Variational EM for Binaural Sound-Source Separation and Localization

Antoine Deleforge, Florence Forbes, Radu Horaud

VARIATIONAL EM FOR BINAURAL SOUND-SOURCE SEPARATION AND LOCALIZATION

Antoine Deleforge, Florence Forbes and Radu Horaud

INRIA Grenoble Rhône-Alpes and Université de Grenoble, France

ABSTRACT

The sound-source separation and localization (SSL) problems are addressed within a unified formulation. Firstly, a mapping between white-noise source locations and binaural cues is estimated. Secondly, SSL is solved via Bayesian inversion of this mapping in the presence of multiple sparse-spectrum emitters (such as speech), noise and reverberations. We propose a variational EM algorithm which is described in detail together with initialization and convergence issues. Extensive real-data experiments show that the method outperforms the state-of-the-art both in separation and localization (azimuth and elevation).

1. INTRODUCTION

In this paper we address the problem of sound-source separation and localization (SSL) using two microphones plugged into the ears of a dummy head. Recently it was suggested that the ILD (interaural level difference) spectrogram carries information about the relationship between the binaural observation space and the two-dimensional (2D) localization space (azimuth and elevation) and that the latter can be retrieved via an unsupervised manifold learning method [1, 2]. Within this framework, the general SSL problem is more challenging for several reasons. Firstly, the mapping from a sound-source location to an ILD observation is unknown and non-linear due to the head-related transfer function (HRTF) which cannot be easily modeled. Secondly, auditory data are corrupted by noise and reverberations. Thirdly, an ILD frequency value is relevant only if the source is actually emitting at that frequency: Natural sounds such as speech are known to be extremely sparse, with often 80% of the frequencies actually missing at a given time. Finally, when several sources emit simultaneously, the assignment of a time-frequency point of the ILD spectrogram to one of the sources is not known.

Binaural-based SSL methods often rely on the assumption that a single source is active at each time-frequency point [3]. Hence, a number of methods combine time-frequency masking with localization-based clustering [3–6]. Binaural-based localization requires to map interaural cues to source positions. Most existing approaches approximate this mapping based on simplifying assumptions, such as direct-path source-to-microphone propagation [3], a sine interpolation of ILD data from a human HRTF dataset [7], or a spiral ear model [8]. These approaches have the disadvantage of requiring extra parameters, *e.g.*, the distance between the microphones, the head dimensions, or an ear model, and quite often they are not valid in real world conditions. We note that the vast majority of current SSL approaches mainly focus on a rough estimation of the azimuth, or *one-dimensional* (1D) localization [5,7,9,10], and that very few perform 2D localization [8]. Alternatively, some approaches [6, 11, 12] bypass the explicit mapping model and perform 2D localization using an exhaustive search in an HRTF look-up table. However, this process is unstable and hardly scalable in practice as the number of required associations yields too prohibitive memory and computational costs.

Recently, we proposed a generative probabilistic framework for characterizing the mapping from the space of sound-source locations to the space of binaural cues. Indeed, the computational experiments reported in [1, 2] suggest the existence of a locally-linear bijection from the space of source locations to the space of binaural cues, and that the high-dimensional space spanned by the latter forms a low-dimensional manifold embedded in the former (source locations). In practice, the source-location-to-binaural-cue mapping can be approximated by a *probabilistic piecewise affine mapping* (PPAM) model whose parameters are learned via an EM procedure. This learning stage may be viewed as a *system calibration* task. Then, accurate 2D localization of a *single* sound source may be inferred from the *inverse* posterior distribution of the PPAM model [2].

This paper generalizes single-source localization [2] to SSL, *e.g.*, the perceived binaural signals are generated from multiple sources with unknown azimuth and elevation. As in [2] the PPAM model is inferred from a training data set of input-output variable pairs, where the input is the known 2D location of a white-noise emitter and the output is the perceived ILD spectrogram. The proposed runtime algorithm estimates separation and localization in the presence of multiple sparse-spectrum sounds. The problem will be viewed as the one of inverting PPAM where the observed signals, generated from multiple latent variables, are both mixed and corrupted by noise. We show that this problem can be cast into a variational EM framework [13]. We propose a factorization of

RESEARCH SUPPORTED BY THE EU PROJECT FP7-ICT-247525

the model's posterior probability that decomposes the E-step into two localization and separation sub-steps. The algorithm yields a fully Bayesian estimation of the 2D locations and time-frequency masks of all the sources.

