Laufer-Goldshtein *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:5
https://doi.org/10.1186/s13636-021-00195-7

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Audio source separation by activity probability detection with maximum correlation and simplex geometry

Bracha Laufer-Goldshtein[1], Ronen Talmon[2] and Sharon Gannot[1*]

Abstract

Two novel methods for speaker separation of multi-microphone recordings that can also detect speakers with infrequent activity are presented. The proposed methods are based on a statistical model of the probability of activity of the speakers across time. Each method takes a different approach for estimating the activity probabilities. The first method is derived using a linear programming (LP) problem for maximizing the correlation function between different time frames. It is shown that the obtained maxima correspond to frames which contain a single active speaker. Accordingly, we propose an algorithm for successive identification of frames dominated by each speaker. The second method aggregates the correlation values associated with each frame in a correlation vector. We show that these correlation vectors lie in a simplex with vertices that correspond to frames dominated by one of the speakers. In this method, we utilize convex geometry tools to sequentially detect the simplex vertices. The correlation functions associated with single-speaker frames, which are detected by either of the two proposed methods, are used for recovering the activity probabilities. A spatial mask is estimated based on the recovered probabilities and is utilized for separation and enhancement by means of both spatial and spectral processing. Experimental results demonstrate the performance of the proposed methods in various conditions on real-life recordings with different reverberation and noise levels, outperforming a state-of-the-art separation method.

Keywords: Blind audio source separation (BASS), Relative transfer function (RTF), Correlation analysis, Linear programming (LP), Simplex, Convex geometry, Beamforming

1 Introduction

Blind audio source separation (BASS) is a prominent task in the field of audio processing, dealing with the analysis of audio streams comprising several speakers. BASS aims at extracting the individual speech signals of each of the sources present in an audio mixture [1]. Most methods for BASS usually assume that the speakers are concurrently active for most of the time, and a little attention was paid to the case of infrequent speakers.

BASS has been a topic of extensive research for the last decades, leading to a large variety of separation algorithms. The measured signals in an array of microphones represent convolutive mixtures of the clean source signals with the corresponding acoustic channels [2–5]. Commonly, the signals are analyzed in the short time Fourier transform (STFT) domain, in which the convolutive mixtures are approximated by multiplicative mixtures. Various approaches for BASS exist, such as the independent component analysis (ICA) and independent vector analysis (IVA) separation methods [6–10], non-negative matrix factorization (NMF) [11–15], and, more recently, deep neural network (DNN)-based separation methods [16–23]. A related problem to acoustic source separation was recently investigated in the field of structural health monitoring based on acoustic emission, dealing with onset detection of overlapped acoustic emission waves [24] for accurate time of arrival estimation [25].

*Correspondence: sharon.gannot@biu.ac.il
[1]Faculty of Engineering, Bar-Ilan University, 5290002, Ramat-Gan, Israel
Full list of author information is available at the end of the article

A vast number of BASS algorithms rely on the sparsity of speech signals in the STFT domain, assuming that speech components of simultaneously active speakers are non-overlapping [26]. One approach is to compute the time or the phase differences at each time-frequency (TF) bin, and then to jointly cluster all TF bins [27–29]. Alternatively, dual-stage methods perform a frequency-wise clustering, followed by a permutation alignment of the identities of the speakers across all frequency bands [30–32]. The resulting algorithms consist of iterative methods, such as the well-known expectation maximization (EM) algorithm, which require a careful initialization, are susceptible to converge to local maxima, and commonly impose high computational load.

Source separation can also be achieved by applying beamformers [1], which are multichannel spatial filters designed by certain criteria, such as the linearly constrained minimum variance (LCMV) beamformer [33]. These algorithms are not fully blind, as their design requires some knowledge on the signal statistics or the acoustic systems. In [33, 34], the beamformer parameters were estimated assuming some prior knowledge on the activity of the speakers, such as assuming the existence of known time intervals in which each of the desired speakers is separately active [33], or, alternatively, assuming a scenario in which speakers become successively active [34]. In [35], a blind approach for learning the beamformer parameters was proposed using variational inference framework, which is initialized based on the speakers' time difference of arrivals (TDOAs).

In many conversations, an unbalanced activity of the different speakers is common, when several speakers may participate frequently, while others only seldom speak. This is often the case in interviews, police interrogations, and counseling, just to name a few. In this type of scenarios, one speaker presents short questions or comments, while the other speaker provides long answers or descriptions, which may be then followed by short expressions of agreement or disagreement from the first speaker. Identifying a speaker with low activity is extremely challenging, since it is based on a very limited amount of data. To the best of our knowledge, this scenario has not been considered in the literature, although it is very common and of great importance in many applications.

In this paper, we present a source separation method based on an LCMV beamformer followed by a postfilter with parameters that are learned in a completely blind manner by recovering the probability of activity of the speakers across time. This method is flexible and can be applied to a wide range of scenarios from conversational speech with only a limited amount of overlapping speech to audio mixtures of simultaneously active speakers with possibly large overlap between them. Furthermore, it can also detect speakers with low activity. The proposed method relies on a probabilistic model that is built upon the speech sparsity in the STFT domain and describes the probability of activity of the speakers across time. Based on the assumed statistical model, it is shown that the correlation between each two time frames along the measured signal equals the multiplication of the associated speaker's probabilities. A correlation function is defined for each frame, consisting of its correlation with all other frames. For frames dominated by one of the speakers, the value of the correlation function equals the probability of this speaker in each frame, and can be used as an estimator for the activity probability. We present two methods for detecting the set of frames dominated by a single speaker. The first method relies on the theory of linear programming (LP) that states that single-speaker frames are the maximum points of the correlation function associated with any other frame. Based on this observation, we present a sequential algorithm to detect frames dominated by each speaker. In the second method, we aggregate the correlations in a correlation matrix. We show that the columns of this matrix lie in a simplex. The vertices of the simplex correspond to frames dominated by a single speaker, and can be recovered by convex geometry tools. Based on the activity probabilities estimated by either method, we recover the spectral mask that assigns each TF bin with the dominant speaker, and exploit it for performing spatial multichannel separation followed by spectral single-channel post-processing.

The contribution of the current work is twofold, both in terms of presenting novel methodologies for source separation which have several advantages over existing separation methods, as well as in terms of achieving high-quality performance in various scenarios as shown in the experimental part. Compared to existing separation algorithms, the proposed methods do not require any initialization processes, do not include iterative search mechanisms, circumvent frequency-permutation problems, and do not require training data, resulting in efficient high-performance methods. Compared to our previously proposed simplex-based method [36, 37], which relies on a similar probabilistic model, here we present new methodologies for solving the problem, yielding simplified processing methods and better supporting infrequent speakers. In addition, one of the presented methods has lower computational complexity with respect to [36]. It is worthwhile noting that the proposed maximum correlation method was utilized for a psychological research on vocal emotional dynamics during psychotherapy sessions [38]. The method was used for detecting time intervals in which the therapist and the patient are active, from which vocal features were then extracted and analyzed. An example demonstrating the capabilities of the proposed method for this task is given in the experimental part. This utilization of the proposed method

demonstrates its applicability to the extraction of speakers with unbalanced activity patterns as in psychotherapy sessions, where patients often speak for longer periods, while therapists frequently respond with shorter utterances.

2 Methods

We start by presenting in Section 2.1 the problem formulation and preliminary concepts underlying the signal model, the statistical model of the activity probabilities, the feature extraction process, and the analysis of the correlation between the defined feature vectors. This study serves as the basis for the derivation of two methods for activity probability estimation, presented in Section 2.2. These methods are then incorporated in the actual separation scheme, presented in Section 2.3.

2.1 Problem formulation and preliminary concepts

2.1.1 Signal model

Consider J static speakers in a reverberant enclosure. The signals emitted by the speakers are measured by an array of M microphones and are analyzed in the short time Fourier transform (STFT) domain. Here, $f \in \{1,\dots,K\}$ is the frequency bin and $l \in \{1,\dots,L\}$ is the frame index. 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 $Y_j^m(l,f) = A_j^m(f) S_j(l,f)$ is the signal of the jth speaker measured by the mth microphone, 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, and $N^m(l,f)$ is the non-directional noise signal measured by the mth microphone. Directional noises can be treated as additional sources, increasing J accordingly.

Our goal is to apply separation, namely to extract the individual source signals $\{Y_j^1(l,f)\}_{l,j}$ from the mixture while reducing the noise. Note that instead of estimating the original source signals, we provide an estimate of the source signals as they are measured by the first microphone that serves as a reference microphone.

2.1.2 Sparsity-based statistical model

Relying on the assumption of the speech sparsity in the STFT domain (a.k.a. W-disjoint orthogonality) [26], each TF bin is dominated by either one of the speakers or consists of noise. We define the *categorical* spectral mask $\{M(l,f)\}_{l,f}$ that assigns each TF bin with its dominating *component*, either one of the J speakers, $1 \le M(l,f) \le J$ or the noise $M(l,f) = J+1$. When the power of the jth speaker is considerably larger compared to the power of the other speakers and the noise, in the (l,f)th TF bin, we have $M(l,f) = j$, $1 \le j \le J$. TF bins that are not dominated by either of the speakers are considered noise, i.e., $M(l,f) = J+1$. Accordingly, we can restate (1):

$$Y^m(l,f) = \begin{cases} A_j^m(f) S_j(l,f) & \text{if } M(l,f) = j, \\ N^m(l,f) & \text{if } M(l,f) = J+1. \end{cases} \tag{2}$$

We assume that the index of the dominating component $M(l,f)$ has a categorical distribution with

$$\Pr\big(M(l,f) = j\big) = p_j(l) \tag{3}$$

and $\sum_{j=1}^{J} p_j(l) \le 1$ for each frame l. The *activity probabilities* $\{p_j(l)\}_{j,l}$ are independent of the frequency bin, and represent the activity patterns of the speakers across time, namely with respect to the frame index l.

