Global and Local Simplex Representations for Multichannel Source Separation

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

***Abstract*—The problem of blind audio source separation (BASS) in noisy and reverberant conditions is addressed by a novel approach, termed Global and LOcal Simplex Separation (GLOSS), which integrates full- and narrow-band simplex representations. We show that the eigenvectors of the correlation matrix between time frames in a certain frequency band form a simplex that organizes the frames according to the speaker activities in the corresponding band. We propose to build two simplex representations: one global based on a broad frequency band and one local based on a narrow band. In turn, the two representations are combined to determine the dominant speaker in each time-frequency (TF). Using the identified dominating speakers, a spectral mask is computed and is utilized for extracting each of the speakers using spatial beamforming followed by spectral postfiltering. The performance of the proposed algorithm is demonstrated using real-life recordings in various noisy and reverberant conditions.**

***Index Terms*—BASS, multichannel, simplex, spectral mask, relative transfer function (RTF), beamformer.**

I. Introduction

Multichannel blind audio source separation (BASS) aims at extracting the individual source signals from multi-microphone recordings of a mixture with several concurrently active speakers [1]. Audio separation capabilities are required in various multi-microphone devices, such as: smart-phones, smart voice assistants and hearing aids, thus leading to an intense research in the field over the last decades [2]–[4]. However, existing algorithms typically suffer from artifacts and distortions, performance degradation in noisy and reverberat conditions, and high computational burden.

In general, in BASS no prior knowledge is available on the speakers and their positions. Most methods rely on some assumptions on the characteristics of the source signals. Assuming independence of the original source signals facilitates the use of independent component analysis (ICA) and independent vector analysis (IVA) separation methods [5]–[9]. Separation algorithms based on non-negative matrix factorization (NMF) rely on the assumption that the spectrum of the speakers can be decomposed as a multiplication of two nonnegative components, namely, a dictionary of recurring patterns and activation coefficients [10]–[14].

Another widely used assumption is that speech is W-disjoint orthogonal in the spectral-temporal domain, specifying that each TF bin in the short time Fourier transform (STFT) representation of the audio mixture is dominated by a single speaker [15]. Following this assumption, a large variety of algorithms were developed, aiming to form a *spectral mask*, which consists of the dominant speaker in each TF bin. Commonly, the mask is estimated using clustering methods that group together TF bins belonging to each speaker. Several methods perform clustering of all frequencies at once, using the phase difference or the time difference of arrival (TDOA) as a common thread that ties together the different frequencies to a specific speaker in a certain position [15]–[17]. These methods are usually greatly affected by reverberation, since additional reflections from various directions hide the direct path between the speakers and the microphones. Robustness to reverberation can be achieved by performing clustering according to the full reflection pattern associated with each speaker [18]–[20]. However, for these methods the clustering is implemented in each frequency separately, thus the identity of the extracted speakers may be different in different frequencies. Several methods exist for solving this permutation ambiguity problem, for instance, based on the TDOA or based on similar time-activity patterns of neighbouring frequencies [18], [21]. Unfortunately, such an additional permutation alignment stage increases the computational complexity, and may lead to sub-optimal performance when a perfect alignment is not achieved.

Recently, new deep neural network (DNN) techniques were harnessed to the BASS problem [22]. While most papers focus on single-channel separation, there are a few methods utilizing multi-microphone settings. In [23] a mask is learned using a neural-network that receives features based on estimated source positions, and is applied to the output of a delay-and-sum beamformer. In [24] a multichannel Wiener filter is derived combining source spectra that are estimated by DNNs, and spatial information that is inferred from a classical multichannel Gaussian model. Deep-clustering [25] is a single-channel method in which a multi-layered bidirectional long short term memory network (BLSTM) learns an embedding for each TF, such that embeddings belonging to the same speaker are close. In [26] the embedding generated by deep-clustering was extended to a multichannel scenario by adding spatial features as an additional input to the network. This approach was further improved in [27], where the separation is initially performed by a network designed for a two-channel input, and the outputs of this network for all pairs of microphones are merged into a single feature that is fed to an additional enhancement network. In [28] deep cluster-

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).

ing embeddings were integrated into a statistical model that combines the spectral information provided by the extracted embeddings with spatial information. In [29], [30] it was proposed to train a DNN to identify the number of concurrent speakers in each frame. The separation is performed by a beamformer constructed using a noise correlation matrix that is estimated during noise-only frames, and steering vectors that are estimated during single-speaker frames. Despite showing promising high-quality separation, most DNN-based approaches require a large amount of examples for training the respective networks. In addition, there is usually no guarantee that DNNs can generalize well to different scenarios, i.e. different number of speakers, acoustic conditions, array geometries, etc., beyond the examples that were used during training.

In this paper we present a novel algorithm for BASS, which aims at estimating a spectral mask, relying on the speech sparsity in the TF domain. The estimation is performed by a dual-stage approach. In the first stage we extract features for each frame based on a concatenation of several frequencies in a broad frequency band, and compute a correlation matrix between the features of all frames. The eigenvalue decomposition (EVD) of the correlation matrix is computed, and the eigenvectors spanning the matrix column space are shown to form a *global* representation that is organized in the shape of a simplex. Based on this representation, the global probabilities of activity of the speakers in each frame can be identified. In the next stage, we repeat the same process for each frequency bin separately, namely, we compute the correlation matrix between all frames based on each single frequency and extract the corresponding EVD. The eigenvectors computed based on a certain frequency form a *local* representation that organizes the frames according to the dominance of the speakers in this frequency. We combine the distances between frames in the local representation with the global probabilities, computed in the first stage, to infer the dominant speaker in each TF bin, which forms the estimated spectral mask. Due to the fact that for each frequency the dominant speaker is identified using the same global probabilities, we avoid the permutation ambiguity of speaker identities across different frequencies. For the separation, we first exploit the spatial diversity of the speakers by applying a beamformer, which is constructed based on the different acoustic systems of the speakers that are estimated using the extracted spectral mask. The separation is further enhanced by multiplying the beamformer output by a single-channel postfilter that is determined by the spectral mask. The proposed algorithm is termed Global and LOcal Simplex Separation (GLOSS).

The current contribution is closely related to our recently proposed simplex-based separation algorithm [31]. The algorithm in [31] is built upon the global simplex representation of all frequencies together, and therefore only provides a frame-wise decision about the activity of the speakers, and detects entire time frames dominated by each speaker. As the number of speakers increases, most frames consist of multiple active speakers and only few frames are dominated by a single speaker. Note also that since only frame-wise dominance is estimated in [31], the spectral sparsity of the speakers cannot be utilized during separation, and therefore only a beamforming-based separation is applied. In addition, in [31] only a noiseless scenario is considered. Furthermore, the approach for detecting the number of speakers in [31] is based on applying a threshold decision to the eigenvalues of the correlation matrix, where the threshold value is set manually. Determining the threshold value is not discussed in [31], and is often sensitive to the reverberation level and to the number of speakers.

In this paper we rely on the theoretical foundations presented in [31], while extendeding them and deriving an improved algorithm. In addition to the global representation of [31], we compute a local representation for each frequency separately, and present a new approach to tie together all the local representations according to the global probabilities in order to estimate the full TF mask. Furthermore, here we address a more practical setting that includes noise, and extend the simplex representation derived in [31] accordingly. In addition, we derive an improved procedure for detecting the number of speakers by a support vector machine (SVM) classifier [32] that is applied to the eigenvalues of the correlation matrix. In the experimental study, we use a dataset of real-life recordings with a large variety of both female and male speakers. We show that the proposed GLOSS algorithm significantly outperforms the algorithm in [31], obtaining higher separation scores when tested under various conditions.

The remainder of the paper is organized as follows. Section II presents the problem formulation. Section III describes the probabilistic model and the derivation of the simplex representation, which serve as a basis for the spectral mask estimation procedure by the GLOSS algorithm that is described in Section IV. The utilization of the estimated spectral mask for separation and enhancement is presented in Section V. The estimation of the number of speakers is discussed in Section VI. Section VII demonstrates the performance of the proposed GLOSS algorithm based on both real-life and simulated data with comparison to several baseline methods. The paper is concluded in Section VIII.

II. Problem Formulation

Consider J concurrent speakers located in a reverberant and noisy enclosure. The signals are measured by an array of M microphones, and are analysed in the STFT domain. The signal measured by the mth microphone is given by:

$$\begin{aligned} Y^m(l,f) &= \sum_{j=1}^{J} Y_j^m(l,f) + N^m(l,f) \\ &= \sum_{j=1}^{J} A_j^m(f) S_j(l,f) + N^m(l,f) \end{aligned} \tag{1}$$

where $A_j^m(f)$ is the acoustic transfer function (ATF) relating the jth speaker and the mth microphone, $S_j(l,f)$ is the signal emitted by the jth speaker, and $N^m(l,f)$ is the noise signal at the mth microphone. Here $f \in \{1,\ldots,K\}$ is the frequency bin and $l \in \{1,\ldots,L\}$ is the frame index. Note that in a typical reverberant enclosure, the ATF $A_j^m(f)$ consists of the direct path between the source and the microphone as well as

the reflections from the different surfaces and objects in the enclosure.

