AN EM ALGORITHM FOR JOINT SOURCE SEPARATION AND DIARISATION OF MULTICHANNEL CONVOLUTIVE SPEECH MIXTURES

Dionyssos Kounades-Bastian[1], *Laurent Girin*[1,2], *Xavier Alameda-Pineda*[3], *Sharon Gannot*[4], *Radu Horaud*[1]

[1]INRIA Grenoble Rhône-Alpes, [2]GIPSA-Lab, Univ. Grenoble Alpes
[3]University of Trento, [4]Faculty of Engineering, Bar-Ilan University

ABSTRACT

We present a statistical model for joint source separation and diarization of multichannel convolutive speech mixtures. We build upon the framework of local Gaussian model (LGM) with non-negative matrix factorization (NMF). The diarization is introduced as a temporal labeling of each source in the mix as active or inactive at the short-term frame level. We devise an EM algorithm in which the source separation process is aided by the diarization state, since the latter indicates the sources actually present in the mixture. The diarization state is tracked with a Hidden Markov Model (HMM) with emission probabilities calculated from the estimated source signals. The proposed EM has separation performance comparable with a state-of-the-art LGM NMF method, while outperforming a state-of-the-art speaker diarization pipeline.

Index Terms— Audio source separation, speaker diarization, local Gaussian model.

1. INTRODUCTION

Multichannel audio source separation (MASS) aims at recovering unobserved source signals from observed mixtures [1]. MASS is mainly concerned with mixtures of speech, music, and ambient noise. Speaker diarization is the segmentation and labelling of an audio signal emitted during multi-party conversations [2, 3]. In short, speaker diarization is answering to the question "who is talking, and when?" while MASS answers "what is the emitted signal?". Both processes are crucial front-ends for higher-level processes such as speech recognition, human-computer or human-robot interaction.

There has been extensive research addressing independently either MASS [4, 5, 6, 7], or speaker diarization [2, 3] tasks. Currently, most of MASS methods implicitly assume all sources as continuously emitting. Besides, state-of-the-art methods on diarization [8] consist of a pipeline starting with feature extraction from the audio mixture, e.g. Mel frequency cepstral coefficients (MFCC) or spatial parameters, and proceed with speech/non-speech segmentation of the mixture and clustering of the speech segments into individual speakers.

This research has received funding from the EU-FP7 STREP project EARS (#609465) and ERC Advanced Grant VHIA (#340113).

Obviously, both problems are highly inter-related. Indeed, knowing the separated sources of an audio mixture helps assessing when each source is active/inactive. Reciprocally, knowing the diarization of the sources within the mixture determines how many sources need to be separated and when. Hence a joint formulation of MASS and diarization can increase the performance on both sides. Except for a series of papers by Higuchi and colleagues, a framework addressing jointly these two problems seems overlooked in the literature; In [9, 10] the active/inactive state of each source is independently modeled by a (factorial) HMM within a MASS framework. Considering the state of a source independently to the state of the other sources may be unrealistic for multi-party conversations. In [11] the source activity detection is combined with a Direction-of-Arrival-dependent HMM for the propagation model. A variational Expectation Maximization (EM) is proposed to estimate the model parameters, but this model is limited by the assumption of a single active source per time-frequency bin.

In the present paper we propose a probabilistic model for simultaneous diarization and MASS for multichannel audio mixtures that does not have these limitations. We consider all possible combinations of simultaneous active sources, and process the activity state of all sources in a joint manner. An EM algorithm is designed for model parameter estimation. We benchmark the performance of the proposed model on mixtures of speech, with [5] in terms of separation, and with [8] in terms of diarization.

The proposed model is presented in Section 2. The EM algorithm that estimates the model parameters, and infers the source signals and the diarization, is described in Section 3. Experimental evaluation reported in Section 4 shows competitive performance on source separation and diarization.