We compute the following bin-wise ratio between the measurement at the mth microphone and the measurement at the reference microphone:

$$R^m(l,f) \equiv \frac{Y^m(l,f)}{Y^1(l,f)}, \quad m = 2,\dots,M. \tag{4}$$

According to the sparsity assumption (2), we have:

$$R^m(l,f) = \begin{cases} H_j^m(f) & \text{if } M(l,f) = j,\ 1 \le j \le J \\ \eta(l,f) & \text{if } M(l,f) = J+1 \end{cases} \tag{5}$$

where

$$H_j^m(f) = \frac{A_j^m(f)}{A_j^1(f)}. \tag{6}$$

is the relative transfer function (RTF) [39, 40] 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. We obtain that the ratio in (5) equals the RTF of one of the speakers or a noise term. 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 [36]. For the sake of simplicity, we assume a unit variance for the real and the imaginary parts of the RTFs and the noise terms in each TF bin. Note that the following derivation also holds for non-unit and non-constant variance by applying a proper normalization.

2.1.3 Feature extraction and correlation analysis

Based on the computed ratios, a feature vector $\mathbf{r}(l)$ of length $D = 2 \cdot F \cdot (M-1)$ is defined for each frame as a concatenation of the real and the imaginary parts of the

ratios (4), in F frequency bins and in $M-1$ microphones:

$$\begin{aligned}\mathbf{r}(l,f) &= \left[R^2(l,f), R^3(l,f), \ldots, R^M(l,f)\right]^T \\ \mathbf{r}^{\mathrm{c}}(l) &= \left[\mathbf{r}^T(l,f_1), \mathbf{r}^T(l,f_2), \ldots, \mathbf{r}^T(l,f_F)\right]^T \\ \mathbf{r}(l) &= \left[\mathrm{real}\left\{\mathbf{r}^{\mathrm{c}}(l)\right\}^T, \mathrm{imag}\left\{\mathbf{r}^{\mathrm{c}}(l)\right\}^T\right]^T. \end{aligned} \tag{7}$$

where $f_1, \ldots, f_F \in \{1, ..., K\}$.

Based on the presented statistical assumptions, consider the expected bin-wise correlation between two different frames $l \neq n, 1 \leq l, n \leq L$, given the identity of the dominating components:

$$\begin{aligned} &E\left\{[\,\mathbf{r}(l)]_k \cdot [\,\mathbf{r}(n)]_k \,|\tilde{M}(l,k), \tilde{M}(n,k)\right\} \\ &\quad = \begin{cases} 1 & \tilde{M}(l,k) = \tilde{M}(n,k) \neq J+1 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{8}$$

where $[\,\mathbf{r}(l)]_k$, $k \in \{1, \ldots, D\}$ denotes the kth entry of $\mathbf{r}(l)$ and $\tilde{M}(l,k) = M(l, \lfloor (k-1)/(M-1) \rfloor \bmod (M-1)+1)$. Equation (8) states that the conditional correlation equals "1" if the same speaker is active in both TF bins, and "0" if there are different dominating speakers or that one of the TF bins is dominated by noise. Thus, according to the law of total expectation, we have:

$$E\left\{[\,\mathbf{r}(l)]_k \cdot [\,\mathbf{r}(n)]_k\right\} = \sum_{j=1}^{J} p_j(l) p_j(n), \tag{9}$$

implying that the bin-wise correlation between the features equals the multiplication of the corresponding activity probabilities. In practice, we can obtain an approximation to this expected correlation by averaging the product of the features over a large amount of frequency bins. To this end, the bin-wise correlations can be treated as a sequence of uncorrelated random variables for different values of k (different frequencies or different microphones). Therefore, according to the strong law of the large numbers, their sample mean converges almost surely to their mean value given in (9):

$$\begin{aligned} &\frac{1}{D}\mathbf{r}^T(l)\mathbf{r}(n) = \frac{1}{D}\sum_{k=1}^{D}[\,\mathbf{r}(l)]_k \cdot [\,\mathbf{r}(n)]_k \\ &\overset{a.s.}{\rightarrow} E\left\{[\,\mathbf{r}(l)]_k \cdot [\,\mathbf{r}(n)]_k\right\} = \sum_{j=1}^{J} p_j(l) p_j(n), \end{aligned} \tag{10}$$

for $D \rightarrow \infty$. Note that for the same frame $l = n$, the expected value does not obey (9) and (10), but instead $E\{[\,\mathbf{r}(l)]_k^2\} = 1$, and therefore $\frac{1}{D}\mathbf{r}^T(l)\mathbf{r}(l) \overset{a.s.}{\rightarrow} 1$. In the following, we show how we can exploit the relation in (10) between the feature-wise correlations and the probability products to estimate the activity probabilities, by exploiting either linear programming theory or convex geometry tools.

2.2 Proposed methods

We present two methods for estimating the activity probabilities $\{p_j(l)\}_{j,l}$ of the different speakers at each frame. The estimation is based on the statistical model presented in Section 2.1, and specifically on the correlation between frames (10). The two methods first identify frames that are dominated by one of the speakers, and then infer the probabilities associated with each speaker using the correlations with respect to the identified frames. The first method is described in Section 2.2.1 and is based on the detection of the maximum points of the correlation functions defined for each frame. The second method is described in Section 2.2.2 and is based on the detection of the vertices of the simplex that consists of the correlation vectors defined for each frame.

2.2.1 Maximum correlation method

A function of J variables $\mathbf{q} = [\,q_1, q_2, \ldots, q_J]^T$ is defined for each frame l:

$$t_l(q_1, q_2, \ldots, q_J) = \sum_{j=1}^{J} p_j(l) q_j, \tag{11}$$

where the probabilities $\{p_j(l)\}_{j=1}^{J}$ associated with the lth frame are the parameters defining the function. Consider the following optimization problem:

$$\begin{aligned} \underset{q_1, q_2, \ldots, q_J}{\text{maximize}} \quad & t_l(q_1, q_2, \ldots, q_J) = \sum_{j=1}^{J} p_j(l) q_j \\ \text{subject to} \quad & q_1 + q_2 + \ldots + q_J \leq 1. \\ & q_j \geq 0, \ \forall\, 1 \leq j \leq J \end{aligned} \tag{12}$$

This is a linear programming (LP) problem, where the constraints, defining the feasible region, specify the J-D probability simplex. The vertices of the simplex are the standard unit vectors $\{\mathbf{e}_j\}_{j=1}^{J}$, where $\mathbf{e}_j = [\,0, \ldots, 1, \ldots, 0]$, with one in the jth entry and zeros elsewhere, and there is an additional vertex at the origin since the sum $\sum_{j=1}^{J} q_j$ can be lower than "1" but is bound to be positive. Based on the theory of LP [41], every local maximum is a global maximum, and the maximum is attained at either of the simplex vertices. Note that the function value at the origin is "0"; therefore, the maximum must be attained at one of the other vertices of the simplex $\{\mathbf{e}_j\}_{j=1}^{J}$.

According to (10), a set of possible values of the function t_l is given by the correlations between the lth frame and all other frames, i.e.:

$$\begin{aligned} t_l(n) &\equiv t_l(q_1(n), q_2(n), \ldots, q_J(n)) \\ &= \sum_{j=1}^{J} p_j(l) q_j(n) \Big|_{q_j(n) = p_j(n)} = \frac{1}{D}\mathbf{r}^T(l)\mathbf{r}(n) \end{aligned} \tag{13}$$

representing the function value at the point $\mathbf{q}(n) = [\,q_1(n), q_2(n), \ldots, q_J(n)]^T$, with $q_j(n) \equiv p_j(n)$. Note

that for a specific frame l, the probabilities $\mathbf{p}(l) = [p_1(l), p_2(l), \ldots, p_J(l)]^T$ are treated as fixed parameters determining the structure of the function t_l, while the probabilities of the other frames $\mathbf{q}(n)$, $1 \leq n \leq L, n \neq l$ are treated as points at which the function is evaluated. Here, we utilize the equivalence between the feature-wise correlations and the probability products implied by (10) to obtain samples from the function t_l, where the sample $t_l(n)$ is given by the correlation between the features associated with the lth and the nth frames. Note that we only have the function value $t_l(n)$, but not the function parameters $\mathbf{p}(l)$ and the point $\mathbf{q}(l)$, in which the function was evaluated.

The formulation of the LP problem (12) can be used to detect frames dominated by a single speaker. It is important to clarify that we do not solve the optimization in (12), since the parameters defining the function t_l are unknown. Instead, we utilize the fact that the maximum is attained at a point corresponding to a frame with a single speaker, i.e., with probability 1 for one of the speakers, and with probability 0 for the other speakers. According to (13), the correlation between the lth frame and all other frames provides $L-1$ values of the function t_l. Therefore, we search for the maximum among the given values of t_l. The maximum value is attributed to a frame that is exclusively dominated by a single speaker. In practice, we define the set of L correlation functions $\{t_l\}_{l=1}^{L}$, by appending to each frame its correlation to all other frames. Next, we count the number of times that each frame is detected as a maximum point of the correlation functions defined by the other frames. This way, we obtain a score function conveying how often each frame serves as maximum point. Frames that achieve a high score are assumed to be dominated by one of the speakers. To obtain only a single representative frame for each speaker, we select these frames sequentially and eliminate in each step the frames that correspond to speakers that have already been identified.

