A Latently Constrained Mixture Model for Audio Source Separation and Localization

Antoine Deleforge, Radu Horaud

A Latently Constrained Mixture Model for Audio Source Separation and Localization

Antoine Deleforge and Radu Horaud

INRIA Grenoble Rhône-Alpes, FRANCE

Abstract. We present a method for audio source separation and localization from binaural recordings. The method combines a new generative probabilistic model with time-frequency masking. We suggest that device-dependent relationships between point-source positions and interaural spectral cues may be learnt in order to constrain a mixture model. This allows to capture subtle separation and localization features embedded in the auditory data. We illustrate our method with data composed of two and three mixed speech signals in the presence of reverberations. Using standard evaluation metrics, we compare our method with a recent binaural-based source separation-localization algorithm.

1 Introduction

We address the problem of simultaneous separation and localization of sound sources mixed in an acoustical environment and recorded with two microphones. Time-frequency masking is a technique allowing the separation of an arbitrary number of sources with only two microphones by assuming that a single source is active at every time-frequency point – the *W-disjoint orthogonality* (W-DO). It was shown that this assumption holds, in general, for simultaneous speech signals [8]. The input signal is represented in a time-frequency domain and points corresponding to the target source are weighted with 1 and otherwise with 0. The masked spectrogram is then converted back to a temporal signal. A number of methods combine time-frequency masking with localization-based clustering ([8],[3],[2]), e.g., DUET [8] which allows to separate anechoic mixtures when each source reaches the microphones with a single attenuation coefficient and delay. This mixing model is well suited for "clean" binaural recordings. In practice, more complex filtering effects exist, namely the *head-related transfer function* (HRTF) and the *room impulse response* (RIR). These filters lead to frequency-dependent attenuations and delays between the two microphones, respectively called the *interaural level difference* (ILD) and the *interaural phase difference* (IPD). Some approaches attempted to account for these dependencies by learning a mapping between azimuth, frequencies and interaural cues [7, 5, 3]. These mappings usually consist in finding a functional relationship that best fits data obtained from an HRTF dataset. To improve robustness to RIR variations, these interaural cues can also be integrated in a mixture model, e.g., [2].

In this paper we propose to directly learn a discrete mapping between a set of 3D point sources and IPD/ILD spectral cues. We will refer to such mappings as *Source-Position-to-Interaural-Cues* maps (SPIC maps). Unlike what is done in [7, 5, 3], the proposed mapping is built point-wise, does not rely on azimuth only and is device-dependent. We explicitly incorporate it into a novel *latently constrained mixture model* for point sound sources. Our model is specifically designed to capture the richness of binaural data recorded with an acoustic dummy head, and this to improve both localization and separation performances. We formally derive an EM algorithm that iteratively performs separation (E-step) followed by localization and source-parameter estimation (M-step). The algorithm is supervised by a training stage consisting in learning a mapping between potential source positions and interaural cues, i.e., SPIC maps. We believe that a number of methods could be used in practice to learn such maps. In particular we propose an *audio-motor mapping* approach. The results obtained with our method compare favorably with the recently proposed MESSL algorithm [2].

2 Binaural Sound Representation

Spectrograms associated with each one of the two microphones are computed using short-term FFT analysis. We use a 64ms time-window with 8ms window overlap, thus yielding $T = 126$ time windows for a 1s signal. Since sounds were recorded at a sample rate of 16,000Hz, each time window contains 1,024 samples. Each window is then transformed via FFT to obtain complex coefficients of $F = 513$ positive frequency channels between 0 and 8,000Hz. We denote with $s_{f,t}^{(k)} \in \mathbb{C}$ the (f,t) point of the spectrogram emitted by sound-source k, and with $s_{f,t}^{(\mathrm{L})}$ and $s_{f,t}^{(\mathrm{R})}$ the spectrogram points perceived by the left- and right-microphone respectively. The W-DO assumption implies that a single sound source k emits at a given point (f,t). The relationships between the emitted and the left and right perceived spectrogram points are:

