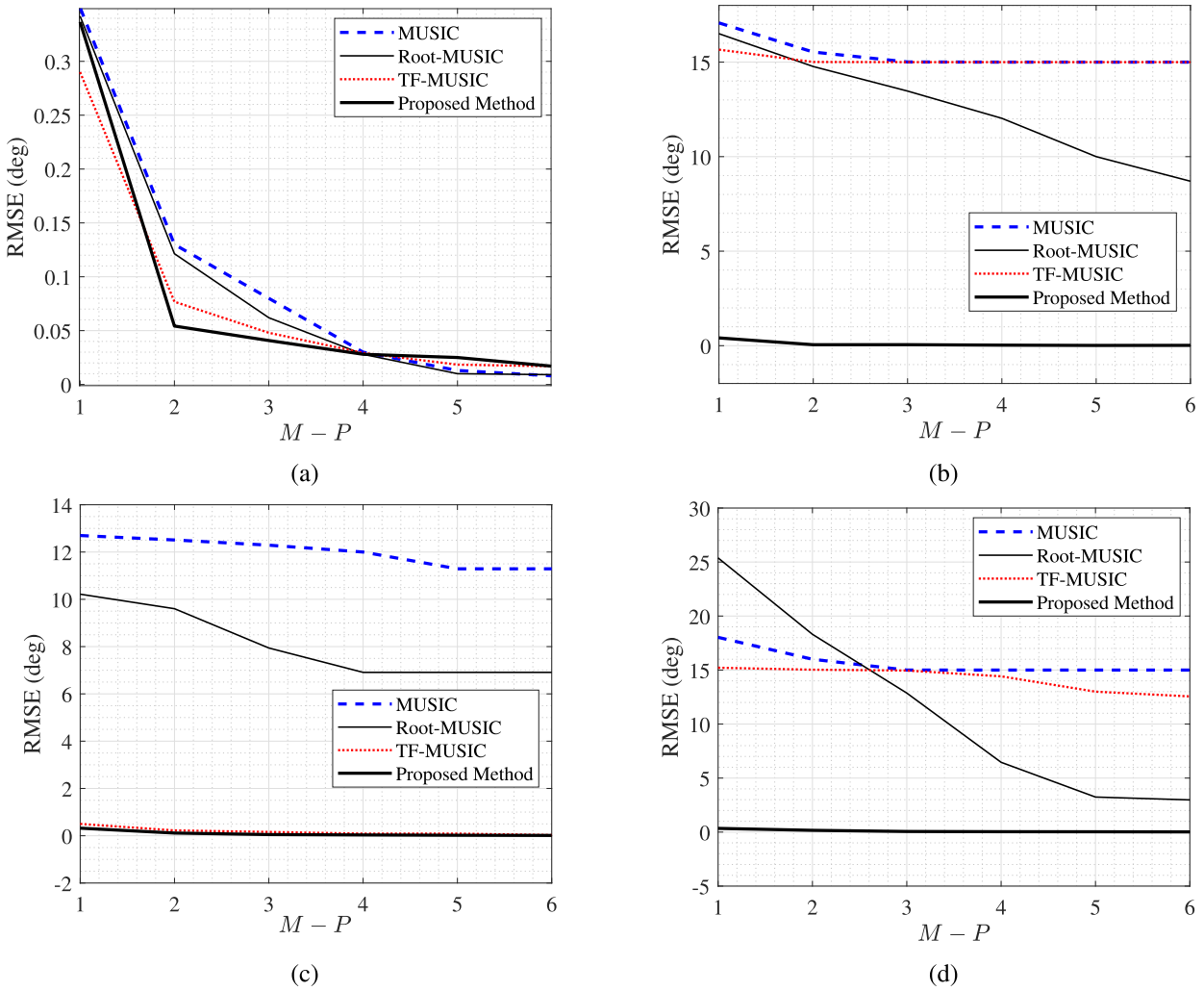


FIGURE 9. Performance of the methods for various numbers of microphones in Case 3 for: a) non-coherent and stationary sinusoidal signals; b) coherent and stationary sinusoidal signals; c) non-coherent and non-stationary chirp signals; d) coherent and non-stationary chirp signals.