2. MODELS

2.1. Audio Mixtures with diarization

As in many source separation methods, the observed signal is modeled as a multichannel time-invariant convolutive noisy mixture of the source signals. We work with the short time Fourier transform (STFT) representation of the input audio, where $\mathbf{x}_{f\ell} = [x_{1,f\ell} \dots x_{I,f\ell}]^\top \in \mathbb{C}^I$ is the I-channel vec-

tor of coefficients at frequency bin $f \in [1, F]$ and time frame $\ell \in [1, L]$. Relying on the narrow-band assumption (i.e. the channel impulse response is shorter than the STFT analysis window) $\mathbf{x}_{f\ell}$ writes [12]: $\mathbf{x}_{f\ell} = \mathbf{A}_f \mathbf{s}_{f\ell} + \mathbf{b}_{f\ell}$, where $\mathbf{s}_{f\ell} = [s_{1,f\ell} \dots s_{J,f\ell}]^\top \in \mathbb{C}^J$ is the latent vector of source coefficients, $\mathbf{A}_f \in \mathbb{C}^{I \times J}$ is the mixing matrix, and $\mathbf{b}_{f\ell} = [b_{1,f\ell} \dots b_{I,f\ell}]^\top \in \mathbb{C}^I$ is a residual noise. In the present study we want to express $\mathbf{x}_{f\ell}$ with a formulation that explicitly encodes the activity of the sources, in a binary way. We have $N = 2^J$ possible combinations, or "states", for source activity. Let us represent a state $n \in [1, N]$ with a diagonal $[J \times J]$ matrix $\mathbf{D}_n$ with j^{th} entry $D_{jj,n}$ set to:

$$\begin{aligned} &D_{jj,n} = 1 \text{ if the } j^{\text{th}} \text{ source is active in state } n, \\ &D_{jj,n} = 0 \text{ otherwise.} \end{aligned} \tag{1}$$

For example, with $J = 2$, the $N = 4$ possible matrices are:

$$\mathbf{D}_1 = \begin{bmatrix} 0 & \\ & 0 \end{bmatrix}, \mathbf{D}_2 = \begin{bmatrix} 0 & \\ & 1 \end{bmatrix}, \mathbf{D}_3 = \begin{bmatrix} 1 & \\ & 0 \end{bmatrix}, \mathbf{D}_4 = \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}. \tag{2}$$

Incorporating $\mathbf{D}_n$ we rewrite $\mathbf{x}_{f\ell}$ as:

$$\mathbf{x}_{f\ell} = \sum_{j=1}^{J} (D_{jj,n} s_{j,f\ell}) \mathbf{a}_{j,f} + \mathbf{b}_{f\ell} = \mathbf{A}_f \mathbf{D}_n \mathbf{s}_{f\ell} + \mathbf{b}_{f\ell}, \tag{3}$$

with $\mathbf{a}_{j,f} \in \mathbb{C}^I$ the j^{th} column of $\mathbf{A}_f$. By choosing a state n at a given time frame ℓ, we actually select which of the J sources compose the mixture at that time frame. In other words $\mathbf{D}_n$ zeroes out the inactive sources.

2.2. Selecting the diarization

The activity of each sound source varies over time, hence the state is to be estimated for each frame ℓ. To do that we define a latent categorical variable $Z_\ell = n, n \in [1, N]$, if state n is selected at frame ℓ. Z_ℓ follows a first-order HMM with:

$$p(Z_1 = n) = \lambda_n, \quad p(Z_\ell = n | Z_{\ell-1} = r) = T_{nr}, \tag{4}$$

with $\lambda_n, T_{nr} \in \mathbb{R}_+, n, r \in [1, N]$ being the prior and transition parameters to be estimated. Let us assume that $\mathbf{b}_{f\ell}$ follows a zero-mean proper complex-Gaussian distribution[1]. We can now express the mixture of (3) probabilistically:

$$p(\mathbf{x}_{f\ell} | \mathbf{s}_{f\ell}, Z_\ell = n) = \mathcal{N}_c\left(\mathbf{x}_{f\ell}; \mathbf{A}_f \mathbf{D}_n \mathbf{s}_{f\ell}, \mathrm{v}_f \mathbf{I}_I\right), \tag{5}$$

where $\mathbf{A}_f$ and $\mathrm{v}_f \in \mathbb{R}_+$ are the parameters to be estimated and $\mathbf{I}_I$ is the identity matrix of dimension I.