$$s_{f,t}^{(\mathrm{L})} = h^{(\mathrm{L})}(\boldsymbol{x}_k, f)\; s_{f,t}^{(k)} \;\text{ and }\; s_{f,t}^{(\mathrm{R})} = h^{(\mathrm{R})}(\boldsymbol{x}_k, f)\; s_{f,t}^{(k)} \tag{1}$$

where $\boldsymbol{x}_k \in \mathbb{R}^3$ is the 3D position of sound source k in a listener-centered coordinate frame and $h^{(\mathrm{L})}$ and $h^{(\mathrm{R})}$ denote the left and right HRTFs. The *interaural transfer function* (ITF) is defined by the ratio between the two HRTFs, i.e., $I(\boldsymbol{x}_k, f) = h^{(\mathrm{R})}(\boldsymbol{x}_k, f)/h^{(\mathrm{L})}(\boldsymbol{x}_k, f) \in \mathbb{C}$. The interaural spectrogram is defined by $\hat{I}_{f,t} := s_{f,t}^{(\mathrm{R})}/s_{f,t}^{(\mathrm{L})}$, so that $\hat{I}_{f,t} \approx I(\boldsymbol{x}_k, f)$. Note that the last approximation only holds if there is a source k emitting at frequency-time point (f,t), and if the time delay between microphones (≈ 0.75ms) is much smaller than the Fourier transform time-window that is used (64ms). Under these conditions, at a given frequency-time point, the interaural spectrogram value $\hat{I}_{f,t}$ does not depend on the emitted spectrogram value $s_{f,t}^{(k)}$ but only on the emitting source position $\boldsymbol{x}_k$. We finally define the *ILD spectrogram* α and the *IPD spectrogram* ϕ as the

log-amplitude and phase of the complex interaural spectrogram $\hat{I}_{f,t}$:

$$\alpha_{f,t} = 20 \log |\hat{I}_{f,t}| \in \mathbb{R}, \quad \phi_{f,t} = \arg(\hat{I}_{f,t}) \in]-\pi, \pi] \tag{2}$$

As already outlined in Section 1 our method makes use of a SPIC map that is learnt during a training stage. Let $\mathcal{X} = \{\boldsymbol{x}_n\}_{n=1}^N$ be a set of 3D sound-source locations in a listener-centered coordinate frame. Let a sound-source n, located at $\boldsymbol{x}_n$ emit white noise and let $\{\alpha_{f,t}^n\}_{f=1,t=1}^{F,T}$ and $\{\phi_{f,t}^n\}_{f=1,t=1}^{F,T}$be the perceived ILD and IPD spectrograms. The *mean ILD* $\boldsymbol{\mu}(\boldsymbol{x}_n) = (\mu_1^n \dots \mu_f^n \dots \mu_F^n)^\top \in \mathbb{R}^F$ and *mean IPD* $\boldsymbol{\xi}(\boldsymbol{x}_n) = (\xi_1^n \dots \xi_f^n \dots \xi_F^n)^\top \in]-\pi, \pi]$ *vectors* associated with n are defined by taking the temporal means of α^n and ϕ^n at each frequency channel:

$$\mu_f^n = 1/T \sum_{t=1}^T \alpha_{f,t}^n \text{ and } \xi_f^n = \arg(1/T \sum_{t=1}^T e^{j\phi_{f,t}^n}) \tag{3}$$

Vector $\boldsymbol{\xi}$ is estimated in the complex domain in order to avoid problems due to phase circularity [4]. White noise is used because it contains equal power within a fixed bandwidth at any center frequency: The source n is therefore the only source emitting at each point (f, t); μ_f^n and ξ_f^n are thus approximating the log-amplitude and phase of $I(\boldsymbol{x}_k, f)$. The set $\mathcal{X}$ of 3D source locations as well as the mappings $\boldsymbol{\mu}$ and $\boldsymbol{\xi}$ will be referred to as the training data to be used in conjunction with the separation-localization algorithm described below.

