Diarization and Separation Based on a Data-Driven Simplex

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

[1] Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002 , Israel.
Email: bracha.laufer@biu.ac.il,Sharon.Gannot@biu.ac.il
[2] The Viterbi Faculty of Electrical Engineering , Technion - Israel Institute of Technology, Haifa, 3200003, Israel.
Email: ronen@ee.technion.ac.il

***Abstract*—Separation of underdetermined speech mixtures, where the number of speakers is greater than the number of microphones, is a challenging task. Due to the intermittent behaviour of human conversations, typically, the instantaneous number of active speakers does not exceed the number of microphones, namely the mixture is locally (over-)determined. This scenario is addressed in this paper using a dual stage approach: diarization followed by separation. The diarization stage is based on spectral decomposition of the correlation matrix between different time frames. Specifically, the spectral gap reveals the overall number of speakers, and the computed eigenvectors form a simplex of the activity of the speakers across time. In the separation stage, the diarization results are utilized for estimating the mixing acoustic channels, as well as for constructing an unmixing scheme for extracting the individual speakers. The performance is demonstrated in a challenging scenario with six speakers and only four microphones. The proposed method shows perfect recovery of the overall number of speakers, close to perfect diarization accuracy, and high separation capabilities in various reverberation conditions.**

***Index Terms*—Blind audio source separation (BASS), diarization, relative transfer function (RTF), simplex.**

I. Introduction

Blind audio source separation (BASS) methods aim at extracting the source signals of the individual speakers from their measured mixtures without any prior knowledge of the sources or the mixing acoustic channels [1]. Numerous BASS methods were proposed over the last decades, exploiting various assumptions regarding the sources and the mixing systems, such as: statistical independence [2], admitting sparse representations [3], decomposability to non-negative components [4], W-disjoint orthogonality (WDO) in the time-frequency (TF) domain [5]–[7], etc.

Diarization methods focus on identifying the speakers' identity in short time segments of audio signals [8], [9]. Most diarization algorithms apply some preprocessing to the data, typically including feature extraction and segmentation to speech/non-speech intervals. Various features are commonly used, where the mel frequency cepstral coefficients (MFCCs) and time difference of arrival (TDOA) estimates are widely-spread. Then, the detected speech segments are clustered into individual speakers using, for example, bottom-up or top-down hierarchical clustering methods [10].

Despite their clear relation, only a few approaches address joint diarization and separation [11]–[15]. More commonly, BASS and diarization are treated as separate problems and are devised with different assumptions regarding the mutual activities of the different speakers. Most BASS methods implicitly assume that all the speakers are continuously and concurrently active, whereas most diarization methods consider non-overlapping speakers. However, neither of these assumptions hold in many real-world scenarios, for example, in meeting rooms, where partial overlap of several speakers is common.

In this paper, we propose a dual-stage approach of diarization and separation. The separation stage relies on the dirarization outcomes, namely, in each time segment an umixing procedure is applied to a small set of active speakers, identified in the preceding diarization stage. Our approach can support challenging scenarios of underdetermined mixtures, with more speakers than microphones, where at each point in time the instantaneous mixture is (over-)determined, i.e. the number of speakers does not exceed the number of microphones.

The diarization method relies on spectral decomposition of the correlation matrix defined between different time frames. The justification of the method is based on a probabilistic model, in which the column space of the correlation matrix is spanned by the probabilities of the various speakers across time. Accordingly, the spectrum decay reveals the overall number of speakers, and the computed eigenvectors form a simplex of the speakers' activity probabilities. Based on the diarization outcomes, the mixing acoustic systems are estimated using time frames, which are highly dominated by a single speaker. Then, an unmixing scheme is constructed to extract the individual speakers. The performance is validated in a separation task of two simultaneous conversations with three speakers each, using an array of four microphones. High diarization and separation scores are demonstrated in various reverberation conditions, as well as perfect recovery of the overall number of speakers.