By the assumption of the speech sparsity in the STFT domain [15], each TF bin is dominated by either one of the speakers or consists of noise i.e.:

$$Y^m(l,f) = \begin{cases} A_j^m(f) S_j(l,f) & \text{if } (l,f) \in \mathcal{S}_j \\ N^m(l,f) & \text{if } (l,f) \in \mathcal{N} \end{cases} \tag{2}$$

where $\mathcal{S}_j$ denotes the collection of TF bins dominated by the jth speaker:

$$\mathcal{S}_j = \left\{ (l,f) \;\middle|\; |Y_j^1(l,f)|^2 \gg \left(\sum_{\substack{i=1 \\ i \neq j}}^{J} |Y_i^1(l,f)|^2 + |N^1(l,f)|^2 \right) \right\} \tag{3}$$

and $\mathcal{N}$ denotes the set of remaining TFs that are not dominated by any of the speakers, hence are considered noisy:

$$\mathcal{N} = \left\{ (l,f) \;\middle|\; (l,f) \notin \bigcup_{j=1}^{J} \mathcal{S}_j \right\}. \tag{4}$$

Let $\{M(l,f)\}_{l,f}$ denote the *spectral mask* that assigns each TF bin with its dominating *component*, either one of the J speakers or the noise:

$$M(l,f) = \begin{cases} j & \text{if } (l,f) \in \mathcal{S}_j \\ J+1 & \text{if } (l,f) \in \mathcal{N}. \end{cases} \tag{5}$$

Our goal is to recover the number of speakers J and to extract each of the individual speakers while suppressing the other speakers and the background noise. The separation will be performed based on a spectral mask, whose estimation is at the core of the proposed method. The signals of the different speakers will be extracted using a dual-stage process: first applying a multichannel beamformer that optimally combines the signals measured by the M microphones, and then implementing a single-channel postfilter. The multichannel beamformer, as well as the single-channel postfilter, are both implemented using the estimated spectral mask.

III. Simplex-Shaped Representation of Speaker and Noise Activities

In this section we describe how to extract a representation that organizes time-frames in a simplex that encodes the probability of activity of the speakers within each frame in a certain frequency band. We start with some preliminary probabilistic assumptions. Next, we define features that are extracted for each frame, and describe how to exploit the extracted features for deriving the simplex-shaped representation, while relying on the assumed probabilistic model.

A. Probabilistic Model

We assume that the dominant component is independently randomly selected in each TF bin according to:

$$M(l,f) = \begin{cases} j,\ 1 \leq j \leq J & \text{with probability } p_j(l) \\ J+1 & \text{with probability } 1 - \sum_{j=1}^{J} p_j(l) \end{cases} \tag{6}$$

where $\sum_j p_j(l) \leq 1$. Note that the probabilities $\{p_j(l)\}_{j=1}^{J}$ depend only on the frame index and not on the frequency index.

For each TF bin, consider the following ratio between the mth microphone and the first microphone that serves as a reference microphone:

$$R^m(l,k) = \frac{Y^m(l,f)}{Y^1(l,f)} \tag{7}$$

Based on the sparsity assumption (2) we get:

$$R^m(l,f) = \begin{cases} H_j^m(f) & \text{if } M(l,f) = j,\ 1 \leq j \leq J \\ \eta(l,f) & \text{if } M(l,f) = J+1 \end{cases} \tag{8}$$

where

$$H_j^m(f) = \frac{A_j^m(f)}{A_j^1(f)}. \tag{9}$$

is the relative transfer function (RTF) [33], [34] defined 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. Here $\eta(l,f) = N^m(l,f)/N^1(l,f)$ is a noise term that is both frequency and frame dependent. According to (8), the ratio $R^m(l,f)$ is the RTF of one of the speakers $H_j^m(f)$ or a noise term $\eta(l,f)$.

We assume that the RTFs and the noise terms are independent zero-mean random variables. The RTFs of different speakers, frequencies or microphones are assumed to be independent, and the same holds for the noise terms of different frequencies or frames. Further discussion on the validity of these assumptions can be found in [31]. For the sake of simplicity, we assume that the variance of the real and the imaginary parts of each RTF equals 1, i.e. $E\left\{\text{real}\{H_j^m(f)\}^2\right\} = E\left\{\text{imag}\{H_j^m(f)\}^2\right\} = 1$. Note that the following derivation also holds for non-unit and non-constant variance by applying a proper normalization. We comment that the noise is assumed to be non-directional. Directional noises can be treated as additional sources, increasing J accordingly.

B. Feature Extraction

Based on the computed ratios (7) in each TF bin, we extract a feature vector $\mathbf{r}(l)$ for each frame l by concatenating the ratios of all microphones in a specific frequency band. The vector $\mathbf{r}(l)$ consists of $D = 2 \cdot (M-1) \cdot F$ elements of the real and imaginary parts of the ratios, in $1 \leq F \leq K$ frequency bins and in $M-1$ microphones (except for the reference microphone):

$$\begin{aligned} \mathbf{r}^m(l) &= [R^m(l,f_1), R^m(l,f_2), \ldots, R^m(l,f_F)]^T \\ \mathbf{r}^{\text{c}}(l) &= \left[\mathbf{r}^{2,T}(l), \mathbf{r}^{3,T}(l), \ldots, \mathbf{r}^{M,T}(l)\right]^T \\ \mathbf{r}(l) &= \left[\text{real}\left\{\mathbf{r}^{\text{c}}(l)\right\}^T, \text{imag}\left\{\mathbf{r}^{\text{c}}(l)\right\}^T\right]^T. \end{aligned} \tag{10}$$

where $\mathcal{F} = \{f_1, f_2, \ldots, f_K\}$ is the chosen frequency band.

We compute (7) and (10) for each frame $1 \leq l \leq L$, and form the set $\{\mathbf{r}(l)\}_{l=1}^{L}$ of all features. The feature vectors can be related to a set of J unknown RTF vectors of the same dimension, which consists of the RTF values of each of the speakers. Let $\mathbf{h}_j$ denote an RTF vector associated with the jth speaker. Similarly to the definition of (10), each RTF vector consists of the real and imaginary parts of the RTF values, in F frequency bins and in $M-1$ microphones:

$$\begin{aligned} \mathbf{h}_j^m &= \left[H_j^m(f_1), H_j^m(f_2), \ldots, H_j^m(f_F)\right]^T \\ \mathbf{h}_j^{\mathrm{c}} &= \left[\mathbf{h}_j^{2,T}, \mathbf{h}_j^{3,T}, \ldots, \mathbf{h}_j^{M,T}\right]^T \\ \mathbf{h}_j &= \left[\mathrm{real}\left\{\mathbf{h}_j^{\mathrm{c}}\right\}^T, \mathrm{imag}\left\{\mathbf{h}_j^{\mathrm{c}}\right\}^T\right]^T. \end{aligned} \tag{11}$$

According to (8), each entry of the feature vector $\mathbf{r}(l)$ is associated with an entry of one of the RTF vectors (11) or to a noise term. The expected number of entries in the vector $\mathbf{r}(l)$ corresponding to a particular RTF $\mathbf{h}_j$ is given by the probability $p_j(l)$ of the jth speaker:

$$\begin{aligned} &E\left\{\frac{1}{D}\sum_{k=1}^{D}[r(l,k) = h_j(k)]\right\} \\ &= \frac{1}{D}\sum_{k=1}^{D} E\left\{[r(l,k) = h_j(k)]\right\} = p_j(l) \end{aligned} \tag{12}$$

where $r(l,k)$ and $h_j(k)$ denote the kth entry of the vectors $\mathbf{r}(l)$ and $\mathbf{h}_j$, respectively.

In Fig. 1, we illustrate a mixture of $J = 2$ speakers. In the illustration, there are two RTF vectors of the two speakers, colored by shades of red and blue. An example of a feature vector $\mathbf{r}(l^*)$, associated with frame l^*, is also presented. The entries of $\mathbf{r}(l^*)$ are in red shade, blue shade or gray texture if dominated by the first speaker, the second speaker or noise, respectively. The probabilities associated with frame l^* are written above the feature vector: $p_1(l^*) = 0.5$ for the first speaker, and $p_2(l^*) = 0.2$ for the second speaker. Hence, in the feature vector $\mathbf{r}(l^*)$, 50% of the entries are identical to entries in $\mathbf{h}_1$, 20% of the entries are identical to entries in $\mathbf{h}_2$, and 30% of the entries are noisy. Note that in this example, the number of entries in $\mathbf{r}(l^*)$ corresponding to each RTF vector exactly matches the associated probability, while in practice it is only approximately satisfied due to randomness in the selection of the most dominant component in each entry.

C. Simplex-Shaped Representation

We describe the derivation of the simplex-shaped representation based on the derived features (10). According to the probabilistic model presented in Section III-A, the inner-product between each two features $\mathbf{r}(l)$ and $\mathbf{r}(n)$, $1 \leq l, n \leq L$ is given by:

$$\frac{1}{D}\mathbf{r}^T(l)\mathbf{r}(n) = \begin{cases} \sum_{j=1}^{J} p_j(l)p_j(n) & \text{if } l \neq n \\ \sum_{j=1}^{J} p_j(l) & \text{if } l = n \end{cases}. \tag{13}$$

The derivation of (13) is given in Appendix A, showing that this inner-product between the features approximates the correlation between each two entries of the feature vectors.

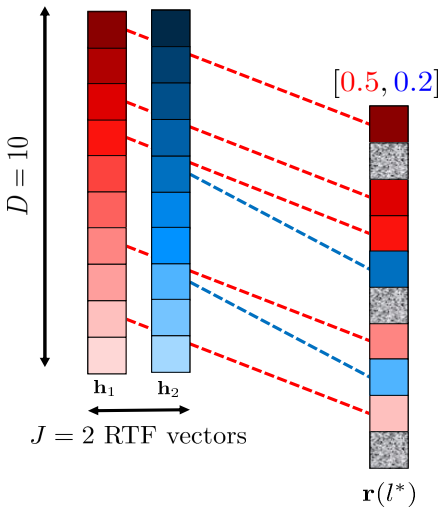


Fig. 1. An illustration of the presented statistical mixture model. In this example there are $J = 2$ speakers associated with 2 unknown RTF vectors $\{\mathbf{h}_j\}_{j=1}^{2}$. Each RTF vector consists of $D = 10$ elements, characterized by varying shades of red or blue. An example of a feature vector associated with frame l^* is also illustrated. Dashed lines are drown between each entry in $\mathbf{r}(l^*)$ and the associated entry of the RTF vector from which it originated. Entries dominated by noise are colored by gray texture. The set of probabilities $[p_1(l), p_2(l)]$ used to construct $\mathbf{r}(l^*)$ is written above it. Note that in this example, the number of entries corresponding to each RTF exactly matches the associated probability, while in practice it is only approximately satisfied.