3 Constrained Mixtures for Separation and Localization

Let's suppose now that there are K simultaneously emitting sounds sources from unknown locations $\{\boldsymbol{x}_k\}_{k=1}^K \subset \mathcal{X}$ and with unknown spectrograms. Using the listener's microphone pair it is possible to build the ILD and IPD observed spectrograms $\{\alpha_{f,t}\}_{f=1,t=1}^{F,T}$ and $\{\phi_{f,t}\}_{f=1,t=1}^{F,T}$. The goal of the sound-source separation and localization algorithm described in this section is to associate each observed point (f, t) with a single source and to estimate the 3D location of each source.

As mentioned in section 2, the observations $\alpha_{f,t}$ (ILD) and $\phi_{f,t}$ (IPD) are significant only if there is a sound source emitting at (f, t). To identify such *significant observations* we estimate the *sound intensity level* (SIL) spectrogram at the two microphones, and retain only those frequency-time points for which the SIL is above some threshold. One empirical way to choose the thresholds (one for each frequency) is to average the SILs at each f in the absence of any emitting source. These thresholds are typically very low compared to SILs of natural sounds, and allow to filter out frequency-time points corresponding to "room silence". Let $M_f \leq T$ be the number of significant observations at f and let $\alpha_{f,m}$ and $\phi_{f,m}$ be the m-th significant ILD and IPD observations at f. Let $\mathbf{A} = \{\alpha_{f,m}\}_{f=1,m=1}^{F,M_f}$ and $\boldsymbol{\Phi} = \{\phi_{f,m}\}_{f=1,m=1}^{F,M_f}$ be the *observed data*.

Let $\boldsymbol{z}_{f,m} \in \{0,1\}^K$ be the *missing data*, i.e., the data-to-source assignment variables, such that $z_{f,m,k} = 1$ if observations $\alpha_{f,m}$ and $\phi_{f,m}$ are generated by source k, and $z_{f,m,k} = 0$ otherwise. The W-DO assumption yields $\sum_{k=1}^{K} z_{f,m,k} = 1$ for all (f,m). $\mathcal{M}_k = \{z_{f,m,k}\}_{f=1,m=1}^{F,M_f}$ is the binary spectral mask of the k-th source. Finally, $\mathbf{Z} = \{\boldsymbol{z}_{f,m}\}_{f=1,m=1}^{F,M_f}$ denotes the set of all missing data. The problem of simultaneous localization and separation amounts to estimate the masking variables $\mathbf{Z}$ and the locations $\{\boldsymbol{x}_k\}_{k=1}^{K}$ conditioned by $\mathbf{A}$ and $\boldsymbol{\Phi}$, given the number of sources K. We assume that observed data are perturbed by Gaussian noise. Hence, the probability of observing $\alpha_{f,m}$ conditioned by source k ($z_{f,m,k} = 1$) located at $\boldsymbol{x}_k$ is drawn from a normal distribution, and the probability of observing $\phi_{f,m}$ is drawn from a circular normal distribution. The source position $\boldsymbol{x}_k$ acts here as a *latent constraint* on ILD and IPD means:

$$P(\alpha_{f,m}|z_{f,m,k} = 1, \boldsymbol{x}_k, \sigma_{f,k}) = \mathcal{N}(\alpha_{f,m}|\mu_f(\boldsymbol{x}_k), \sigma_{f,k}^2) \text{ and} \tag{4}$$

$$P(\phi_{f,m}|z_{f,m,k} = 1, \boldsymbol{x}_k, \rho_{f,k}) = \mathcal{N}(\Delta(\phi_{f,m}, \xi_f(\boldsymbol{x}_k))|0, \rho_{f,k}^2) \tag{5}$$