2.3. The Source Model

Let $\{\mathcal{K}_j\}_{j=1}^J$ denote a non-trivial partition of $\{1 \dots K\}$, with $K \geq J$ the number of *latent components* that is known in advance. Following [5, 14, 15], $s_{j,f\ell}$ is modeled as the sum of the latent components $c_{k,f\ell}$, $k \in \mathcal{K}_j$:

$$s_{j,f\ell} = \sum_{k \in \mathcal{K}_j} c_{k,f\ell} \Leftrightarrow \mathbf{s}_{f\ell} = \mathbf{G}\mathbf{c}_{f\ell}, \tag{6}$$

where $\mathbf{G} \in \mathbb{N}^{J \times K}$ is a matrix with binary entries $G_{jk} = 1$ if $k \in \mathcal{K}_j$ and $G_{jk} = 0$ otherwise, and $\mathbf{c}_{f\ell} = [c_{1,f\ell}, \dots, c_{K,f\ell}]^\top \in \mathbb{C}^K$ is the vector of component coefficients. Each component $c_{k,f\ell}$ is assumed to follow a zero-mean proper complex Gaussian distribution with variance $w_{fk} h_{k\ell}$, where $w_{fk}, h_{k\ell} \in \mathbb{R}_+$. The components are assumed to be mutually independent and individually independent across frequency and time. Thus the component vector probability density function (pdf) writes:

$$p(\mathbf{c}_{f\ell}) = \mathcal{N}_c\Big(\mathbf{c}_{f\ell}; \mathbf{0}, \mathrm{diag}_K\left(w_{fk} h_{k\ell}\right)\Big), \tag{7}$$

where $\mathbf{0}$ denotes the zero-vector, $\mathrm{diag}_K(d_k)$ denotes the $K \times K$ diagonal matrix with entries $[d_1 \dots d_k \dots d_K]^\top$. Eq. (7) corresponds to a non-negative matrix factorization (NMF) model for the $F \times L$ source power spectral density (PSD) matrix, which is now common practice in audio signal processing, e.g. [16, 17, 18].

3. EM FOR JOINT MASS & DIARIZATION

We derived an EM algorithm [19] to infer the hidden variables $\mathcal{H} = \{\mathbf{c}_{f\ell}, Z_\ell\}_{f,\ell=1}^{F,L}$ and estimate the parameters $\theta = \{\mathbf{A}_f, \mathrm{v}_f, T_{nr}, \lambda_n, w_{fk}, h_{k\ell}\}_{f,\ell,k,n,r=1}^{F,L,K,N,N}$. The E-step and M-step are given below. The complete EM is given in Algorithm1. The iteration index is omitted for clarity.

3.1. E-step

The E-step consists of computing the joint posterior probability: $p(\mathbf{c}_{f\ell}, Z_\ell = n | \{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L})$. This is done by first computing the posterior over the component coefficients: $p(\mathbf{c}_{f\ell} | Z_\ell = n, \{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L})$ and then computing the posterior over the states: $\eta_{\ell n} = p(Z_\ell = n | \{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L})$. After a few manipulations, it turns out that for every state $n \in [1, N]$ we obtain a different complex Gaussian pdf for the components:

$$\begin{aligned} p\left(\mathbf{c}_{f\ell} | Z_\ell = n, \{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}\right) &\propto p(\mathbf{c}_{f\ell}) p(\mathbf{x}_{f\ell} | \mathbf{c}_{f\ell}, Z_\ell = n) \\ &= \mathcal{N}_c\left(\mathbf{c}_{f\ell}; \hat{\mathbf{c}}_{f\ell n}, \mathbf{\Sigma}^c_{f\ell n}\right), \end{aligned} \tag{8}$$

with $\hat{\mathbf{c}}_{f\ell n} \in \mathbb{C}^K$ and $\mathbf{\Sigma}^c_{f\ell n} \in \mathbb{C}^{K \times K}$ given by:

$$\begin{aligned} \mathbf{\Sigma}^c_{f\ell n} &= \left[\mathrm{diag}_K\left(\frac{1}{w_{fk} h_{k\ell}}\right) + \mathbf{G}^\top \mathbf{D}_n \frac{\mathbf{A}_f^{\mathrm{H}} \mathbf{A}_f}{\mathrm{v}_f} \mathbf{D}_n \mathbf{G}\right]^{-1}, \\ \hat{\mathbf{c}}_{f\ell n} &= \mathbf{\Sigma}^c_{f\ell n} \mathbf{G}^\top \mathbf{D}_n \mathbf{A}_f^{\mathrm{H}} \frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f}. \end{aligned} \tag{9}$$