TABLE 5. The numerical results of the methods for coherent and non-stationary sources under noise of SNR = −5 dB in Case 2.

Method	$S_1(-30^o)$	$S_2(-25^o)$	$S_3(40^o)$	$S_4(80^o)$
MUSIC	-	-25	-	80
Root-MUSIC	-	-25	33.6	80
TF-MUSIC	-	-25	-	80
Proposed Method	-30	-25	40	80

However, the methods discussed in this paper operate only when the number of microphones is larger than the number of sources $M > P$, which is known as the over-determined case. Thus, the RMSE values for different values of $M - P$ are calculated and illustrated in Fig. 9. The methods were tested with four different signals: 1) non-coherent and stationary sinusoidal signals; 2) coherent and stationary sinusoidal signals; 3) non-coherent and non-stationary chirp signals; 4) coherent and non-stationary chirp signals. The common trend for all subplots is the decrease in error with the increase of $M-P$ value. Considering the details, Fig. 9(a) suggests that all methods perform similarly for the non-coherent and stationary sinusoidal signal. The highest error does not exceed 0.5^o even when the number of microphones is more than that of sources by one. From Fig. 9(b), it can be observed that the proposed method continues to show excellent performance for coherent and stationary sinusoidal signals, whereas other methods have errors around 15^o mainly due to the lack of spatial smoothing. When non-coherent and stationary signals are applied, we notice from Fig. 9(c) that MUSIC and Root-MUSIC methods do not perform well in contrast to the proposed method and TF-MUSIC. In the case of coherent and non-stationary chirp signals, the results in Fig. 9(d) are similar to Fig. 9(b), where the proposed method outperforms the other three methods.

D. CASE 4: EFFECT OF SNR

Although the case study 2 considered the effect of noise to a certain extent, the performances under various SNR values