where $\sigma_{f,k}^2$ and $\rho_{f,k}^2$ are the ILD and IPD variances associated with source k at frequency f and the Δ function is defined by $\Delta(x,y) = \arg(e^{j(x-y)}) \in]-\pi, \pi]$. As in [2], (5) approximates the normal distribution on the circle $]-\pi, \pi]$ when $\rho_{f,k}$ is small relative to 2π. Preliminary experiments on IPD spectrograms of white noise showed that this assumption holds in the general case. As emphasized in [2], the well known correlation between ILD and IPD does not contradict the assumption that Gaussian noises corrupting the observations are independent. The conditional likelihood of the observed data $(\alpha_{f,m}, \phi_{f,m})$ is therefore given by the product of (4) and (5). We also define the priors $\pi_{f,k} = P(z_{f,m,k})$ which model the proportion of the observed data generated by source k at frequency f. In summary, the model parameters are $\Theta = \{\{\boldsymbol{x}_k\};\ \{\pi_{f,k}\};\ \{\sigma_{f,k}^2\};\ \{\rho_{f,k}^2\}\}_{f=1,k=1}^{F,K}$. The problem can now be expressed as the maximization of the observed-data log-likelihood conditioned by Θ. In order to keep the model as general as possible, there is no assumption on the emitted sounds as well as the way their spectra are spread across the frequency-time points. Therefore, we assume that all the observations are statistically independent, yielding the following expression for the observed-data log-likelihood:

$$\mathcal{L}(\mathbf{A}, \boldsymbol{\Phi}; \Theta) = \log P(\mathbf{A}, \boldsymbol{\Phi}; \Theta) = \sum_{f=1}^{F} \sum_{m=1}^{M_f} \log P(\alpha_{f,m}, \phi_{f,m}; \Theta) \tag{6}$$

We address this maximum-likelihood with missing-data problem within the framework of expectation-maximization (EM). In our case, the E-step computes the posterior probabilities of assigning each spectrogram point to a sound source k (separation) while the M-step maximizes the expected complete-data log-likelihood with respect to the model parameters Θ and, most notably, with the source locations $\{\boldsymbol{x}_k\}_{k=1}^{K}$ (localization). The MAP criterion provides binary spectral masks $\mathcal{M}_k$ associated with each source k while the final parameters $\{\boldsymbol{x}_k\}_{k=1}^{K}$ provide estimates for the source locations. The expected complete-data log-likelihood writes ($^{(p)}$ denotes the p-th iteration):

$$Q(\Theta|\Theta^{(p-1)}) = \sum_{f=1}^{F} \sum_{m=1}^{M_f} \sum_{k=1}^{K} r_{f,m,k}^{(p)} \log \pi_{f,k} P(\alpha_{f,m}, \phi_{f,m}|\boldsymbol{z}_{f,m}; \Theta) \tag{7}$$

The *E-step* updates the responsibilities according to the standard formula:

$$r_{f,m,k}^{(p)} = \frac{\pi_{f,k} P(\alpha_{f,m}, \phi_{f,m} | \boldsymbol{z}_{f,m}; \Theta^{(p-1)})}{\sum_{i=1}^{K} \pi_{f,i} P(\alpha_{f,m}, \phi_{f,m} | \boldsymbol{z}_{f,m}; \Theta^{(p-1)})} \tag{8}$$

The *M-step* maximizes (7) with respect to Θ. By combining (4) and (5) with (7) the equivalent minimization criterion writes:

$$\sum_{f=1}^{F} \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} \left(\log \left(\frac{\sigma_{f,k}^2 \rho_{f,k}^2}{\pi_{f,k}^2} \right) + \frac{(x_{f,m} - \mu_f(\boldsymbol{x}_k))^2}{\sigma_{f,k}^2} + \frac{\Delta(\phi_{f,m}, \xi_f(\boldsymbol{x}_k))^2}{\rho_{f,k}^2} \right) \tag{9}$$

which can be differentiated with respect to $\{\pi_{f,k}\}_f$, $\{\sigma_{f,k}\}_f$ and $\{\rho_{f,k}\}_f$ to obtain closed-form expressions for the optimal parameter values conditioned by $\boldsymbol{x}_k$:

$$\widetilde{\pi}_{f,k} = \frac{\overline{r}_{f,k}}{M_f}, \text{ with } \overline{r}_{f,k} = \sum_{m=1}^{M_f} r_{f,m,k} \tag{10}$$

$$\widetilde{\sigma}_{f,k}^2(\boldsymbol{x}_k) = \frac{1}{\overline{r}_{f,k}} \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} (x_{f,m} - \mu_f(\boldsymbol{x}_k))^2 \tag{11}$$

$$\widetilde{\rho}_{f,k}^2(\boldsymbol{x}_k) = \frac{1}{\overline{r}_{f,k}} \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} \Delta(\phi_{f,m}, \xi_f(\boldsymbol{x}_k))^2 \tag{12}$$