Let $\mathbf{W}$ be the $L \times L$ correlation matrix, with $W_{ln} = \frac{1}{D}\mathbf{r}^T(l)\mathbf{r}(n)$. According to (13) the correlation matrix can be recast as:

$$\mathbf{W} = \mathbf{P}\mathbf{P}^T + \Delta\mathbf{W} \tag{14}$$

where $\mathbf{P}$ is an $L \times J$ matrix, whose rows consist of the probabilities associated with each frame, i.e. $P_{lj} = p_j(l)$, and $\Delta\mathbf{W}$ is a diagonal matrix with $\Delta W_{ll} = \sum_{j=1}^{J}(1 - p_j(l))p_j(l) \leq 1$ on its main diagonal and zero elsewhere. The matrix $\Delta\mathbf{W}$ has a negligible effect on the spectral decomposition of $\mathbf{W}$ (which can be shown using similar considerations to the ones presented in Appendix B in [31]). Therefore, henceforth we omit $\Delta\mathbf{W}$ from our derivations, and resort to the approximation of the correlation matrix by:

$$\mathbf{W} \approx \mathbf{P}\mathbf{P}^T. \tag{15}$$

Applying EVD to the correlation matrix $\mathbf{W}$, sets of L eigenvectors $\{\mathbf{u}_j\}_{j=1}^{L}$ and L eigenvalues $\{\lambda_j\}_{j=1}^{L}$ are obtained. For $L > J$ we deduce from the above decomposition that:

$$\mathrm{rank}(\mathbf{W}) = \mathrm{rank}(\mathbf{P}) = J \tag{16}$$

indicating that the rank of the correlation matrix is directly determined by the number of speakers. We address the inference of the number of speakers based on the information implied by the rank of the correlation matrix in Section VI, and proceed for now by assuming that the value of J is known.

Let $\mathbf{U}_J = [\mathbf{u}_1, \ldots, \mathbf{u}_J]$ denote an $L \times J$ matrix, which consists of the first J eigenvectors of $\mathbf{W}$ associated with its J largest eigenvalues. Based on (15), the column spaces of

$\mathbf{U}_J$ and $\mathbf{P}$ are equal, i.e. the columns are related by a linear invertible transformation:

$$\mathbf{U_J} = \mathbf{P}\mathbf{G}^{\mathbf{T}} \tag{17}$$

where $\mathbf{G}$ is a $J \times J$ invertible matrix.

For each frame $1 \leq l \leq L$, we define the following *mapping* vector:

$$\boldsymbol{\nu}(l) = [u_1(l), u_2(l), \ldots, u_J(l)]^T \tag{18}$$

which corresponds to the lth row of the matrix $\mathbf{U}_J$. Let $\mathbf{p}(l) = [p_1(l), p_2(l), \ldots, p_J(l)]^T$ denote the probability vector associated with the lth frame, corresponding to the lth row of the matrix $\mathbf{P}$. According to (17), the mapping vector $\boldsymbol{\nu}(l)$ is a linear transformation of the probability vector $\boldsymbol{p}(l)$:

$$\boldsymbol{\nu}(l) = \mathbf{G}\mathbf{p}(l). \tag{19}$$

By (19) we infer that the mapping $\boldsymbol{\nu}(l)$ is closely related to the probability vector $\mathbf{p}(l)$, which contains an important information about the activity of the speakers. However, the transformation from the given mappings to the unknown probabilities is non-trivial, since the transformation matrix $\mathbf{G}$ is also unknown. In order to solve the problem, we investigate the geometrical structure of both $\{\mathbf{p}(l)\}_{l=1}^L$ and $\{\boldsymbol{\nu}(l)\}_{l=1}^L$, and show how it can be utilized for recovering the unknown transformation between the two sets.

Let $\mathbf{e}_j = [0, \ldots, 1, \ldots, 0]^T \in \mathbb{R}^J, 1 \leq j \leq J$ denote the standard unit vectors, with one at the jth entry and zeros elsewhere. The probability vectors $\{\mathbf{p}(l)\}_{l=1}^L$ are confined to a J-dimensional simplex defined by the standard unit vectors:

$$\Theta = \left\{ \theta_1\mathbf{e}_1 + \ldots + \theta_J\mathbf{e}_J \,\middle|\, \sum_{j=1}^J \theta_j \leq 1, \theta_j \geq 0 \right\}. \tag{20}$$

The simplex structure (20) is formed due to the fact that each probability vector can be expressed as a convex combination of $\{\mathbf{e}_j\}_{j=1}^J$:

$$\mathbf{p}(l) = \sum_{j=1}^J p_j(l)\mathbf{e}_j, \quad \sum_{j=1}^J p_j(l) \leq 1. \tag{21}$$

Note that the simplex (20) has $J+1$ *vertices*: J vertices at the standard unit vectors $\{\mathbf{e}_j\}_{j=1}^J$ and an additional vertex at the origin (due to the fact that the sum of the probabilities can be less than 1).

According to the linear transformation between $\boldsymbol{\nu}(l)$ and $\mathbf{p}(l)$ implied by (19), the mapping vectors $\{\boldsymbol{\nu}(l)\}_{l=1}^L$ occupy a simplex, which is a rotated and scaled version of the simplex in (20):

$$\Theta^* = \left\{ \theta_1\mathbf{e}_1^* + \ldots + \theta_J\mathbf{e}_J^* \,\middle|\, \sum_{j=1}^J \theta_j \leq 1, \theta_j \geq 0 \right\} \tag{22}$$

where

$$\mathbf{e}_j^* = \mathbf{G}\mathbf{e}_j = \mathbf{g}_j \tag{23}$$

with $\mathbf{g}_j$ denoting the jth column of the matrix $\mathbf{G}$. The transformed simplex (22) also has $J+1$ vertices, one in the origin and additional J vertices at $\{\mathbf{e}_j^*\}_{j=1}^J$. Accordingly, frames with high probability of the jth speaker are concentrated near the jth vertex $\mathbf{e}_j^*$, and noisy frames with low probability of each of the speakers are concentrated near the origin.

Note that (23) implies that the J columns of the transformation matrix G coincide with the J vertices of the transformed simplex Θ^* (except for the trivial vertex at the origin). Therefore, if the vertices $\{\mathbf{e}_j^*\}_{j=1}^J$ of the transformed simplex (22) can be recovered among the available set of mapping vectors $\{\boldsymbol{\nu}(l)\}_{l=1}^L$, then they can be used to form the transformation matrix $\mathbf{G}$. Using the inverse of $\mathbf{G}$, the mapping vectors $\boldsymbol{\nu}(l)$ can be transformed back to the probability vectors $\mathbf{p}(l)$, following the relation in (19). This principle will be utilized for recovering the probability vectors in the proposed method that is described in Section IV.

IV. Spectral Mask Estimation Combining Global and Local Simplex Representations

Our aim is to use the simplex representation derived by the eigenvectors of the correlation matrix to obtain a spectral mask. For this purpose, we form full- and narrow-band simplex representations that recover the activity of the speakers in different scales. We use a *global* mapping based on a broad range of frequencies, as well as *local* mappings based on single frequencies. We combine the global and the local representations to extract information on the most dominant component in each TF bin, and accordingly, we form the spectral mask.

A. Global Mapping

We start with a global simplex mapping based on a broad range of frequencies, which provides a global organization of the frames according to the overall activity of the speakers in each frame. To form the mapping we first compute the ratios (7) and the feature vectors (10) based on a wide frequency range $\mathcal{F}_G$. Then, the correlation matrix between all feature vectors is constructed, and its EVD is computed. Based on the derived eigenvectors, the global mapping denoted by $\boldsymbol{\nu}^{\mathrm{G}}(l)$ is defined based on (18).

We use the global mapping to estimate the speaker probabilities associated with each frame. Note that these probabilities represent the global activity of the speakers in each frame, namely, they reflect the relative dominance of each of the speakers in the entire frame. The simplex vertices corresponding to the different speakers and assigned with indexes $\{l_j\}_{j=1}^J$, are identified using a *successive projection* algorithm [35], [36]. This algorithm, summarized in Algorithm 2, is based on successively identifying the simplex vertices by maximizing the projection onto the orthogonal complement of the space spanned by the previously identified vertices. The recovered vertices can be used to transform the mapping $\boldsymbol{\nu}^{\mathrm{G}}(l)$ to the speaker probabilities $\mathbf{p}^{\mathrm{G}}(l)$, by applying the inverse transformation of (19):

$$\mathbf{p}^{\mathrm{G}}(l) = \hat{\mathbf{G}}^{-1}\boldsymbol{\nu}^{\mathrm{G}}(l) \tag{24}$$

where the matrix $\hat{\mathbf{G}}$ is constructed according to (23), based on the identified vertices:

$$\hat{\mathbf{G}} = \left[\boldsymbol{\nu}^{\mathrm{G}}(l_1), \boldsymbol{\nu}^{\mathrm{G}}(l_2), \ldots, \boldsymbol{\nu}^{\mathrm{G}}(l_J)\right]. \tag{25}$$

The noise probability is given by $p^{\mathrm{G}}_{J+1}(l) = 1 - \sum_{j=1}^{J} p^{\mathrm{G}}_{j}(l)$. A demonstration of the global simplex representation is given in Fig. 2. The demonstration corresponds to a mixture of $J = 2$ speakers, which is generated by the setup described in Section VII. Frames are colored according to the oracle probability of (a) the first speaker, (b) the second speaker (c) the noise. In this case, the simplex is of dimension 2, consisting of three vertices: two vertices corresponding to frames with high probability of one of the speakers, and one vertex at the origin corresponding to noisy frames.

