Source Counting and Separation Based on Simplex Analysis

Bracha Laufer-Goldshtein, *Student Member, IEEE*, Ronen Talmon, *Member, IEEE*, and Sharon Gannot, *Senior Member, IEEE*

***Abstract*—Blind source separation (BSS) is addressed, using a novel data-driven approach, based on a well-established probabilistic model. The proposed method is specifically designed for separation of multichannel audio mixtures. The algorithm relies on spectral decomposition of the correlation matrix between different time frames. The probabilistic model implies that the column space of the correlation matrix is spanned by the probabilities of the various speakers across time. The number of speakers is recovered by the eigenvalue decay, and the eigenvectors form a simplex of the speakers' probabilities. Time frames dominated by each of the speakers are identified exploiting convex geometry tools on the recovered simplex. The mixing acoustic channels are estimated utilizing the identified sets of frames, and a linear unmixing is performed to extract the individual speakers. The derived simplexes are visually demonstrated for mixtures of 2, 3 and 4 speakers. We also conduct a comprehensive experimental study, showing high separation capabilities in various reverberation conditions.**

***Index Terms*—blind audio source separation (BASS), relative transfer function (RTF), spectral decomposition, simplex.**

I. Introduction

Blind source separation (BSS) is a core problem in signal processing with numerous applications in various fields, such as: biomedical data processing, audio processing, digital communication, and image processing [1]. In BSS problems, only the output observations are given, whereas neither the original sources nor the mixing systems are known. Separation methods usually rely on some *a priori* hypothesis regarding the characteristics of the original sources or the obtained mixtures. Assuming that the sources are independent and have non-Gaussian distributions leads to independent component analysis (ICA) methods based on probabilistic or information theoretic criteria [2]–[4]. Non-negative matrix factorization (NMF) methods can be employed for signals which admit factorization to non-negative components [5]. Sparsity of the signals is also often assumed, allowing a representation as a linear combination of few elementary signals [6].

In audio applications, the measured signals in an array of microphones represent convolutive mixtures of the source signals [7]–[9]. The measured signals are obtained by filtering the clean source signals with the corresponding acoustic channels relating the sources and the microphones. The acoustic channels, in a typical reverberant environment, consist of various reflections from the objects and surfaces defining the acoustic enclosure. The measured signals are commonly analysed in the short time Fourier transform (STFT) domain, where the convolutive mixtures are transformed into multiplicative mixtures at each frequency bin.

ICA-based methods can be applied, subject to scale-ambiguity and source permutation problems [10], [11]. Alternatively, numerous separation methods rely on the sparsity of speech sources in the STFT domain, assuming that each time-frequency (TF) bin is occupied by a single source [12]. In algorithms based on NMF, the speech spectrum is decomposed to a multiplication of non-negative basis and activation functions [13], [14]. Due to joint estimation of source parameters and mixing coefficients, these methods are free from permutation alignment problems. Other full-band approaches cluster the measurements according to time difference of arrival (TDOA) estimates or phase difference levels with respect to several microphones [15]–[17]. However, these models cannot be successfully applied in the presence of high reverberation, when the TDOA estimates are of poor quality. Robustness to room reverberations can be attained by performing bin-wise clustering, in the cost of adding a second stage of permutation alignment procedure [18]–[20]. The TIFROM algorithm [21] avoids the TF sparsity assumption. It inspects the variations of computed instantaneous ratios, and detects small regions in the TF plane with a single active speaker.

Source separation can also be achieved by applying beamformers [22], which are multichannel spatial filters designed by certain criteria, such as the linearly constrained minimum variance (LCMV) beamformer [23]. These algorithms are not completely blind, as their design requires some knowledge on the signal statistics or on their associated acoustic channels. In [23] the acoustic channels were estimated assuming the availability of known time intervals with interferences only and known time intervals comprising each of the desired speakers separately.

In this paper, we present a novel source separation algorithm, which is specifically applicable to speech mixtures. The key point lies in the spectral decomposition of the correlation matrix between different observations. The justification of the method is based on a probabilistic model, in which each observation consists of different portions of the hidden sources. The relative portion of each source is randomly generated according to the sources' probabilities, which vary from one observation to another. Based on this model, we show that the column space of the correlation matrix is spanned by the probabilities of the different sources. Accordingly, the rank of

Bracha Laufer-Goldshtein and Sharon Gannot are with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel (e-mail: Bracha.Laufer@biu.ac.il, Sharon.Gannot@biu.ac.il); Ronen Talmon is with the Viterbi Faculty of Electrical Engineering, The Technion-Israel Institute of Technology, Technion City, Haifa 3200003, Israel, (e-mail: ronen@ee.technion.ac.il).

the correlation matrix equals the number of sources, and its eigenvectors form a simplex of the sources' activity probabilities. The vertices of the simplex correspond to observations dominated by a single source with high probability, facilitating the estimation of the hidden sources.

The applicability of the presented model for blind separation of speech mixtures relies on two main attributes of multichannel audio mixtures. The first is the sparsity of the speech in the STFT domain, implying that different time-frames contain different portions of speech components of the different speakers. The second is the fact that in a multichannel framework each speaker is associated with a unique *spatial signature*, manifested in the associated acoustic channel. Applying the above procedure and exploiting convex geometry tools, we can identify frames dominated by a single speaker, enabling estimation of the corresponding acoustic channels. Given the estimated acoustic channels, the individual speakers are extracted using the pseudo-inverse of the acoustic mixing system. It is important to note that our model relies on the sparsity assumption for the acoustic channel estimation, yet, the separation is based on standard unmixing rather than masking techniques, thus avoiding artifacts often attributed to masking-based approaches.

A. Related Work

Our method recovers a simplex of the probability of activity of the different sources. Convex geometry tools are more commonly utilized for hyperspectral unmixing (HU) in the emerging field of hyperspectral remote sensing [24], [25]. In those studies, the goal is to identify materials in a scene, using hyperspectral images with high spectral resolution. The work relies on a linear mixing model, where each pixel is modelled as a linear sum of the radiated energy curves of the materials contained in this pixel. The nature of the problem entails a positivity constraint on the weights of the different materials. In addition, the weights must sum to one due to energy conservation. The latter constraint violates the statistical independence assumption, making the application of many standard BSS algorithms inappropriate. Alternatively, the above constraints lay the ground for the application of convex geometry tools for HU. There was also an attempt to borrow these principles for quasi-stationary sources separation such as speech [26], [27]. In general, it is clear that speech mixtures are not formed as convex mixtures. In [26], a certain normalization followed by a pre-processing procedure for cross-correlation mitigation, were proposed in order to enforce bin-wise convexity. In [27], the output of a phase-normalized steered response beamformer was used as a feature.

It is important to emphasize that the mixture model presented in this paper is fundamentally different from the one used for HU. In our model, we recover a simplex of the probability of activity of the different sources, while in HU the simplex is formed in the original (often high-dimensional) domain of the mixing systems. In addition, our method also inherently identifies the number of sources in the mixture, whereas HU methods generally assume that the number of sources is known. Moreover, in contrast to [26], we present a full-band approach based on averaging over a large number of frequency bins, which enhances robustness and avoids permutation problems.

It should be noted that convex hulls and simplex shapes arise also in other contexts, whenever dealing with objects whose weights sum to one, such as histograms. Probabilistic latent component analysis (PLCA), which is a probabilistic extension of NMF, is employed for source separation by modeling the mixture with independent distributions that lie in a simplex. This framework was combined with sparsity constraints in [28] for learning the latent variables of histogram data. The method learns an overcomplete set of bases, which form a convex hull surrounding the data distributions. Another extension of PCLA was presented in [29] to deal with separation of sources that have some common basis vectors. In [30] a new nearest-subspace representation for sound mixtures was derived. The authors formulated the search as a sparse coding problem with l_2 regularization, leading to a solution that lies on a vertex of the weight simplex. In contrast to these models, the proposed method recovers a simplex of the probability of activity of the speakers across time.

B. Contributions

The proposed method has several advantages over most of the above-mentioned separation algorithms. First, most separation schemes assume that the number of speakers is known, while other methods focus solely on the task of speaker counting. The proposed method carries out a combined speaker counting and separation task. In addition, this method has a low computational cost since it does not contain iterations, as opposed to iterative methods such as EM-based approaches. It is also more efficient than bin-wise clustering methods, since it is based on a full-band approach, which does not require a permutation alignment stage. The method is also free of any initialization procedures, which are often required by separation algorithms, such as NMF-based methods. In terms of performance, we show in the experimental part that the proposed method obtains high separation scores in various reverberation levels, and has an advantage over an NMF-based method [14] and a DNN-based concurrent speaker detector [31]. The proposed method is also shown to be much more computationally efficient with respect to an independent vector analysis (IVA) algorithm [32].

The paper is organized as follows. The probabilistic model and its analysis by convex geometry principles are presented in Section II. The model is applied to speech mixtures and an algorithm for speaker counting and separation is derived in Section III. Section IV contains an extensive experimental study demonstrating the performance of the proposed method in comparison to several competing methods. Section V concludes this paper.

II. Statistical Mixture Model and Analysis

We present a general statistical model describing the generation of a collection of observations as mixtures of a set of hidden sources. The observations consist of different portions of each of the sources, where each source occurs with a certain

probability. The separation is based on the computation of the correlation matrix defined over the given observations. Based on the spectral decomposition of the correlation matrix, we can identify the number of hidden sources and derive a simplex representation, which relates each observation with its corresponding probabilities. In Section III, we discuss the relation between this general model and the problem of blind separation of speech mixtures. We use the analogy between the two to derive an algorithm for estimating the number of active speakers and separating them.

A. Mixture Generation

Consider J unknown *hidden sources* $\{\mathbf{h}_j\}_{j=1}^{J}$. The hidden sources are i.i.d. random vectors consisting of D *coordinates*, i.e. $\mathbf{h}_j \in \mathbb{R}^D$, where the kth coordinate of the jth source is denoted by $h_j(k)$. The hidden sources follow a multivariate distribution with zero-mean and identity covariance matrix, i.e.:

$$E\left\{\mathbf{h}_j \mathbf{h}_j^T\right\} = \mathbf{\Upsilon}_D \tag{1}$$

where $\mathbf{\Upsilon}_D$ is the identity matrix of size $D \times D$. The diagonal covariance matrix implies that the coordinates of the hidden sources are assumed to be uncorrelated. It should be noted that the unit variance assumption is used here for the sake of simplicity, and that the following derivation also holds for non-unit and non-constant variance by applying a proper normalization.

Suppose we are given a set of $L \gg J$ observations $\{\mathbf{a}(l)\}_{l=1}^{L}$, also in $\mathbb{R}^D$. Each observation $\mathbf{a}(l)$ is formed by the J hidden sources. Each coordinate in $\mathbf{a}(l)$ equals the corresponding entry of $\mathbf{h}_j$ with probability $p_j(l)$, where the probabilities of all sources sum to one $\sum_{j=1}^{J} p_j(l) = 1$. Accordingly, the kth coordinate of the lth observation can be written as:

$$a(l,k) = \sum_{j=1}^{J} I_j(l,k) h_j(k) \tag{2}$$

where $I_j(l,k) \sim Br(p_j(l))$ is an indicator function following a Bernoulli distribution with parameter $p_j(l)$, i.e. $I_j(l,k)$ equals 1 if the jth source occupies the kth coordinate of $\mathbf{a}(l)$ and 0 otherwise. The indicator functions satisfy:

$$\begin{aligned} \sum_{j=1}^{J} I_j(l,k) &= 1 \\ I_j(l,k) I_i(l,k) &= I_j(l,k)\delta_{ij} \end{aligned} \tag{3}$$

where $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ otherwise. We further assume that the indicator functions of different coordinates and of different frames are mutually independent.

According to this statistical model, for each l, the probability $p_j(l)$ corresponds to the *relative portion* of the jth source in the construction of the observation $\mathbf{a}(l)$. An illustration of the presented mixture model is depicted in Fig. 1 for $J = 3$ sources, $D = 10$ coordinates and $L = 6$ observations. Consider for example the first observation $\mathbf{a}(1)$, with associated probabilities: $p_1(1) = 0.5$, $p_2(1) = 0.3$ and $p_3(1) = 0.2$. In the vector $\mathbf{a}(1)$, 5 coordinates are taken from $\mathbf{h}_1$, 3 coordinates are taken from $\mathbf{h}_2$, and 2 coordinates are taken from $\mathbf{h}_3$. In practice, the relative portion of each source only approximately matches the corresponding probability for D large enough.