By substituting (11) and (12) into (9) the optimal location $\widetilde{\boldsymbol{x}}_k$ is obtained by minimizing the following expression with respect to $\boldsymbol{x}_k$:

$$\sum_{f=1}^{F} \overline{r}_{f,k} \left(\log \left(1 + \frac{(\overline{\alpha}_{f,k} - \mu_f(\boldsymbol{x}_k))^2}{V_{f,k}} \right) + \log \left(1 + \frac{\Delta(\overline{\phi}_{f,k}, \xi_f(\boldsymbol{x}_k))^2}{W_{f,k}} \right) \right) \tag{13}$$

with: $\overline{\alpha}_{f,k} = \frac{1}{\overline{r}_{f,k}} \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} \alpha_{f,m}$; $\quad V_{f,k} = \frac{1}{\overline{r}_{f,k}} \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} (\alpha_{f,m} - \overline{\alpha}_{f,k})^2$
$\overline{\phi}_{f,k} = \arg \left(\frac{1}{\overline{r}_{f,k}} \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} e^{j\phi_{f,m}} \right)$; $W_{f,k} = \frac{1}{\overline{r}_{f,k}} \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} \Delta(\phi_{f,m}, \overline{\phi}_{f,k})^2$

(13) is evaluated for each source location in the training dataset $\mathcal{X}$ (Section 2) in order to find an optimal 3D location $\widetilde{\boldsymbol{x}}_k$. This is then substituted back in (11) and (12) to estimate $\widetilde{\sigma}_{f,k}$ and $\widetilde{\rho}_{f,k}$ and repeated for each unknown source k.

In general, EM converges to a local maximum of (6). The non-injectivity nature of the interaural functions μ_f and ξ_f and the high cardinality of Θ leads to a very large number of such maxima, especially when the training set $\mathcal{X}$ is large. This makes our algorithm very sensitive to initialization. One way to avoid being trapped in local maxima is to initialize the mixture's parameters at random several times. This cannot be easily applied here since there is no straightforward way to initialize the model's variances. Alternatively, one may randomly initialize the assignment variables $\mathbf{Z}$ and then proceed with the M-step. However, extensive simulated experiments revealed that this solution fails to converge to the ground-truth solution in most of the cases. We therefore propose to combine these strategies by randomly perturbing both the source locations and the source assignments during the first stages of the algorithm. We developed a *stochastic initialization* procedure similar in spirit to SEM [1].

The SEM algorithm includes a stochastic step (S) between the E- and the M-step, during which random samples $R_{f,m,k} \in \{0,1\}$ are drawn from the responsibilities (8). These samples are then used instead of (8) during the M-step. To initialize our algorithm, we first set $r^{(0)}_{f,m,k} = 1/K$ for all k and then proceed through the sequence S M* E S M, where M* is a variation of M in which the source positions are drawn randomly from $\mathcal{X}$ instead of solving (13). In practice, ten such initializations are used to enforce algorithm convergence, and only the one providing the best log-likelihood after two iterations is iterated twenty more times. A second technique was used to overcome local maxima issues due to the large number of parameters. During the first ten steps of the algorithm only, a unique pair of variances (σ_k^2, ρ_k^2) is estimated for each source. This is done by calculating the means $\overline{\sigma}_k^2(\boldsymbol{x}_k)$ and $\overline{\rho}_k^2(\boldsymbol{x}_k)$ of frequency-dependent variances (11) and (12) weighted by $\overline{r}_{f,k}$. The optimal value $\widetilde{\boldsymbol{x}}_k$ is the one minimizing $\overline{\sigma}_k^2(\boldsymbol{x})\overline{\rho}_k^2(\boldsymbol{x})$ evaluated over all $\boldsymbol{x} \in \mathcal{X}$. Intensive experiments showed that the proposed method converges to a global optimum in most of the cases.