2. PROBABILISTIC PIECEWISE AFFINE MAPPING

This section briefly summarizes the PPAM model presented in detail in [2]. Let $\mathcal{X} \subset \mathbb{R}^L$ be the space of sound-source positions and $\mathcal{Y} \subset \mathbb{R}^D$ be the ILD space. The observed ILDs are denoted by $\boldsymbol{Y} = \{\boldsymbol{Y}_t\}_{t=1}^T \in \mathcal{Y}$, where $\boldsymbol{Y}_t = [Y_{1t} \dots Y_{Dt}]^\top$ is a vector of frequency-dependent values, the source positions are denoted by $\boldsymbol{X} = \{\boldsymbol{X}_m\}_{m=1}^M \in \mathbb{R}^L$, and the source assignment variables are denoted by $\boldsymbol{W} = \{W_{dt}\}_{d=1,t=1}^{D,T}$, *i.e.*, $\boldsymbol{W}_{dt} = m$ means that Y_{dt} is generated from source m.

The PPAM model parameters are estimated from a training data-set of known input-output pairs $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N \subset \mathcal{X} \times \mathcal{Y}$. We thus have $T = M = N$ and $\boldsymbol{W}$ variables are known. Assuming that $\mathcal{Y}$ is an $L-$dimensional manifold embedded in $\mathbb{R}^D$, a mapping $g : \mathcal{X} \to g(\mathcal{X}) = \mathcal{Y}$ is estimated using this data-set. The local linearity of manifolds suggests that each $\boldsymbol{y}_n$ is the image of a source location $\boldsymbol{x}_n \in \mathcal{R}_k$ by an affine transformation t_k plus an error term, where $\{\mathcal{R}_k\}_{k=1}^K$ is a partitioning of $\mathcal{X}$. Assuming that there are K such affine transformations t_k, one for each region $\mathcal{R}_k$, a piecewise-affine approximation of g can be recovered from the training data-set: $\boldsymbol{y}_n = \sum_{k=1}^K \mathbb{I}_{\{Z_n=k\}}(\boldsymbol{A}_k \boldsymbol{x}_n + \boldsymbol{b}_k) + \boldsymbol{e}_n$ where $Z_n \in \{1 \dots K\}$ is associated with $(\boldsymbol{x}_n, \boldsymbol{y}_n)$ such that $Z_n = k$ if $\boldsymbol{y}_n$ is the image of $\boldsymbol{x}_n \in \mathcal{R}_k$ by t_k. Each t_k is defined by a matrix $\boldsymbol{A}_k$ and a vector $\boldsymbol{b}_k$ while $\boldsymbol{e}_n$ captures the reconstruction error. Assuming that the $\boldsymbol{e}_n$'s are independent of $\boldsymbol{Y}_n$, $\boldsymbol{X}_n$ and Z_n, and that they are i.i.d. realizations of a centered Gaussian variable with diagonal covariance $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2 ... \sigma_D^2)$, we obtain: $p(\boldsymbol{y}_n | \boldsymbol{X}_n = \boldsymbol{x}_n, Z_n = k; \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{A}_k \boldsymbol{x}_n + \boldsymbol{b}_k, \boldsymbol{\Sigma})$ where $\boldsymbol{\theta}$ designates the model parameters. To make the transformations t_k local we define a Gaussian mixture prior on $(\boldsymbol{X}_n, Z_n)$, *i.e.*, $p(\boldsymbol{X}_n = \boldsymbol{x}_n | Z_n = k; \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)$ and $p(Z_n = k; \boldsymbol{\theta}) = \pi_k$. The closed-form EM algorithm proposed in [2] maximizes $\log p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta} = \left\{\{\boldsymbol{\Gamma}_k, \boldsymbol{c}_k, \boldsymbol{A}_k, \boldsymbol{b}_k, \}_{k=1}^K, \boldsymbol{\Sigma}\right\}$.