Examples of the function t_l are given in Fig. 1, for two mixtures of $J = \{2, 3\}$ speakers. Further details on the generation of the mixtures are given in Section 3. Note that for $J = 2$, the constraint $q_1 + q_2 \leq 1$ specifies a triangle, and for $J = 3$, the constraint $q_1 + q_2 + q_3 \leq 1$ specifies a corner of a cube. The coloring of the points is according to the function value at each point. It can be observed that the maximum is attained at a vertex, which corresponds to the speaker with maximum probability. Note that the maximum is not necessarily unique, for example, a flat function is obtained for a frame with equal speakers' probabilities. We avoid these cases by considering only correlation functions that their maximum value is above a certain threshold.

Based on these observations, we propose an algorithm for sequential recovery of J frames, each of them dominated by one of the J speakers. We assume that for each speaker, there is at least one frame, with index l_j, which is entirely dominated by this speaker, i.e., $\mathbf{p}(l_j) = \mathbf{e}_j$. For the simplicity of the notation, we ignore possible permutation in the order of the identified speakers.

We define a function g which assigns to any frame index l a frame index $g(l)$ with maximum correlation to frame l, i.e.:

$$g(l) = \underset{\eta \in \mathcal{S}_1 \setminus l}{\operatorname{argmax}}\, t_l(\eta) \tag{14}$$

where $\mathcal{S}_1 = \{1, \ldots, L\}$. We define by $c(l')$ the number of frames with maximal correlation attained at the l'th frame, i.e.:

$$c(l') = \left| g^{-1}(l') \right| \tag{15}$$

where $g^{-1}(l') = \left\{ l \in \mathcal{S}_1 | g(l) = l' \right\}$ is the inverse image of g, and $|\cdot|$ denotes the set cardinality. The frame associated with the first speaker is the one most frequently detected as a maximum point:

$$l_1 = \max_{\eta \in \mathcal{S}_1} c(\eta). \tag{16}$$

The probabilities associated with frame l_1 satisfy $\mathbf{q}(l_1) = \mathbf{p}(l_1) = \mathbf{e}_1$; hence, in (11), we have:

$$t_{l_1}(n) = \mathbf{p}^T(l_1)\mathbf{q}(n) = \mathbf{e}_1^T \mathbf{q}(n) = q_1(n) = p_1(n). \tag{17}$$

Next, we define a smaller subset of frames with low probability of activity of the first speaker:

$$\mathcal{S}_2 = \left\{ l \in \mathcal{S}_1, t_{l_1}(l) < \varepsilon \right\} \tag{18}$$

with ε a threshold parameter. A second frame, dominated exclusively by the second speaker, is chosen using the same criterion as in (16), $l_2 = \max_{\eta \in \mathcal{S}_2} c(\eta)$, where the search runs now over $\mathcal{S}_2$. Limiting our search to frames in $\mathcal{S}_2$ prevents choosing a frame dominated by the first speaker.

Assuming that $r-1$ speakers have already been detected, a frame dominated by the rth speaker is identified by:

$$l_r = \max_{\eta \in \mathcal{S}_r} c(\eta) \tag{19}$$

where

$$\mathcal{S}_r = \left\{ l \in \mathcal{S}_{r-1}, t_{l_{r-1}}(l) < \varepsilon \right\} \tag{20}$$

The process is stopped when $r = J$. The probabilities of the speakers in each frame are estimated by their correlation to the identified frames $\{l_j\}_{j=1}^{J}$:

$$p_j(n) = t_{l_j}(n),\ \forall n \in \{1, \ldots, L\}, j \in \{1, \ldots, J\}. \tag{21}$$

In the case the set $\mathcal{S}_r$ is empty and $r < J$, we replace the search rule of (19) by:

$$l_r = \min_{\eta \in \mathcal{S}_1} \sum_{j=1}^{r-1} t_{l_j}(\eta). \tag{22}$$

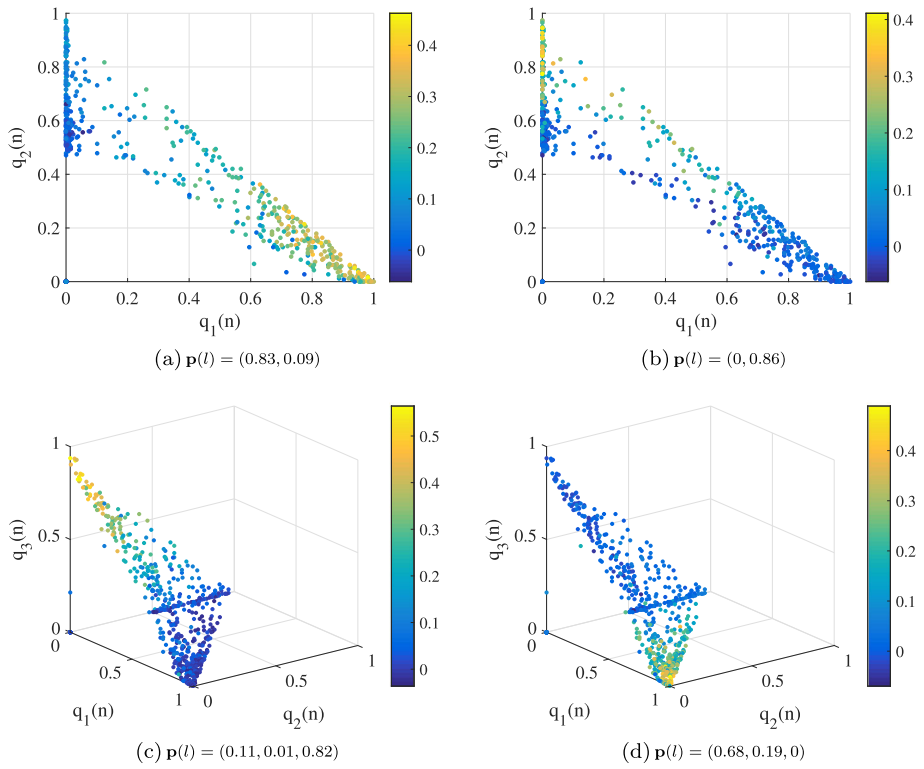


Fig. 1 Scatter plot of **a**, **b** $t_l(q_1, q_2)$ and **c**, **d** $t_l(q_1, q_2, q_3)$. The position of each point is according to **a**, **b** $[q_1, q_2]$ or **c**, **d** $[q_1, q_2, q_3]$, and the color coding is according to the function value at each point

Note that according to (21), $\sum_{j=1}^{r-1} t_{l_j}(\eta) = \sum_{j=1}^{r-1} p_{l_j}(\eta)$; hence, the criterion in (22) amounts to selecting the frame with lowest activity probability of the speakers that were already detected. The proposed maximum correlation method is summarized in Algorithm 1.

2.2.2 *Correlation simplex method*

For the second method, we aggregate the values of the correlation function defined for each frame in an $L \times 1$ correlation vector $\mathbf{t}_l$ defined as:

Algorithm 1: Maximum Correlation Method

- Define $g(l)$ by detecting the maximum point for each correlation function (14).
- Define $c(l)$ by counting the number of times each frame serves as a maximum point (15).
- Detect frames dominated by each speaker:
  - $\mathcal{S}_1 = \{1, \ldots, L\}$
  - **for** $r = 1 : J$ **do**
    - $l_r = \max_{\eta \in \mathcal{S}_r} c(\eta)$
    - $\mathcal{S}_{r+1} = \left\{ l \in \mathcal{S}_r, t_{l_r}(l) < \varepsilon \right\}$
- Obtain the activity probabilities $p_j(n) = t_{l_j}(n)$ (21).

$$\mathbf{t}_l = [\, t_l(1), t_l(2), \ldots, t_l(L)]^T . \tag{23}$$

Based on (13), we have:

$$\mathbf{t}_l = \sum_{j=1}^{J} p_j(l)\mathbf{p}_j + \Delta\mathbf{t}_l \approx \sum_{j=1}^{J} p_j(l)\mathbf{p}_j \tag{24}$$

where $\mathbf{p}_j = [\, p_j(1), p_j(2), \ldots, p_j(L)]^T$ is the probability vector of size $L \times 1$, which consists of the probabilities of the jth speaker in each frame. Here, $\Delta\mathbf{t}_l = (1 - \sum_{j=1}^{J} p_j^2(l))\mathbf{e}_j$ represents a small difference vector that stands for the deviation in the jth entry due to the fact that the self-correlation of the frame with itself equals "1." This deviation has a negligible effect and is therefore ignored.