4 Experiments, Results, and Conclusions

In order to evaluate and compare our method, a specific data set of binaural records was built[1] using a Sennheiser MKE 2002 acoustic dummy-head mounted onto a robotic system with two rotational degrees of freedom, namely pan (ψ) and tilt (θ). This device, specifically designed to perform accurate and reproducible motions, allows us to collect both a very dense SPIC map for the training set (section 2) and a large test set of mixed speech point sources. The emitter (a loud speaker) is placed at approximately 2.5 meters in front of the listener. Under these conditions the HRTF mainly depends on the sound-source direction: Hence, the location is parameterized by the angles ψ and θ. All the experiments were carried out in a reverberant room and in the presence of background noise. For recording purposes, the robot is placed in 90 pan angles $\psi \in [-90°, 90°]$ (left-right) and 60 tilt angles $\theta \in [-60°, 60°]$ (top-down), i.e., $N = 5,400$ uniformly distributed *motor states* in front of the *static* emitter, forming the set $\mathcal{X}$. Five binaural recordings are available with each motor state: Sound #0 corresponds to a 1s "room silence" used to estimate the SIL thresholds (section 3). Sound #1 corresponds to 1s white-noise used to build the training set (section 2). Sounds #2, #3 and #4 form the test set and correspond to "They never met you know" by a female (#2), "It was time to go up myself" by a male (#3), and "As we ate we talked" by a male (#4). The three sounds are about 2s long and were randomly chosen from the TIMIT database. Each record was associated to its ground-truth motor-state, thus allowing to create signals of mixed sound sources from different direction with 2° resolution.

We generated 1000 mixtures of two and three speech signals emitted by randomly located sources. 97.7% of the individual sources were correctly mapped to

[1] Online at: `http://perception.inrialpes.fr/~Deleforge/CAMIL_Dataset`

	2 Sources		3 Sources	
	SDR	SIR	SDR	SIR
Oracle Mask	11.73	19.23	9.20	16.16
Our Approach (Loc)	5.28	8.91	2.44	3.92
Our Approach (All)	5.19	8.84	1.72	2.74
MESSL-G	2.83	5.74	1.48	1.47
Original Mixture	0.00	0.45	-3.50	-2.82

Table 1. Comparing the mean source-to-distortion ratio (SDR) and source-to-interference ratio (SIR), in dB, for 1000 mixtures of 2 and 3 sources. Mean separation results with our approach are calculated over all sources (All) and over correctly localized sources only (Loc).

their associated position (i.e. $\leq 2^\circ$ error for both ψ and θ) in the two-source case, and 63.8% in the three-source case. The performance of separation was evaluated with the standard SDR and SIR metrics [6]. We compared our results to those obtained with the original mixture (no mask applied), with the ground truth or *Oracle mask* [8], and with the recently proposed MESSL[2] algorithm [2]. The Oracle mask is set to 1 at every spectrogram point in which the target signal is at least as loud as the combined other signals and 0 everywhere else. The version MESSL-G used includes a garbage component and ILD priors to better account for reverberations and is reported to outperform four methods in reverberant conditions, including [8] and [3]. Table 1 shows that our method yields significantly better results than MESSL-G on an average, although both algorithm require similar computational times. Notice how the localization correctness critically affects the separation performances, and decreases in the three-source case, as the number of observations per source becomes lower and the number of local maxima in (6) becomes higher. Our SDR scores strongly outperform MESSL-G in most cases, while SIR results are only slightly better when sources are more than 70° apart in azimuth (pan angle), e.g., Fig. 1. However, they become much higher when sources are nearby in azimuth, or share the same azimuthal plane with different elevations (tilt angles). This is because MESSL relies on the estimation of a probability density in a discretized ITD space for each source, and thus does not account for more subtle spatial cues induced by the HRTF.