3. SOUND SEPARATION AND LOCALIZATION

The SSL problem can now be formulated as a piecewise affine inversion problem, where observed signals generated from multiple sources (modeled as latent variables) are both mixed and corrupted by noise. We propose to use a variational expectation-maximization (VEM) framework [13] to deal with the missing data. In more detail, given the mapping parameters estimated with the PPAM algorithm applied to the training data set, we are now addressing the problem of separating and localizing M sound sources. The VEM algorithm described below will be referred to as *variational EM sound separation and localization* (VESSL). Typical examples of the algorithm's inputs and outputs are shown in Fig. 1.

The observed data correspond to a time series of T noisy ILD cues $\boldsymbol{Y} = \{\boldsymbol{Y}_t\}_{t=1}^T$ while all the other variables, namely the source assigments $\boldsymbol{W} \in \mathcal{W}$, the source positions $\boldsymbol{X} \in \mathcal{X}$, and the transformation assignments $\boldsymbol{Z} \in \mathcal{Z}$ are unknown. Typically the number of simultaneously emitting sources M is much smaller than T and D, and several observed time-frequency points Y_{dt} can be assigned to the same source. To account for an unknown $\boldsymbol{W}$, the observation model is reformulated as $p(\boldsymbol{y}_t \,|\, \boldsymbol{w}_t, \boldsymbol{x}, \boldsymbol{z}) = \prod_d p(y_{dt} \,|\, w_{dt}, \boldsymbol{x}_{w_{dt}}, z_{w_{dt}})$ where $p(y_{dt} \,|\, W_{dt} = m, \boldsymbol{X}_m = \boldsymbol{x}_m, Z_m = k) = \mathcal{N}(y_{dt}; \boldsymbol{a}_{dk}^{\mathrm{t}} \boldsymbol{x}_m + \boldsymbol{b}_{dk}, \sigma_d^2)$. We assume that the different source positions are independent, yielding $p(\boldsymbol{x}, \boldsymbol{z}) = \prod_{m=1}^M p(\boldsymbol{x}_m, z_m)$. Source assignments are also assumed to be independent over both time (t) and frequency (d), so that $p(\boldsymbol{w}) = \prod_{d,t} p(w_{dt})$ with $p(W_{dt} = m) = \lambda_{dm}$, where λ_{dm} are positive numbers representing the relative presence of each source in each frequency channel (sources' weights), so that $\sum_{m=1}^M \lambda_{dm} = 1$ for all d. We will write $\boldsymbol{\lambda} = \{\lambda_{dm}\}_{d=1,m=1}^{D,M}$. The complete-model parameter set $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ is $\boldsymbol{\psi} = \{\{\boldsymbol{\Gamma}_k, \boldsymbol{c}_k, \boldsymbol{A}_k, \boldsymbol{b}_k\}_{k=1}^K, \boldsymbol{\Sigma}, \boldsymbol{\lambda}\}$. Notice that among these parameters, the values of $\{\boldsymbol{\Gamma}_k, \boldsymbol{c}_k, \boldsymbol{A}_k, \boldsymbol{b}_k\}_{k=1}^K$ have been estimated during the training stage using PPAM. Therefore, only the parameters $\{\boldsymbol{\Sigma}, \boldsymbol{\lambda}\}$ need to be estimated. $\boldsymbol{\Sigma}$ is re-estimated to account for possibly higher noise levels in the mixed observed signals compared to training.