[1] The proper complex Gaussian distribution is defined as $\mathcal{N}_c(\mathbf{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) = |\pi\mathbf{\Sigma}|^{-1} \exp\left(-[\mathbf{x} - \boldsymbol{\mu}]^{\mathrm{H}} \mathbf{\Sigma}^{-1} [\mathbf{x} - \boldsymbol{\mu}]\right)$, with $\mathbf{x}, \boldsymbol{\mu} \in \mathbb{C}^I$ and $\mathbf{\Sigma} \in \mathbb{C}^{I \times I}$ being the argument, mean vector, and covariance matrix respectively [13].

From (6), we easily deduce that the source posterior distribution is a complex Gaussian with mean vector $\hat{\mathbf{s}}_{f\ell n} \in \mathbb{C}^J$, and covariance matrix $\mathbf{\Sigma}^s_{f\ell n} \in \mathbb{C}^{J\times J}$ calculated with:

$$\hat{\mathbf{s}}_{f\ell n} = \mathbf{G}\hat{\mathbf{c}}_{f\ell n}, \quad \mathbf{\Sigma}^s_{f\ell n} = \mathbf{G}\mathbf{\Sigma}^c_{f\ell n}\mathbf{G}^\top. \tag{10}$$

From (4), the posterior probability $\eta_{\ell n}$ of the n^{th} state is calculated by decoding a first-order HMM over the diarization states $n \in [1, N]$, with emission probabilities $\iota_{\ell n} = p(\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}|Z_\ell = n)$ given by:

$$\iota_{\ell n} \propto \exp\left(\sum_{f=1}^{F}\left[\log\left|\mathbf{\Sigma}^s_{f\ell n}\right| + \frac{\mathbf{x}_{f\ell}{}^{\mathrm{H}}}{\mathrm{v}_f}\mathbf{A}_f\mathbf{D}_n\hat{\mathbf{s}}_{f\ell n}\right]\right), \tag{11}$$

where $|.|$ denotes the determinant function. The decoding is done using the forward-backward algorithm [19]. The forward and the backward recursions respectively compute the $\phi_{\ell n}$ and $\beta_{\ell n}$ probabilities:

$$\phi_{\ell n} \propto \iota_{\ell n}\sum_{r=1}^{N} T_{nr}\phi_{(\ell-1)r}, \quad \beta_{\ell n} \propto \sum_{r=1}^{N} T_{rn}\iota_{(\ell+1)r}\beta_{(\ell+1)r}, \tag{12}$$

and we finally combine $\phi_{\ell n}$ and $\beta_{\ell n}$ to obtain:

$$\eta_{\ell n} \propto \phi_{\ell n}\beta_{\ell n}. \tag{13}$$

The forward-backward algorithm requires initialization of ϕ_{1n} and β_{Ln}. We observed faster convergence by, at each EM iteration, setting $\phi_{1n} = \iota_{1n}\lambda_n$, running the forward recursion, and then setting $\beta_{Ln} = \phi_{Ln}$ to initialize the backward recursion.

3.2. M-step

In the M step, the parameters maximizing the expected complete-data log-likelihood are computed.

M-$w_{fk}, h_{k\ell}$ step: The parameters w_{fk} and $h_{k\ell}$ are coupled in the objective function and an alternation strategy is required, i.e. fixing one parameter to estimate the other. Similar to [5, 14] the updates for w_{fk} and $h_{k\ell}$ are:

$$w_{fk} = \frac{1}{L}\sum_{\ell=1}^{L}\frac{u_{k,f\ell}}{h_{k\ell}}, \quad h_{k\ell} = \frac{1}{F}\sum_{f=1}^{F}\frac{u_{k,f\ell}}{w_{fk}}, \tag{14}$$

with $u_{k,f\ell} \in \mathbb{R}_+$ being here the posterior PSD of $c_{k,f\ell}$ averaged with the posterior probability of the states:

$$u_{k,f\ell} = \sum_{n=1}^{N}\eta_{\ell n}\left(\Sigma^c_{kk,f\ell n} + |\hat{c}_{k,f\ell n}|^2\right), \tag{15}$$

with $\Sigma^c_{kk,f\ell n} \in \mathbb{R}_+$ being the k^{th} diagonal entry of $\mathbf{\Sigma}^c_{f\ell n}$ and $\hat{c}_{k,f\ell n} \in \mathbb{C}$ being the k^{th} entry of $\hat{\mathbf{c}}_{f\ell n}$.