TABLE I
ANALOGY BETWEEN GENERAL MODEL AND SPEECH MIXTURES

General Model	Speech Mixtures
J hidden sources	Unknown relative transfer function (RTF) vectors of the J speakers
A collection of L observations	L instantaneous RTFs computed for each frame in the STFT representation
D vector coordinates	Concatenation of RTF values in different frequencies and all microphones (excluding the reference microphone)
Observations consists of different portions of the hidden sources	Sparsity of speech in the STFT representation - each TF bin is occupied by a different speaker

The motivation for this model comes from separation of speech mixtures. According to the sparsity assumption of speech sources in the STFT domain [12], each TF bin is dominated by a single speaker. Given the spectrogram of the mixed signal, we can define a column vector for each frame index, consisting of the STFT values in a certain frequency band. Relying on the sparsity assumption, each frequency bin in this vector contains a signal from a single speaker. The challenge in speech mixtures, is that they are time-varying. In Section II-A we mitigate this problem by proposing features based on the acoustic channels, which are approximately fixed as long as the environment and the source positions do not change dramatically. Specifically, we use features based on the relative transfer functions (RTFs), which are defined as the ratio between the transfer functions of each of the microphones and the reference microphone. Accordingly, for speech mixtures, the hidden vectors correspond to the RTF vectors, and their dimension D is proportional to the length of the chosen frequency band multiplied by the number of microphones (excluding the reference microphone). The analogy between the general model of Section II-A and the model of speech mixtures of Section III is summarized in Table I.

B. Analysis of the Correlation Matrix

Our goal is to recover the number J of hidden sources $\{\mathbf{h}_j\}_{j=1}^{J}$ and to estimate them based on the given set of observations $\{\mathbf{a}(l)\}_{l=1}^{L}$. The key to our separation scheme lies in the spectral decomposition of the correlation matrix defined over the different observations, which is analysed in this section.

Based on the assumed statistical model (1, 2, 3), the correlation between each two observations $\mathbf{a}(l)$ and $\mathbf{a}(n)$, $1 \le l, n \le L$ is given by (for details refer to Appendix A):

$$E\left\{\frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)\right\} = \begin{cases} \sum_{j=1}^{J} p_j(l)p_j(n) & \text{if } l \neq n \\ 1 & \text{if } l = n \end{cases}. \tag{4}$$

Let $\mathbf{W}$ be the $L \times L$ correlation matrix, with $W_{ln} = E\left\{\frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)\right\}$. According to (4) the correlation matrix can be recast as:

$$\mathbf{W} = \mathbf{P}\mathbf{P}^T + \Delta\mathbf{W} \tag{5}$$

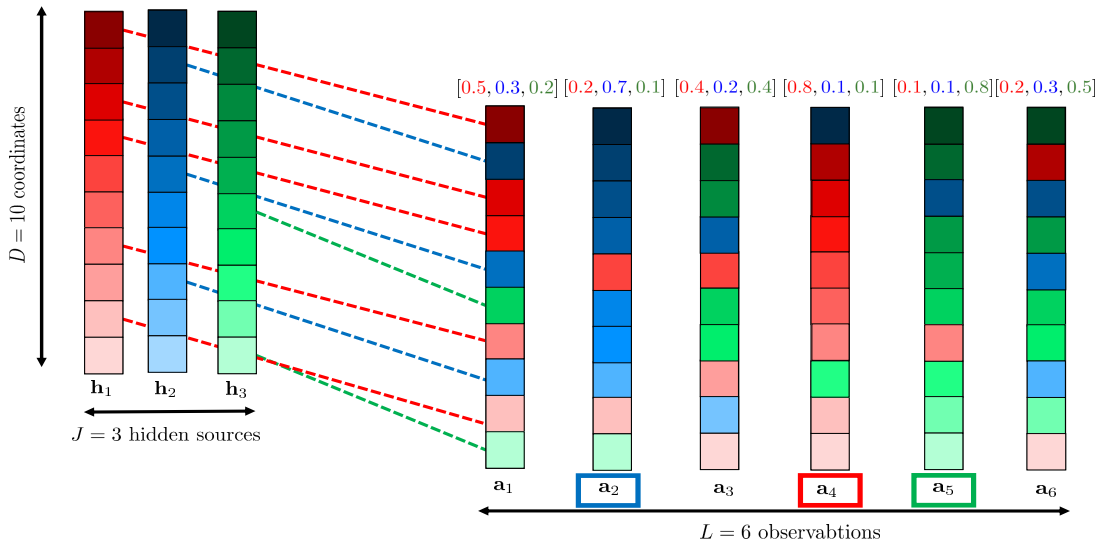


Fig. 1. An illustration of the presented statistical mixture model. In this example there are $J = 3$ hidden sources $\{\mathbf{h}_j\}_{j=1}^3$ consisting of $D = 10$ coordinates, characterized by varying shades of red, blue and green, respectively. The hidden sources are used to construct $L = 6$ observations $\{\mathbf{a}(l)\}_{l=1}^6$, where each coordinate is taken from a different source. For the first observation $\mathbf{a}(1)$ dashed lines are drown between each coordinate and the associated coordinate of the source from which it was taken. The set of probabilities $[p_1(l), p_2(l), p_3(l)]$ used to construct each observation is written above it. Note that in this example for each observation the number of coordinates taken from each source exactly matches the corresponding probability, while in practice it is only approximately satisfied. Note also that three observations out of the six are highly dominated by a specific source (occupies at least 70% of the observation coordinates). The second observation $\mathbf{a}(2)$ is dominated by the second source $\mathbf{h}_2$. The fourth observation $\mathbf{a}(4)$ is dominated by the first source $\mathbf{h}_1$. The fifth observation $\mathbf{a}(5)$ is dominated by the third source $\mathbf{h}_3$.

where $\mathbf{P}$ is a $L \times J$ matrix with $P_{lj} = p_j(l)$, and $\Delta\mathbf{W}$ is a diagonal matrix with $\Delta W_{ll} = 1 - \sum_{j=1}^{J} p_j^2(l)$. We show in Appendix B, that $\Delta\mathbf{W}$ has a negligible effect on the spectral decomposition of $\mathbf{W}$. Therefore, henceforth we omit $\Delta\mathbf{W}$ from our derivations and consider the correlation matrix as $\mathbf{W} \approx \mathbf{P}\mathbf{P}^T$.

Following the mutual independence assumption of the sources, the columns of $\mathbf{P}$ are linearly independent, i.e. the rank of $\mathbf{P}$ equals the number of sources J. Hence, the rank of $\mathbf{W}$ also equals J, i.e. it has J nonzero eigenvalues. We apply an eigenvalue decomposition (EVD) $\mathbf{W} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, with $\mathbf{U}$ an orthonormal matrix consisting of the eigenvectors $\{\mathbf{u}_j\}_{j=1}^L$, and $\mathbf{D}$ a diagonal matrix with the eigenvalues $\{\lambda_j\}_{j=1}^L$ on its diagonal. The eigenvalues $\{\lambda_j\}_{j=1}^L$ are sorted by their values in a descending order. According to (5), the first J eigenvectors $\{\mathbf{u}_j\}_{j=1}^J$, associated with the J nonzero eigenvalues $\{\lambda_j\}_{j=1}^J$, form a basis for the column space of the matrix $\mathbf{P}$. Accordingly, the following identity holds:

$$\mathbf{U}_\mathrm{J} = \mathbf{P}\mathbf{Q}^\mathbf{T} \tag{6}$$

where $\mathbf{U}_\mathrm{J} = [\mathbf{u}_1, \ldots, \mathbf{u}_J]$, and $\mathbf{Q}$ is a $J \times J$ invertible matrix.

Each observation can be represented as a point in $\mathbb{R}^J$, defined by the corresponding set of probabilities: $\mathbf{p}(l) = [p_1(l), p_2(l), \ldots, p_J(l)]^T$. Note that each point $\mathbf{p}(l)$ is a convex combination of the standard unit vectors:

$$\mathbf{p}(l) = \sum_{j=1}^{J} p_j(l)\mathbf{e}_j, \ \sum_{j=1}^{J} p_j(l) = 1. \tag{7}$$

where $\mathbf{e}_j = [0, \ldots, 1, \ldots, 0]^T$ with one in the jth coordinate and zeros elsewhere. Accordingly, the collection of probability sets $\{\mathbf{p}(l)\}_{l=1}^L$ lies in a $(J-1)$-simplex in $\mathbb{R}^J$. This is a *standard simplex*, whose *vertices* are the standard unit vectors $\{\mathbf{e}_j\}_{j=1}^J$. Note that in this representation, points for which the probability of the jth source is dominant over the probabilities of the other sources, i.e. $p_j(l) \gg p_i(l)$, $\forall i \neq j$, $1 \leq i \leq J$, satisfy: $\mathbf{p}(l) \approx \mathbf{e}_j$, namely these points are concentrated nearby the jth vertex.

We can use the eigenvectors of $\mathbf{W}$ to form an equivalent representation in $\mathbb{R}^J$, defined by: $\boldsymbol{\nu}(l) = [u_1(l), u_2(l), \ldots, u_J(l)]^T$. According to (6), this representation is related to the former representation by the following transformation:

$$\boldsymbol{\nu}(l) = \mathbf{Q}\mathbf{p}(l). \tag{8}$$

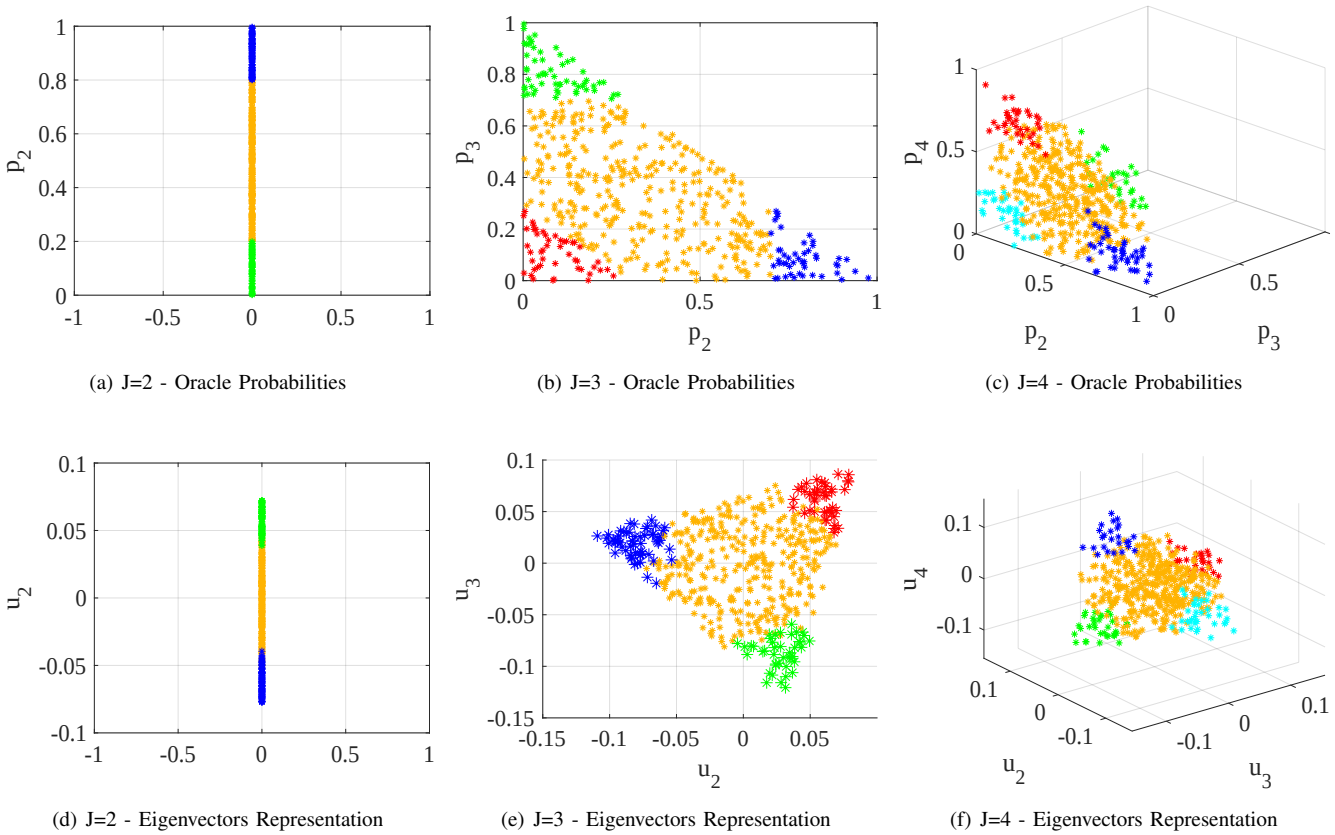
Hence, the set $\{\boldsymbol{\nu}(l)\}_{l=1}^L$ occupies a simplex, which is a rotated and scaled version of the standard simplex defined by the standard unit vectors. The new simplex is the convex hull of the following J vertices:

$$\mathbf{e}_j^* = \mathbf{Q}\mathbf{e}_j = \mathbf{Q}_j \tag{9}$$

where $\mathbf{Q}_j$ is the jth column of the matrix $\mathbf{Q}$.