B. Local Mappings

We construct local mappings based on each single frequency $f \in \{1, \ldots, K\}$. For each frequency we extract features (11) of length $D = 2 \cdot (M-1)$, in which the ratio values (7) are concatenated in all microphones except for the reference microphone. Next, we construct the correlation matrix associated with the fth frequency, compute its EVD, and derive the mapping $\boldsymbol{\nu}^{\mathrm{L}}(l,f)$ based on (18).

We would like to recover the dominant component, either one of the speakers or the noise, in each frame. The index of the dominant component in each TF is chosen by combining the relations between the local mappings $\boldsymbol{\nu}^{L}(l,f)$ with the global probabilities $\mathbf{p}^{G}(l)$ estimated on the basis of the global mapping. For each frame, the assignment is determined based on the following weighted nearest-neighbour rule:

$$M(l,f) = \underset{j \in \{1,\ldots,J+1\}}{\operatorname{argmax}} \frac{1}{\pi_j} \sum_{n=1}^{L} \omega^{L}_{ln}(f) \cdot p^{G}_{j}(n) \tag{26}$$

where the weight $\omega^{L}_{ln}(f)$ of each frame n with respect to the inspected frame l is inversely proportional to distances in the space defined by the local representation $\{\boldsymbol{\nu}^{L}(l,f)\}_{l=1}^{L}$. Particularly, we use the following Gaussian weighting based on Mahalanobis distance:

$$\omega^{L}_{ln}(f) = \exp\Big\{ -\big(\boldsymbol{\nu}^{L}(l,f) - \boldsymbol{\nu}^{L}(n,f)\big)^{T} \boldsymbol{\Sigma}^{-1} \times \big(\boldsymbol{\nu}^{L}(l,f) - \boldsymbol{\nu}^{L}(n,f)\big) \Big\}. \tag{27}$$

where $\boldsymbol{\Sigma}$ is the sample covariance of $\{\boldsymbol{\nu}^{L}(l,f)\}_{l=1}^{L}$. Note that alternative weights that decay with respect to distances between the local mappings can be applied as well. In (26), π_j serves as a class normalization, and is given by:

$$\pi_j = \sum_{n=1}^{L} p^{G}_{j}(n) \tag{28}$$

Note that in (26) we obtain the estimated spectral mask, where the global mapping serves as the common thread that ties all local mappings of each individual frequency. Since the local mappings are aligned using the same global mapping, we circumvent possible permutation ambiguity across frequencies. In this sense, the proposed method presents a different approach from common separation schemes that perform clustering in each frequency bin [18]–[20]. In contrast, in the proposed method the association of TF bins with the different speakers is formulated as a classification task, where the global probabilities serve as labels.

The global mapping was used in [31] to identify frames with a single active speaker. The identified single-speaker frames were utilized for estimating the RTF of each speaker. Frames with multiple active speakers were not utilized in [31]. Here, we utilize all measured frames, and obtain a finer decision on each frequency bin individually.

V. Separation and Enhancement

We use the estimated spectral mask (26) to extract each of the speakers while suppressing the other speakers and reducing the noise level. To this end, we use a two-step approach: we first apply a multichannel beamformer that takes advantage of the spatial diversity of the speakers in the different positions, and then apply a single-channel spectral masking to further utilize the spectral diversity of the speakers in the TF domain.

In the first stage, we apply a linearly constrained minimum variance (LCMV) beamformer, defined by:

$$\mathbf{b}^{\mathrm{LCMV}}(f) = \boldsymbol{\Phi}^{-1}_{nn}(f)\mathbf{C}(f)\left(\mathbf{C}^{H}(f)\boldsymbol{\Phi}^{-1}_{nn}(f)\mathbf{C}(f)\right)^{-1}\mathbf{g} \tag{29}$$

where $\boldsymbol{\Phi}_{\mathrm{vv}}(f)$ is the noise power spectral density (PSD) matrix of size $M \times M$ and $\mathbf{C}(f)$ is an $M \times J$ matrix, which consists of the RTFs of all speakers, i.e.:

$$\begin{aligned} \mathbf{C}(f) &= [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_J] \\ \mathbf{c}_j(f) &= [H^1_j(f), H^2_j(f), \ldots, H^M_j(f)]^T. \end{aligned} \tag{30}$$

In (29) $\mathbf{g} \in \mathbb{R}^J$ extracts the jth speaker, with one in the jth entry and zeros elsewhere. Both the RTFs of each of the speakers and the noise covariance matrix can be estimated based on the derived spectral mask. The estimation of $\boldsymbol{\Phi}_{\mathrm{vv}}(f)$ is based on frequencies dominated by the noise:

$$\hat{\boldsymbol{\Phi}}_{nn}(f) = \frac{1}{\sum_{l=1}^{L} \mathbb{I}_{J+1}(l,f)} \sum_{l=1}^{L} \mathbb{I}_{J+1}(l,f)\mathbf{y}(l,f)\mathbf{y}^{H}(l,f) \tag{31}$$

where

$$\mathbb{I}_j(l,f) = \begin{cases} 1 & \text{if } M(l,f) = j \\ 0 & \text{if } M(l,f) \neq j \end{cases}. \tag{32}$$

and

$$\mathbf{y}(l,f) = \left[Y^1(l,f), Y^2(l,f), \ldots, Y^M(l,f)\right]^T. \tag{33}$$

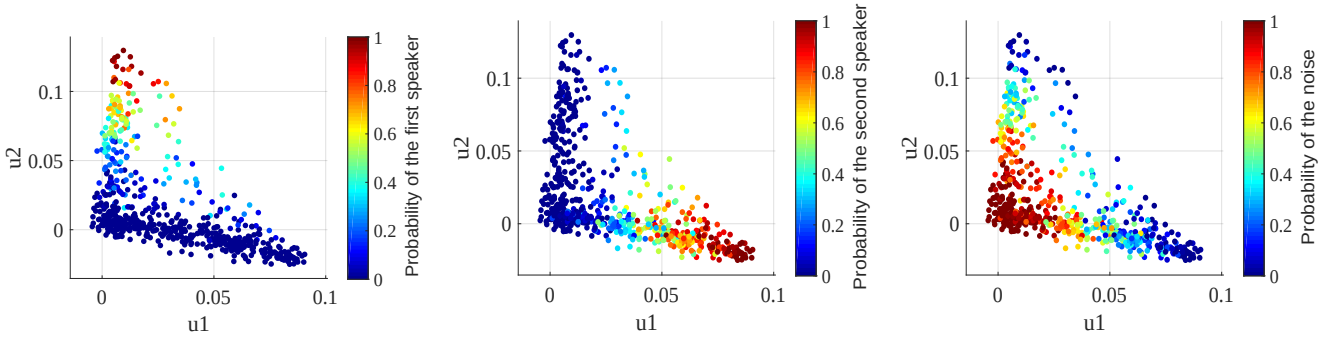
An estimator of the RTF of the jth speaker, is obtained by solving the following generalized eigenvalue decomposition (GEVD) problem [33]:

$$\hat{\boldsymbol{\Phi}}_{jj}(f)\boldsymbol{\psi}_j(f) = \mu\hat{\boldsymbol{\Phi}}_{nn}(f)\boldsymbol{\psi}_j(f) \tag{34}$$

where $\hat{\boldsymbol{\Phi}}_{jj}(f)$ is computed based on TF bins dominated by the jth speaker:

$$\hat{\boldsymbol{\Phi}}_{jj}(f) = \frac{1}{\sum_{l=1}^{L} \mathbb{I}_{j}(l,f)} \sum_{l=1}^{L} \mathbb{I}_{j}(l,f)\mathbf{y}(l,f)\mathbf{y}^{H}(l,f) \tag{35}$$

Assuming $\boldsymbol{\psi}_j(f)$ is the eigenvector associated with the largest eigenvalue μ, the vector $\tilde{\mathbf{c}}_j = \hat{\boldsymbol{\Phi}}_{nn}(f)\boldsymbol{\psi}_j(f)$ is a scaled version of the RTF of the jth speaker. Since, by definition



(a) (b) (c)

Fig. 2. Scatter plots of the global mappings $\{\boldsymbol{\nu}^G(l)\}_{l=1}^L$ obtained from the eigenvectors of the correlation matrix $\mathbf{W}$ between different time frames. Scatter plots correspond to a mixture of $J=2$ speakers. Frames are colored according to the oracle probability of (a) the first speaker, (b) the second speaker (c) the noise.

Algorithm 1: Successive Projection Algorithm [36]

- $\mathbf{P}^{\perp}=\mathbf{I}$
- **for** $j=1:J$ **do**
  - $\diamond$ $l_j=\text{argmax}_{l\in\{1,\ldots,L\}}\,\|\mathbf{P}^{\perp}\boldsymbol{\nu}(l)\|_2^2$
  - $\diamond$ $\mathbf{d}_j=\mathbf{P}^{\perp}\boldsymbol{\nu}(l_j)$
  - $\diamond$ $\mathbf{P}^{\perp}=\left(\mathbf{I}-\mathbf{d}_j\mathbf{d}_j^T/\|\mathbf{d}_j\|_2^2\right)\mathbf{P}^{\perp}$
- **end**

Algorithm 2: Simplex Mapping

Feature Extraction:

- Construct ratio vectors $\{\mathbf{r}(l)\}_{l=1}^L$ (10) based on the frequencies in the set $\mathcal{F}$.
- Compute the correlation matrix $\mathbf{W}$ by $W_{ln}=\frac{1}{D}\mathbf{r}^T(l)\mathbf{r}(n)$.
- Apply EVD to $\mathbf{W}$ to obtain $\{\mathbf{u}_j\}_{j=1}^L$.
- Construct $\boldsymbol{\nu}(l)=[\mathbf{u}_1(l),\mathbf{u}_2(l),\ldots,\mathbf{u}_J(l)]$.