II. Problem Formulation

Consider J speakers measured by an array of M microphones in a reverberant environment. The signals are analysed

[1] This work is supported by the Adams Foundation of the Israel Academy of Sciences and Humanities.

in the short time Fourier transform (STFT) domain, where $l \in \{1, \ldots, L\}$ and $f \in \{1, \ldots, F\}$ are used to denote the frame and the frequency indices, respectively. We use the narrowband approximation entailing that the jth signal measured by the mth microphone is given by $Y_j^m(l,f) = A_j^m(f) S_j(l,f)$, where $S_j(l,f)$ denotes the signal of the jth speaker, and $A_j^m(f)$ denotes the acoustic transfer function (ATF) relating the jth speaker and the mth microphone. Accordingly, the mixed signal $Y^m(l,f)$ measured by the mth microphone is given by:

$$Y^m(l,f) = \sum_{j=1}^{J} \mathcal{I}_j(l) Y_j^m(l,f) = \sum_{j=1}^{J} \mathcal{I}_j(l) H_j^m(f) Y_j^1(l,f) \tag{1}$$

where $H_j^m(f) = \frac{A_j^m(f)}{A_j^1(f)}$ is the relative transfer function (RTF) of the jth speaker, defined between the mth microphone and the first microphone, which serves as a reference microphone.

Here, $\mathcal{I}_j(l)$ is an activity indicator which equals one if the jth source is active in the lth frame and zero otherwise. We assume that the overall mixture is underdetermined, i.e. $J > M$. However, the mixture is locally (over-)determined, i.e. in each frame the number of active speakers does not exceed the number of microphones:

$$J_l \equiv \sum_{j=1}^{J} \mathcal{I}_j(l) \leq M, \forall 1 \leq l \leq L \tag{2}$$

Our goal is to recover the number of speakers J, and to perform both diarization and separation of the measurements. In the diarization stage, each frame is assigned with its corresponding set of active speakers, namely the value of the indicator functions are estimated for each frame $1 \leq l \leq L$ and speaker $1 \leq j \leq J$. In the separation stage, the individual speakers are extracted from the measured mixtures.

III. Separation Method

We propose a separation scheme, which assumes that the activities of the speakers across time and the estimated RTFs are available. Let $\mathcal{S}_l$ denote the set of J_l active speakers in the lth frame, i.e. $\mathcal{S}_l = \{j \mid \mathcal{I}_j(l) = 1,\ j \in \{1, \ldots, J\}\}$. The jth source (as measured by the reference microphone) is extracted from the measurements (1) using the following unmixing scheme:

$$\hat{Y}_j^1(l,f) = \begin{cases} \mathbf{b}_j^H(l,f)\mathbf{y}(l,f) & j \in \mathcal{S}_l, J_l > 1 \\ Y^1(l,f) & j \in \mathcal{S}_l, J_l = 1 \\ 0 & j \notin \mathcal{S}_l \end{cases} \tag{3}$$

where

$$\mathbf{y}(l,f) = \left[Y^1(l,f), Y^2(l,f), \ldots, Y^M(l,f)\right]^T \tag{4}$$

and $\mathbf{b}_j(l,f)$ consists of the pseudo-inverse of the instantaneous mixing system at frame l:

$$\mathbf{b}_j(l,f) = \mathbf{C}(l,f)(\mathbf{C}(l,f)^H \mathbf{C}(l,f))^{-1} \mathbf{g}_j(l) \tag{5}$$

where the estimated RTFs of the active speakers in $\mathcal{S}_l$ constitute the columns of $\mathbf{C}(l,f) \in \mathbb{C}^{M \times J_l}$. The vector $\mathbf{g}_j(l)$ is a J_l dimension vector that extracts the jth speaker, with one in the corresponding entry of the jth speaker and zeros elsewhere.

In order to compute (5) an estimation of the RTFs of the different speakers is required. We define the set $\mathcal{L}_j$, which consists of frame indices for which only the jth speaker is active, i.e. $\mathcal{I}_j(l) = 1$ and $\mathcal{I}_i(l) = 0$, $\forall i \neq j$. Exploiting the set $\mathcal{L}_j$, the RTF of the jth speaker can be estimated by:

$$\hat{H}_j^m(f) = \frac{\sum_{l \in \mathcal{L}_j} Y^m(l,f) Y^{1*}(l,f)}{\sum_{l \in \mathcal{L}_j} Y^1(l,f) Y^{1*}(l,f)} \tag{6}$$

The proposed unmixing scheme (3) requires an identification of the sets $\{\mathcal{S}_l\}_l$ of active speakers in each frame l, as well as an RTF estimation (6) based on the sets $\{\mathcal{L}_j\}_j$. The remainder of this paper presents a diarization method, in which the value of the indicator functions $\{\mathcal{I}_j(l)\}_{l,j}$ is determined, facilitating the estimation of the sets $\{\mathcal{S}_l\}_l$ and $\{\mathcal{L}_j\}_j$.