Algorithm 1 Separation & diarization of J sound sources

input $\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}$, binary matrix $\mathbf{G}$, initial parameters θ.
construct: The 2^J matrices $\mathbf{D}_n, n \in [1, 2^J]$ with (2).
repeat
  E step C: Compute $\mathbf{\Sigma}^c_{f\ell n}$ and $\hat{\mathbf{c}}_{f\ell n}$ with (9).
  E step S: Compute $\mathbf{\Sigma}^s_{f\ell n}, \hat{\mathbf{s}}_{f\ell n}$ with (10).
  E step Z: Compute $\iota_{\ell n}$ with (11), set $\phi_{1n} = \lambda_n\iota_{1n}$,
    for ℓ : 2 to L. Compute $\phi_{\ell n}$ with (12). **end**.
    Set $\beta_{Ln} = \phi_{Ln}$.
    for ℓ : $L-1$ to 1. Compute $\beta_{\ell n}$ with (12). **end**.
    Compute $\eta_{\ell n}$ with (13).
  M-$w_{fk}, h_{k\ell}$ step: Update w_{fk}, and then $h_{k\ell}$, with (14).
  M-T_{nr}, λ_n step: Update $\xi_{\ell,nr}$, then T_{nr}, λ_n, with (16).
  M-$\mathbf{A}_f$ step: Compute $\mathbf{o}_{f\ell}, \mathbf{R}_{f\ell}$ with (17), $\mathbf{A}_f$ with (18).
  M-v_f step: Update v_f with (19).
until convergence
return The source images, the diarization $\mathbf{D}_{\hat{n}_\ell}\forall\ell$.

M-T_{nr}, λ_n step: Classically, for the HMM parameters, we calculate $\xi_{\ell,nr} = p(Z_\ell = n, Z_{\ell-1} = r|\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L})$ and then update T_{nr} and λ_n with:

$$\xi_{\ell,nr} \propto \beta_{(\ell+1)n}\iota_{(\ell+1)n}T_{nr}\phi_{\ell r},$$
$$T_{nr} \propto \sum_{\ell=1}^{L-1}\xi_{\ell,nr}, \quad \lambda_n = \eta_{1n}. \tag{16}$$

M-$\mathbf{A}_f$ step: As the mixing matrix is common for all diarization states, we need to first define the "final" source estimate $\mathbf{o}_{f\ell} \in \mathbb{C}^J$, i.e. average source estimate over the states, and the corresponding second-order statistic $\mathbf{R}_{f\ell} \in \mathbb{C}^{J\times J}$:

$$\mathbf{o}_{f\ell} = \sum_{n=1}^{N}\eta_{\ell n}\mathbf{D}_n\hat{\mathbf{s}}_{f\ell n},$$
$$\mathbf{R}_{f\ell} = \sum_{n=1}^{N}\eta_{\ell n}\mathbf{D}_n\left(\mathbf{\Sigma}^s_{f\ell n} + \hat{\mathbf{s}}_{f\ell n}\hat{\mathbf{s}}^{\mathrm{H}}_{f\ell n}\right)\mathbf{D}_n. \tag{17}$$

Then the optimal value for $\mathbf{A}_f$ is:

$$\mathbf{A}_f = \left(\sum_{\ell=1}^{L}\mathbf{x}_{f\ell}\mathbf{o}^{\mathrm{H}}_{f\ell}\right)\left(\sum_{\ell=1}^{L}\mathbf{R}_{f\ell}\right)^{-1}, \tag{18}$$

which is a standard form of least square estimator [5, 7].

M-v_f step: The optimal noise variance is:

$$\mathrm{v}_f = \frac{1}{LI}\sum_{\ell=1}^{L}\Big(\mathbf{x}^{\mathrm{H}}_{f\ell}\mathbf{x}_{f\ell} - 2\mathfrak{Re}\left\{\mathbf{x}^{\mathrm{H}}_{f\ell}\mathbf{A}_f\mathbf{o}_{f\ell}\right\} + \mathrm{tr}\left\{\mathbf{A}_f\mathbf{R}_{f\ell}\mathbf{A}^{\mathrm{H}}_f\right\}\Big). \tag{19}$$

where $\mathrm{tr}\{.\}$ denotes the trace and $\mathfrak{Re}\{.\}$ denotes the real part.