$H_j^1(f)=1$, $\tilde{\mathbf{c}}_j$ can be normalized to yield a proper estimate of the RTF of the jth speaker:

$$\hat{\mathbf{c}}_j=\frac{\tilde{\mathbf{c}}_j}{[\tilde{\mathbf{c}}_j]_1} \tag{36}$$

where $[\tilde{\mathbf{c}}_j]_1$ denotes the first element of the vector $\tilde{\mathbf{c}}_j$.

The LCMV (29) implemented based on the estimated noise PSD matrix (31) and the estimated RTFs (36) is applied to the multichannel recordings as:

$$\hat{Y}_j^{\text{LCMV}}(l,f)=\left(\mathbf{b}^{\text{LCMV}}(f)\right)^H\mathbf{y}(l,f) \tag{37}$$

The output of the LCMV beamformer contains residual noise and interferences, which can be further suppressed applying single-channel masking that takes advantage of the assumed TF sparsity of the speech signals. To this end, we exploit again the estimated spectral mask by:

$$\begin{aligned}\hat{Y}_j^{\text{LCMV+MASK}}(l,f)&=\mathbb{I}_j(l,f)\hat{Y}_j^{\text{LCMV}}(l,f)\\&+\beta(f)\left(1-\mathbb{I}_j(l,f)\right)\hat{Y}_j^{\text{LCMV}}(l,f)\end{aligned} \tag{38}$$

where $\beta(f)$ is a frequency-dependent attenuation factor, that is applied in noisy TF bins to obtain a smoother frequency behaviour, thus mitigating musical noise effects. The proposed method is termed Global and LOcal Simplex Separation (GLOSS) and is summarized in Algorithm 3.

VI. Speaker Counting

In this section we address the problem of estimating the unknown number of speakers J. According to (16) the rank of the correlation matrix $\mathbf{W}$ equals J, therefore directly indicates the number of speakers. However, in practice, due to estimation inaccuracies the matrix is full-rank. The decay of the eigenvalues should reflect the matrix expected theoretical rank. In [31] we proposed to determine the number of speakers using a threshold rule on the eigenvalues. This approach has a disadvantage since it is not clear how the threshold should be determined. Moreover, we have shown in [31] that the optimal threshold is influenced by the reverberation time, preferring higher threshold values as the reverberation time increases. In addition, as the number of speakers increases, the spread of the eigenvalues also increases, i.e. lower threshold values are required for increasing number of speakers.

Here we propose a more robust approach for determining the number of speakers. We formulate the problem as a classification task in which each class represents a different number of speakers. The classifier is built based on a training set of prerecorded measurements with varying number of speakers in different conditions. The feature given as an input to the classifier is a vector consisting of the first τ eigenvalues of the correlation matrix:

$$\boldsymbol{\psi}=[\lambda_1,\lambda_2,\ldots,\lambda_\tau]^T \tag{39}$$

where τ equals (at least) to the maximum possible number of speakers. The problem is linearly separable as the value of λ_j (the jth coordinate $\boldsymbol{\psi}$) distinguishes between classes with number of speakers smaller than j to classes with number of speakers larger than j. This property is illustrated in Fig. 3. The illustration corresponds to recordings with $J\in\{1,2,3,4\}$ speakers. The full setup for genertaing this

Algorithm 3: GLOSS Algorithm

Feature Extraction:

- Compute ratios $\{R^m(l,f)\}_{l,f,m}$ (7).

Global Simplex Mapping:

- Construct global mapping $\{\boldsymbol{\nu}^G(l)\}_{l=1}^L$ applying Algorithm 2 on a global frequency range $\mathcal{F}^G$.
- Recover simplex vertices $\{l_j\}_{j=1}^J$ by Algorithm 1.
- Estimate global probabilities $\{\mathbf{p}^G(l)\}_{l=1}^L$ (24).

Local Simplex Mapping:

- **for** each frequency $f = 1 : K$:
  - ◇ Construct the local mapping $\{\boldsymbol{\nu}^L(l,f)\}_{l=1}^L$ applying Algorithm 1 on frequency f.
  - ◇ Compute weights $\{\omega_{ln}^L(f)\}_{l,n}$ (27).
  - ◇ Estimate the mask $M(l,f)$ (26) for each frame.

  end

Separation and Enhancement

- Estimate noise PSD $\hat{\mathbf{\Phi}}_{nn}(f)$ from $M(l,f)$, $\mathbf{y}(l)$ (31).
- Estimate RTFs of all speakers $\{\hat{\mathbf{c}}_j\}_{j=1}^J$ from $M(l,f)$, $\mathbf{y}(l)$, $\hat{\mathbf{\Phi}}_{nn}(f)$ (34)-(36).
- Compute the LCMV beamformer $\mathbf{b}^{\text{LCMV}}(f)$ from $\hat{\mathbf{\Phi}}_{nn}(f)$, $\{\hat{\mathbf{c}}_j\}_{j=1}^J$ (29)
- Obtain $\hat{Y}_j^{\text{LCMV}}(l,f)$ by applying the LCMV beamformer $\mathbf{b}^{\text{LCMV}}(f)$ on $\mathbf{y}(l)$ (37).
- Obtain $\hat{Y}_j^{\text{LCMV+MASK}}(l,f)$ by applying a single-channel spectral masking on $\hat{Y}_j^{\text{LCMV}}(l,f)$ (38).

illustration will be described in Section VII. Each point in the figure represents a single recording, colored according to the corresponding number of speakers. Since it is difficult to present a demonstration with more than two dimensions, we split the presentation into three figures, where in Fig. 3 (a) the axes correspond to λ_1 and λ_2, in Fig. 3 (b) the axes correspond to λ_2 and λ_3, and in Fig. 3 (c) the axes correspond to λ_3 and λ_4. In Fig. 3 (a) we observe that the value of λ_2 distinguishes between recordings with a single speaker to recordings with two or more speakers. Fig. 3 (b) shows that the value of λ_3 distinguishes between recordings with two speakers to recordings with three or four speakers, while recordings with a single speaker are concentrated near the origin. Similarly, in Fig. 3 (c) the value of λ_4 distinguishes between recordings with three speakers to recordings with four speakers, while recordings with one or two speakers are concentrated near the origin. Motivated by this observation, we propose to use a linear classifier such as the multiclass SVM classifier [32].

Note that as opposed to the proposed separation algorithm, which is entirely unsupervised, the proposed speaker counting scheme is supervised, and requires a set of prerecorded or simulated measurements. For these measurements we need to know only the number of participating speakers, and not the signals of the individual speakers. Therefore, this type of training data can be easily collected.

VII. Experimental Study

In this section we present a performance evaluation of the proposed GLOSS algorithm using several datasets with both simulated and real-life measurements. We first present the separation performance assuming that the correct number of speakers is known, and evaluate the source counting accuracy in a separate experiment.

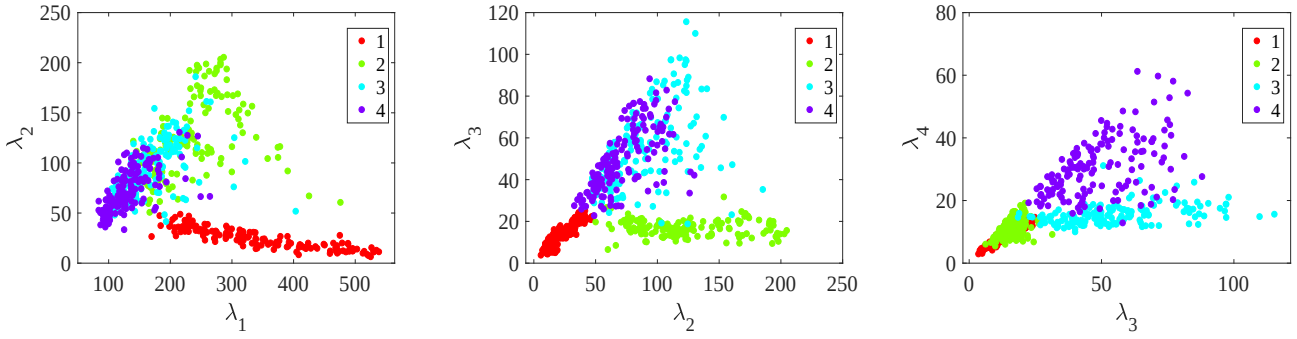
First, a comprehensive performance evaluation is carried out using real-life measurements recorded at the Bar-Ilan University (BIU) acoustic lab. We start by describing the real-life recording setup. Next, we list the performance measures used for assessment of the separation performance and describe the baseline methods that are used for comparison. Finally, the separation results are presented under various conditions. In addition, we compare the proposed method to two recent DNN-based separation algorithms [26], [37] using the same datasets that were used by the authors for evaluating their proposed methods. Finally, we present the performance of the proposed procedure for source counting.

A. Experimental Setup

The recordings are carried out at the BIU acoustic lab, which is a $6 \times 6 \times 2.4$ room covered by two sided controllable panels that can be configured in different ways to imitate various reverberation conditions. Two lab configurations of two reverberation levels were tested, namely low reverberation with $T_{60} \approx 150$ ms and high reverberation with $T_{60} \approx 550$ ms. In the room, a rectangular table was placed with six chairs around it. Signals were recorded by 32 microphones. On the table 24 microphones were placed: 6 microphones were set in a uniform linear array (ULA), 18 microphones were set in 6 nodes of 3 microphones each, and the remaining 8 microphones were placed around the table in 4 nodes of 2 microphones each. A top view of the recording setup is depicted in Fig. 4(a) and a photograph of the room configuration is presented in Fig. 4(b).