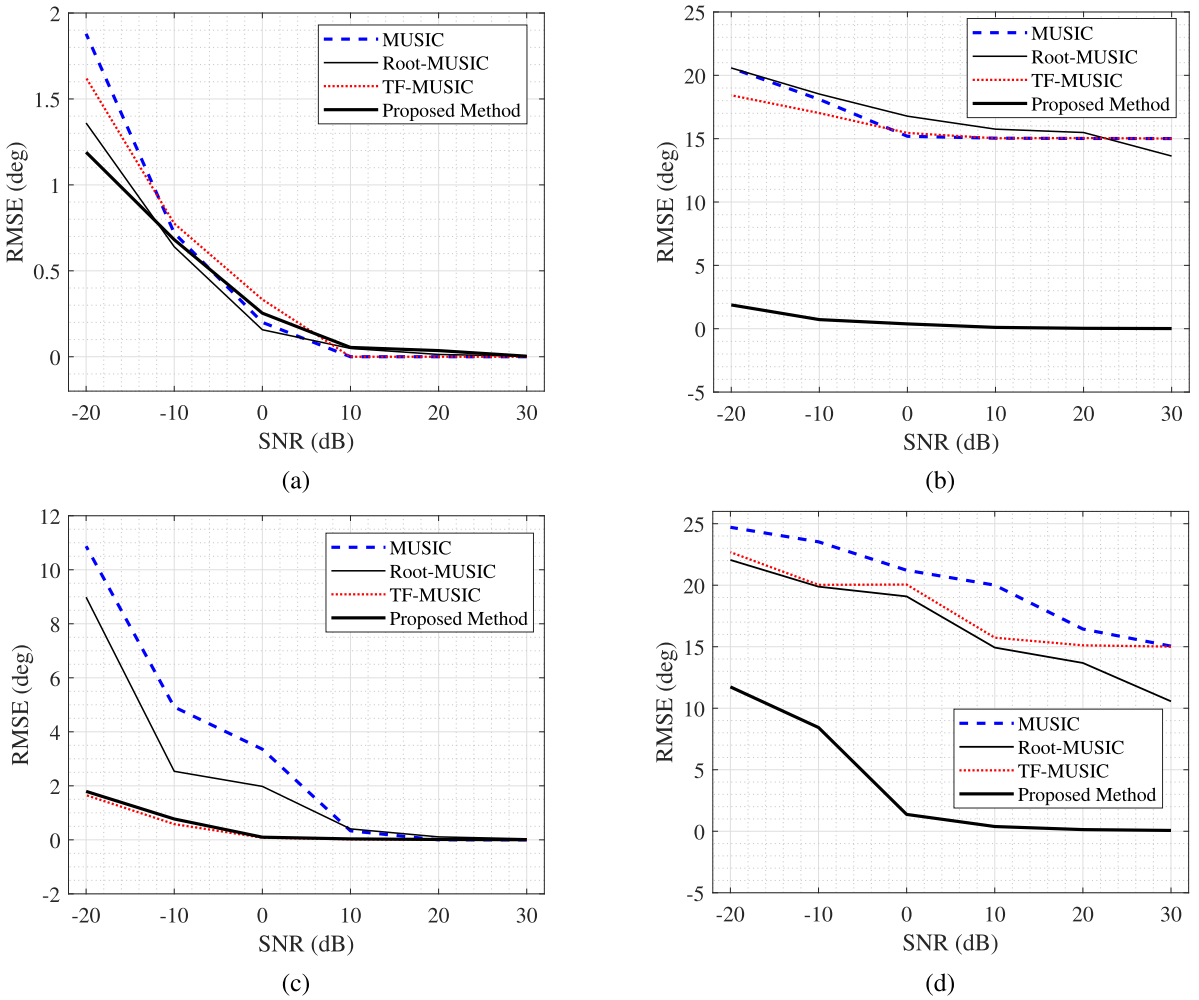


FIGURE 10. Performance of the methods at different levels of noise in Case 3 for: a) non-coherent and stationary sinusoidal signals; b) coherent and stationary sinusoidal signals; c) non-coherent and non-stationary chirp signals; d) coherent and non-stationary chirp signals.

should be rigorously addressed. The number of microphones is set as $M = 6$ and the threshold value is kept at 25%. The results for SNR values between -20 dB and 30 dB with a step of 10 dB are illustrated in Fig. 10. It can be observed from Fig. 10(a) that for non-coherent and stationary sinusoidal signals all methods have similar performance. All methods achieve significantly low error values at high SNR. However, the error values start increasing starting from SNR $= 0$ dB. At SNR $= -20$ dB, the proposed method has the lowest error of 1.19^o, followed by Root-MUSIC with 1.36^o, TF-MUSIC with 1.62^o, and MUSIC with the highest error of 1.879^o. Overall, due to the simplicity of the signal model, all four methods perform comparably. The methods show different results when processing coherent and stationary sinusoidal signals as illustrated in Fig. 10(b). The major observation is that the proposed spatially smoothed TF-Root-MUSIC has the lowest errors throughout all SNR values thanks to the application of spatial smoothing technique. Whereas the other methods have errors around 15^o even at SNR $= 30$ dB. For non-coherent and non-stationary chirp signals, the results are given in Fig. 10(c). One can observe that TF-MUSIC and TF-Root-MUSIC perform similarly, almost reaching RMSE value of 2^o under these conditions. Whereas the values for MUSIC and Root-MUSIC start increasing significantly after SNR $= 10$ dB, reaching the error value of 10.87^o and 8.98^o. Such a high difference in performance is owing to the TF preprocessing steps present in both TF-MUSIC and TF-Root-MUSIC methods. However, the proposed method surpasses TF-MUSIC and other methods when working with coherent and non-stationary chirp signals as shown in Fig. 10(d). The trends are similar to the results in Fig. 10(b), where the proposed method have the lowest errors at all SNR values. However, the difference is that the error increases notably after SNR $= 0$ dB for the proposed method, resulting in an error of 11.73^o at SNR $= -20$ dB.