IV. Diarization Method

A. Statistical Model

In a multichannel static setup, each speaker can be uniquely identified by its spatial signature, manifested in the associated RTF values. For each speaker, we define an RTF vector $\mathbf{h}_j$ with $D = 2 \cdot (M-1) \cdot K$ elements for the real and the imaginary parts of the RTF values, in K frequency bins and in $M - 1$ microphones:

$$\begin{aligned} \mathbf{h}_j^m &= \left[H_j^m(f_1), H_j^m(f_2), \ldots, H_j^m(f_K)\right]^T \\ \mathbf{h}_j^{\mathrm{c}} &= \left[\mathbf{h}_j^{2^T}, \mathbf{h}_j^{3^T}, \ldots, \mathbf{h}_j^{M^T}\right]^T . \\ \mathbf{h}_j &= \left[\text{real}\left\{\mathbf{h}_j^{\mathrm{c}}\right\}^T, \text{image}\left\{\mathbf{h}_j^{\mathrm{c}}\right\}^T\right]^T . \end{aligned} \tag{7}$$

Note that $\mathbf{h}_j^1$ is an all-ones vector for all $1 \leq l \leq L$, hence excluded from $\mathbf{h}_j$ in (7). We assume that the RTF vectors are i.i.d. random vectors with zero-mean and a unit covariance function:

$$E\left\{\mathbf{h}_j \mathbf{h}_j^T\right\} = \mathbf{I}_D. \tag{8}$$

Further discussion on these assumptions can be found in [16].

The RTFs are *hidden* vectors, which represent the different speakers in a multichannel scenario. We would like to extract instantaneous observation vectors from the measured mixtures (1), and show their relation to the defined hidden vectors. We assume that low-energy frames do not contain speech components, and hence these frames are excluded from our analysis. We adopt the WDO assumption [5], stating that each TF bin is exclusively dominated by a single speaker. Accordingly, we assume that the (l,f)th TF bin is occupied by either of the speakers with probabilities $\{p_j(l)\}_{j=1}^J$, satisfying $\sum_{j=1}^J p_j(l) = 1$. Hence, we have:

$$Y^m(l,f) = Y_{\eta_l}^m(l,f) \tag{9}$$

where η_l has a categorical distribution with $\Pr(\eta_l = j) = p_j(l)$. We compute the following instantaneous ratio:

$$R^m(l,f) = \frac{Y^m(l,f)}{Y^1(l,f)} = \frac{Y^m_{\eta_l}(l,f)}{Y^1_{\eta_l}(l,f)} = H^m_{\eta_l}(l,f) \tag{10}$$

which due to the sparsity assumption, equals the RTF of one of the speakers. Based on the computed ratios, we define the observation vector $\mathbf{a}(l)$, which consists of the real and the imaginary parts of the ratio values, in K frequency bins and in $M-1$ microphones (Recall (7)):

$$\begin{aligned} \mathbf{a}^m(l) &= [R^m(l,f_1), R^m(l,f_2), \ldots, R^m(l,f_K)]^T \\ \mathbf{a}^{\mathrm{c}}(l) &= \left[\mathbf{a}^{2^T}(l), \mathbf{a}^{3^T}(l), \ldots, \mathbf{a}^{M^T}(l)\right]^T \\ \mathbf{a}(l) &= \left[\mathrm{real}\left\{\mathbf{a}^{\mathrm{c}}(l)\right\}^T, \mathrm{image}\left\{\mathbf{a}^{\mathrm{c}}(l)\right\}^T\right]^T . \end{aligned} \tag{11}$$