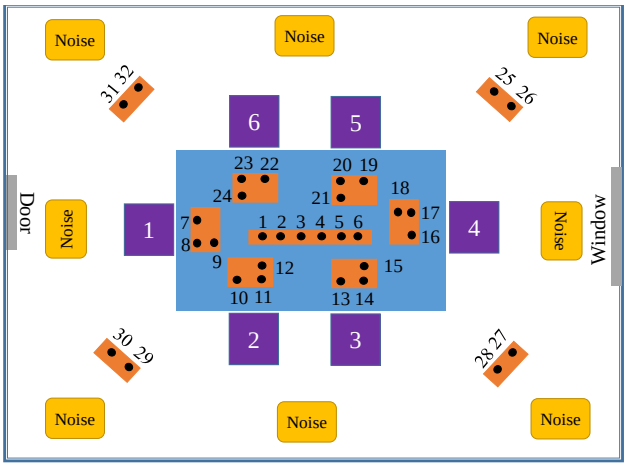
Twenty native English speakers were recorded while sitting on the chairs around the table: 10 males and 10 females. Each speaker was recorded separately, in order to allow flexibility in forming different mixtures with different number of speakers, at various locations. Each speaker was recorded 12 times while sitting in each one of the 6 chairs, and repeated twice for the two reverberation levels. Each recording is 55 s length, in which the speaker was recorded reading 5 sentences of about 5 s long each, and a pause of 5 s between two successive sentences. For each speaker and for each recording, different sentences were used. The sampling rate was 48kHz.

In addition, two noise types were recorded separately. The first is an air-conditioner noise, which is the sound produced by the air-conditioner, located at the top of one of the walls in the lab, as shown in Fig. 4 (b). This noise was recorded by all the microphones, while the air-conditioner was on. The second type of noise imitates a diffuse noise field, arriving evenly from all directions. To generate this noise, 8 loudspeakers were placed near the lab's walls (one at every corner of the room and one at the middle of each of the walls, as illustrated in Fig. 4 (a)). The loudspeakers were directed towards the walls



(a) (b) (c)

Fig. 3. Scatter plots corresponding to recordings with $J \in \{1, 2, 3, 4\}$ speakers. Each point represents a single recording, colored according to the corresponding number of speakers. In (a) the axes correspond to λ_1 and λ_2, in (b) the axes correspond to λ_2 and λ_3, and in (c) the axes correspond to λ_3 and λ_4.



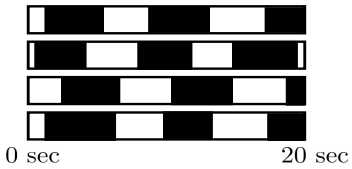
(a)

(b)

Fig. 4. The room layout (a) and a photo of the room setup (b) of the experiments conducted at the BIU acoustic lab.

and were recorded by the microphones while playing babble noises.

For the evaluation in the current experimental study, we use a subset of 8 microphones with indexes $\{25, 26, \ldots, 32\}$. The mixtures are generated by choosing J speakers at J different chairs. To ensure that the signals are not fully aligned by their speech and silence segments, we choose for each signal a random initial starting point, uniformly drawn between $1-3$s, and sum all the signals to form the mixture. At the beginning



(a)

Fig. 5. Representative temporal activities of a mixture of $J = 4$ speakers, with random initial point of each speaker.

of each mixture signal there is a 0.5s long segment without speech, and the signals are truncated to a fixed length of 20s. A typical activity timeline of the speakers for $J = 4$ is depicted in Fig. 5. For each reverberation level, a total number of 50 mixtures is generated with various random combinations of speakers and chairs. The signals were downsampled to 16kHz, and were analysed using the STFT with window size of 2048 samples, and 75% overlap between adjacent frames.

B. Performance Measures & Competing Algorithms

The separation performance is assessed by the signal to interference ratio (SIR) and the signal to distortion ratio (SDR) measures, computed using the BSS-Eval Toolbox [38].

The results of the proposed GLOSS algorithm are compared to the IVA algorithm [8], [39] and to our previously proposed simplex-based algorithm [31] ('Global'), which extracts a single global simplex representation, and identifies frames dominated by each speaker. Based on these frames, the RTFs and the noise covariance matrix are estimated, which, in turn, are utilized for implementing the LCMV beamformer (29). This method consists of only the beamforming (37) without the spectral masking (38), since the method does not include spectral mask estimation. For both the the GLOSS algorithm and [31] we extract the global mapping based on the frequency range $1,000 - 2,000$ kHz. Note that for extracting the global mapping we do not include low frequencies, which mostly contain noise, as well as high frequencies, in which there is typically only low speech energy. The local mapping is

performed for each of the 2048 frequency bins in the range $0-8,000$ kHz.

We further comment on the choice of the frequency range $1-2$kHz for the global mapping. This range was chosen since it provided good performance in all examined scenarios, i.e. different noise types and microphone constellations. Note that in the case of babble noise the bandwidths of the noise and the speech spectra are similar, but there is a difference in their characteristics. Speech signals are sparse in the time-frequency domain, namely, there are high-power components only in several frequencies, while in the rest of the frequencies the power is very low. In contrast, the babble noise consists of several speech signals, hence its power is more spread over the different frequencies. Therefore, in frequencies where speech components are present, their power usually dominates the power of the babble noise. Following this distinction, the chosen frequency range can be used to detect the global activity of the speakers within each frame.

In addition, we compare the results to an ideal separator that implements (37) and (38) based on an ideal binary mask, defined as:

$$M_{\mathrm{i}}(l,f) = \begin{cases} \max\limits_{1\leq j\leq J} |Y_j^1(l,f)| & \text{if } \max\limits_{1\leq j\leq J} |Y_j^1(l,f)| > |N(l,f)| \\ J+1 & \text{Otherwise} \end{cases}. \quad (40)$$

For both the ideal separator and the proposed method we implement the single-channel masking (38) using a fixed attenuation factor of $\beta(f) = 0.1$, $\forall 1 \leq f \leq 2048$.

C. Experimental Results

In this section, we present the separation performance with respect to the noise level and the number of speakers, where in all examinations the correct number of speakers is assumed to be known. We first present the results obtained for mixtures of $J = 4$ speakers and for both reverberation levels with various noise levels. Figure 6 presents the SIR and the SDR scores obtained by all algorithms as a function of the input signal to noise ratio (SNR) for diffuse noise (a)-(b) and for air-conditioner noise (c)-(d). Similar trends are observed for both noise types. It can be seen that all methods obtain lower separation scores as the level of reverberation increases. Regarding the influence of SNR levels, there is a noticeable decrease in the performance only in low SNR of 5dB. As expected, the ideal separator always achieves the best results. The proposed GLOSS algorithm obtains superior results compared to both the global approach [31] and the IVA algorithm for all SNR and reverberation levels. Sound samples can be found on the lab website.[1]

In addition, we examine the performance with respect to the number of speakers in the mixture. Figure 7 presents the SIR and the SDR scores obtained by all algorithms as a function of the number of speakers for both reverberation levels, in the presence of diffuse noise with 20dB SNR. It can be observed that the performance measures of all algorithms decrease as the number of speakers increases. Here as well, the proposed GLOSS algorithm outperforms both competing algorithms.

D. Comparison to DDESS Algorithm

The proposed GLOSS algorithm is also compared to a recently proposed deep direction estimation for speech separation (DDESS) algorithm [37]. The DDESS algorithm is built upon a U-net architecture that receives the phase of the instantaneous RTFs and infers the direction of arrival (DOA) of each TF bin. Separation is obtained by multiplying the reference microphone by the masks associated with the different DOAs. The performance is evaluated on the dataset used in [37], which consists of mixtures of $J = 2$ speakers, formed by clean signals from the Wall Street Journal (WSJ) database [40] that are convolved with room impulse responses (RIRs) recorded at the BIU acoustic lab [41]. The RIRs correspond to a ULA of $M = 8$ microphones with inter-distances of $(8, 8, 8, 3, 8, 8, 8)$ cm. The speakers are positioned at radius of 1 m or 2 m with respect to the array center, at a relative angle from the set $\{0°, 15°, \ldots, 180°\}$. Further details on the examined setup can be found in [37].