Regarding the computation of the matrix $\mathbf{W}$, we do not have access to the expected values $\frac{1}{D}E\{\mathbf{a}^T(l)\mathbf{a}(n)\}, \forall 1 \leq l, n \leq L$, hence we use instead the typical values $\widehat{W}_{ln} = \frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)$. In Appendix A, we show that the variance of $\frac{1}{D}E\{\mathbf{a}^T(l)\mathbf{a}(n)\}$ is proportional to $1/D$, hence approaches zero for D large enough, implying that the typical value is close to the expected value.

We demonstrate the above derivation using three examples with $J = 2$, $J = 3$ and $J = 4$ sources. We generate J independent sources of dimension $D = 1000$ with $h_j(k) \sim \mathcal{N}(0, 1)$ Next, we generate $L = 500$ observations, $\{\mathbf{a}(l)\}_{l=1}^L$ according to (2). To generate the probabilities $\{p_j(l)\}_{j=1}^J$ for each l, we



(a) J=2 - Oracle Probabilities (b) J=3 - Oracle Probabilities (c) J=4 - Oracle Probabilities

(d) J=2 - Eigenvectors Representation (e) J=3 - Eigenvectors Representation (f) J=4 - Eigenvectors Representation

Fig. 2. (a)-(c) Scatter plots of oracle probabilities $\{\mathbf{p}(l)\}_{l=1}^{L}$ representing the expected relative portion each source in the construction of each observation. (d)-(f) Scatter plots of the mappings $\{\boldsymbol{\nu}(l)\}_{l=1}^{L}$ obtained from the eigenvectors of the estimated correlation matrix $\widehat{\mathbf{W}}$ between different observations. Scatter plots correspond to mixtures of $J=2$ (a) (d), $J=3$ (b) (e) and $J=4$ (c) (f) sources. Red, blue, green and cyan points stand for observations dominated by a single source, whereas orange points stand for observations with multiple sources. A line shape is obtained for $J=2$ (a), a triangle shape is obtained for $J=3$ (b) and a tetrahedron shape is obtained for $J=4$ (c). Rotated and scaled versions of theses shapes are formed in the scatter plots (d)-(f) of the mappings $\{\boldsymbol{\nu}(l)\}_{l=1}^{L}$, as implied by (8).

draw $J-1$ uniform variables between $[0,1]$ and sort them in an ascending order: $\rho_1(l) < \rho_2(l) < \ldots < \rho_{J-1}(l)$. Accordingly, for each l, we define the probability of each source by: $p_1(l) = \rho_1(l)$, $p_j(l) = \rho_j(l) - \rho_{j-1}(l)$, $\forall 2 \leq j \leq J-1$ and $p_J(l) = 1 - \rho_{J-1}(l)$. Next, we construct the matrix $\widehat{\mathbf{W}}$ with $\widehat{W}_{ln} = \frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)$, and apply an EVD.

Figure 2 (a)-(c) depicts $\{\mathbf{p}(l)\}_{l=1}^{L}$, for $J=2$ (a), $J=3$ (b) and $J=4$ (c). To enable visualization also for $J=4$ we omit the first coordinate of $\mathbf{p}(l)$, and represent the simplexes in $\mathbb{R}^{J-1}$. The colouring of the points is as follows: blue, green, red and cyan for observations dominated by the first, the second, the third, and the fourth source, respectively (for $J=3$ only blue, green and red, and for $J=2$ only blue and green). Orange points depict frames with mixture of sources. We observe that in each plot the points form a $(J-1)$-simplex, i.e. a line segment (a), a triangle (b) and a tetrahedron (c).

Figure 2 (d)-(f) depicts $\{\boldsymbol{\nu}(l)\}_{l=1}^{L}$, for $J=2$ (d), $J=3$ (e) and $J=4$ (f). The coloring of the points is the same as in Fig. 2 (a)-(c). We observe that the scattering of the points in (d)-(f) represents a linear transformation of the scattering in (a)-(c), as implied by (8).

Figure 3 depicts the computed eigenvalues of the estimated correlation matrix $\widehat{\mathbf{W}}$, sorted in a descending order. We

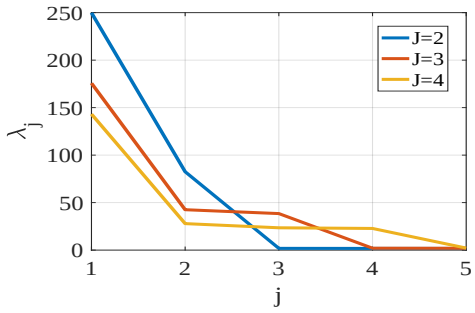


Fig. 3. The values of the first 5 eigenvalues $\{\lambda_j\}_{j=1}^{5}$ of the estimated correlation matrix $\widehat{\mathbf{W}}$ between different observations, obtained for mixtures with $J=\{2,3,4\}$ sources.

observe that the number eigenvalues with significant value above zero, exactly matches the number of sources J.

We conclude with the practical aspects of the new representation derived by the EVD of the matrix $\mathbf{W}$. By examining the rank of the obtained decomposition, we can estimate the number of sources involved in the construction of the set $\{\mathbf{a}(l)\}_{l=1}^{L}$. Furthermore, the eigenvectors $\{\mathbf{u}_j\}_{j=1}^{J-1}$ form a simplex that corresponds to the probability of activity of each

source along the observation index $1 \leq l \leq L$. We can use this representation to identify observations, which are highly dominated by a certain source, i.e. with $p_j(l) \gg p_i(l), \forall i \neq j$, implying $\mathbf{a}(l) \approx \mathbf{h}_j$. The identified observations can be used for estimating the original J hidden sources $\{\mathbf{h}_j\}_{j=1}^J$.

III. Speaker Counting and Separation

In this section, we devise a statistical model for speech mixtures, which resembles the model presented in Section II-A. Next, we use the analysis of Section II-B to derive an algorithm for speaker counting and separation.

A. Speech Mixtures

Consider J concurrent speakers, located in a reverberant enclosure. The signals are measured by an array of M microphones. The measured signals are analysed in the STFT domain with a window of length N samples and overlap of η samples:

$$Y^m(l,f) = \sum_{j=1}^{J} Y_j^m(l,f) = \sum_{j=1}^{J} A_j^m(f) S_j(l,f) \tag{10}$$

where $A_j^m(f)$ is the acoustic transfer function (ATF) relating the jth speaker and the mth microphone, and $S_j(l,f)$ is the signal of the jth speaker. Here $f \in \{1,\ldots,K\}$ is the frequency bin and $l \in \{1,\ldots,L\}$ is the frame index.

The first microphone ($m = 1$) is considered as the reference microphone. We define the relative transfer function (RTF) [33], [34] as the ratio between the ATF of the mth microphone and the ATF of the reference microphone, both of which are associated with the jth speaker:

$$H_j^m(f) = \frac{A_j^m(f)}{A_j^1(f)}. \tag{11}$$

In order to transform the measurements (10) into features that correspond to the model presented in Section II-A, we rely on two main assumptions. The first assumption regards the fact that each speaker has a unique spatial signature, which is manifested in the associated RTF (11). The second assumption regards the sparsity of speech signals in the STFT domain.

For speech mixtures, the J hidden sources are defined by the RTFs of each of the speakers. Each hidden source $\mathbf{h}_j$ consists of $D = 2 \cdot (M-1) \cdot F$ coordinates for the real and imaginary parts of the RTF values, in $F \leq K$ frequency bins and in $M-1$ microphones:

$$\begin{aligned} \mathbf{h}_j^m &= \left[H_j^m(f_1), H_j^m(f_2), \ldots, H_j^m(f_F)\right]^T \\ \mathbf{h}_j^{\mathrm{c}} &= \left[\mathbf{h}_j^{2^T}, \mathbf{h}_j^{3^T}, \ldots, \mathbf{h}_j^{M^T}\right]^T \\ \mathbf{h}_j &= \left[\mathrm{real}\left\{\mathbf{h}_j^{\mathrm{c}}\right\}^T, \mathrm{imag}\left\{\mathbf{h}_j^{\mathrm{c}}\right\}^T\right]^T. \end{aligned} \tag{12}$$

Note that $\mathbf{h}_j^1$ is an all-ones vector for all $1 \leq l \leq L$, hence is excluded from $\mathbf{h}_j$ in (12). We assume that the RTF vectors have a diagonal covariance matrix (1). The attributes of the Fourier transform prescribe that the real and the imaginary parts of the RTF values, as well as the different frequency bins, are uncorrelated. For F large enough, the model can tolerate slight correlations between adjacent frequency bins, or between neighbouring microphones. In addition, we assume that the RTFs of the different speakers are mutually independent.

After defining the J the hidden vectors associated with each of the speakers, we have to extract related observations from the measured signals (10). We assume that low-energy frames do not contain speech components, and hence these frames are excluded from our analysis. We use the assumption of the speech sparsity in the TF domain [12], which is widely employed in the STFT analysis of speech mixtures, and is often applied for localization [17], [35], [36] and separation tasks [14], [19], [37]. According to [12], each TF bin is exclusively dominated by a single speaker. Let $I_j(l,f)$ denote an indicator function with expected value $p_j(l)$, which equals 1 if the jth speaker is active in the (l,f)th bin, and equals 0, otherwise. The assumption that the probability $p_j(l)$ is dependent on l but independent of f, reflects that the frequency components of a speech signal tend to be activated synchronously [19], [38]. According to the TF sparsity assumption, the following holds for each TF bin (recall (3)):

$$\begin{aligned} \sum_{j=1}^{J} I_j(l,f) &= 1 \\ I_j(l,f) I_i(l,f) &= I_j(l,f)\delta_{ij} \end{aligned} \tag{13}$$

Hence, (10) can be recast as:

$$Y^m(l,f) = \sum_{j=1}^{J} I_j(l,f) A_j^m(f) S_j(l,f). \tag{14}$$

We compute the following instantaneous ratio between the mth microphone and the reference microphone:

$$R^m(l,k) = \frac{Y^m(l,f)}{Y^1(l,f)} = \frac{\sum_{j=1}^{J} I_j(l,f) A_j^m(f) S_j(l,f)}{\sum_{j=1}^{J} I_j(l,f) A_j^1(f) S_j(l,f)}. \tag{15}$$

According to (11), (13) and (15), we get (recall (2)):

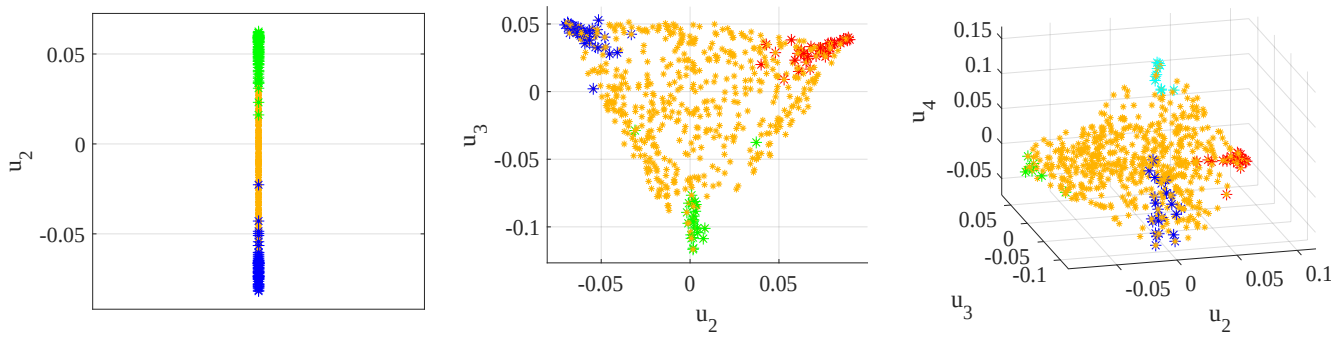
$$R^m(l,f) = \sum_{j=1}^{J} I_j(l,f) H_j^m(f) \tag{16}$$

implying that the ratio in the (l,f)th TF bin equals the RTF of one of the speakers. To obtain robustness, we replace the ratio in (16) by power spectra estimates averaged over $T+1$ frames around l [33]:

$$\tilde{R}^m(l,f) \equiv \frac{\hat{\Phi}_{y^m y^1}(l,f)}{\hat{\Phi}_{y^1 y^1}(l,f)} \equiv \frac{\sum_{n=l-T/2}^{l+T/2} Y^m(n,f) Y^{1*}(n,f)}{\sum_{n=l-T/2}^{n+T/2} Y^1(n,f) Y^{1*}(n,f)}. \tag{17}$$

Let $\mathbf{a}(l)$ denote the observed RTF of frame l, which consists of the real and imaginary parts of the RTF values, in F frequency bins and in $M-1$ microphones (recall (12)):

$$\begin{aligned} \mathbf{a}^m(l) &= \left[\tilde{R}^m(l,f_1), \tilde{R}^m(l,f_2), \ldots, \tilde{R}^m(l,f_F)\right]^T \\ \mathbf{a}^{\mathrm{c}}(l) &= \left[\mathbf{a}^{2^T}(l), \mathbf{a}^{3^T}(l), \ldots, \mathbf{a}^{M^T}(l)\right]^T \\ \mathbf{a}(l) &= \left[\mathrm{real}\left\{\mathbf{a}^{\mathrm{c}}(l)\right\}^T, \mathrm{imag}\left\{\mathbf{a}^{\mathrm{c}}(l)\right\}^T\right]^T. \end{aligned} \tag{18}$$



(a) J=2 - Eigenvectors Representation (b) J=3 - Eigenvectors Representation (c) J=4 - Eigenvectors Representation

Fig. 4. Scatter plots of the mappings $\{\boldsymbol{\nu}(l)\}_{l=1}^{L}$ obtained from the eigenvectors of the estimated correlation matrix $\widehat{\mathbf{W}}$ between different time frames. Scatter plots correspond to mixtures of $J=2$ (a), $J=3$ (b) and $J=4$ (c) speakers. Red, blue, green and cyan points stand for frames dominated by a single speaker, whereas orange points stand for frames with multiple active speakers. A line shape is obtained for $J=2$ (a), a triangle shape is obtained for $J=3$ (b) and a tetrahedron shape is obtained for $J=4$ (c). There is a good correspondence between the mappings obtained for speech mixtures and the mappings obtained in the synthetic example, presented in Fig. 2.