We denote with $\mathrm{E}_q\big[.\big]$ the expectation with respect to a probability distribution q. Denoting current parameter values by $\boldsymbol{\psi}^{(i)}$, the proposed VEM algorithm provides, at each iteration (i), an approximation $q^{(i)}(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{z})$ of the posterior probability $p(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{z} \,|\, \boldsymbol{y}; \boldsymbol{\psi}^{(i)})$ that factorizes as $q^{(i)}(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{z}) = q_W^{(i)}(\boldsymbol{w})\; q_{X,Z}^{(i)}(\boldsymbol{x}, \boldsymbol{z})$ where $q_W^{(i)}$ and $q_{X,Z}^{(i)}$ are probability distributions on $\mathcal{W}$ and $\mathcal{X} \times \mathcal{Z}$ respectively. Such a factorisation may seem drastic but its main beneficial effect is to replace stochastic dependencies between latent variables with deterministic dependencies between relevant moments of the two sets of variables. It follows that the E-step becomes an approximate E-step that can be further decomposed into two sub-steps whose goal is to update $q_{X,Z}$ and q_W in turn. Closed-form expressions for these sub-steps at iteration (i), extension to missing observations, initialization strategies, and maximum a posteriori (MAP) estimations are detailed below.

E-XZ: $q_{X,Z}^{(i)}(\boldsymbol{x}, \boldsymbol{z}) \;\propto\; \exp \mathrm{E}_{q_W^{(i-1)}}\big[\log p(\boldsymbol{x}, \boldsymbol{z} \,|\, \boldsymbol{y}, \boldsymbol{W}; \boldsymbol{\psi}^{(i)})\big]$. It follows from standard algebra that $q_{X,Z}^{(i)}(\boldsymbol{x}, \boldsymbol{z}) = \prod_{m=1}^M q_{X_m,Z_m}^{(i)}(\boldsymbol{x}_m, z_m)$ where $q_{X_m,Z_m}^{(i)}(\boldsymbol{x}, k) = \alpha_{km}^{(i)} \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{km}^{(i)}, \boldsymbol{S}_{km}^{(i)})$ and $\boldsymbol{\mu}_{km}^{(i)}, \boldsymbol{S}_{km}^{(i)}, \alpha_{km}^{(i)}$ are given in (1), (2). One can see this as the *localization* step, since it estimates a mixture of Gaussians over the latent space $\mathcal{X}$ for each source.

E-W: $q_W^{(i)}(\boldsymbol{w}) \;\propto\; \exp \mathrm{E}_{q_{X,Z}^{(i)}}\big[\log p(\boldsymbol{w} \,|\, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\psi}^{(i)}\big]$. It comes that $q_W^{(i)}(\boldsymbol{w}) = \prod_{d,t} q_{W_{dt}}^{(i)}(w_{dt})$ where $q_{W_{dt}}^{(i)}$ is given in