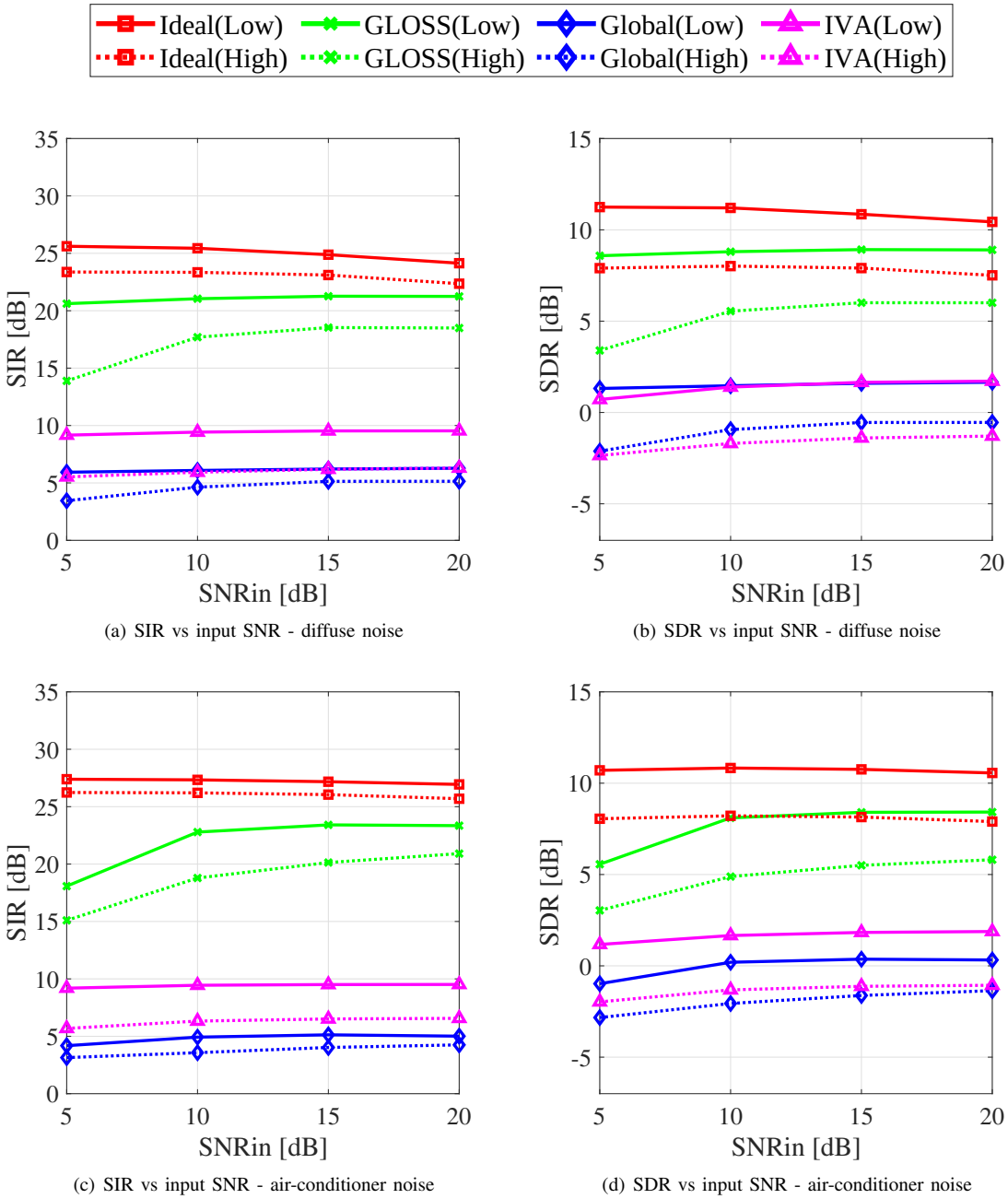
Table I summarizes the results for both radii: 1 m or 2 m, and for two reverberation levels: 160 ms or 360 ms. Note that in the DDESS method only spectral masking is applied (38) without the preceding beamforming step (37). Therefore, here we present the scores of the GLOSS algorithm with spectral masking only ('GLOSS Mask'), with beamforming only ('GLOSS LCMV') and with both beamforming and spectral masking ('GLOSS Full'). We observe a clear advantage of the GLOSS algorithm compared to both the IVA and the DDESS methods, even when using spectral-masking or beamforming only. The SIR scores are further improved using the proposed two-stage separation of multichannel beamforming followed by single-channel spectral postfiltering.

E. Comparison to Multichannel Deep Clustering Algorithm

We also compare the proposed GLOSS algorithm to the multichannel deep clustering algorithm [26], using the multichannel corpus presented in their paper. Since the proposed GLOSS algorithm is unsupervised, only the test set, which consists of 3000 ($\sim$ 5h) utterances, is used. This set is generated using the code provided by the authors,[2] by first creating mixtures of 2 speakers drawn from WSJ0, and then, convolving them with various simulated room impulse responses. To this end, two randomly selected utterances of speakers in the WSJ0 development and evaluation sets are mixed with signal to interference ratio randomly drawn from -5dB and $+5$dB. The chosen re-scaled utterances are convolved with RIRs simulated by the image method [42], using the simulator in [43] with the following random characteristics: the aperture sizes sampled from 15cm to 25cm; the reverberation time drawn from 0.2s to 0.6s; the average distance between speaker and array center is 1.3m with 0.4m standard deviation; and the average direct-to-reverberant energy ratio is 2.5dB with 3.8dB standard deviation. The proposed GLOSS algorithm is

[1] www.eng.biu.ac.il/gannot/speech-enhancement/

[2] www.merl.com/demos/deep-clustering/spatialize_wsj0-mix.zip



(a) SIR vs input SNR - diffuse noise

(b) SDR vs input SNR - diffuse noise

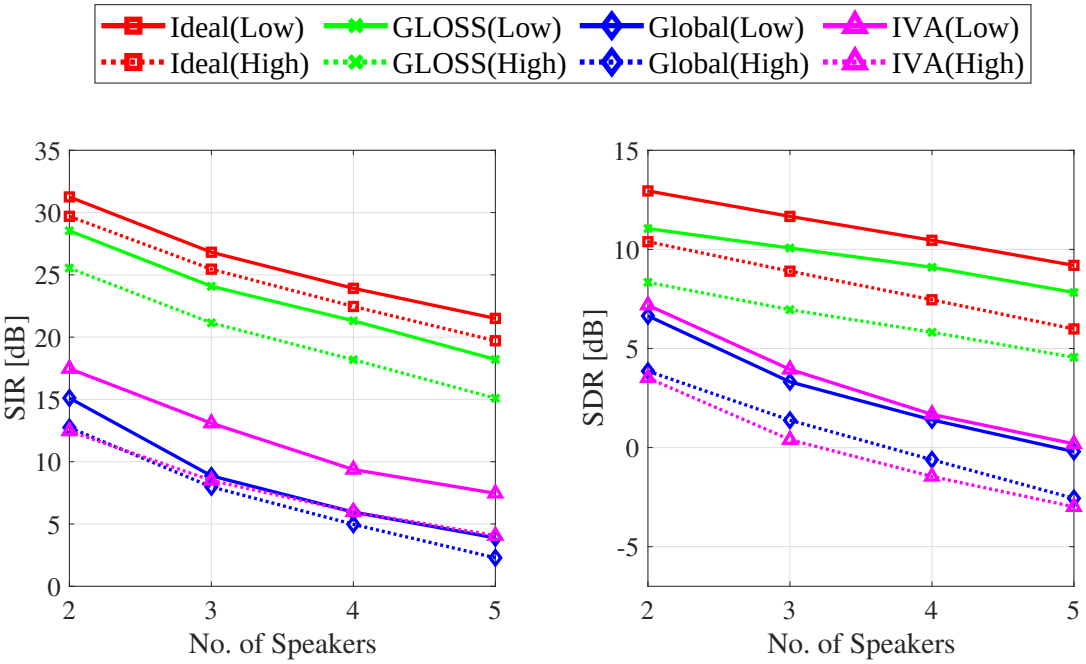
(c) SIR vs input SNR - air-conditioner noise

(d) SDR vs input SNR - air-conditioner noise

Fig. 6. SIR and SDR scores as a function of the input SNR for diffuse (a),(b) and air-conditioner (c),(d) noises. Performance is evaluated for both low and high reverberation conditions, marked by solid and dashed lines, respectively.

TABLE I
SEPARATION PERFORMANCE DEPENDING ON THE DISTANCE FROM THE ARRAY CENTER AND THE REVERBERATION TIME ($J = 2$ SPEAKERS). THE PROPOSED GLOSS ALGORITHM IS COMPARED TO IVA [8] AND DDESS [37] ALGORITHMS.

Method	1m				2m			
	160ms		360ms		160ms		360ms	
	SIR (dB)	SDR (dB)	SIR (dB)	SDR (dB)	SIR (dB)	SDR (dB)	SIR (dB)	SDR (dB)
IVA	13.5	7.3	9.3	4.1	10.4	4.5	7.6	3.0
DDESS	10.4	1.8	9.3	1.6	9.9	2.8	7.5	0.3
GLOSS Mask	14.2	8.9	13.1	7.9	13	8.1	12.9	8.1
GLOSS LCMV	17.5	11.6	14.2	7.8	13.6	8.6	12.0	6.7
GLOSS Full	23.0	10.2	19.6	8.3	20.0	9.0	18.3	7.9



(a) SIR vs No. of speakers - diffuse noise

(b) SDR vs No. of speakers - diffuse noise

Fig. 7. SIR and SDR scores as a function of the number of speakers for diffuse noise with 20dB SNR. Performance is evaluated for both low and high reverberation conditions, marked by solid and dashed lines, respectively.

implemented over the generated test set using a 4-microphone array. The obtained average SIR and SDR scores are 19.1 dB and 9.3 dB, respectively. The obtained SDR is similar to the score reported for the multichannel deep clustering algorithm in [26], which is also 9.3dB. This comparison shows that the performance of the GLOSS algorithm is comparable to that of state-of-the-art deep-learning based separation algorithm, while being completely unsupervised, given that the number of speakers is known.

F. Source Counting Performance

We examine the performance of the proposed source counting scheme derived in Section VI. We focus on estimating the number of speakers in the range $J \in \{1, 2, 3, 4\}$. We used mutually exclusive sets of data for training and testing with different array geometries and source positions. The training data is generated using recorded RIRs of an 8-microphone ULA as described in Section VII-D, corresponding to three reverberation levels 160 ms, 360 ms and 610 ms. For each number of speakers $J \in \{1, 2, 3, 4\}$ and for each reverberation level $T_{60} \in \{160, 360, 610\}$ms, we generated 50 examples with different speakers at different locations, resulting in a total amount of $4 \times 3 \times 50 = 600$ training examples. No noise was added to the training data. For testing the performance we used real-recorded data contaminated by diffuse noise, corresponding to the setup described in Section VII-A. For each number of speakers and reverberation level, we generated 200 examples with different speakers in different locations, resulting in a total amount of $4 \times 2 \times 200 = 1600$ test examples. We trained a multiclass SVM classifier with input vectors corresponding to the first 4 eigenvalues of the correlation matrix computed for each recording.