We compute (10) and (11) for each frame $1 \le l \le L$, and form the set $\{\mathbf{a}(l)\}_{l=1}^L$. Note that according to the defined model, the observation vectors $\{\mathbf{a}(l)\}_{l=1}^L$ consist of different portions of the RTF vectors $\{\mathbf{h}_j\}_{j=1}^J$. For each l, the number of entries in the vector $\mathbf{a}(l)$ corresponding to a particular RTF $\mathbf{h}_j$, is proportional to the probability $p_j(l)$ of the jth speaker. Consider for example a case where $J=2,\ p_1(l^*) = 0.7,\ p_2(l^*) = 0.3$ for a particular time frame l^*. In the vector $\mathbf{a}(l^*)$, approximately 70% of the entries are taken from the vector $\mathbf{h}_1$, and approximately 30% of the entries are taken from the vector $\mathbf{h}_2$.

Consider the set of probabilities associated with each frame l: $\mathbf{p}(l) = [p_1(l), p_2(l), \ldots, p_J(l)]^T$. The collection of all probability sets $\{\mathbf{p}(l)\}_{l=1}^L$ occupies the standard $(J-1)$-simplex in $\mathbb{R}^J$. 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. Note that recovering the probabilities associated with each frame reveals the activity patterns of the different speakers across time, which resolves the diarization problem.

B. Analysis of the Correlation Matrix

We show that a spectral decomposition of the correlation matrix between different time frames can be used to form a new representation, which corresponds to the simplex of the speakers' probabilities across time. Based on the assumed statistical model (8),(9),(10), the correlation between each two observations $\mathbf{a}(l)$ and $\mathbf{a}(n)$, $1 \le l,n \le L$ is given by (for details refer to [16]):

$$E\left\{\frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)\right\} = \left\{ \begin{array}{ll} \sum_{j=1}^J p_j(l)p_n(l) & l \ne n \\ 1 & l = n \end{array} \right. . \tag{12}$$

Let $\mathbf{W}$ be the $L \times L$ correlation matrix, with $W_{ln} = E\left\{\frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)\right\}$. According to (12) the correlation matrix can be recast as:

$$\mathbf{W} = \mathbf{P}\mathbf{P}^T + \Delta\mathbf{W} \tag{13}$$

where $\mathbf{P}$ is a $L \times J$ matrix with $P_{lj} = p_j(l)$, and $\Delta\mathbf{W}$ is a diagonal matrix with $\Delta W_{ll} = 1 - \sum_{j=1}^J p_j^2(l)$. We show in [16], that $\Delta\mathbf{W}$ has a negligible effect on the spectral decomposition of $\mathbf{W}$. Therefore, henceforth we omit $\Delta\mathbf{W}$ from our derivations, and assume that the correlation matrix is given by $\mathbf{W} \approx \mathbf{P}\mathbf{P}^T$. Accordingly, the rank of the correlation matrix $\mathbf{W}$ is identical to the rank of the matrix $\mathbf{P}$, which equals the overall number of speakers J (due to their independence assumption).

We apply an eigenvalue decomposition (EVD) $\mathbf{W} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, where $\mathbf{U}$ is an orthonormal matrix consisting of the eigenvectors $\{\mathbf{u}_j\}_{j=1}^L$, and $\mathbf{D}$ is a diagonal matrix with the eigenvalues $\{\lambda_j\}_{j=1}^L$ on its diagonal. The eigenvalues $\{\lambda_j\}_{j=1}^L$ are sorted in a descending order. According to (13), the first J eigenvectors $\{\mathbf{u}_j\}_{j=1}^J$, form a basis for the column space of the matrix $\mathbf{P}$, leading to the following relation:

$$\mathbf{U}_{\mathrm{J}} = \mathbf{P}\mathbf{Q}^T \tag{14}$$

where $\mathbf{U}_{\mathrm{J}} = [\mathbf{u}_1, \ldots, \mathbf{u}_J]$, and $\mathbf{Q}$ is a $J \times J$ invertible matrix. Regarding the computation of the matrix $\mathbf{W}$, we substitute the unavailable expected values by their typical values $\widehat{W}_{ln} = \frac{1}{D}\mathbf{a}^T(l)\mathbf{a}(n)$. This approximation is justified in [16].