According to (24), the correlation vectors $\{\mathbf{t}_l\}_{l=1}^{L}$ are obtained as convex combinations of $\{\mathbf{p}_j\}_{j=1}^{J}$; hence, they lie in the following simplex in $\mathbb{R}^L$:

$$\Theta = \left\{ \theta_1\mathbf{p}_1 + \ldots + \theta_J\mathbf{p}_J \,\middle|\, \sum_{j=1}^{J} \theta_j \le 1, \theta_j \ge 0 \right\}, \tag{25}$$

where the simplex vertices are $\{\mathbf{p}_j\}_{j=1}^{J}$, and there is an additional vertex at the origin, since the sum of the weights can also be lower than one. Therefore, recovering

Algorithm 2: Correlation Simplex Method

- Define the simplex obtained by the correlation vectors $\{\mathbf{t}_j\}_{l=1}^{L}$ of (23).
- Detect frames dominated by each speaker by recovering the simplex vertices using SPA [43]:
  - $\mathbf{P}^{\perp} = \mathbf{I}$
  - **for** $j = 1 : J$ **do**
    - $l_j = \underset{l \in \{1,\ldots,L\}}{\operatorname{argmax}} \|\mathbf{P}^{\perp}\mathbf{t}_l\|_2^2$
    - $\mathbf{d}_j = \mathbf{P}^{\perp}\mathbf{t}_{l_j}$
    - $\mathbf{P}^{\perp} = \left(\mathbf{I} - \mathbf{d}_j\mathbf{d}_j^T/\|\mathbf{d}_j\|_2^2\right)\mathbf{P}^{\perp}$
  - **end**
- Obtain the activity probabilities $\mathbf{p}_j = \mathbf{t}_{l_j}$ (26).

the simplex vertices with indexes $\{l_j\}$ provides an estimate of the columns of the probability matrix $\mathbf{P}$, i.e.:

$$\mathbf{p}_j = \mathbf{t}_{l_j}. \tag{26}$$

The simplex vertices are detected by means of convex geometry tools using the successive projection algorithm (SPA) [42, 43]. In this algorithm, the vertices are sequentially detected by maximum norm criterion, when each vector is first projected to the orthogonal complement of the subspace spanned by the already identified vertices. The proposed correlation simplex method is summarized in Algorithm 2.

Note that the two proposed methods are based on iterative procedures for detecting frames dominated by a single speaker. In the maximum correlation method, the detection criterion is based on how frequently the frame serves as a maximum point, while the simplex correlation method is based on maximum norm criterion. These two criteria are related to each other since when a frame is frequently detected as a maximum, it implies that it has high correlation with other frames, indicating that its correlation vector has high norm. The main difference is that for the maximum correlation method, the detection criterion is computed only once, and the set of possible frames is reduced in each step by eliminating frames with non-negligible correlation to previously identified frames, while for the simplex correlation method the detection criterion is computed in each iteration by first projecting the correlation vectors to the orthogonal complement of the subspace spanned by the correlation vectors of the previously identified frames. Following this difference, we show in the experimental part that the maximum correlation method is much more computationally efficient.

2.2.3 Relation to Simplex-EVD method

In this section, we discuss the relation of the proposed methods for activity probability estimation and our previously proposed simplex method [36]. Both the simplex algorithm [36] and the proposed methods rely on the statistical model presented in Section 2.1.2, and specifically on the correlation between frames (10). In [36], we define a correlation matrix with L columns that correspond to the correlation vectors defined in (23). We apply eigenvalue decomposition (EVD) to the correlation matrix and obtain a simplex representation embedded in $\mathbb{R}^J$. Next, we detect the simplex vertices and use them to transform the simplex representation that is based on the computed eigenvectors to the probability simplex. This method shares some similarity with the correlation simplex method derived in Section 2.2.2. The difference is that in [36], we first apply an EVD of the correlation matrix, while here we use the correlation vectors directly. Accordingly, in [36], we obtain a simplex in $\mathbb{R}^J$, while here we have a simplex in $\mathbb{R}^L$. The maximum correlation method, derived in Section 2.2.1, takes an entirely different approach that is based on LP theory. Note also that both proposed methods estimate the probabilities directly from the correlations, while in [36] they are estimated based on the computed eigenvectors. A diagram summarizing the three methods is depicted in Fig. 2a.

We discuss the computational complexity of the three methods:

1. *Maximum correlation method*: (i) Search for the maximum of each correlation function $\mathcal{O}(L^2)$. (ii) Count the number of times each frame is detected as a maximum $\mathcal{O}(L)$. (iii) Sequentially select frames most frequently detected as maximum and with low correlation to previously identified frames $\mathcal{O}(J \cdot L)$. Total: $\mathcal{O}(L^2 + J \cdot L)$; for $J \ll L$ we have $\mathcal{O}(L^2)$.
2. *Correlation simplex method*: (i) SPA. For each source: project correlation vectors $\mathcal{O}(L^3)$, compute the norm $\mathcal{O}(L^2)$, find maximum norm $\mathcal{O}(L)$. For all J sources: $\mathcal{O}(J \cdot L^3)$.
   Total: $\mathcal{O}(J \cdot L^3)$.
3. *EVD simplex method* [36]: (i) EVD $\mathcal{O}(L^3)$. Note that since only the first J eigenvectors are required, the computational complexity can be reduced by using iterative methods and by computing only the required eigenvectors, for example with Arnoldi iterations [44]. (ii) SPA. For each source: project correlation vectors $\mathcal{O}(J^2 \cdot L)$, compute the norm $\mathcal{O}(J \cdot L)$, find maximum norm $\mathcal{O}(L)$. For all J sources: $\mathcal{O}(J^3 \cdot L)$. (iii) Probability estimation: inverse of vertices matrix $\mathcal{O}(J^3)$, transform to probability vectors $\mathcal{O}(J^2 \cdot L)$
   Total: $\mathcal{O}(L^3 + J^3 \cdot L)$

We compare the performance and the computational time of the three methods in Section 3.

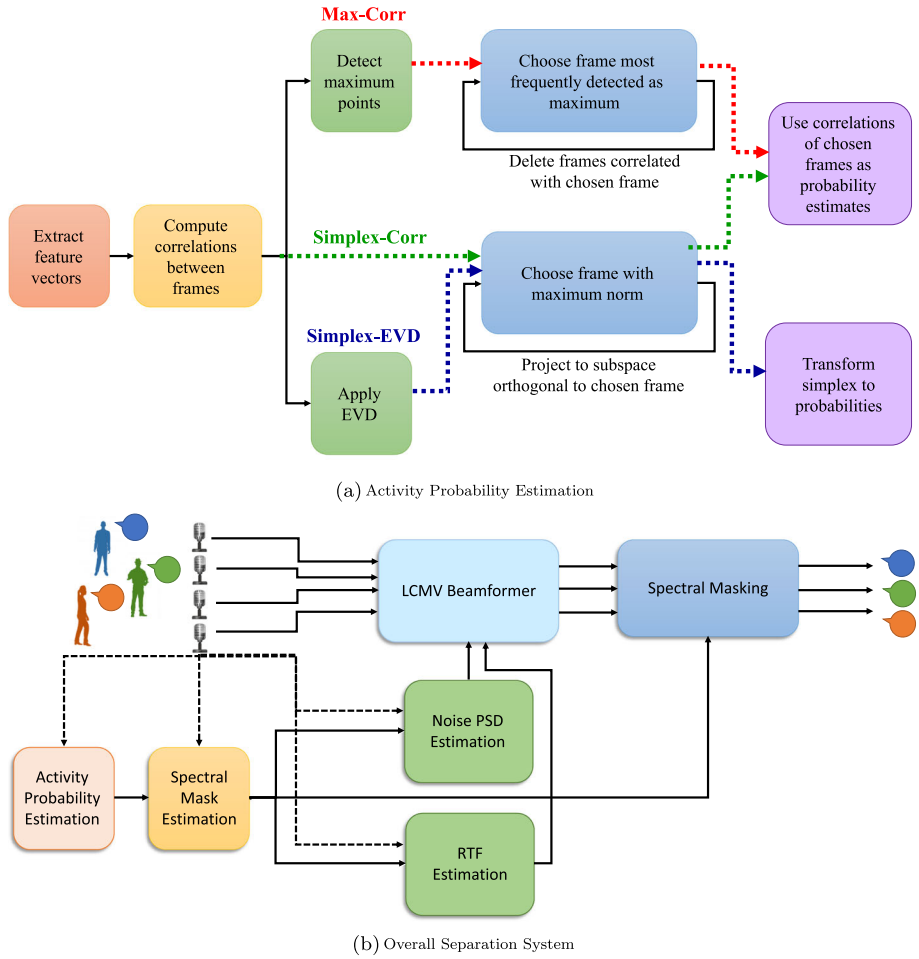


Fig. 2 Block-diagrams of **a** the methods for activity probability detection ("Max-Corr" derived in Section 2.2.1, "Simplex-Corr" derived in Section 2.2.2, and "Simplex-EVD" proposed in [36]) and **b** the overall separation system

2.3 Separation based on activity probability

We present a separation scheme that relies on the estimated activity probabilities. In the first stage, we estimate the spectral mask $M(l,f)$ based on the activity probabilities. In the second stage, we utilize the estimated spectral mask for the actual separation by applying a multichannel beamformer followed by a single-channel postfilter.

2.3.1 Spectral mask estimation

The spectral mask is estimated per frequency by combining the local relations between the bin-wise ratio features $\boldsymbol{r}(l,f)$ defined in (7) with the activity probabilities $\mathbf{p}(l)$ estimated by one of methods derived in Section 2.2. Specifically, the value of the spectral mask is determined for each TF bin based on the following weighted nearest-neighbor rule:

$$M(l,f) = \underset{j\in\{1,\ldots,J+1\}}{\operatorname{argmax}} \frac{1}{\pi_j} \sum_{n=1}^{L} \omega_{ln}(f) \cdot p_j(n) \tag{27}$$

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

$$\omega_{ln}(f) = \exp\left\{-||\boldsymbol{r}(l,f) - \boldsymbol{r}(n,f)||\right\}. \tag{28}$$

In (27), π_j serves as a class normalization and is given by:

$$\pi_j = \sum_{n=1}^{L} p_j(n). \tag{29}$$

Note that since the local mappings are aligned using the same global probabilities, the proposed method does not suffer from permutation ambiguity of the identity of the speakers across the different frequencies. The mask estimation procedure was adopted from [45], with a change in the definition of the local feature vectors. In [45], the feature vectors were defined based on a local simplex representation extracted from the EVD of the correlation

matrix defined for each frequency, while here we directly use the ratio values of all microphones.