J	No. of sources/speakers, $j \in \{1,\ldots,J\}$
M	No. of microphones, $m \in \{1,\ldots,M\}$
L	No. of observations/frames in the STFT, $l \in \{1,\ldots,L\}$
F	No. of frequency bins in the chosen band, $f \in \{f_1,\ldots,f_F\}$
D	No. of coordinates $D = 2\times(M-1)\times F$, $k \in \{1,\ldots,D\}$
$\mathbf{h}_j$	Hidden sources defined by RTF values of each speaker
$\mathbf{a}(l)$	Observations defined by instantaneous RTFs of each frame
$\mathbf{p}(l)$	Probability of activity of the speakers in each frame
$\mathbf{W}$	Correlation matrix with $W_{ln} = \frac{1}{D}E\{\mathbf{a}^T(l)\mathbf{a}(n)\}$
$\{\lambda_j\}_{j=1}^{L}$	Eigenvalues of the correlation matrix $\mathbf{W}$
$\{\mathbf{u}_j\}_{j=1}^{L}$	Eigenvectors of the correlation matrix $\mathbf{W}$
$\boldsymbol{\nu}(l)$	A transformation of $\mathbf{p}(l)$, obtained by the eigenvectors of $\mathbf{W}$
$\{\mathbf{e}_j\}_{j=1}^{J}$	Vertices of the standard simplex occupied by $\{\mathbf{p}(l)\}_{l=1}^{L}$
$\{\mathbf{e}_j^*\}_{j=1}^{J}$	Vertices of the transformed simplex occupied by $\{\boldsymbol{\nu}(l)\}_{l=1}^{L}$

TABLE II
NOMENCLATURE

Note that for a certain frequency bin, the same speaker (both the real and the imaginary parts) is captured by all the microphones. However, this does not affect the relative portions of the different speakers in $\mathbf{a}(l)$, and has a negligible effect on the variance of the correlation (34) provided $F \gg M$. There is a trade-off choosing the frequency band $\{f_1,\ldots,f_F\}$. On the one hand, we should focus on the frequency band in which most of the speech components are concentrated in order to avoid TF bins with low-energy speech components. On the other hand, a sufficient broad frequency band should be used in order to reduce the effect of TF bins occupied by several speakers, and to obtain a better averaging with smaller variance (34).

We compute (17) and (18) for each $1 \le l \le L$, and form the set $\{\mathbf{a}(l)\}_{l=1}^{L}$. We conclude that the obtained set is constructed from the RTF vectors of the different sources (12), and has similar properties to the set of observations defined in Section II-A. A nomenclature listing the different symbols and their meanings is given in Table II.

B. Forming a Data-Driven Simplex

After we have shown that the speech separation problem can be formulated using the model in Section II-A, we would like to use the analysis of Section II-B to derive an algorithm for speaker counting and separation.

Following the derivation of Section II-B, we construct an $L \times L$ matrix $\widehat{\mathbf{W}}$ with $\widehat{W}_{ln} = \frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)$, and apply EVD. Based on the computed eigenvectors, we form a representation in $\mathbb{R}^J$, defined by: $\boldsymbol{\nu}(l) = [u_1(l), u_2(l), \ldots, u_J(l)]^T$.

We provide a similar demonstration for speech mixtures as we have presented in the syntactic case in Section II-B. We present three examples with $J=2$, $J=3$ and $J=4$ speakers. The generation of the mixtures and the associated parameters are described in details in the experimental part, in Section IV. Figure 4 depicts the points $\{\boldsymbol{\nu}(l)\}_{l=1}^{L}$, for $J=2$ (a), $J=3$ (b) and $J=4$ (c). Here to, we omit one coordinate of $\boldsymbol{\nu}(l)$ to enable visualization also for $J=4$. The plots in Fig. 4 are generated in a similar way to the plots in Fig. 2. We observe a good correspondence between Fig. 4 and Fig. 2, which gives evidence to the applicability of the general model of Section II to the case of speech mixtures.

Figure 5 depicts the computed eigenvalues sorted in a descending order, and normalized by the value of the maximum eigenvalue. As in Fig. 3, the number of eigenvalues with significant value above zero matches the number of sources J. Hence, we can estimate the number of sources in the mixture by:

$$\hat{J} = \left(\underset{j}{\operatorname{argmin}} \frac{\lambda_j}{\lambda_1} < \alpha\right) - 1 \tag{19}$$

where α is a threshold parameter.

C. Recovering The Activity of Speakers

We use the obtained representation $\{\boldsymbol{\nu}(l)\}_{l=1}^{L}$ to recover the probabilities of the speakers. Next, we detect frames, which are dominated by one of the speakers, and utilize them for estimating the corresponding RTFs. As discussed in Section II-B, the vertices of the simplex defined by $\{\boldsymbol{\nu}(l)\}_{l=1}^{L}$ correspond

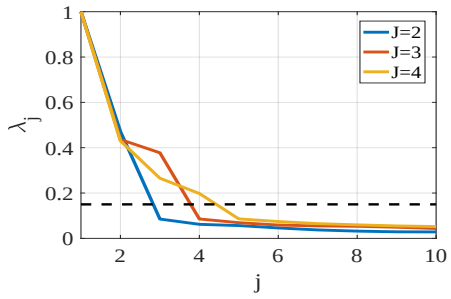


Fig. 5. The values of the first 5 eigenvalues $\{\lambda_j\}_{j=1}^5$ of the estimated correlation matrix $\widehat{\mathbf{W}}$ between different time frames, obtained for mixtures with $J = \{2, 3, 4\}$ speakers.

to single-speaker points. We recover the simplex vertices, and then utilize them to transform the obtained representation $\{\boldsymbol{\nu}(l)\}_{l=1}^L$ to the original probabilities $\{\mathbf{p}(l)\}_{l=1}^L$.

We assume that for each speaker there is at least one frame, with index l_j, which contains only this speaker, i.e. $\mathbf{p}(l_j) = \mathbf{e}_j$. The single-speaker frames are the simplex vertices, i.e. $\boldsymbol{\nu}(l_j) = \mathbf{e}_j^*$. Note that single-speaker frames are tantamount to pure pixels in HU. Several algorithms for identifying the vertices of a simplex were developed in the context of HU [39]–[41]. We use a simple approach based on the family of *successive projection* algorithms [42]. We first identify two vertices of the simplex, and then successively identify the remaining vertices by maximizing the projection onto the orthogonal complement of the space spanned by the previously identified vertices. We start with the first vertex, which is chosen as the point with the maximum norm:

$$\hat{\mathbf{e}}_1^* = \boldsymbol{\nu}(l_1),\ l_1 = \underset{1\le l\le L}{\operatorname{argmax}} \|\boldsymbol{\nu}(l)\|_2 \tag{20}$$

Then, the second vertex is chosen as the point with maximum distance with respect to the first identified vertex:

$$\hat{\mathbf{e}}_2^* = \boldsymbol{\nu}(l_2),\ l_2 = \underset{1\le l\le L}{\operatorname{argmax}} \|\boldsymbol{\nu}(l) - \hat{\mathbf{e}}_1^*\|_2. \tag{21}$$

Next, we identify the remaining vertices of the simplex. Let $\bar{\boldsymbol{\nu}}(l) = \boldsymbol{\nu}(l) - \hat{\mathbf{e}}_1^*$ and $\hat{\bar{\mathbf{e}}}_j^* = \hat{\mathbf{e}}_j^* - \hat{\mathbf{e}}_1^*$. Suppose we have already identified $r-1$ vertices $\{\hat{\mathbf{e}}_j^*\}_{j=1}^{r-1}$ with $r > 1$. We define the matrix $\mathbf{E}_{r-1} = \left[\hat{\bar{\mathbf{e}}}_2^*, \ldots, \hat{\bar{\mathbf{e}}}_{r-1}^*\right]$, from which we construct its orthogonal complement projector $\mathbf{P}_{r-1}^{\perp} \equiv \mathbf{I}_J - \mathbf{E}_{r-1}(\mathbf{E}_{r-1}^T\mathbf{E}_{r-1})^{+}\mathbf{E}_{r-1}^T$, where $^+$ denotes the matrix pseudoinverse. The rth vertex is chosen as the point with maximum projection to the column space of $\mathbf{P}_{r-1}^{\perp}$:

$$\hat{\mathbf{e}}_r^* = \boldsymbol{\nu}(l_r),\ l_r = \underset{1\le l\le L}{\operatorname{argmax}} \|\mathbf{P}_{r-1}^{\perp}\bar{\boldsymbol{\nu}}(l)\|_2. \tag{22}$$

We successively repeat (22) for $3 \le r \le J$, and recover all the simplex vertices $\{\hat{\mathbf{e}}_j^*\}_{j=1}^J$. For simplicity of notation, we ignore possible permutation of the indices of the vertices with respect to the actual identity of the speakers.

Based on (9), an approximation of the matrix $\mathbf{Q}$ is formed by the identified vertices: $\hat{\mathbf{Q}} = [\hat{\mathbf{e}}_1^*, \hat{\mathbf{e}}_2^*, \ldots, \hat{\mathbf{e}}_J^*]$. Using the recovered matrix $\hat{\mathbf{Q}}$ we can map the new representation to the original probabilities by (recall (8)):

$$\hat{\mathbf{p}}(l) = \hat{\mathbf{Q}}^{-1}\boldsymbol{\nu}(l) \tag{23}$$

Let $\mathcal{L}_j$ denote the set of frames dominated by the jth speaker. Based on the recovered probabilities, we define the set $\mathcal{L}_j$ by:

$$\mathcal{L}_j = \{l \mid \hat{p}_j(l) > \beta,\ l \in \{1, \ldots, L\}\} \tag{24}$$

where β is a probability threshold.

D. Unmixing Procedure

Given the set $\mathcal{L}_j$, an RTF estimator for the jth speaker, is given by:

$$\hat{H}_j^m(f) = \frac{\sum_{l\in\mathcal{L}_j} Y^m(l,f)Y^{1*}(l,f)}{\sum_{l\in\mathcal{L}_j} Y^1(l,f)Y^{1*}(l,f)} \tag{25}$$

Based on the estimated RTFs $\hat{H}_j^m(k)$ of each of the speakers $1 \le j \le J$, the mixture can be unmixed applying the pseudo-inverse of the matrix containing the estimated RTFs:

$$\mathbf{z}(l,f) = \mathbf{B}^H(f)\mathbf{y}(l,f) \tag{26}$$

where

$$\begin{aligned}\mathbf{y}(l,f) &= \left[Y^1(l,f), Y^2(l,f), \ldots, Y^M(l,f)\right]^T \\ \mathbf{B}(f) &= \mathbf{C}(f)(\mathbf{C}(f)^H\mathbf{C}(f))^{-1}\end{aligned} \tag{27}$$

and $[\mathbf{C}(f)]_{(m,j)} = \hat{H}_j^m(f)$. The time-domain separated signals are obtained by applying the inverse-STFT. The proposed method is summarized in Algorithm 1, and a flow diagram illustrating its main steps is given in Fig. 6.

Note that the proposed method exploits the estimated speaker probabilities to detect frames dominated by each of the speakers and uses them to estimate the corresponding RTFs of each of the speakers. Exploring other alternatives for utilizing the estimated probabilities in order to provide better separation capabilities is an important issue, which is beyond the scope of the current contribution and is left for future work.

The method presented in this paper is utilized in [43] for a diarization task, namely, to determine the set of active speakers in each time segment. Exploiting the diarization results, a separation of undetermined mixtures with $M < J$ is demonstrated in [43], where at each time instance the number of active speakers does not exceed the number of microphones.