These results clearly demonstrate the efficiency of our method, but they somehow favor our algorithm because of the absence of RIR variations both in the training and the test data sets. The aim of experimenting with these relatively simple data has been to show that our method can conceptually separate and accurately locate both in azimuth and elevation a binaural mixture of 2 to 3 sound sources. The prerequisite is a training stage: the interaural cues associated with source positions need to be learnt in advance using white noise, and we showed that the algorithm performs well even for a very large and dense set of learnt positions. Preliminary results obtained while changing the position of the test sound source in the room suggested that our constrained mixture model coupled with frequency-dependent variances presented some robustness to RIR variations. Alternatively, one could build a training set on different premises such as seat locations in a conference room or musician locations in a concert hall, and thus directly learn the RIR during the training stage.

[2] `http://blog.mr-pc.org/2011/09/14/messl-code-online/`.

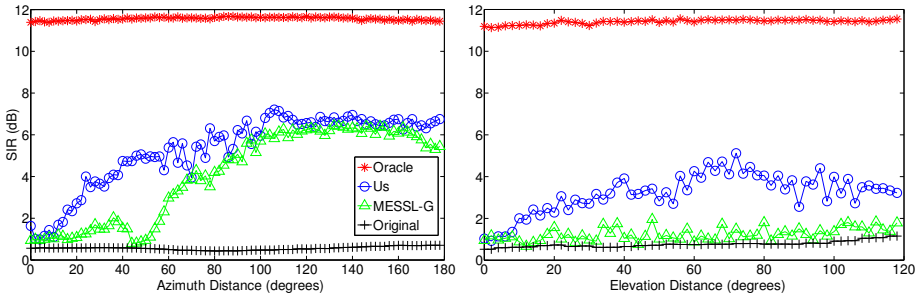

Fig. 1. SIR as a function of azimuth (pan) and elevation (tilt) separation between two sources. Left: one source fixed at $(-90°, 0°)$ while the other takes 90 positions between $(-90°, 0°)$ and $(+90°, 0°)$. Right: one source fixed at $(0°, -60°)$ while the other takes 60 positions between $(0°, -60°)$ and $(0°, +60°)$. SIRs are averaged over 6 mixtures of 2 sources (12 targets). Top-to-down: Oracle ($*$), our method ($\circ$), MESSL-G ($\triangle$), and original mixture ($+$).

To conclude, we proposed a novel audio source separation and localization method based on a mixture model constrained by a SPIC map. Experiments and comparisons showed that our algorithm performs better than a recently published probabilistic spectral masking technique in terms of separation and yields very good multi-source localization results. The combination of a SPIC map with a mixture model is a unique feature. In the future, we plan to study more thoroughly the behavior of our algorithm to RIR variations, and improve its robustness by extending our model to a continuous and probabilistic mapping between source positions and interaural parameters. Dynamic models incorporating moving sound sources and head movements could also be included.

References

1. G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Comp. Stat. & Data An.*, 14(3):315–332, 1992.
2. M. I. Mandel, R. J. Weiss, and D. P. W. Ellis. Model-based expectation-maximization source separation and localization. *IEEE TASLP*, 18:382–394, 2010.
3. J. Mouba and S. Marchand. A source localization/separation/respatialization system based on unsupervised classification of interaural cues. In *Proceedings of the International Conference on Digital Audio Effects*, pages 233–238, 2006.
4. J. Nix and V. Hohmann. Sound source localization in real sound fields based on empirical statistics of interaural parameters. *JASA*, 119(1):463–479, 2006.
5. N. Roman, D. Wang, and G. J. Brown. Speech segregation based on sound localization. *JASA*, 114(4):2236–2252, 2003.
6. E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE TASLP*, 14(4):1462–1469, 2006.
7. H. Viste and G. Evangelista. On the use of spatial cues to improve binaural source separation. In *Proc. Int. Conf. on Digital Audio Effects*, pages 209–213, 2003.
8. O. Yılmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52:1830–1847, 2004.