E. CASE 5: EFFECT OF THRESHOLD ϕ

The performance of the proposed spatially smoothed TF-Root-MUSIC depends on the value of the user-defined threshold. This case study aims to observe the impact of changing its value. The signal model is maintained the same

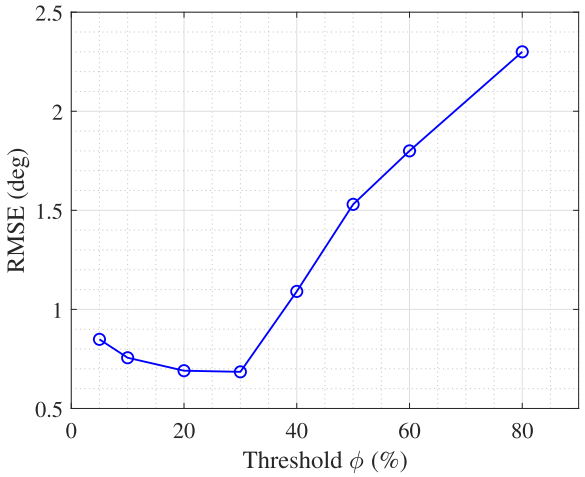


FIGURE 11. Performance of the proposed method at different threshold values in Case 5..

as in the second experiment of the case study 2. The results are depicted in Fig. 11. It is observed that the lowest error corresponds to the threshold value around 25% of the maximum TF point. At this value, most of the contributions of the noise and cross terms are eliminated. Hence, the noise performance of the method is increased in contrast to the classical MUSIC.

VI. CONCLUSION

A method for efficient sound source DOA estimation is proposed in the paper. From the simulation results, it can be observed that it performs comparably or better than other existing methods when given signals such as non-coherent sinusoidal and LFM signals. However, it outperforms them when handling both coherent and non-stationary sources under severe noise conditions. The paper discussed the details of three main constituent parts of the method, which are the derivation of STFD matrix, application of the spatial smoothing technique, and solving the polynomial to obtain numerical estimates of DOA values. The basic simulations showed the method's robustness in various types of source signals and different configurations of other parameters such as the threshold value and the number of microphones. The case studies illustrated solid RMSE performances for varying number of microphones as well as SNR levels. An optimal threshold value was empirically found for the given signal settings. The possible application of machine learning algorithms to select the optimal value of the user-defined threshold is planned to be investigated in the future.

REFERENCES

[1] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robot. Auton. Syst.*, vol. 96, pp. 184–210, Oct. 2017.

[2] L. Wan, G. Han, L. Shu, S. Chan, and T. Zhu, "The application of DOA estimation approach in patient tracking systems with high patient density," *IEEE Trans. Ind. Informat.*, vol. 12, no. 6, pp. 2353–2364, Dec. 2016.

[3] Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilistic neural network," *IEEE Trans. Ind. Electron.*, vol. 65, no. 8, pp. 6403–6413, Aug. 2018.

[4] N. Bnilam, D. Joosens, R. Berkvens, J. Steckel, and M. Weyn, "AoA-based localization system using a single IoT gateway: An application for smart pedestrian crossing," *IEEE Access*, vol. 9, pp. 13532–13541, 2021.

[5] I. A. H. Adam and M. R. Islam, "Perfomance study of direction of arrival (DOA) estimation algorithms for linear array antenna," in *Proc. Int. Conf. Signal Process. Syst.*, May 2009, pp. 268–271.

[6] P. Handel, P. Stoica, and T. Soderstrom, "Capon method for doa estimation: Accuracy and robustness aspects," in *Proc. IEEE Winter Workshop Nonlinear Digit. Signal Process.*, Jan. 1993, pp. 1–7.

[7] F. Akbari, S. S. Moghaddam, and V. Tabataba Vakili, "MUSIC and MVDR DOA estimation algorithms with higher resolution and accuracy," in *Proc. 5th Int. Symp. Telecommun.*, Dec. 2010, pp. 76–81.

[8] Z.-M. Liu, C. Zhang, and P. S. Yu, "Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections," *IEEE Trans. Antennas Propag.*, vol. 66, no. 12, pp. 7315–7327, Dec. 2018.

[9] T. B. Lavate, V. K. Kokate, and A. M. Sapkal, "Performance analysis of MUSIC and ESPRIT DOA estimation algorithms for adaptive array smart antenna in mobile communication," in *Proc. 2nd Int. Conf. Comput. Netw. Technol.*, Apr. 2010, pp. 308–311.