We represent each frame using the obtained set of eigenvectors by: $\boldsymbol{\nu}(l) = [u_1(l), u_2(l), \ldots, u_J(l)]^T$. According to (14), this representation is obtained as a linear transformation of the set of the speakers' probabilities:

$$\boldsymbol{\nu}(l) = \mathbf{Q}\mathbf{p}(l). \tag{15}$$

Hence, the collection $\{\boldsymbol{\nu}(l)\}_{l=1}^L$ occupies a simplex, which is a rotated and scaled version of the standard probability simplex occupied by $\{\mathbf{p}(l)\}_{l=1}^L$. Note that the columns of $\mathbf{Q}$ consist of the transformed simplex vertices:

$$\mathbf{Q}\mathbf{e}_j = \mathbf{Q}_j \tag{16}$$

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

C. Speaker Counting and Diarization

We recover the overall number of speakers and perform diarization based on the computed spectral decomposition of the correlation matrix $\mathbf{W}$. Recall that the rank of $\mathbf{W}$ equals J, implying that it has exactly J non-zero eigenvalues. Consequently, the overall number of speakers can be estimated by the following thresholding rule on the normalized eigenvalues:

$$\hat{J} = \left(\underset{j}{\operatorname{argmin}} \frac{\lambda_j}{\lambda_1} < \alpha\right) - 1 \tag{17}$$

where α is a threshold parameter.

Let l_j denote an index of a frame consisting of only the jth speaker, i.e: $\mathbf{p}(l_j) \approx \mathbf{e}_j$. To perform dirazation, we first identify the vertices of the simplex with indices $\{l_j\}_{j=1}^J$. For this purpose we use a *successive projection* algorithm [17]. We first identify two vertices of the simplex, and then successively identify the remaining vertices by maximizing the projection onto the orthogonal complement of the space spanned by the previously identified vertices.

Based on (16), we construct the matrix $\hat{\mathbf{Q}}$ using the identified set of vertices $\hat{\mathbf{Q}} = [\boldsymbol{\nu}(l_1), \ldots, \boldsymbol{\nu}(l_J)]^T$. We utilize $\hat{\mathbf{Q}}$ to

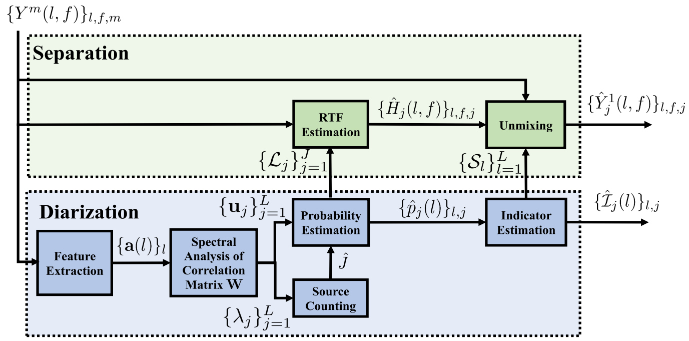

Fig. 1. Block diagram of proposed method.

map the obtained representation $\boldsymbol{\nu}(l)$ to the associated set of probabilities $\mathbf{p}(l)$ by (Recall (15)):

$$\hat{\mathbf{p}}(l) = \hat{\mathbf{Q}}^{-1}\boldsymbol{\nu}(l) \tag{18}$$

The values of the indicator functions of the different speakers are determined by a hard-thresholding over the estimated probabilities:

$$\hat{\mathcal{I}}_j(l) = \begin{cases} 1 & \hat{p}_j(l) > \beta \\ 0 & \hat{p}_j(l) < \beta \end{cases} \tag{19}$$

where β is a threshold probability. Accordingly, the instantaneous number of active speakers is given by: $\hat{J}_l = \sum_{j=1}^{J} \hat{\mathcal{I}}_j(l)$.

In addition, we exploit the estimated probabilities to form the sets of frame indices $\{\mathcal{L}_j\}_{j=1}^J$ dominated by each of the speakers. We define the set $\mathcal{L}_j$ by:

$$\mathcal{L}_j = \{l \,|\, p_j(l) > \gamma,\, l \in \{1,\ldots,L\}\} \tag{20}$$

where γ is a threshold probability, larger than β.