2.3.2 Separation and enhancement

The separation is performed based on the estimated spectral mask (27) and is carried out in two stages by applying a multichannel beamformer followed by a single-channel spectral masking. The beamformer utilizes the diversity in the spatial characteristics of each speaker location and the noise, while the spectral masking utilizes the spectral diversity of the signals in the TF domain.

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

$$\hat{Y}_j^{\mathrm{LCMV}}(l,f) = \left(\mathbf{b}_j^{\mathrm{LCMV}}(f)\right)^H \mathbf{y}(l,f) \tag{30}$$

where the LCMV beamformer is defined by:

$$\mathbf{b}_j^{\mathrm{LCMV}}(f) = \mathbf{\Phi}_{nn}^{-1}(f)\mathbf{C}(f)\left(\mathbf{C}^H(f)\mathbf{\Phi}_{nn}^{-1}(f)\mathbf{C}(f)\right)^{-1}\mathbf{g}_j \tag{31}$$

where $\mathbf{\Phi}_{\mathrm{nn}}(f)$ is the noise power spectral density (PSD) matrix of size $M \times M$ and $\mathbf{C}(f)$ is an $M \times J$ matrix, comprising the RTFs of all speakers, i.e., $[\mathbf{C}(f)]_{m,j} = H_j^m(f)$. The estimation of the noise PSD matrix and the RTF matrix is summarized in Algorithm 3. The vector $\mathbf{g}_j \in \mathbb{R}^J$ extracts the jth speaker, with one in the jth entry and zeros elsewhere.

In the second stage, residual noise and interference signals at the output of the LCMV beamformer can be further suppressed by applying the estimated spectral mask:

$$\hat{Y}_j^{\mathrm{LCMV+MASK}}(l,f) = \mathbb{I}_j(l,f)\hat{Y}_j^{\mathrm{LCMV}}(l,f) + \beta\left(1-\mathbb{I}_j(l,f)\right)\hat{Y}_j^{\mathrm{LCMV}}(l,f) \tag{32}$$

where $\mathbb{I}_j(l,f)$ is the indicator function defined in Algorithm 3 and β is an attenuation factor. The entire separation process is summarized in Algorithm 4, and a block-diagram is depicted in Fig. 2b.

Algorithm 3: Beamformer Parameter Estimation

Define the indicator function, $1 \le j \le J$:

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

Noise PSD estimation:

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

RTF estimation:

- Speech PSD estimation:

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

- Solve the following generalized eigenvalue decomposition (GEVD) problem [33]:

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

- RTF estimate of the jth speaker:

$$\hat{H}_j^m(f) = \frac{[\hat{\mathbf{\Phi}}_{nn}(f)\boldsymbol{\psi}_j(f)]_m}{[\hat{\mathbf{\Phi}}_{nn}(f)\boldsymbol{\psi}_j(f)]_1}$$

3 Results and discussion

The algorithm performance was tested on a dataset that was self-recorded at the Bar-Ilan University (BIU) acoustic lab. We first describe the competing methods, the performance measures used for evaluation, the experimental setup, and the examined scenarios. Next, we present and discuss the results obtained by the proposed methods and the baseline methods.

3.1 Competing methods

In all experiments, we compared the proposed methods ("Max-Corr" and "Simplex-Corr") to the simplex method ("Simplex-EVD") [36]. As a baseline method, we used independent low-rank matrix analysis (ILRMA) algorithm [15], which is a state-of-the-art blind source separation (BSS) method that unifies IVA and NMF. Moreover, we compared to an ideal separator, which uses an ideal mask computed from the individual signals of each of the speakers.

3.2 Performance measures

The separation performance was evaluated in terms of signal to interference ratio (SIR) and signal to distortion ratio (SDR) measures as defined in [46]. The SIR measure reflects the suppression of interfering speech components with respect to the desired estimated speaker. The SDR measure reflects the preservation of the original speech components of the desired estimated speaker with respect to the corresponding true reference signal. Both measures were evaluated using the BSS-Eval toolbox [46].

3.3 Experimental setup

We describe the setup for the recordings carried out at the BIU acoustic lab. The room of size $6 \times 6 \times 2.4$ is equipped with controllable panels mounted over the ceiling, the floor, and the walls, which are used to adjust the reverberation time. In this experiment, the panels were adjusted to create two reverberation levels: a low reverberation level set to $T_{60} \approx 150$ ms, and a high reverberation level set to $T_{60} \approx 550$ ms. The recordings included 20

Algorithm 4: The entire Separation Algorithm

Feature Extraction:

- Compute ratios $\left\{R^m(l,f)\right\}_{l,f,m}$ (4).
- Construct ratio vectors $\{\mathbf{r}(l)\}_{l=1}^{L}$ (7).

Spectral Mask Estimation:

- Estimate the activity probabilities by Algorithm 1 or 2.
- **for** each frequency $f = 1 : K$:
  - Compute weights $\{\omega_{ln}^{L}(f)\}_{l,n}$ (28).
  - Estimate the mask $M(l,f)$ (27) for each frame.

  end

Separation and Enhancement

- Estimate the beamformer parameters using Algorithm 3
- Compute the LCMV beamformer $\mathbf{b}_j^{\mathrm{LCMV}}(f)$ (31)
- Obtain $\hat{Y}_j^{\mathrm{LCMV}}(l,f)$ by applying the LCMV beamformer $\mathbf{b}_j^{\mathrm{LCMV}}(f)$ on $\mathbf{y}(l)$ (30).
- Obtain $\hat{Y}_j^{\mathrm{LCMV+MASK}}(l,f)$ by applying a single-channel spectral masking on $\hat{Y}_j^{\mathrm{LCMV}}(l,f)$ (32).

human participants (10 females and 10 males), sitting in 6 possible seats around a rectangular table. The participants were recorded individually one at a time to enable the generation of various multi-party scenarios. Each participant was recorded 12 times for each seat and for each reverberation level. In each recording, the speaker uttered five different sentences of about 5 s each with a pause of about 5 s between two subsequent sentences. The sentences were unique for each speaker and for each seat. The signals were measured by 24 CK32 omnidirectional microphones of AKG, placed in the room and mounted on the table. The room layout with the positions of the speakers and the microphones is depicted in Fig. 3a, and a photo of the room setup is presented in Fig. 3b.

In addition, a babble noise was recorded, imitating a diffuse noise field that arrives evenly from all directions. To generate this noise, 8 loudspeakers were placed at each corner of the room and at the middle of each of the walls, as illustrated in Fig. 3a. The loudspeakers were pointed towards the walls and were recorded while playing babble noise signals.

In the experiments presented here, we utilized the measurements of a subset of $M = 8$ microphones with indexes 25–32 in the setup depicted in Fig. 3a. The signals were acquired with 24-bit resolution and 48 kHz sampling rate, and were then downsampled to 16 kHz for further processing. The signals were analyzed in the STFT domain with window length of $K = 2048$ samples with 75% overlap between adjacent frames. The feature vectors (7) consist of $F = 257$ frequency bins, corresponding to 1–3 kHz. Note that we focus on a range that contains most of the speech power, and exclude low frequencies that mostly contain noise, as well as high frequencies, in which there is typically only low speech energy. The following parameter values were used: $\varepsilon = 0.2$, $\beta = 0.3$. The parameters were chosen empirically to obtain good and stable results. The parameter ε (18) is a probability threshold used for excluding frames correlated with sources that have already been identified, and therefore should be low enough to exclude all the frames associated with the detected sources but high enough so that the remaining sources would not be missed. The parameter β (32) presents a trade-off between speech distortion and noise and interference suppression, such that as β increases we obtain better SIR but lower SDR, and vice versa.

3.4 Examined scenarios

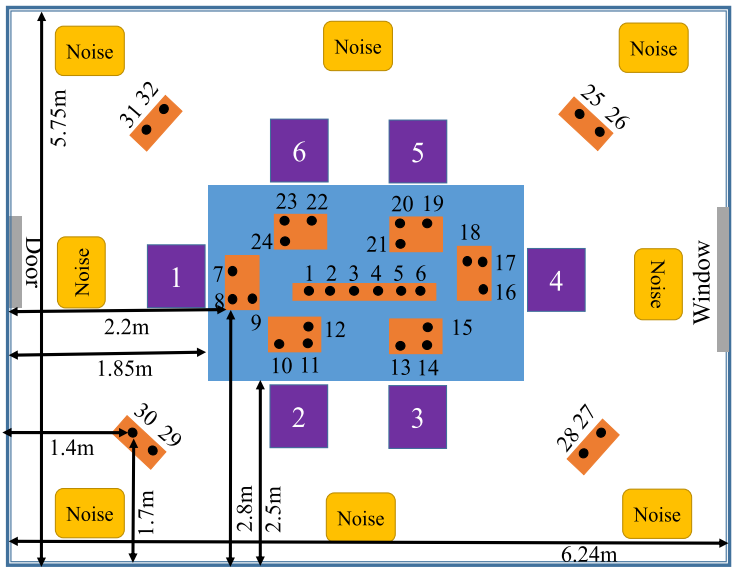
The performance was examined on two scenarios as summarized in Table 1. The first scenario consists of mixtures of 4 speakers, where the first speaker has infrequent activity while the other three speakers have balanced activity. The first speaker is active solely for a short duration, and then, the other 3 speakers start speaking one after the other. In the second scenario, there are mixtures of J speakers with balanced activity.