IV. Experimental Study

In this section, we evaluate the performance of the proposed method in various test scenarios of both simulated data and real-life recordings.

A. Simulation Setup

The measured signals are generated using concatenated TIMIT sentences. The clean signals are convoluted with acoustic impulse responses (AIRs), which are drawn from an open database [44]. The AIRs in the database were measured in a reverberant room of size 6m×6m×2.4m with reverberation times of 160ms, 360ms and 610ms. We use a uniform linear array of $M = 8$ microphones with 8cm inter-microphone spacing. The different speaker positions are located on a spatial

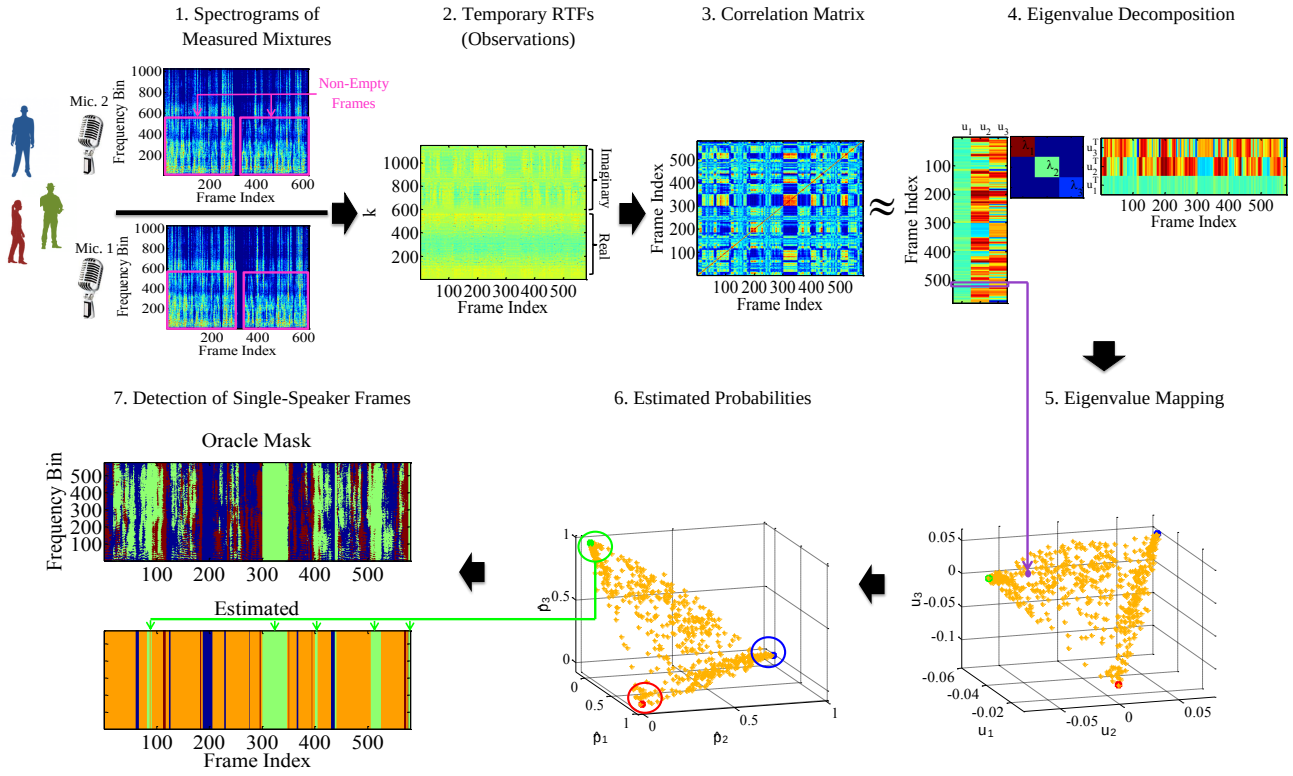


Fig. 6. A flow diagram of the proposed method. The method is illustrated on a 20s long mixture of $J = 3$ concurrent speakers measured by an array of $M = 8$ microphones (only two microphones are presented). 1) We use the STFT representation of the measured signals in two microphones with $F = 2048$ frequency bins and $L = 622$ frames. We focus on non-empty frames with non-negligible energy (depicted by magenta rectangles). For each non-empty frame, we compute the instantaneous ratio between the measurements of the second microphone and the measurements of the first microphone, which serves as a reference microphone. 2) An observation of length $D = 1152$ is associated with each frame, which comprises the real and the imaginary parts of the computed ratios corresponding to the frequency bins $1 - 576$ ($0 - 4.5$kHz). 3) The correlation matrix W between observations is approximated by the inner product between the observations of the different frames. 4) An eigenvalue decomposition is applied to the correlation matrix, resulting in a low-rank representation with 3 significant eigenvalues associated with 3 eigenvectors, which span the column space of the probability matrix P. 5) The computed eigenvectors form a 3-D mapping for each frame. In the subfigure of Step 5, each orange point represents a specific frame, and is embedded in the new 3 dimensional coordinate system according to the corresponding entries of the 3 eigenvectors. It is evident that the points lie on a 2-D simplex (triangle). The vertices of the simplex (marked by red, blue and green dots) are detected, corresponding to 3 frames, where each frame is dominated by a single speaker. 6) The mapping is transformed to estimated probability of activity of the different speakers. Here, all points lie on the standard probability simplex. 7) Points lying near each vertex are associated with frames dominated by a single speaker (colored by red, blue and green). The algorithm detection is compared to the oracle mask.

grid of angles ranging from $-90°$ to $90°$ in $15°$ steps with 1m and 2m distance from the microphone array.

The signal duration is 20s, with sampling rate of 16kHz. The window length of the STFT is set to $N = 2048$ with $\eta = 75\%$ overlap between adjacent frames, which corresponds to a total amount of $L = 622$ frames. For each frame, the instantaneous RTF of each frequency bin in (17), is estimated by averaging the signals in 3 adjacent frames ($T = 2$). The instantaneous RTF vectors in (18) consist of $F = 576$ frequency bins, corresponding to $0 - 4.5$kHz, in which most of the speech components are concentrated. The obtained concatenated vectors of length $D = 2 \cdot (M-1) \cdot F = 8064$ are normalized to have a unit-norm. The results are demonstrated for mixtures of $J = 2$, $J = 3$ and $J = 4$ speakers in different locations.

B. Speaker Counting

We first examine the ability of the proposed method to estimate the number of speakers in the mixture. Here, we use a smaller frequency range between $0.5 - 1.5$kHz, which yields better results for the task of counting the number of speakers. We conduct 100 Monte-Carlo trials for each $J \in \{1, 2, 3\}$, in which the angles and the distances of the speakers, as well as their input sentences, are randomly selected. Figure 7 depicts the average counting accuracy as a function of the threshold parameter α (19) in the range between 0.09 and 0.15 for different reverberation levels $T_{60} = \{160, 360, 610\}$ms. It can be seen that the best results are achieved for moderate reverberation time of 360ms. For low reverberation time of 160ms better results are obtained for lower threshold values, while for high reverberation time of 610ms better results are obtained for higher threshold values. The mean counting accuracy averaged over all reverberation levels is also presented in magenta solid line. It can be deduced that the mean counting accuracy is robust to the choice of the threshold value with above 90% accuracy in the range between 0.1 and 0.14.

We also examine the counting accuracy with respect to the

Algorithm 1: Separation Algorithm

Feature Extraction:

- Estimate instantaneous RTFs $\left\{\tilde{R}^m(l,f)\right\}_{l,f,m}$ (17).
- Construct observation vectors $\{\mathbf{a}(l)\}_{l=1}^{L}$ (18).

Forming a Data-Driven Simplex:

- Estimate the correlation matrix $\widehat{\mathbf{W}}$ with $\widehat{W}_{ln} = \frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)$.
- Compute EVD of $\widehat{\mathbf{W}}$ and obtain $\{\mathbf{u}_j, \lambda_j\}_{j=1}^{L}$.
- Estimate the number of speakers $\hat{J}$ (19).
- Construct $\boldsymbol{\nu}(l) = [\mathbf{u}_1(l), \mathbf{u}_2(l), \ldots, \mathbf{u}_J(l)]$.
- Form the set $\{\boldsymbol{\nu}(l)\}_{l=1}^{L}$ lying in a simplex.

Recovering Activity of Speakers:

- Recover simplex vertices $\{\hat{\mathbf{e}}_j^*\}_{j=1}^{J}$ (20),(21),(22).
- Estimate speakers' probabilities $\{\mathbf{p}(l)\}_{l=1}^{L}$ (23).
- Identify single-speaker frames $\{\mathcal{L}_j\}_{j=1}^{J}$ (24).

Unmixing Procedure:

- Estimate the RTFs $\left\{\hat{H}_j^m(f)\right\}_{f,j,m}$.
- Separate the individual speakers (26).

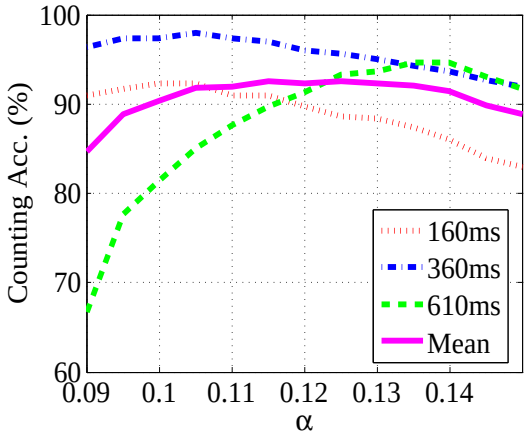


Fig. 7. Counting Accuracy as a function of the threshold parameter α for different reverberation times of 160ms, 360ms and 610ms. Each point in the graphs is obtained by an average over 300 trials with different speakers in different locations: 100 mixtures of $J = 2$ speakers, 100 mixtures of $J = 3$ speakers and 100 mixtures of $J = 4$ speakers. The mean counting accuracy averaged over all reverberation levels is also presented in a solid magenta line.

relative positions of the speakers. For this purpose, mixtures of J=3 speakers are generated with one speaker in a fixed position, while the other two speakers are located in different positions. The first speaker is positioned 1m from the array at a relative angle of $-60°$. The two other speakers are positioned 2m from the array with different relative angles ranging from $-90°$ to $90°$ in steps of $15°$. The reverberation time is set to 360ms. For each angle combination, the performance is averaged over 10 trials with different speakers and sentences. Figure 8 illustrates the counting accuracy obtained for each angle combination of the two speakers for fixed threshold value $\alpha = 0.12$. Note that the values on the diagonal of the matrix are meaningless since they represent a scenario with

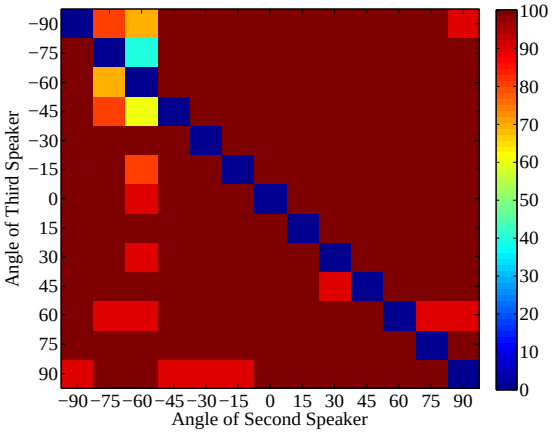


Fig. 8. Counting Accuracy for a mixture of $J = 3$ speakers with one speaker in a fixed position at a relative angle of $-60°$, while the other two speakers are located in various relative angles ranging from $-90°$ to $90°$ in steps of $15°$. The reverberation time is set to 360ms. Each entry in the matrix correspond to a specific angle combination of the two speakers, and is colored according to the obtained counting accuracy, averaged over 10 trials with different speakers.

two speakers located at the same spot. It can be seen that most of the counting errors occur when all three speakers are very close to each other. Nevertheless, the correct number of speakers is recovered most of the time with average accuracy of 97.8%, even when two speakers are located one in front of the other (i.e., with the same relative angle but in different distances).

C. Detection of Single-Speaker Frames

Next, we examine the ability of the proposed method to identify the set of frames $\{\mathcal{L}_j\}_{j=1}^{J}$ dominated by each speaker. Figure 9 illustrates the time-domain signals of each of the speakers for a mixture of $J = 2$ speakers (a), and for a mixture of $J = 4$ speakers (b). The shaded areas stand for time instances, which were found to be dominated by each of the speakers, using (24). It can be seen that the proposed algorithm successfully identifies time-periods for which one speaker is dominant over the other speakers. Comparing Fig. 9(a) and (b), we observe that as more speakers are involved in the mixture, then less time-periods are dominated by a single speaker.