Based on the identified sets of frames $\{\mathcal{L}_j\}_{j=1}^J$ and the estimated indicator functions $\left\{\hat{\mathcal{I}}_j(l)\right\}_{l,j}$, a separation is carried out applying the unmixing scheme presented in Section III. A flow diagram of the proposed algorithm is presented in Fig. 1.

V. Experimental Study

We examine the performance of the proposed method in a challenging underdetermined scenario with $J = 6$ speakers and $M = 4$ microphones. We assume that the speakers are divided into two groups of three speakers each, where each group is holding a separated conversation. Within each group only one member is speaking at a time. The two conversations are held simultaneously, hence at each time instance there are at most two concurrent speakers ($J_l \leq 2$). The overall conversation of each group lasts 24s, with two sentences of 4s for each speaker. The sentences, sampled at 16kHz, are drawn from the TIMIT database, and are normalized to a fixed energy level (input signal to interference ratio (SIR) is approximately 0dB). The speakers and the microphones are positioned in a 6m×6m×2.4m room, as illustrated in Fig. 2. The corresponding acoustic impulse responses are drawn from the database presented in [18]. The four microphones (out of the eight available channels in the database) are organized in a uniform linear array with 8cm inter-microphone spacing.

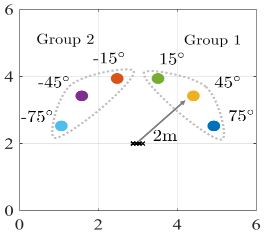

Fig. 2. Room setup.

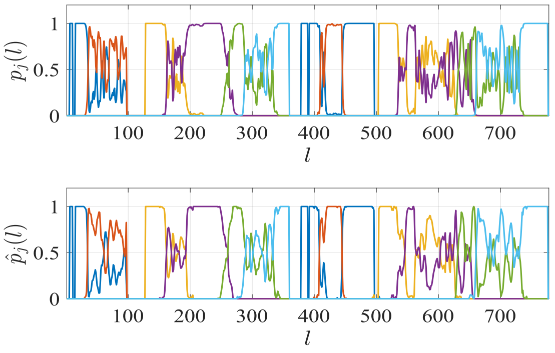

Fig. 3. True and estimated probabilities of the different speakers. The members of the first group are colored by: blue, orange and green, and the members of the second group are colored by: red, purple and light blue.

The signals are analysed in the STFT domain using windows of length 2048 samples and with 75% overlap, leading to a total amount of $L = 778$ frames. The ratios computed by (10) are averaged over 5 successive frames. The RTF vector in (11) consists of $K = 576$ frequency bins, corresponding to $0 - 4.5$kHz frequency band, in which most of the speech content is concentrated, and is normalized to have a unit norm. The estimated probabilities of each speaker are zeroed in regions where low activity is detected, and are normalized such that their sum is one. The threshold parameters are set to $\alpha = 0.11$, $\beta = 0.08$ and $\gamma = 0.95$, which empirically yield good and stable results.

We first demonstrate the diarization performance, when the reverberation time is 360ms. Figure 3 depicts the true probabilities (top plot), computed using the individual speakers, and the estimated probabilities (bottom plot), computed by (18), as a function of the frame index. We observe that the estimated probabilities represent similar trends to the true probabilities, even in time segments with two simultaneous speakers. For the same case, Fig. 4 presents the true and the estimated indicators (19) with respect to the time-domain waveform measured by the reference microphone. We see that the proposed algorithm successfully recovers the activities of the speakers across time, and performs almost perfect diarization.

The diarization and the separation performance are further evaluated for different reverberation levels of 160ms, 360ms

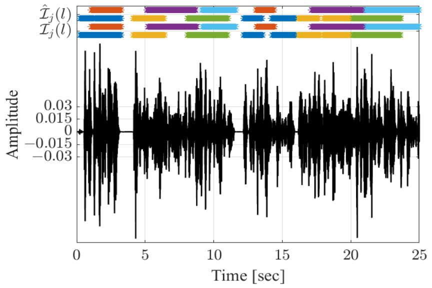

Fig. 4. Waveform of reference microphone. The true and the estimated indicators are illustrated by asterisks in compatible colors (same as in Fig. 3), denoting active time instances of each speaker.