At the beginning of each mixture signal, there is a 1-s-long segment with noise only. The total length of the signals is 20 s. For each condition in each scenario, we conducted 50 Monte Carlo (MC) trials. In each trial, a random subset of speakers and seats was chosen. Representative timelines of the two scenarios are given in Fig. 4.

3.5 Results

For the first scenario, we examined two levels of activity of the first speaker, i.e., 5% and 10% activity percentages with respect to the entire signal duration. The signal to noise ratio (SNR) was set to 20 dB. The scores of the first speaker were averaged over 50 trials, and the scores of the three balanced speakers were averaged together in all 50 trials. The average SIR scores are given in Table 2, and the average SDR scores are given in Table 3 for the two reverberation levels.

We observe the superiority of the proposed methods over ILRMA in all cases, and especially for the infrequent speaker. The "Simplex-EVD" and "Simplex-Corr" methods obtain similar scores for the balanced speakers, which



(a)

(b)

Fig. 3 a The room layout, containing a table (blue rectangle), 6 chairs where speakers are sitting (purple squares), 32 microphones (black circles), and 8 loudspeakers emitting noise (yellow rectangles). The microphones utilized for the presented experiments are 25–32 or 1–6. **b** A photo of the room setup of the experiments conducted at the BIU acoustic lab

are higher than the scores obtained for the "Max-Corr" method. Both "Simplex-Corr" and "Max-Corr" methods obtain better performance compared to the "Simplex-EVD" method with respect to the infrequent speaker. For this speaker, we observe an advantage of the "Max-Corr" method compared to "Simplex-Corr" method, except for the case of high reverberation and 10% activity percentage.

Table 1 Examined scenarios

Scenario	Infrequent Spks	Balanced Spks	SNR [dB]	T_{60} [ms]
Unbalanced	1	3	20	150, 550
Balanced	None	3	5–20	150, 550
	None	2–5	20	150, 550

For the second scenario, we first examined the performance with respect to the noise level for mixtures of $J = 3$ speakers with balanced activity. We carried out 50 MC trials for each SNR level and averaged the obtained scores over all trials and over the three speakers. The average SIR and SDR scores are depicted in Fig. 5. It can be seen that the performance of the "Simplex-Corr" method is comparable to that of the "Simplex-EVD" method, and both of them are superior with respect to the "Max-Corr" method. In addition, both "Simplex-Corr"

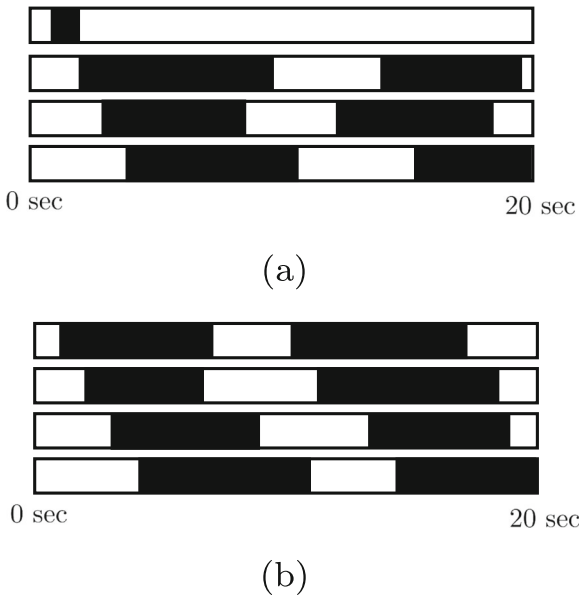


Fig. 4 Representative temporal activity of **a** the first scenario with a mixture of $J = 4$ participants, consisting of one speaker with low activity, and **b** the second scenario with a mixture of $J = 4$ participants with balanced activity

and "Max-Corr" methods achieve better results compared to the ILRMA algorithm.

In order to show that the obtained performance trends are not tailored to a specific array configuration, we repeated this experiment with a different array constellation. We used a uniform linear array (ULA), located on the room table, which consists of six microphones, indexed as 1–6 in Fig. 3a. The results obtained for mixtures of 3 speakers are depicted in Fig. 6. The performance trends obtained for the ULA are similar to those obtained for the distributed array in Fig. 5, and here too the proposed methods outperform ILRMA. In general, the SIR scores are lower and the SDR scores are higher for the ULA compared to the distributed configuration.

For the second scenario of balanced speakers, we also carried out an evaluation of the performance with respect to the number of speakers, based on microphones 25–32. The obtained scores are depicted in Fig. 7, where each point in the figure represents an average over 50 MC trials and over all speakers. The SNR was set to 20 dB. We observe a decrease in the separation scores obtained by all methods as the number of speakers increases. Similar trends are observed here as for the performance evaluation with respect to the noise levels, namely the comparable performance of "Simplex-Corr" and "Simplex-EVD" methods that is superior to that of the "Max-Corr" method. Regarding ILRMA algorithm, it outperforms the "Max-Corr" method in terms of the SDR score only for high reverberation conditions and large number of speakers, and achieves lower scores in all other cases as compared with both proposed methods.

In addition, we conducted an experiment to evaluate the running time of the three methods for activity probability estimation. The algorithms were implemented in Matlab on a standard PC (CPU Intel Core2 Quad 3.7 GHz, RAM 8 GB). The running times of each algorithm for mixtures of $J = 3$ speakers are summarized in Table 4 for different recording lengths. Each running time in the table is obtained by an average over 4 trials. We observe that the "Max-Corr" method achieves the lowest running times.

3.6 Activity detection of a counseling session

Finally, we demonstrate the performance of the proposed methods on real recordings of a session of a psychological counseling recorded at the BIU Psychotherapy Research Lab. For this purpose, we used a two lapel microphone recording of a client, which speaks most of the time, and a therapist, who is involved only during short time segments. Figure 8 depicts the two-channel measured waveforms. On top, asterisks denote true and estimated time instances with activity of each speaker. The true annotation was *manually* determined with Praat software [47]. We observe that the "Simplex-EVD" method detects only the client, while the proposed methods successfully detect the activities of both the client and the therapist almost all the time, even when they overlap.

3.7 Discussion

We conclude that the proposed methods obtain high separation scores for both balanced and infrequent speakers

Table 2 Distributed array: SIR scores—mixtures with unbalanced activity for "Low"/"High" reverberation and for "5%"/"10%" activity of the 1st speaker

Reverb.	Activity prec.	Infrequent speaker					Balanced speakers				
		Ideal	Max-Corr	Simplex-Corr	Simplex-EVD	ILRMA	Ideal	Max-Corr	Simplex-Corr	Simplex-EVD	ILRMA
Low	5%	19.87	**16.82**	12.11	8.64	1.54	23.77	21.03	**22.23**	22.04	11.89
	10%	21.61	**19.79**	16.98	16.72	6.00	23.65	20.86	21.92	**22.22**	10.38
High	5%	17.46	**14.03**	12.34	6.26	−1.34	22.38	19.05	**21.11**	21.10	10.63
	10%	19.18	14.13	**16.22**	15.50	3.17	22.33	18.29	20.38	**20.83**	9.82

Table 3 Distributed array: SDR scores—mixtures with unbalanced activity for "Low"/"High" reverberation and for "5%"/"10%" activity of the 1st speaker

Reverb.	Activity prec.	Infrequent speaker					Balanced speakers				
		Ideal	Max-Corr	Simplex-Corr	Simplex-EVD	ILRMA	Ideal	Max-Corr	Simplex-Corr	Simplex-EVD	ILRMA
Low	5%	9.64	**6.95**	3.08	0.25	−3.36	9.98	7.53	8.37	**8.68**	5.80
	10%	9.86	**8.22**	6.02	6.19	0.60	9.85	7.62	8.32	**8.96**	4.43
High	5%	7.47	**5.12**	3.54	−1.74	−5.47	7.66	5.22	6.63	**6.96**	4.11
	10%	7.96	3.91	**5.73**	5.38	−1.73	7.69	4.94	6.39	**7.06**	3.58

and outperform the ILRMA algorithm for most cases in various noise and reverberation conditions.

The robustness of the proposed methods to reverberation may be attributed to the fact that for both parts of the activity detection and separation, we use the RTF, which consists of the full reflection pattern. For the activity detection, we differentiate between the speakers using features based on the RTFs and thus obtain robustness to reverberation compared to methods that rely on the direct-path only, which may be masked by reflections in high reverberation conditions. For the separation, we apply a beamformer that is constructed with a steering vector based on the RTF rather than the direct-path only, which results in milder distortion of the speech signal due to the preservation of the entire speech power coming from both the direct and reflected paths, as was shown in [39]. Note also that relying on the RTFs, which provide a richer spatial information compared to the direct-path only, has also the potential of separating sources that are located one behind the other, as was demonstrated in [48] and [36] (see Fig. 8 therein).