[10] H. Abeida and J.-P. Delmas, "Gaussian Cramer-Rao bound for direction estimation of noncircular signals in unknown noise fields," *IEEE Trans. Signal Process.*, vol. 53, no. 12, pp. 4610–4618, Dec. 2005.

[11] J. L. Navarro-Mesa, M. J. Millan-Munoz, and E. Herndndez-Perez, "An approach to DOA estimation of wide-band sources based on ar signal modeling," in *Proc. Process. Workshop Sensor Array Multichannel Signal*, Barcelona, Spain, 2004, pp. 323–326.

[12] N. Hu, Z. Ye, D. Xu, and S. Cao, "A sparse recovery algorithm for DOA estimation using weighted subspace fitting," *Signal Process.*, vol. 92, no. 10, pp. 2566–2570, Oct. 2012.

[13] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 36, no. 4, pp. 532–544, Mar. 1988.

[14] B. D. Rao and K. V. S. Hari, "Performance analysis of root-music," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1939–1949, Dec. 1989.

[15] J. Zhao, R. Gui, X. Dong, and S. Wu, "Time-varying DOA tracking algorithm based on generalized labeled multi-Bernoulli," *IEEE Access*, vol. 9, pp. 5943–5950, 2021.

[16] J. Wen, B. Liao, and C. Guo, "Spatial smoothing based methods for direction-of-arrival estimation of coherent signals in nonuniform noise," *Digit. Signal Process.*, vol. 67, pp. 116–122, Aug. 2017.

[17] Q. Chen and R. Liu, "On the explanation of spatial smoothing in MUSIC algorithm for coherent sources," in *Proc. Int. Conf. Inf. Sci. Technol.*, Mar. 2011, pp. 699–702.

[18] M. G. Amin and Y. Zhang, "Direction finding based on spatial time-frequency distribution matrices," *Digit. Signal Process.*, vol. 10, no. 4, pp. 325–339, Oct. 2000.

[19] A. Belouchrani and M. G. Amin, "Time-frequency MUSIC," *IEEE Signal Process. Lett.*, vol. 6, no. 5, pp. 109–110, May 1999.

[20] A. Belouchrani, M. G. Amin, N. Thirion-moreau, and Y. D. Zhang, "Source separation and localization using time-frequency distributions: An overview," *IEEE Signal Process. Mag.*, vol. 30, no. 6, pp. 97–107, Nov. 2013.

[21] R. Zhagypar, K. Zhagyparova, and M. T. Akhtar, "Exploiting the rules of the TF-MUSIC and spatial smoothing to enhance the DOA estimation for coherent and non-stationary sources," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Auckland, New Zealand, Dec. 2020, pp. 236–241.

[22] H. Tang, "DOA estimation based on MUSIC algorithm," M.S. thesis, Dept. Phys. Elect. Eng., Fac. Technol., Linnaeus Univ., Kalmar, Sweden, May 2015.

[23] S. Kung, C. Lo, and R. Foka, "A toeplitz approximation approach to coherent source direction finding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 11, Apr. 1986, pp. 193–196.

[24] J. K. Hammond and P. R. White, "The analysis of non-stationary signals using time-frequency methods," *J. Sound Vib.*, vol. 190, no. 3, pp. 419–447, 1996.

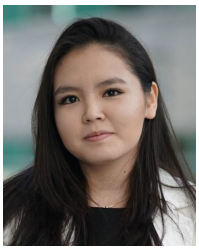
[25] F. Gustafsson and F. Gunnarsson, "Positioning using time-difference of arrival measurements," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Hong Kong, Apr. 2003, pp. VI-553.

[26] V. Ingle, S. Kogon, and D. Manolakis, *Statisical and Adaptive Signal Processing*. Norwood, MA, USA: Artech House, 2005.