The results obtained for both reverberation levels and 20dB SNR are summarized in a confusion matrix, depicted in Fig. 8. We observe that the precision values (the bottom summarizing table) are quite similar for the different classes, with average value of 88.75%. With respect to the recall values (the right summarizing table), we have the best scores for mixtures with a single speaker or with four speakers. The average recall among all classes is 88.7%. The classifier largest error is for 3 speakers that are misclassified as 4 speakers. This can be explained using the demonstration in Fig. 3 (c), where we can see that recordings with 3 and 4 speakers are not fully separated. Overall, we conclude that we obtain high scores despite the fact that the training and the testing are performed in completely different conditions.

In addition, we examine the speaker counting performance as a function of the noise level and the reverberation time. Figure 9 illustrates the counting accuracy, defined as the percentage of cases in which the correct number of speakers was recovered by the algorithm, as a function of the input SNR. Each point in the graph represents the average accuracy obtained for 800 examples, which consists of 200 examples for each number of speakers in $J \in \{1, 2, 3, 4\}$. We observe that in low reverberation conditions the counting accuracy is high for all SNR levels. The performance of the proposed counting method decreases in high noise and reverberation conditions.

VIII. Conclusions

A novel separation algorithm is presented based on simplex representations. First, a global simplex representation is constructed based on a broad frequency range, and is used

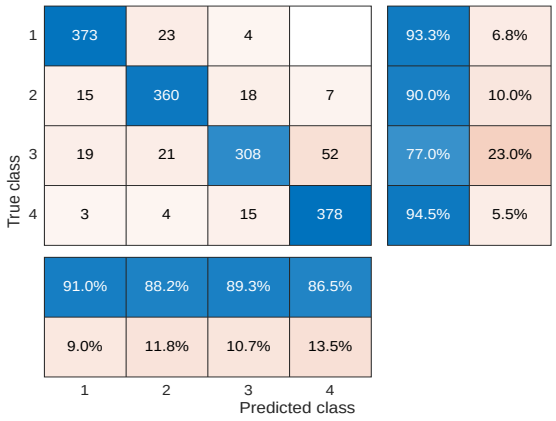


Fig. 8. Confusion matrix for detecting the number of speakers in the range $J \in \{1, 2, 3, 4\}$.

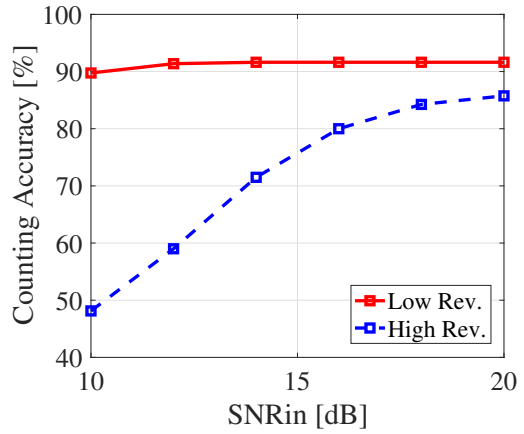


Fig. 9. Counting Accuracy versus input SNR for low and high reverberation levels.

for estimating the global probability of activity of each of the speakers in each time frame. Second, we derive a local simplex representation of each frequency individually. Combining the global probabilities with the local relations in each frequency, a spectral mask is estimated. The separation is performed using the estimated spectral mask, using a two-stage approach of multichannel beamforming followed by single-channel post-filtering. The proposed GLOSS algorithm is shown to achieve high separation scores in various conditions, outperforming state-of the art methods.

APPENDIX A

Based on the statistical assumptions presented in Section III-A, the expected bin-wise correlation of two different frames $l \neq n, 1 \leq l, n \leq L$ at the same frequency and microphone (same entry k), given the identity of the dominating component, is:

$$E\left\{r(l,k)r(n,k)|M(l,k),M(n,k)\right\} = \begin{cases} 1 & M(l,k) = M(n,k) \neq J+1 \\ 0 & \text{otherwise} \end{cases} \quad (41)$$

Namely, the correlation is one if the same speaker is active in both TF bins, and zero if there are different dominating speakers or that one of the TF bins is dominated by noise. Thus, following the law of total expectation, we have that the expected value of the bin-wise correlation is given by:

$$\begin{aligned} &E\left\{r(l,k)r(n,k)\right\} \\ &= E_{M(l,k),M(n,k)}\left\{E\left\{r(l,k)r(n,k)|M(l,k),M(n,k)\right\}\right\} \quad (42) \\ &= \sum_{j=1}^{J+1}\sum_{i=1}^{J+1}\Big(E\left\{r(l,k)r(n,k)|M(l,k)=j,M(n,k)=i\right\} \\ &\cdot \Pr\big(M(l,k)=j\big)\cdot\Pr\big(M(n,k)=i\big)\Big). \end{aligned}$$

Since $E\left\{r(l,k)r(n,k)|M(l,k)=j,M(n,k)=i\right\} = 0$ for $i = J+1$ or $j = J+1$, we have:

$$\begin{aligned} &E\left\{r(l,k)r(n,k)\right\} \\ &= \sum_{j=1}^{J}\sum_{i=1}^{J}\Big(E\left\{r(l,k)r(n,k)|M(l,k)=j,M(n,k)=i\right\} \quad (43) \\ &\cdot \Pr\big(M(l,k)=j\big)\cdot\Pr\big(M(n,k)=i\big)\Big) \\ &= \sum_{j=1}^{J}\sum_{i=1}^{J}\delta_{ij}p_j(l)p_i(n) \\ &= \sum_{j=1}^{J}p_j(l)p_j(n) \end{aligned}$$

where $\delta_{ij} = 1$ for $i = j$ and 0 otherwise. Based on the strong law of the large numbers, we have:

$$\begin{aligned} \frac{1}{D}\mathbf{r}^T(l)\mathbf{r}(n) &= \frac{1}{D}\sum_{k=1}^{D}r(l,k)r(n,k) \\ &\overset{a.s.}{\rightarrow} E\left\{r(l,k)r(n,k)\right\} = \sum_{j=1}^{J}p_j(l)p_j(n). \quad (44) \end{aligned}$$

Similarly, for the same frame $l = n$:

$$E\left\{r(l,k)r(l,k)|M(l,k)\right\} = \begin{cases} 1 & M(l,k) \neq J+1 \\ 0 & \text{otherwise} \end{cases} \quad (45)$$

Therefore, in this case we have

$$\begin{aligned} E\left\{r(l,k)\right\} &= E_{M(l,k)}\left\{E\left\{r(l,k)r(l,k)|M(l,k)\right\}\right\} \\ &= \sum_{j=1}^{J+1}E\left\{r^2(l,k)|M(l,k)=j\right\}\Pr\big(M(l,k)=j\big) \quad (46) \\ &= \sum_{j=1}^{J}E\left\{r^2(l,k)|M(l,k)=j\right\}\Pr\big(M(l,k)=j\big) \\ &= \sum_{j=1}^{J}p_j(l) \end{aligned}$$

implying that

$$\frac{1}{D}\mathbf{r}^T(l)\mathbf{r}(l) \overset{a.s.}{\to} \sum_{j=1}^{J} p_j(l). \tag{47}$$

REFERENCES

[1] S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[2] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*. Springer, 2007, vol. 615.

[3] S. Makino, *Audio source separation*. Springer, 2018.

[4] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[5] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[6] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE transactions on speech and audio processing*, vol. 13, no. 1, pp. 120–134, 2005.

[7] S.-Y. Lee, "Blind source separation and independent component analysis: A review," *Neural Information Processing-Letters and Reviews*, vol. 6, no. 1, pp. 1–57, 2005.

[8] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2006, pp. 165–172.

[9] Z. Koldovsky and P. Tichavsky, "Time-domain blind separation of audio sources on the basis of a complete ICA decomposition of an observation space," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 406–416, 2011.

[10] 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.

[11] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 3437–3440.

[12] 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.

[13] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, 2014.

[14] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1622–1637, 2016.

[15] 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.

[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] N. Madhu and R. Martin, "A versatile framework for speaker separation using a model-based speaker localization approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1900–1912, 2011.

[18] 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.

[19] 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.

[20] B. Schwartz, S. Gannot, and E. A. Habets, "Two model-based EM algorithms for blind source separation in noisy environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2209–2222, 2017.

[21] 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.

[22] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[23] P. Pertilä and J. Nikunen, "Distant speech separation using predicted time–frequency masks from spatial features," *Speech communication*, vol. 68, pp. 97–106, 2015.

[24] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[25] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

[26] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1–5.

[27] Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 457–468, 2019.

[28] L. Drude and R. Haeb-Umbach, "Tight integration of spatial and spectral features for BSS with deep clustering embeddings." in *in Proc. of The Annual Conference of the International Speech Communication Association (Interspeech)*, 2017, pp. 2650–2654.

[29] S. E. Chazan, J. Goldberger, and S. Gannot, "DNN-based concurrent speakers detector and its application to speaker extraction with LCMV beamforming," in *in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6712–6716.

[30] S. E. Chazan, S. Gannot, and J. Goldberger, "Attention-based neural network for joint diarization and speaker extraction," in *in Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 301–305.

[31] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Source counting and separation based on simplex analysis," *IEEE Transactions on Signal Processing*, vol. 66, no. 24, pp. 6458–6473, 2018.

[32] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[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] 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.

[36] 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.

[37] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, and S. Gannot, "Multi-microphone speaker separation based on deep DOA estimation," in *European Signal Processing Conference (EUSIPCO)*, 2019.

[38] 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.

[39] T. Kim, https://github.com/teradepth/iva, Feb. 2018.

[40] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[41] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *in Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 313–317.

[42] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[43] E. A. Habets, "Room impulse response (RIR) generator," https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator, May. 2008.