For speakers with balanced activity, the best performance is achieved by the "Simplex-Corr" and the "Simplex-EVD" methods with a small advantage for latter. For infrequent speakers with low activity, the "Simplex-Corr" and the "Max-Corr" methods are preferable over the "Simplex-EVD" method, where the "Max-corr" method performs the best in the case of very low

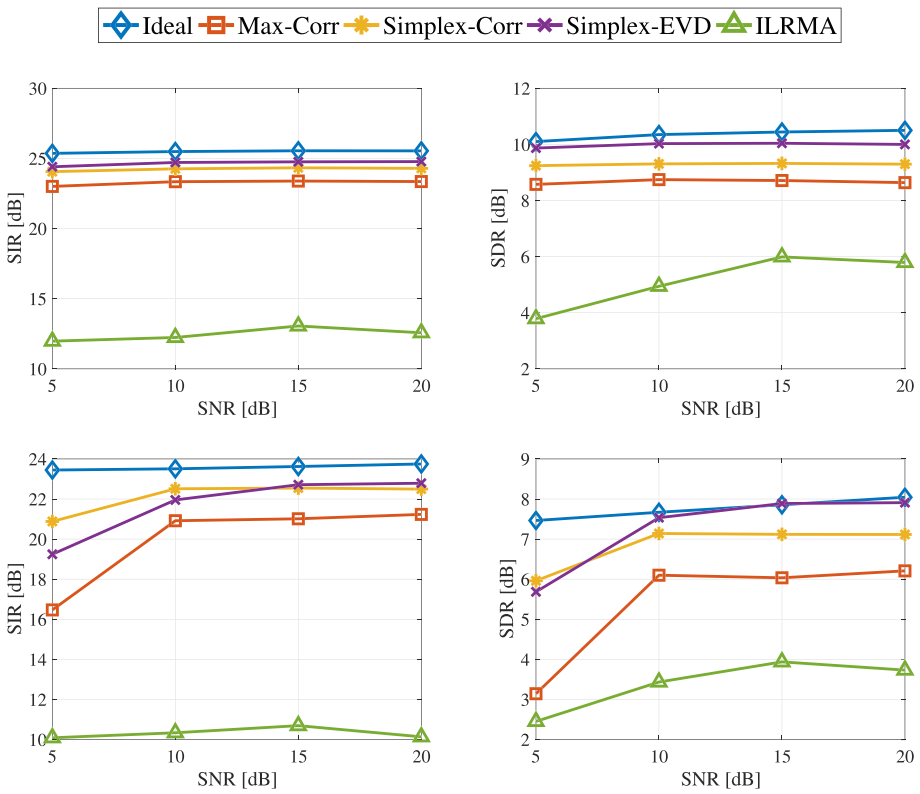


Fig. 5 Distributed array: evaluation for mixtures with $J = 3$ speakers with balanced activity—performance as a function of the SNR for **a**, **b** low and **c**, **d** high reverberation conditions

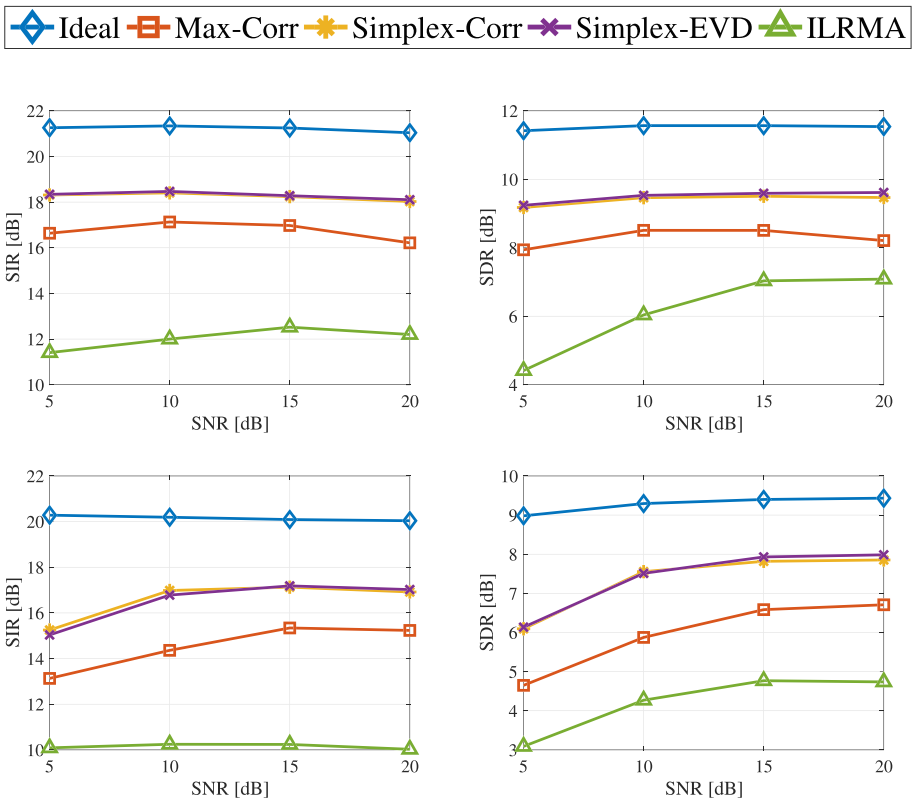


Fig. 6 ULA: evaluation for mixtures with $J = 3$ speakers with balanced activity—performance as a function of the SNR for **a**, **b** low and **c**, **d** high reverberation conditions

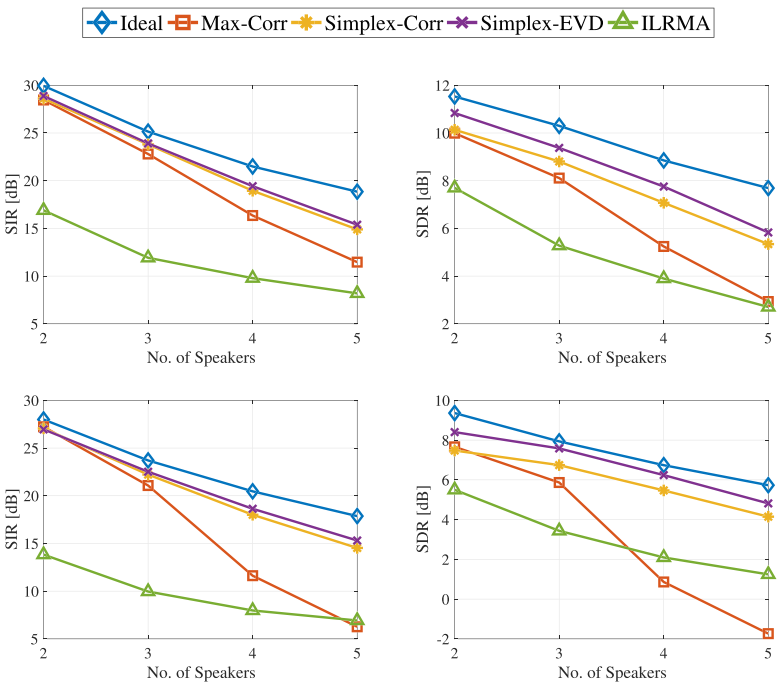


Fig. 7 Distributed array: evaluation for mixtures with balanced activity—performance as a function of the number of speakers for **a**, **b** low and **c**, **d** high reverberation conditions and SNR = 20 dB

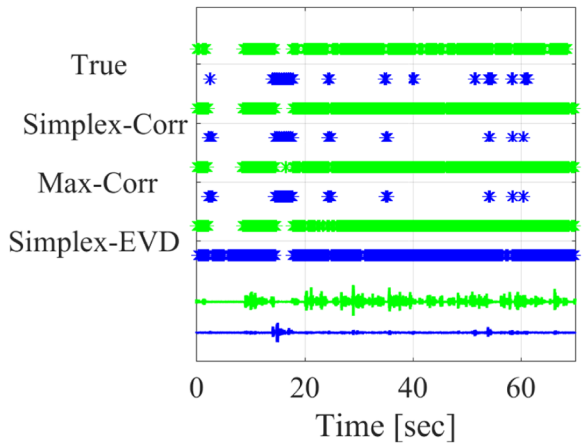


Fig. 8 Blue and green waveforms of microphones attached to a therapist and to a patient, respectively. The true and the estimated time instances with activity of each speaker are marked by asterisks

activity. The lowest computational time is achieved by the "Max-corr" method. Table 5 presents several scenarios and the method that would be preferred in each case.

The reason for the differences in the performance of "Simplex-EVD," "Simplex-Corr," and the "Max-Corr" methods can be explained as follows. The "Simplex-EVD" method performs a global processing using an EVD and hence may have the best performance in standard cases, but often misses low-active speakers, which have a minor contribution to the obtained decomposition. In contrast, the local approach taken by both proposed methods is more sensitive to infrequent participants. It turns out that in most cases, the performance of the "Simplex-Corr" method is preferable; however, the "Max-Corr" method is more sensitive to speakers with very low activity and also has the advantage of lower computational complexity.

4 Conclusions

We presented two novel methods for multichannel speaker separation. For the first method, it is shown that the maxima of the correlation function between different frames correspond to single-speaker frames. Accordingly, we propose an algorithm for sequential recovery of frames dominated by each speaker, and in turn, we use their correlations as an estimator for the activity probabilities. In the second method, single-speaker frames correspond to vertices of the simplex defined by the correlation vectors and are detected by means of convex geometry. A spectral mask is recovered by the estimated probabilities and is utilized for the actual separation of the mixture. Both proposed methods show high separation capabilities in real-life scenarios for different reverberation and noise levels, and especially in the challenging scenario of speakers with low activity. The maximum correlation method has better performance for speakers with very low activity and is also more computationally efficient, while the correlation simplex method performs better for speakers with balanced activity, especially in adverse conditions of high noise and reverberation.