3.3. Estimation of source images & diarization

The separation performance is assessed with the time domain source *images*, i.e. the multichannel source signals at all microphones [6, 20], estimated by applying the inverse STFT with overlap-add on $\{o_{j,f\ell}\mathbf{a}_{j,f}\}_{f,\ell=1}^{F,L}$. The diarization output $\hat{n}_\ell$ is given at each frame by selecting the higher value of $\eta_{\ell n}$, $n \in [1, N]$. Then from $\mathbf{D}_{\hat{n}_\ell}$ we have the active sources at ℓ^{th} frame. Frames where $\eta_{\ell 1}$ is dominant are non-speech frames.

4. EVALUATION

To assess the performance of the proposed method, we simulated the challenging task of separating and diarizing $J = 3$ sources from a convolutive stereo mixture ($I = 2$). Each source signal was a 27-s speech signal made by concatenating utterances chosen from the TIMIT database [21] (one different speaker for each source). As mixing filters, we used binaural room impulse responses (BRIRs) from [22] having $\text{RT}_{60} \approx 0.68$s. We generated two types of mixtures: *Mix-DC* where all sources are emitting continuously. *Mix-8* where each source has balanced portions of speech and silence so that all $N = 8$ states appear.

MASS performance is assessed with the signal-to-distortion (SDR), signal-to-interference (SIR), and signal-to-artefact (SAR) measures (in dB) [23]. Diarization is assessed with Accuracy, which is defined as the percentage of frames for which a source was correctly identified (as either active if actually active, or inactive if actually inactive). As a baseline, we used [5] for source separation and [8] for speaker diarization. Both baselines were provided with the correct number of sources ($J = 3$). Because [8] is designed for non-overlapping audio streams, we considered each of the 2^J source combinations as a virtual speaker, and we translated the result of the clustering over virtual speakers into clustering of individual sources (we tested all possible associations and reported the one giving the highest accuracy). Afterwards, we use a median filter on the estimated label of each source to remove any potential spiking behavior.

The initialization of the parameters is crucial for EM-based methods. To be fair, we initialized the parameters $w_{fk}, h_{k\ell}$ of both our EM and the MASS baseline [5], by applying the KL-NMF algorithm [17] with $|\mathcal{K}_j| = 20$ components per source, on corrupted versions of the true source spectra. All other parameters were initialized randomly. For the STFT analysis we used a sine analysis window with 512 taps and 50% frame overlap, leading to $L = 1697$ frames.

In Table 1 we report detailed MASS and diarization scores. Each value is an average measure over 10 mixture realizations with different speakers. Fig. 1 illustrates one diarization result for *Mix-8*. In terms of MASS, we see that the proposed method performs equally well with [5] on both *Mix-8* and *Mix-DC*. E.g. on *Mix-8* the avg. SDR of the proposed method is 0.2dB higher (8.3dB versus 8.1dB). This

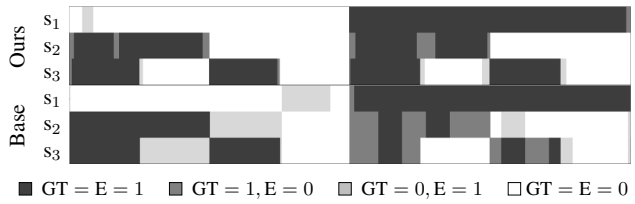

Fig. 1. Timing diagram of per source estimated diarization for *Mix-8* when using the proposed method (top) and the baseline (bottom). The graphic is color-coded as a function of the ground-truth (GT) and the estimate (E) per each segment.

Table 1. Average source separation & diarization scores.