TABLE I
DIARIZATION ACCURACY AND AVERAGE SIR AND SDR MEASURES.

	Diar. Acc.	SIR		SDR	
		Oracle	Proposed	Oracle	Proposed
160ms	99.1	26.9	24.3	14.8	12.4
360ms	99.1	26.2	23.5	10.5	9.4
610ms	98.8	24.8	20.1	6.5	5.7

and 610ms. We compare the proposed method to an oracle separator based on (3), in which the RTFs and the indicator values are computed using the individually measured signals. Diarization accuracy is assessed by the percentage of correctly estimated indictor functions (out of $L \times J$). The separation performance is reported in terms of SIR and signal to distortion ratio (SDR) measures, evaluated by the BSS-Eval toolbox [19].

The measures are averaged over 20 trials with different combinations of speakers. The overall number of speakers ($J = 6$) was perfectly recovered in all trials using (17). Diarization and separation measures are summarized in Table I. We observe that the proposed method achieves high diarization accuracy of more than 98% for all reverberation levels. Regarding the separation performance, the difference between our method and the oracle is attributed to RTF estimation inaccuracies, which are estimated using only a few frames with possibly low components of other speakers. Nevertheless, using a completely blind approach the proposed method achieves high separation scores with average 22.6dB SIR and 9.2dB SDR.

VI. CONCLUSIONS

We have presented a method for combined diarization and separation of underdetermined mixtures with unknown number of speakers. The diarization stage is based on spectral decomposition of the correlation matrix between different time frames. The spectrum decay reveals the overall number of speakers, and the computed eigenvectors form a simplex that facilitates the estimation of probabilities of speakers. The separation stage exploits the diarization results. Frames, which are highly dominated by a single speaker, are utilized for estimating the corresponding RTFs, and an unmixing scheme is carried out to extract the individual speakers. The performance of the proposed method is demonstrated in a challenging scenario with six speakers and only four microphones in various reverberation conditions.

REFERENCES

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[2] I. Lee, T. Kim, and T.-W. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Processing*, vol. 87, no. 8, pp. 1859–1871, Aug. 2007.

[3] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural computation*, vol. 13, no. 4, pp. 863–882, Apr. 2001.

[4] 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, Mar. 2010.

[5] 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, Jul. 2004.

[6] 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, Feb. 2010.

[7] 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, Mar. 2011.

[8] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1557–1565, Sep. 2006.

[9] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[10] N. Evans, S. Bozonnet, D. Wang, C. Fredouille, and R. Troncy, "A comparative study of bottom-up and top-down approaches to speaker diarization," *IEEE Transactions on Audio, speech, and language processing*, vol. 20, no. 2, pp. 382–392, Feb. 2012.

[11] T. Higuchi, H. Takeda, T. Nakamura, and H. Kameoka, "A unified approach for underdetermined blind signal separation and source activity detection by multichannel factorial hidden markov models," in *Proc. of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore, Sep. 2014.

[12] T. Higuchi and H. Kameoka, "Unified approach for audio source separation with multichannel factorial HMM and DOA mixture model," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Nice, France, Aug. 2015, pp. 2043–2047.

[13] W. B. Kleijn and F. Lim, "Robust and low-complexity blind source separation for meeting rooms," in *Proc. of International Workshop on Hands-free Speech Communications and Microphone Arrays (HSCMA)*, San Francisco, CA, USA, Mar. 2017, pp. 156–160.

[14] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "An EM algorithm for joint source separation and diarisation of multichannel convolutive speech mixtures," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, Mar. 2017.

[15] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, R. Horaud, and S. Gannot, "Exploiting the intermittency of speech for joint separation and diarization," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2017.

[16] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Data-driven source separation based on simplex analysis," *pre-print*, Feb. 2018, arXiv:1802.03148v1.

[17] 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, Jul. 2001.

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

[19] 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, Jul. 2006.