Table 4 Running times in seconds of methods for activity probability estimation

	40 s	80 s	120 s	160 s	200 s	240 s
Simplex-EVD	0.07	0.09	0.20	0.35	0.55	0.80
Simplex-Corr	0.18	1.09	3.64	8.57	17.50	29.06
Max-Corr	0.02	0.05	0.10	0.18	0.30	0.51

Table 5 Preferred methods in different scenarios

Scenario/method	Max-Corr	Simplex-Corr	Simplex-EVD
High reverberation		✓	✓
High noise		✓	✓
Many speakers		✓	✓
Infrequent speaker	✓	✓	
Lowest complexity	✓		

Authors' contributions
Model development: BLG, RT, and SG. Experimental testing: BLG. Writing paper: BLG, RT, and SG. The authors read and approved the final manuscript.

Funding
This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245. Bracha Laufer-Goldshtein is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities.

Availability of data and materials
N/A

Consent for publication
All authors agree to the publication in this journal.

Competing interests
N/A

Author details
[1]Faculty of Engineering, Bar-Ilan University, 5290002, Ramat-Gan, Israel. [2]Viterbi Faculty of Electrical Engineering, Technology, 3200003, Haifa, Israel.

Received: 31 August 2020 Accepted: 4 January 2021
Published online: 28 January 2021

References

1. S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, A consolidated perspective on multimicrophone speech enhancement and source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(4), 692–730 (2017)
2. S. Makino, T.-W. Lee, H. Sawada, *Blind Speech Separation, vol. 615*. (Springer, New-York, Berlin, Heilderberg, 2007)
3. M. S. Pedersen, J. Larsen, U. Kjems, L. C. Parra, in *Springer Handbook of Speech Processing*. Convolutive blind source separation methods (Springer, New-York, Berlin, Heilderberg, 2008), pp. 1065–1094
4. E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, M. E. Davies, Probabilistic modeling paradigms for audio source separation. Mach Audition Princ. Algoritm. Syst., 162–185 (2010)

5. S. Makino, *Audio Source Separation*. (Springer, New-York, Berlin, Heilderberg, 2018)
6. P. Smaragdis, Blind separation of convolved mixtures in the frequency domain. Neurocomputing. **22**(1-3), 21–34 (1998)
7. H. Buchner, R. Aichner, W. Kellermann, A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. IEEE Trans. Speech Audio Process. **13**(1), 120–134 (2005)
8. S.-Y. Lee, Blind source separation and independent component analysis: a review. Neural Inf. Process.-Lett. Rev. **6**(1), 1–57 (2005)
9. T. Kim, T. Eltoft, T.-W. Lee, in *International Conference on Independent Component Analysis and Signal Separation*. Independent vector analysis: an extension of ICA to multivariate components (Springer-Verlag, Berlin Heidelberg, 2006), pp. 165–172
10. Z. Koldovsky, P. Tichavsky, Time-domain blind separation of audio sources on the basis of a complete ICA decomposition of an observation space. IEEE Trans. Audio Speech Lang. Process. **19**(2), 406–416 (2011)
11. C. Févotte, N. Bertin, J.-L. Durrieu, Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. Neural Comput. **21**(3), 793–830 (2009)
12. H. Kameoka, N. Ono, K. Kashino, S. Sagayama, in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Complex NMF: a new sparse representation for acoustic signals, (New York, 2009), pp. 3437–3440
13. A. Ozerov, C. Févotte, Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Trans Audio Speech Lang. Process. **18**(3), 550–563 (2010)
14. P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, M. Hoffman, Static and dynamic source separation using nonnegative factorizations: a unified view. IEEE Signal Process. Mag. **31**(3), 66–75 (2014)
15. D. Kitamura, N. Ono, H. Sawada, H. Kameoka, H. Saruwatari, Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(9), 1622–1637 (2016)
16. P. Pertilä, J. Nikunen, Distant speech separation using predicted time–frequency masks from spatial features. Speech Commun. **68**, 97–106 (2015)
17. A. A. Nugraha, A. Liutkus, E. Vincent, Multichannel audio source separation with deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(9), 1652–1664 (2016)
18. X.-L. Zhang, D. Wang, A deep ensemble learning method for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(5), 967–977 (2016)
19. Z.-Q. Wang, J. Le Roux, J. R. Hershey, in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Multi-channel deep clustering: discriminative spectral and spatial embeddings for speaker-independent speech separation, (New York, 2018), pp. 1–5
20. D. Wang, J. Chen, Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(10), 1702–1726 (2018)
21. Z.-Q. Wang, D. Wang, Combining spectral and spatial features for deep learning based blind speaker separation. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(2), 457–468 (2019)
22. L. Drude, R. Haeb-Umbach, in *Proc. of The Annual Conference of the International Speech Communication Association (Interspeech)*. Tight integration of spatial and spectral features for BSS with deep clustering embeddings, (2017), pp. 2650–2654
23. S. E. Chazan, J. Goldberger, S. Gannot, in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DNN-based concurrent speakers detector and its application to speaker extraction with LCMV beamforming, (New York, 2018), pp. 6712–6716
24. A. K. Das, C. K. Y. Leung, Icd: a methodology for real time onset detection of overlapped acoustic emission waves. Autom. Constr. **119**, 103341 (2020)
25. A. K. Das, T. T. Lai, C. W. Chan, C. K. Leung, A new non-linear framework for localization of acoustic sources. Struct. Health Monit. **18**(2), 590–601 (2019)
26. O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. **52**(7), 1830–1847 (2004)
27. S. Arberet, R. Gribonval, F. Bimbot, A robust method to count and locate audio sources in a multichannel underdetermined mixture. IEEE Trans. Signal Process. **58**(1), 121–133 (2010)
28. M. I. Mandel, R. J. Weiss, D. P. W. Ellis, Model-based expectation-maximization source separation and localization. IEEE Trans. Audio Speech Language Process. **18**(2), 382–394 (2010)
29. J. Traa, P. Smaragdis, Multichannel source separation and tracking with RANSAC and directional statistics. IEEE Trans. Audio Speech Language Process. **22**(12), 2233–2243 (2014)
30. S. Winter, W. Kellermann, H. Sawada, S. Makino, MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l_1-norm minimization. EURASIP J. Appl. Signal Process. **2007**(1), 81–81 (2007)
31. H. Sawada, S. Araki, S. Makino, Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. IEEE Trans. Audio Speech Language Process. **19**(3), 516–527 (2011)
32. M. Souden, S. Araki, K. Kinoshita, T. Nakatani, H. Sawada, A multichannel MMSE-based framework for speech source separation and noise reduction. IEEE Trans. Audio Speech Language Process. **21**(9), 1913–1928 (2013)
33. S. Markovich, S. Gannot, I. Cohen, Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. IEEE Trans. Audio Speech Language Process. **17**(6), 1071–1086 (2009)
34. D. Cherkassky, S. Gannot, Successive relative transfer function identification using blind oblique projection. IEEE/ACM Trans. Audio Speech Language Process. **28**, 474–486 (2019)
35. Y. Laufer, S. Gannot, in *Proc. of 28th European Signal Processing Conference (EUSIPCO)*. A Bayesian hierarchical model for blind audio source separation (IEEE, New York, 2020), pp. 1–5
36. B. Laufer-Goldshtein, R. Talmon, S. Gannot, Source counting and separation based on simplex analysis. IEEE Trans. Signal Process. **66**(24), 6458–6473 (2018)
37. B. Laufer-Goldshtein, R. Talmon, S. Gannot, in *Proc. of 26th European Signal Processing Conference (EUSIPCO)*. Diarization and separation based on a data-driven simplex (IEEE, 2018), pp. 842–846
38. A. Paz, E. Rafaeli, E. Bar-Kalifa, E. Gilboa-Schectman, S. Gannot, B. Laufer-Goldshtein, S. Narayanan, J. Keshet, D. Atzil-Slonim, Intrapersonal and interpersonal vocal emotional dynamics during psychotherapy. J. Consult. Clin. Psychol. (2020)
39. S. Gannot, D. Burshtein, E. Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Signal Process. **49**(8), 1614–1626 (2001)
40. I. Cohen, Relative transfer function identification using speech signals. IEEE Trans. Speech Audio Process. **12**(5), 451–459 (2004)
41. G. B. Dantzig, M. N. Thapa, *Linear Programming 2: Theory and Extensions*. (Springer, New-York, Berlin, Heilderberg, 2006)
42. M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvao, T. Yoneyama, H. C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometr. Intell. Lab. Syst. **57**(2), 65–73 (2001)
43. W.-K. Ma, J. M. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. J. Plaza, A. Ambikapathi, C.-Y. Chi, A signal processing perspective on hyperspectral unmixing: insights from remote sensing. IEEE Signal Process. Mag. **31**(1), 67–81 (2014)
44. W. E. Arnoldi, The principle of minimized iterations in the solution of the matrix eigenvalue problem. Q. Appl. Math.s. **9**(1), 17–29 (1951)
45. B. Laufer-Goldshtein, R. Talmon, S. Gannot, Global and local simplex representations for multichannel source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **28**(1), 914–928 (2020)
46. E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
47. P. Boersma, D. Weenink, Praat (version 4.5. 25)[software] (2007). Latest version available for download from http://www.praat.org
48. E. Hadad, F. Heese, P. Vary, S. Gannot, in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. Multichannel audio database in various acoustic environments (IEEE, New York, 2014), pp. 313–317

Publisher's Note