D. Performance on Simulated Data

The separation performance is assessed by the signal to interference ratio (SIR) and signal to distortion ratio (SDR) measures, using the BSS-Eval toolbox [45]. We also quantify the improvement in speech intelligibility through the short-time objective intelligibility (STOI) measure [46], where we used the signals of the individual speakers, as received by the first microphone, as the reference signals for evaluating the STOI measure. The measures are averaged over 20 Monte-Carlo trials, in which the angles and the distances of the sources, as well as their input sentences, are randomly selected.

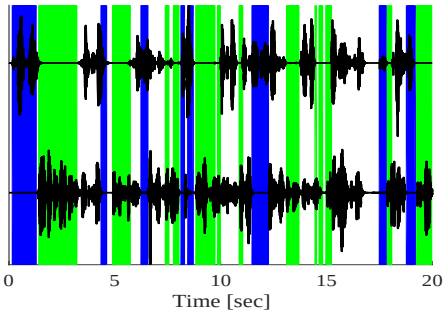
We compare the proposed method to two oracle methods, which are also based on the unmixing scheme of (26). In

TABLE III
SEPARATION AND INTELLIGIBILITY PERFORMANCES DEPENDING ON THE NUMBER OF SPEAKERS (RT=360MS). 'PROPOSED' METHOD IS COMPARED TO TWO ORACLE METHODS 'IDEAL' AND 'SEMI-IDEAL' AND TO A MULTICHANNEL 'NMF' ALGORITHM WITH SEMI-BLIND INITIALIZATION.

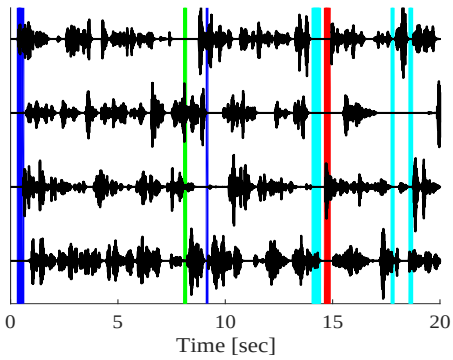
J	SIRin	STOIin	SIR				SDR				STOI			
			Ideal	Semi-Ideal	Proposed	NMF	Ideal	Semi-Ideal	Proposed	NMF	Ideal	Semi-Ideal	Proposed	NMF
2	0	98.1	22.0	21.2	20.7	13.2	10.5	10.3	10.0	7.4	99.9	99.8	99.8	99.7
3	−3	82.7	16.5	14.8	13.9	9.8	8.0	7.2	6.7	4.9	99.6	99.5	99.4	98.9
4	−5	53.2	12.1	9.7	9.3	6.8	6.2	4.5	4.2	2.6	98.8	97.9	97.3	91.7

TABLE IV
SEPARATION AND INTELLIGIBILITY PERFORMANCES DEPENDING ON REVERBERATION TIME (3 SPEAKERS). 'PROPOSED' METHOD IS COMPARED TO TWO ORACLE METHODS 'IDEAL' AND 'SEMI-IDEAL' AND TO A MULTICHANNEL 'NMF' ALGORITHM WITH SEMI-BLIND INITIALIZATION.

T_{60}	SIRin	STOIin	SIR				SDR				STOI			
			Ideal	Semi-Ideal	Proposed	NMF	Ideal	Semi-Ideal	Proposed	NMF	Ideal	Semi-Ideal	Proposed	NMF
160	−3	64.7	17.7	16.5	15.8	8.0	12.8	11.8	11.0	4.1	99.8	99.7	99.7	97.9
360	−3	82.7	16.4	13	11.9	9.3	6.3	5.2	4.6	4.6	99.6	99.5	99.4	98.9
610	−3	64.1	16.1	13.5	13.2	8.1	4.3	3.6	3.6	4.1	98.8	98.5	98.4	98.0



(a)



(b)

Fig. 9. Time-domain waveforms of each of the speakers for mixtures of (a) $J = 2$ and (b) $J = 4$ speakers. Time instances, which were detected to be dominated by each of the speakers, are shaded in compatible colors: blue for the first speaker (top), green for the second speaker, red for the third speaker, and cyan for the fourth speaker (down). The probability threshold β for detecting single speaker frames is set to 0.9.

addition, we compare to a multichannel NMF algorithm [14] representing state of-the-art algorithms of the BSS family. The methods based on (26) use either of the following procedures for estimating the RTFs, used to compute the unmixing matrix:

1) Ideal: The RTFs are estimated using the individually measured signals, i.e.:

$$\hat{H}_j^m(f) = \frac{\sum_{l=1}^{L} Y_j^m(l,f) Y_j^{1*}(l,f)}{\sum_{l=1}^{L} Y_j^1(l,f) Y_j^{1*}(l,f)} \tag{28}$$

2) Semi-Ideal: The RTFs are estimated by (25) based on the measured mixtures (10), where the sets $\{\mathcal{L}_j\}_{j=1}^J$ are determined using the oracle speakers' probabilities computed by:

$$l \in \mathcal{L}_j, \quad \text{if} \quad \frac{\sum_{m,f} \|Y_j^m(l,f)\|^2}{\sum_{j=1}^{J} \sum_{m,f} \|Y_i^m(l,f)\|^2} > \gamma \tag{29}$$

where γ is a threshold set to 0.95, 0.9 or 0.8 for $J = 2$, $J = 3$ or $J = 4$, respectively.

3) Proposed: The RTFs are estimated by (25), where the sets $\mathcal{L}_j$, $1 \leq j \leq J$ are determined using the proposed algorithm, presented in Section III-B, where β is set to 0.9.

The parameters of the NMF algorithm are initialized using the separated speakers, which are artificially mixed with SIR that is improved with respect to the input SIR of the given mixture by 3dB.

We evaluate the performance of all the algorithms depending on the number of speakers and on the reverberation time. The results depending on the number of speakers are depicted in Table III for $J = \{2, 3, 4\}$, with a fixed reverberation time of 360ms. The results depending on the reverberation time are depicted in Table IV for $T_{60} = \{160, 360, 610\}$ms, for mixtures of $J = 3$ speakers.

We observe that the ideal unmixing yields the best results. In fact, it represents an upper bound for the separation capabilities, since it is derived using the separated speakers. The semi-ideal unmixing is inferior with respect to the upper bound, since the ideal unmixing uses the original signals for estimating the RTFs, whereas the semi-ideal unmixing uses non-pure frames from the mixed signals, which may contain also low energy components of other speakers. The proposed estimator determines the frames dominated by each speaker based on the mixed signals. Its performance is comparable to the semi-ideal unmixing with a small gap of $0 - 1.1$dB. The

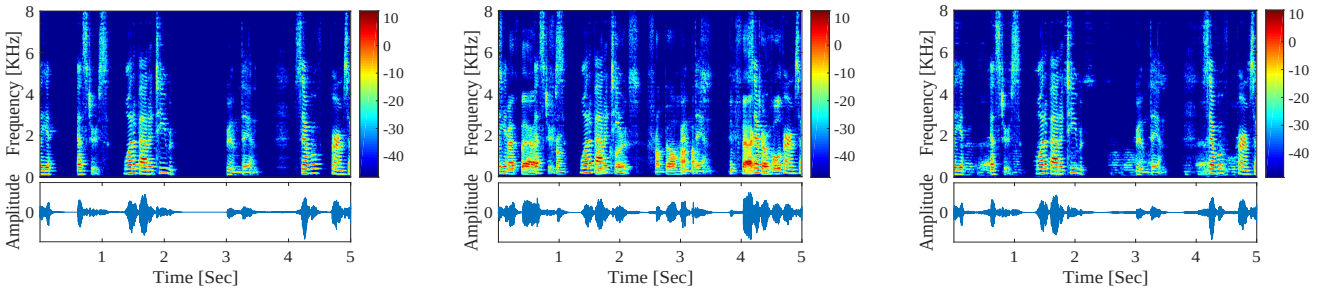


Fig. 10. Example of separation performance on a mixture with $J = 2$ speakers: spectrograms and waveforms of the first speaker at the first microphone (a), the mixture of the two speakers at the first microphone (b), the estimated first speaker (c).

NMF method is inferior with respect to the proposed method in almost all cases. It should be emphasized that the NMF algorithm uses an initialization with improved SIR, whereas the proposed method is completely blind. In terms of intelligibility improvement, all algorithms achieve comparable results. For all algorithms, a performance degradation is observed as the number of speakers increases or as the reverberation time increases. It should be noted that for both the semi-ideal unmixing and the proposed method, an increase in the number of speakers means a decrease in the number of frames dominated by a single speaker, hence, the performance gap between both algorithms and the ideal unmixing increases.

Figure 10 presents an example of the spectrograms and the waveforms of a mixture of $J = 2$ speakers, where the first speaker (a), the mixture (b), and the output signal of the proposed method (c), are depicted. It is evident that the spectral components of the second speaker are significantly attenuated, while preserving most of the spectral components of the first speaker. There is also a good match between the original and the output waveforms.

E. Performance on Real-Life Recordings

We also examined the performance on real-life recordings carried out in a low echoic noisy enclosure as described in [31]. The speakers are located in four optional positions. Their signals are measured by a U-shaped array with $M = 7$ omnidirectional microphones. Six speakers (3 males and 3 females) and a diffuse noise were recorded separately, and were used to construct mixtures of $J = 2$ speakers with different combinations of SIR and signal to noise ratio (SNR). The time-line of signals' activity for all scenarios is described in Table V. We applied the proposed algorithm to mixtures with different SIR levels $\{-15, -10, -5, 0\}$dB and a fixed signal to noise ratio (SNR) of 15dB. Due to the presence of background noise, instead of the unmixing in (26) we used the following linearly constrained minimum variance (LCMV) beamformer, which extracts one of the speakers, while suppressing the other speaker and reducing the noise level:

$$z(l,f) = \mathbf{b}^H(f)\mathbf{y}(l,f)$$
$$\mathbf{b}(f) = \mathbf{\Phi}_{\text{vv}}^{-1}(f)\mathbf{C}(f)(\mathbf{C}(f)^H\mathbf{\Phi}_{\text{vv}}^{-1}(f)\mathbf{C}(f))^{-1}\mathbf{g}$$

TABLE V
TIME-LINE OF SIGNALS' ACTIVITY IN THE REAL-LIFE RECORDINGS

Time [sec]	0-0.5	0.5-3	3-6	6-9	9-16	16-18
First Speaker	0	1	0	0	1	0
Second Speaker	0	0	1	0	1	0
Background Noise	1	1	1	1	1	1

where $\mathbf{g} = [1, 0]^T$ and $\mathbf{\Phi}_{\text{vv}}(f)$ is the noise correlation matrix of size $M \times M$. The estimation of $\mathbf{\Phi}_{\text{vv}}(f)$ is based on noise-only frames, which are detected as frames with correlation below 0.3 with all other frames. We compare the performance of the proposed method with an LCMV beamformer controlled by a DNN-based concurrent speaker detector, presented in [31]. In addition, we also compare our method with an independent vector analysis (IVA) algorithm [32], [47], which is an extension of ICA that circumvents the problem of permutation ambiguity across frequency bins. Figure 11 depicts the STOI measure [46] obtained for the input signal and for the output signals of all algorithms. The STOI scores were evaluated on the time segment between $9-16$s, when the two speakers overlap. We observe the superiority of the proposed method over [31] for all SIR levels, and a slight superiority of the proposed method over [32] for low SIR levels. In terms of computational complexity, the proposed method exhibits remarkable advantage over IVA, as its processing takes on average 1.9s, whereas the IVA processing takes on average 27.4s, when both are implemented in MATLAB on a standard PC (CPU Intel Core Quad 2.8 GHz, RAM 16 GB). Figure 12 depicts an example of the spectrograms and the waveforms of the input and the output signals. Comparing to the activity timeline presented in Table V, it is clearly seen that the second speaker is successfully suppressed at the output signal of the proposed method.

V. CONCLUSIONS

We present a novel framework for speaker counting and separation in a completely blind manner. The separation is based on the sparsity of speech in the STFT domain, as well as the fact that each speaker is associated with a unique spatial signature, manifested by the RTF between the speaker and the microphones. A spectral decomposition of the correlation

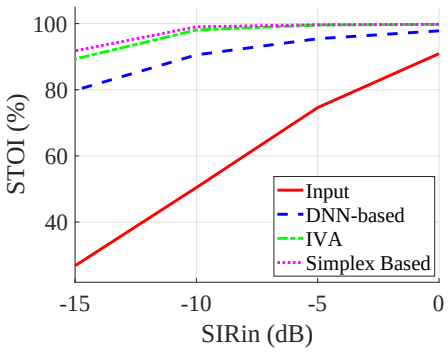
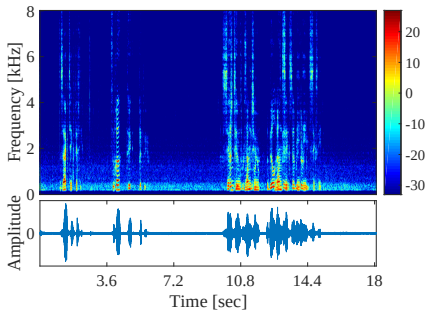
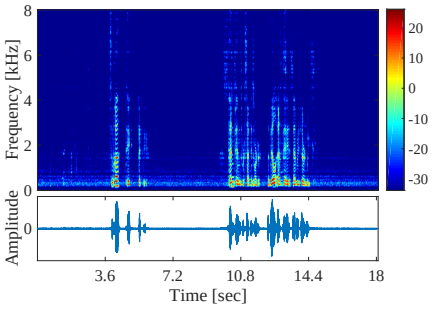


Fig. 11. STOI performance as a function of SIR on real-life noisy recordings with $J = 2$ speakers and SNR set to 15dB. The proposed method is compared to a DNN-based concurrent speaker detector [31] and to an independent vector analysis (IVA) algorithm [32].