		Mix-8				*Mix-DC*			
		SDR	SIR	SAR	Acc.(%)	SDR	SIR	SAR	Acc.(%)
Ours	s_1	7.7	11.6	12.1	93.5	7.8	12.4	12.2	99.5
	s_2	7.9	14.9	16.6	94.3	7.3	14.0	15.1	93.2
	s_3	9.2	13.4	14.1	87.5	8.9	13.3	14.0	99.3
	avg.	8.3	13.3	14.3	91.7	8.0	13.3	13.7	97.3
Base	s_1	7.6	12.6	12.4	89.0	7.7	12.6	12.7	87.8
	s_2	7.6	13.5	15.9	68.4	7.3	13.1	16.0	82.2
	s_3	9.0	13.1	14.8	67.4	8.8	13.0	14.8	61.8
	avg.	8.1	13.1	14.4	74.9	7.9	12.9	14.5	77.3

is encouraging considering that the proposed method has to learn the additional parameters to solve for diarization. As for the performance in diarization, the Accuracy for the proposed method is 16.8% higher than the baseline [8] on *Mix-8* (91.7% versus 74.9%) and 20.0% on *Mix-DC* (97.3% versus 77.3%). This justifies experimentally the joint modeling of the source activity detection and the source signal extraction. From a qualitative perspective, we see from Fig. 1 that the activity pattern is adequately detected with only a few misdetections.

5. CONCLUSION

In this paper we introduced a LGM-based probabilistic framework for joint MASS and diarization of audio sources in a compact formulation. Experiments on underdetermined mixtures of speech showed competitive performance of the proposed method compared to the state-of-the-art, in particular in diarization scores. Future research will investigate properties that can emerge from this model as is the automatic determination of the number of sources J using $\mathbf{D}_{\hat{n}_\ell}$. Importantly, we will explore realistic initialization techniques, allowing to create a fully blind joint MASS and diarization method. We will also explore the joint use of source localization cues to simultaneously improve separation and diarization.

6. REFERENCES

[1] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation - Independent Component Analysis and Applications*. Academic Press, 2010.

[2] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE TASLP*, vol. 14, no. 5, pp. 1557–1565, 2006.

[3] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE TASLP*, vol. 20, no. 2, pp. 356–371, 2012.

[4] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," *Machine Audition: Principles, Algorithms and Systems*, pp. 162–185, 2010.

[5] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE TASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[6] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE TASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[7] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "An inverse-gama source variance prior with factorized parametrization for audio source separation," in *IEEE ICASSP*, 2016.

[8] D. Vijayasenan, F. Valente, and H. Bourlard, "Multistream speaker diarization of meetings recordings beyond mfcc and tdoa features," *Springer handbook on speech processing and speech communication*, vol. 54, no. 1, 2012.

[9] T. Higuchi, H. Takeda, N. Tomohiko, and H. Kameoka, "A unified approach for underdetermined blind signal separation and source activity detection by multichannel factorial hidden markov models," in *Interspeech*, Singapore, 2014.

[10] T. Higuchi and H. Kameoka, "Unified approach for audio source separation with multichannel HMM and DOA mixture model," in *Eusipco*, Nice, France, 2015.

[11] T. Higuchi, N. Takamune, N. Tomohiko, and H. Kameoka, "Underdetermined blind separation and tracking of moving sources based on DOA-HMM," in *IEEE ICASSP*, Florence, Italy, 2014.

[12] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech, Audio Process.*, vol. 8, no. 3, pp. 320–327, 2000.

[13] F. Neeser and J. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Info. Theory*, vol. 39, no. 4, pp. 1293–1302, 1993.

[14] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of time-varying convolutive audio mixtures," *IEEE TASLP*, vol. 24, no. 8, pp. 1408–1423, 2016.

[15] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *ISSPA*, 2010.

[16] L. Benaroya, L. Donagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for Wiener based source separation with a single sensor," in *IEEE ICASSP*, vol. 6, 2003, pp. 613–616.

[17] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[18] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE TASLP*, vol. 21, no. 10, pp. 2140–2151, 2013.

[19] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[20] N. Sturmel, A. Liutkus, J. Pinel, L. Girin, S. Marchand, G. Richard, R. Badeau, and L. Daudet, "Linear mixing models for active listening of music productions in realistic studio conditions," in *Proc. Audio Eng. Soc.*, 2012.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," 1993, linguistic Data Consortium, Philadelphia.

[22] C. Hummersone, R. Mason, and T. Brookes, "A comparison of computational precedence models for source separation in reverberant environments," *J. Audio Eng. Soc*, vol. 61, no. 7/8, pp. 508–520, 2013.

[23] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.