[27] J. Pan, M. Sun, Y. Wang, and X. Zhang, "An enhanced spatial smoothing technique with ESPRIT algorithm for direction of arrival estimation in coherent scenarios," *IEEE Trans. Signal Process.*, vol. 68, pp. 3635–3643, 2020.
[28] K. V. S. Hari and U. Gummadavelli, "Effect of spatial smoothing on the performance of subspace methods in the presence of array model errors," *Automatica*, vol. 30, no. 1, pp. 11–26, Jan. 1994.
[29] H. K. Hwang, Z. Aliyazicioglu, M. Grice, A. Yakovlev, and P. Lu, "Direction of arrival estimation using polynomial root intersection for multidimensional estimation (prime)," in *Proc. IMECS Conf.*, 2008, pp. 1–6.
[30] B. Boashash, *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*, 2nd ed. Cambridge, MA, USA: Academic, 2016, pp. 331–518.
[31] B. Boashash and A. Aïssa-El-Bey, "Robust multisensor time–frequency signal processing: A tutorial review with illustrations of performance enhancement in selected application areas," *Digit. Signal Process.*, vol. 77, pp. 153–186, Jun. 2018.

RUSLAN ZHAGYPAR (Student Member, IEEE) was born in Kyzylorda, Kazakhstan, in 1998. He received the B.Eng. degree in electrical and electronics engineering from the School of Engineering and Digital Sciences, Nazarbayev University, Nur-Sultan, Kazakhstan, in 2021.

Since 2019, he has been working as a Research Assistant with the Applications of Signal Processing Laboratory (ASP-LAB), Nazarbayev University. His research interests include sound source localization (SSL), namely, direction-of-arrival (DOA) estimation, and the applications of machine learning algorithms for this problem. The current manuscript is based on his final-year Capstone Project.



KALAMKAS ZHAGYPAROVA (Student Member, IEEE) was born in Nur-Sultan, Kazakhstan, in 1999. She received the B.Eng. degree in electrical and electronics engineering from the School of Engineering and Digital Sciences, Nazarbayev University, Nur-Sultan, in 2021.

Since 2019, she has been working as a Research Assistant with the Applications of Signal Processing Laboratory, Nazarbayev University. Her research interests include the sound source localization (SSL), namely, distance estimation and the applications of machine learning algorithms for this problem. Her final-year Capstone Project is concerned with studying and implementation of different methods for SSL.

MUHAMMAD TAHIR AKHTAR (Senior Member, IEEE) received the B.Sc. degree in electrical electronics and communication engineering from the University of Engineering and Technology, Taxila, Pakistan, in 1997, the M.Sc. degree in systems engineering from Quaid-i-Azam University, Islamabad, Pakistan, in 1999, and the Ph.D. degree in electronic engineering from Tohoku University, Sendai, Japan, in 2004,

From 2004 to 2005, he was a COE Postdoctoral Fellow with Tohoku University, Sendai. From 2006 to 2008, he worked as an Assistant Professor with United Arab Emirates University, United Arab Emirates. From December 2008 to February 2009, he was a Visiting Researcher with the Institute of Sound and Vibration Research (ISVR), University of Southampton, U.K. From 2008 to 2014, he was an Assistant Professor with the University of Electro-Communications, Tokyo, Japan, and a Special Visiting Researcher with the Tokyo Institute of Technology, Tokyo. From November 2010 to March 2011, he was with the Institute for Neural Computations (INC), University of California, San Diego. From 2014 to 2017, he was an Associate Professor with COMSATS University Islamabad, Pakistan. He is currently working as an Associate Professor with the School of Engineering and Digital Sciences, Nazarbayev University, Nur-Sultan, Kazakhstan. His research interests include adaptive signal processing, active noise control, blind source separation, and biomedical signal processing. He has published about 95 articles in the peer-reviewed international journals and conference proceedings.

Dr. Akhtar is a member of the IEEE Signal Processing Society and IEEE Industrial Electronic Society. He was the Editorial Board Member of *Advances in Mechanical Engineering*. He received the Best Student Paper at the IEEE 2004 Midwest Symposium on Circuits and Systems, Hiroshima, Japan, and the Student Paper Award (with Marko Kanadi) at the 2010 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing. From 2011 to 2013, he has served as a Co-Editor for the Newsletter of the Asia-Pacific Signal and Information Processing Association (APSIPA).