(a)



(b)

Fig. 12. Example of separation performance on real-recordings of a mixture of $J = 2$ speakers, with SNR= 15dB and SIR= -15dB: spectrograms and waveforms of a mixture of two speakers at the first microphone (a), the estimated second speaker (b).

matrix of different time frames reveals the number of speakers, and forms a simplex of the speakers' probabilities across time. Utilizing convex geometry tools, the frames dominated by each speaker are identified. The RTFs of the different speakers are estimated using these identified frames, and an unmixing scheme is implemented to separate the individual speakers. The performance is demonstrated in an experimental study for various reverberation levels.

APPENDIX A

In this section, we compute the expected correlation between observations and evaluate its variance. The computation is based on the statistical model of Section II. Recall the following assumption regarding the hidden sources:

$$E\left\{h_i(k)h_j(\tilde{k})\right\} = \delta_{ij}\cdot\delta_{k\tilde{k}}. \tag{30}$$

which follows from (1), the zero-mean assumption and the mutual independence of the hidden sources. In addition, the indicator functions satisfy (recall (3)):

$$I_j(l,k)I_i(l,k) = I_j(l,k)\delta_{ij}. \tag{31}$$

We compute the correlation for $1 \le l,n \le L, l \ne n$:

$$\begin{aligned}
&E\left\{\frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)\right\} \\
&= \frac{1}{D}E\left\{\sum_{k=1}^{D}\sum_{i,j=1}^{J} I_i(l,k)I_j(n,k)h_i(k)h_j(k)\right\} \qquad (32)\\
&= \frac{1}{D}\sum_{k=1}^{D}\sum_{i,j=1}^{J} E\left\{I_i(l,k)I_j(n,k)\right\}E\left\{h_i(k)h_j(k)\right\} \\
&= \frac{1}{D}\sum_{k=1}^{D}\sum_{i,j=1}^{J} E\left\{I_i(l,k)\right\}E\left\{I_j(n,k)\right\}E\left\{h_i(k)h_j(k)\right\} \\
&= \frac{1}{D}\sum_{k=1}^{D}\sum_{i,j=1}^{J} p_i(l)p_j(n)\delta_{ij} = \sum_{j=1}^{J} p_j(l)p_j(n).
\end{aligned}$$

The second equality follows from the independence of the indicator functions and the sources. The third equality follows from the independence of the indicator functions for $l \ne n$. The fourth equality is due to (30).

For $l = n$ the autocorrelation is given by:

$$\begin{aligned}
&E\left\{\frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(l)\right\} \\
&= \frac{1}{D}\sum_{k=1}^{D}\sum_{i,j=1}^{J} E\left\{I_i(l,k)I_j(l,k)\right\}E\left\{h_i(k)h_j(k)\right\} \qquad (33)\\
&= \frac{1}{D}\sum_{k=1}^{D}\sum_{j=1}^{J} E\left\{I_j(l,k)\right\}E\left\{h_j^2(k)\right\} = \sum_{j=1}^{J} p_j(l) = 1
\end{aligned}$$

where the second equality follows from (31).

We compute the variance of (32):

$$\begin{aligned}
&\mathrm{Var}\left\{\frac{1}{D}\mathbf{a}(l)^T\mathbf{a}(n)\right\} \\
&= \frac{1}{D^2}E\left\{(\mathbf{a}(l)^T\mathbf{a}(n))^2\right\} - \frac{1}{D^2}E^2\left\{\mathbf{a}(l)^T\mathbf{a}(n)\right\}. \qquad (34)
\end{aligned}$$

We show that the variance (34) approaches zero for D large enough, implying that the typical value $\frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)$ approaches the expected value $E\left\{\frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)\right\}$.

The first moment is given in (32). We compute the second moment for $1 \leq l, n \leq L, l \neq n$:

$$\begin{aligned} & E\left\{(\mathbf{a}^T(l)\mathbf{a}(n))^2\right\} \\ & = E\left\{\left(\sum_{k=1}^{D}\sum_{i,j=1}^{J} I_i(l,k)I_j(n,k)h_i(k)h_j(k)\right)^2\right\} \qquad (35) \\ & = \sum_{k,\tilde{k}=1}^{D}\sum_{\substack{i,j,\\ \tilde{i},\tilde{j}=1}}^{J} E\left\{I_i(l,k)I_j(n,k)I_{\tilde{i}}(l,\tilde{k})I_{\tilde{j}}(n,\tilde{k})\right\} \\ & \cdot E\left\{h_i(k)h_j(k)h_{\tilde{i}}(\tilde{k})h_{\tilde{j}}(\tilde{k})\right\}. \end{aligned}$$

Splitting the sum over $\tilde{k}$ into two parts, for $\tilde{k} = k$ and for $\tilde{k} \neq k$, we receive:

$$\begin{aligned} & E\left\{(\mathbf{a}^T(l)\mathbf{a}(n))^2\right\} \\ & = \sum_{k=1}^{D}\sum_{\substack{i,j,\\ \tilde{i},\tilde{j}=1}}^{J} E\left\{I_i(l,k)I_j(n,k)I_{\tilde{i}}(l,k)I_{\tilde{j}}(n,k)\right\} \qquad (36) \\ & \cdot E\left\{h_i(k)h_j(k)h_{\tilde{i}}(k)h_{\tilde{j}}(k)\right\} \\ & + \sum_{\substack{k,\tilde{k}=1\\ \tilde{k}\neq k}}^{D}\sum_{\substack{i,j,\\ \tilde{i},\tilde{j}=1}}^{J} E\left\{I_i(l,k)I_j(n,k)I_{\tilde{i}}(l,\tilde{k})I_{\tilde{j}}(n,\tilde{k})\right\} \\ & \cdot E\left\{h_i(k)h_j(k)h_{\tilde{i}}(\tilde{k})h_{\tilde{j}}(\tilde{k})\right\} \\ & = \sum_{k=1}^{D}\sum_{i,j=1}^{J} E\left\{I_i(l,k)\right\}E\left\{I_j(n,k)\right\}E\left\{h_i^2(k)h_j^2(k)\right\} \\ & + \sum_{\substack{k,\tilde{k}=1\\ \tilde{k}\neq k}}^{D}\sum_{\substack{i,j,\\ \tilde{i},\tilde{j}=1}}^{J} E\left\{I_i(l,k)\right\}E\left\{I_j(n,k)\right\} \\ & E\left\{I_{\tilde{i}}(l,\tilde{k})\right\}E\left\{I_{\tilde{j}}(n,\tilde{k})\right\}E\left\{h_i(k)h_j(k)h_{\tilde{i}}(\tilde{k})h_{\tilde{j}}(\tilde{k})\right\}. \end{aligned}$$

where the second equality follows from (31), and the independence of the indicator functions for $l \neq n$ or $k \neq \tilde{k}$. Evaluating the expectations of the indicators, we get:

$$\begin{aligned} & E\left\{(\mathbf{a}(l)^T\mathbf{a}(n))^2\right\} = \sum_{k=1}^{D}\sum_{i,j=1}^{J} p_i(l)p_j(n)E\left\{h_i^2(k)h_j^2(k)\right\} + \\ & \sum_{\substack{k,\tilde{k}=1\\ \tilde{k}\neq k}}^{D}\sum_{\substack{i,j,\\ \tilde{i},\tilde{j}=1}}^{J} p_i(l)p_j(n)p_{\tilde{i}}(l)p_{\tilde{j}}(n)E\left\{h_i(k)h_j(k)h_{\tilde{i}}(\tilde{k})h_{\tilde{j}}(\tilde{k})\right\} \end{aligned} \qquad (37)$$

Relying on the independence between $h_i(k)$ and $h_{\tilde{i}}(\tilde{k})$ for $k \neq \tilde{k}$, and on the statistical model of (30), the second term of (37) is simplified as:

$$\begin{aligned} & \sum_{\substack{i,j,\\ \tilde{i},\tilde{j}=1}}^{J} p_i(l)p_j(n)p_{\tilde{i}}(l)p_{\tilde{j}}(n)E\left\{h_i(k)h_j(k)\right\}E\left\{h_{\tilde{i}}(\tilde{k})h_{\tilde{j}}(\tilde{k})\right\} \\ & = \sum_{\substack{k,\tilde{k}=1\\ \tilde{k}\neq k}}^{D}\sum_{\substack{i,j,\\ \tilde{i},\tilde{j}=1}}^{J} p_i(l)p_j(n)p_{\tilde{i}}(l)p_{\tilde{j}}(n)\delta_{ij}\delta_{\tilde{i}\tilde{j}} \\ & = D(D-1)\sum_{j,\tilde{j}=1}^{J} p_j(l)p_j(n)p_{\tilde{j}}(l)p_{\tilde{j}}(n) \\ & = D(D-1)\left(\sum_{j=1}^{J} p_j(l)p_j(n)\right)^2 \qquad (38) \end{aligned}$$

Substituting (38) into (37), we get:

$$\begin{aligned} E\left\{(\mathbf{a}^T(l)\mathbf{a}(n))^2\right\} &= \sum_{k=1}^{D}\sum_{i,j=1}^{J} p_i(l)p_j(n)E\left\{h_i^2(k)h_j^2(k)\right\} \\ &+ D(D-1)\left(\sum_{j=1}^{J} p_j(l)p_j(n)\right)^2. \qquad (39) \end{aligned}$$

Substituting (32) and (39) into (34), we receive:

$$\begin{aligned} & \mathrm{Var}\left\{\frac{1}{D}\mathbf{a}(l)^T\mathbf{a}(n)\right\} \\ & = \frac{1}{D^2}E\left\{(\mathbf{a}(l)^T\mathbf{a}(n))^2\right\} - \frac{1}{D^2}E^2\left\{\mathbf{a}(l)^T\mathbf{a}(n)\right\} \qquad (40) \\ & = \frac{1}{D^2}\sum_{k=1}^{D}\sum_{i,j=1}^{J} p_i(l)p_j(n)E\left\{h_i^2(k)h_j^2(k)\right\} \\ & + \frac{D(D-1)}{D^2}\left(\sum_{j=1}^{J} p_j(l)p_j(n)\right)^2 - \left(\sum_{j=1}^{J} p_j(l)p_j(n)\right)^2. \end{aligned}$$

For D large enough, we have $\frac{D-1}{D} \approx 1$, and (40) simplifies to:

$$\begin{aligned} & \mathrm{Var}\left\{\frac{1}{D}\mathbf{a}(l)^T\mathbf{a}(n)\right\} \approx \frac{1}{D^2}\sum_{k=1}^{D}\sum_{j=1}^{J} p_j(l)p_j(n)E\left\{h_j^4(k)\right\} \\ & + \frac{1}{D^2}\sum_{k=1}^{D}\sum_{\substack{i,j=1\\ i\neq j}}^{J} p_i(l)p_j(n)E\left\{h_i^2(k)\right\}E\left\{h_j^2(k)\right\} \qquad (41) \\ & = \frac{1}{D^2}\sum_{k=1}^{D}\sum_{j=1}^{J} p_j(l)p_j(n)E\left\{h_j^4(k)\right\} + \frac{1}{D}\sum_{\substack{i,j=1\\ i\neq j}}^{J} p_i(l)p_j(n). \end{aligned}$$

In the second term in (41), we have:

$$\begin{aligned} \sum_{j=1}^{J}\left(p_j(n)\sum_{\substack{i=1\\ i\neq j}}^{J} p_i(l)\right) &= \sum_{j=1}^{J} p_j(n)\left(1-p_j(l)\right) \\ &= 1 - \sum_{j=1}^{J} p_j(l)p_j(n) \qquad (42) \end{aligned}$$

Let $E\left\{h_j^4(k)\right\} \equiv C_4$, substituting (42) into (41), we get:

$$\begin{aligned}&\mathrm{Var}\left\{\frac{1}{D}\mathbf{a}(l)^T\mathbf{a}(n)\right\}\\ &\approx \frac{C_4}{D}\sum_{j=1}^{J}p_j(l)p_j(n)+\frac{1}{D}\left(1-\sum_{j=1}^{J}p_j(l)p_j(n)\right). \quad (43)\\ &=\frac{C_4-1}{D}\sum_{j=1}^{J}p_j(l)p_j(n)+\frac{1}{D}\le\frac{C_4-1}{D}+\frac{1}{D}=\frac{C_4}{D}.\end{aligned}$$

For zero-mean Gaussian sources $C4 = E\left\{h_j^4(k)\right\} = 3E\left\{h_j^2(k)\right\}$, which under the unit variance assumption amounts to $C4 = 3$. Hence, we can easily set the value of D, satisfying $D \gg C_4$. Accordingly, we get $\mathrm{Var}\left\{\frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)\right\} \approx 0$. We conclude that for D large enough the typical value of $\frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)$ is close to its expected value $E\left\{\frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)\right\}$. Hence, $\frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)$ can be used instead of its expected value.

APPENDIX B

In this section, we discuss the spectral decomposition of the correlation matrix $\mathbf{W}$, and its approximation as $\mathbf{W} \approx \mathbf{P}\mathbf{P}^T$. Recall the following representation of the correlation matrix $\mathbf{W}$ (Eq. (5)):

$$\mathbf{W} = \mathbf{P}\mathbf{P}^T + \Delta\mathbf{W} \quad (44)$$

where $\Delta\mathbf{W}$ is a diagonal matrix with $\Delta W_{ll} = 1-\sum_{j=1}^{J}p_j^2(l)$. Here, we analyse the influence of $\Delta\mathbf{W}$ on the obtained spectral decomposition, and show that it has a negligible affect on the proposed speaker counting and separation method.

For this purpose, we use matrix perturbation theory [48]. Consider the perturbed matrix $\mathbf{W}$ given by:

$$\mathbf{W} = \mathbf{K} + \Delta\mathbf{W} \quad (45)$$

where the matrix $\Delta\mathbf{W}$ represents a small perturbation. According to the matrix perturbation theory [48], the following Theorem relates the EVDs of the matrices $\mathbf{W}$ and $\mathbf{K}$:

Theorem 1. *Let $\{\lambda_j, \mathbf{u}_j\}_j$ be the set of eigenvalues and eigenvectors of the matrix $\mathbf{K}$, and let $\{\tilde{\lambda}_j, \tilde{\mathbf{u}}_j\}_j$ be the set of eigenvalues and eigenvectors of the matrix $\mathbf{W} = \mathbf{K} + \Delta\mathbf{W}$. Then:*

$$\tilde{\lambda}_j = \lambda_j + \mathbf{u}_j^T\Delta\mathbf{W}\mathbf{u}_j + O\left(\|\Delta\mathbf{W}\|^2\right) \quad (46)$$

$$\tilde{\mathbf{u}}_j = \mathbf{u}_j + \sum_{i\neq j}\frac{\mathbf{u}_i^T\Delta\mathbf{W}\mathbf{u}_j}{\lambda_j-\lambda_i}\mathbf{u}_i + O\left(\|\Delta\mathbf{W}\|^2\right) \quad (47)$$

According to Theorem 1, each eigenvalue $\tilde{\lambda}_j$ of the perturbed matrix deviates from the corresponding eigenvalue λ_j of the original matrix by the weighted norm $\mathbf{u}_j^T\Delta\mathbf{W}\mathbf{u}_j$. In addition, each perturbed eigenvector $\tilde{\mathbf{u}}_j$ equals the corresponding original eigenvector $\mathbf{u}_j$ plus a term, which consists of the contributions of the other eigenvectors of the original matrix. The contribution of the other eigenvectors is proportional to the weighted inner product $\mathbf{u}_i^T\Delta\mathbf{W}\mathbf{u}_j$ divided by the difference $\lambda_j - \lambda_i$ between the corresponding eigenvalues.

In our case, the original matrix $\mathbf{K} \equiv \mathbf{P}\mathbf{P}^T$ has a rank-J decomposition. Accordingly, $\mathbf{P}\mathbf{P}^T$ has J nonzero eigenvalues $\Lambda_1 \equiv \{\lambda_j\}_{j=1}^{J}$, associated with J eigenvectors $\mathcal{U}_1 \equiv \{\mathbf{u}_j\}_{j=1}^{J}$ that span the column space of the matrix $\mathbf{P}$. In addition, there are $L-J$ zero eigenvalues $\Lambda_0 \equiv \{\lambda_j\}_{j=J+1}^{L}$, associated with $L-J$ eigenvectors $\mathcal{U}_0 \equiv \{\mathbf{u}_j\}_{j=J+1}^{L}$ that span the null space of $\mathbf{P}$.

The weighted inner product can be written as:

$$\mathbf{u}_i^T\Delta\mathbf{W}\mathbf{u}_j = \mathbf{v}_i^T\mathbf{v}_j = \|\mathbf{v}_i\|\|\mathbf{v}_j\|\cos\theta_{ij} \quad (48)$$

where $\mathbf{v}_j = \mathbf{A}\mathbf{u}_j$ with $\Delta\mathbf{W} = \mathbf{A}^T\mathbf{A}$, and θ_{ij} is the angle between $\mathbf{v}_i$ and $\mathbf{v}_j$. In our case, $\mathbf{A}$ is a diagonal matrix with elements $A_{ll} = \sqrt{1-\sum_{j=1}^{J}p_j^2(l)} \le 1$, implying $\|\mathbf{v}_j\| \le \|\mathbf{u}_j\| = 1$. We assume that multiplication by $\mathbf{A}$ only slightly affect the right angle between the orthonormal vectors $\mathbf{u}_i$ and $\mathbf{u}_j$ for $i \neq j$, implying $\cos\theta_{ij} \approx \epsilon$. Hence, we get the following bound:

$$\left|\mathbf{u}_i^T\Delta\mathbf{W}\mathbf{u}_j\right| \le \begin{cases}\epsilon & \text{if } i\neq j\\ 1 & \text{if } i=j\end{cases}. \quad (49)$$

Accordingly, the eigenvalue perturbation is limited to 1 and the eigenvector perturbation depends on the ratio $\frac{\epsilon}{\lambda_j-\lambda_i}$. An eigenvector $\mathbf{u}_{j^*}$ will have small contribution from the vector $\mathbf{u}_i$, when $|\lambda_{j*} - \lambda_i| \gg \epsilon$.

Note that in the proposed algorithm we are only interested in the eigenvectors in $\mathcal{U}_1$, spanning the column space of $\mathbf{P}\mathbf{P}^T$. For a particular $\mathbf{u}_{j^*} \in \mathcal{U}_1$, there may be some contribution of the other eigenvectors in $\mathcal{U}_1$, depending on the respective eigenvalues decay. The contribution of eigenvectors in $\mathcal{U}_0$, associated with zero eigenvalues, is necessarily smaller and is negligible for $|\lambda_{j*}| \gg \epsilon$.

We demonstrate the conclusions of the above analysis using the example of Section II-B. We compute the eigenvectors of $\mathbf{P}\mathbf{P}^T$ and of $\mathbf{W}$, and measure their correlation for $J = 3$. We present the correlation between the first 3 eigenvectors of $\mathbf{W}$ and the first 5 eigenvectors of $\mathbf{P}\mathbf{P}^T$:

$$\begin{bmatrix}1 & -4e^{-16} & -6e^{-5} & -6e^{-5} & -3e^{-5}\\ 4e^{-5} & 1 & -5e^{-3} & -4e^{-6} & 2e^{-5}\\ -6e^{-5} & -5e^{-3} & 1 & 1e^{-4} & -7e^{-5}\end{bmatrix}.$$

where the (i,j)th element equals $\tilde{\mathbf{u}}_i^T\mathbf{u}_j$. We deduce that $\tilde{\mathbf{u}}_j \approx \mathbf{u}_j$ for $1 \le j \le 3$, i.e. the first J eigenvectors of $\mathbf{W}$ are almost identical to the first J eigenvectors of $\mathbf{P}\mathbf{P}^T$. We also compare between the first 5 eigenvalues of both matrices:

$$\begin{aligned}&\lambda_1 = 167,\ \lambda_2 = 44,\ \lambda_3 = 37,\ \lambda_4 = 8e^{-15},\ \lambda_5 = 8e^{-15}\\ &\tilde{\lambda}_1 = 168,\ \tilde{\lambda}_2 = 44,\ \tilde{\lambda}_3 = 38,\ \tilde{\lambda}_4 = 0.7,\ \tilde{\lambda}_5 = 0.7. \quad (50)\end{aligned}$$

We observe that $|\tilde{\lambda}_j - \lambda_j| < 1$ as expected. Note that the slight differences between the eigenvalues, seem to have a minor impact on the decision rule of (19), for counting the number of sources. We conclude that the derivations in Section II, regarding the spectral decomposition of the matrix $\mathbf{P}\mathbf{P}^T$, apply also for the correlation matrix $\mathbf{W}$.

REFERENCES

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.
[2] T.-W. Lee, "Independent component analysis," in *Independent Component Analysis*. Springer, 1998, pp. 27–66.

[3] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4, pp. 411–430, 2000.
[4] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2004, vol. 46.
[5] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.
[6] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural computation*, vol. 13, no. 4, pp. 863–882, 2001.
[7] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*. Springer, 2007, vol. 615.
[8] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "Convolutive blind source separation methods," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 1065–1094.
[9] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," *Machine Audition: Principles, Algorithms and Systems*, pp. 162–185, 2010.
[10] N. Mitianoudis and M. E. Davies, "Audio source separation of convolutive mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 489–497, 2003.
[11] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.
[12] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.
[13] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.
[14] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.
[15] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 121–133, 2010.
[16] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.
[17] J. Traa and P. Smaragdis, "Multichannel source separation and tracking with RANSAC and directional statistics," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 2233–2243, 2014.
[18] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l 1-norm minimization," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 81–81, 2007.
[19] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.
[20] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online mvdr beamformer based on complex gaussian mixture model with spatial prior for noise robust asr," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 780–793, 2017.
[21] F. Abrard and Y. Deville, "A time–frequency blind signal separation method applicable to underdetermined mixtures of dependent sources," *Signal Processing*, vol. 85, no. 7, pp. 1389–1403, 2005.
[22] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.
[23] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.
[24] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 2, pp. 6–36, 2013.
[25] W.-K. Ma, J. M. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. J. Plaza, A. Ambikapathi, and C.-Y. Chi, "A signal processing perspective on hyperspectral unmixing: Insights from remote sensing," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 67–81, 2014.
[26] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain." *IEEE Transactions Signal Processing*, vol. 63, no. 9, pp. 2306–2320, 2015.
[27] H. Q. H. Dam and S. Nordholm, "Source separation employing beamforming and srp-phat localization in three-speaker room environments," *Vietnam Journal of Computer Science*, vol. 4, no. 3, pp. 161–170, 2017.
[28] M. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete latent variable decomposition of counts data," in *Advances in neural information processing systems*, 2008, pp. 1313–1320.
[29] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Low-artifact source separation using probabilistic latent component analysis," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.
[30] P. Smaragdis, "Approximate nearest-subspace representations for sound mixtures," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 5892–5895.
[31] S. E. Chazan, J. Goldberger, and S. Gannot, "DNN-based concurrent speakers detector and its application to speaker extraction with LCMV beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 6712–6716.
[32] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 1, pp. 70–79, 2007.
[33] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614 –1626, Aug. 2001.
[34] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, 2004.
[35] N. Madhu and R. Martin, "A scalable framework for multiple speaker localization and tracking," in *International Workshop for Acoustic Echo Cancellation and Noise Control*, 2008.
[36] Y. Dorfan and S. Gannot, "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 10, pp. 1692–1703, 2015.
[37] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech source separation and noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2013.
[38] N. Ito, S. Araki, and T. Nakatani, "Permutation-free clustering of relative transfer function features for blind source separation," in *23rd European Signal Processing Conference*, Nice, France, Sep. 2015.
[39] J. W. Boardman, "Automating spectral unmixing of aviris data using convex geometry concepts," in *The 4th Annual JPL Airborne Geoscience Workshop*, 1993, pp. 11–14.
[40] M. E. Winter, "N-findr: An algorithm for fast autonomous spectral end-member determination in hyperspectral data," in *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*. International Society for Optics and Photonics, 1999, pp. 266–275.
[41] J. M. Nascimento and J. M. Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 898–910, 2005.
[42] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvao, T. Yoneyama, H. C. Chame, and V. Visani, "The successive projections algorithm for variable selection in spectroscopic multicomponent analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 57, no. 2, pp. 65–73, 2001.
[43] B. Laufer, R. Talmon, and S. Gannot, "Diarization and separation based on a data-driven simplex," in *26th European Signal Processing Conference*, 2018.
[44] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 313–317.
[45] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
[46] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
[47] T. Kim, https://github.com/teradepth/iva, Feb. 2018.
[48] G. W. Stewart, *Matrix perturbation theory*. Academic Press, 1990.