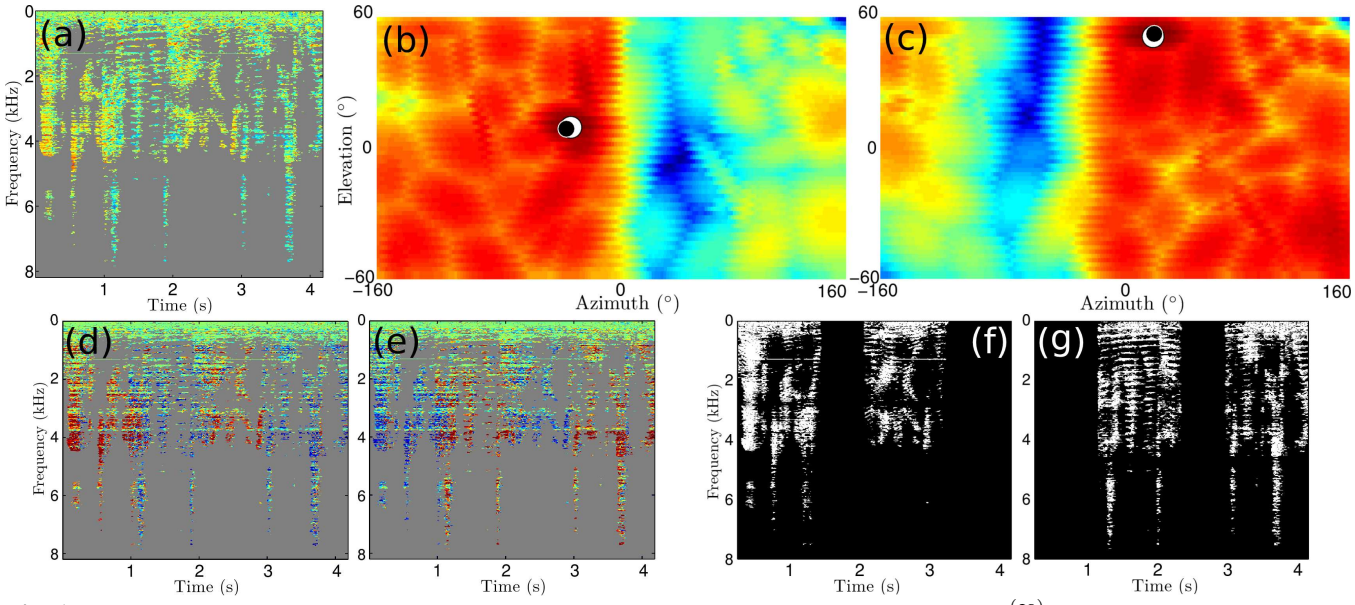

Fig. 1. (a) Input ILD spectrogram. (b,c) Output log-density of each source position as determined by $q^{(\infty)}_{X,Z}$. Ground-truth source positions are noted with a black dot, and the peak of the log-density with a white circle. (d,e) Output source assignment probabilities $q^{(\infty)}_{W}$. (f,g) Ground truth binary masks. Red color denotes high values, blue color low values, and grey colors missing observations.

$$\boldsymbol{\mu}^{(i)}_{km} = \boldsymbol{S}^{(i)}_{km}\left(\boldsymbol{\Gamma}_k^{-1}\boldsymbol{c}_k + \textstyle\sum_{d,t}\sigma_d^{-2}q^{(i-1)}_{W_{dt}}(m)(y_{dt}-b_{dk})\boldsymbol{a}_{dk}\right),\quad \boldsymbol{S}^{(i)}_{km} = \left(\boldsymbol{\Gamma}_k^{-1} + \textstyle\sum_{d,t}\sigma_d^{-2}q^{(i-1)}_{W_{dt}}(m)\boldsymbol{a}_{dk}\boldsymbol{a}_{dk}^\top\right)^{-1}, \tag{1}$$

$$\alpha^{(i)}_{km} = \frac{\pi_k\tilde{\alpha}^{(i)}_{km}}{\sum_{l=1}^K \pi_l\tilde{\alpha}^{(i)}_{lm}},\quad \tilde{\alpha}^{(i)}_{km} = \frac{\exp(\frac{-1}{2}(\boldsymbol{\mu}^{(i)}_{km}-\boldsymbol{c}_k)^{\mathrm{t}}\boldsymbol{\Gamma}_k^{-1}(\boldsymbol{\mu}^{(i)}_{km}-\boldsymbol{c}_k))}{|\boldsymbol{S}^{(i)-1}_{km}\boldsymbol{\Gamma}_k|^{\frac{1}{2}}}\prod_{d,t}\exp\left(\frac{-q^{(i)}_{W_{dt}}(m)(y_{dt}-\boldsymbol{a}^{\mathrm{t}}_{dk}\boldsymbol{\mu}^{(i)}_{km}-b_{dk})^2}{2\sigma_d^2}\right) \tag{2}$$

$$q^{(i)}_{W_{dt}}(m) = \frac{\lambda^{(i)}_{dm}\beta^{(i)}_{dtm}}{\sum_{l=1}^M \lambda^{(i)}_{dl}\beta^{(i)}_{dtl}},\quad \beta^{(i)}_{dtm} = \prod_{k=1}^K \exp\left\{-\frac{\alpha^{(i)}_{km}}{2\sigma_d^2}\left(\mathrm{tr}\left(\boldsymbol{S}^{(i)}_{km}\boldsymbol{a}_{dk}\boldsymbol{a}_{dk}^\top\right) + (y_{dt}-\boldsymbol{a}^{\mathrm{t}}_{dk}\boldsymbol{\mu}^{(i)}_{km}-b_{dk})^2\right)\right\}, \tag{3}$$

$$\lambda^{(i)}_{dm} = \frac{1}{T}\sum_{t=1}^T q^{(i)}_{W_{dt}}(m),\ \sigma_d^{2(i)} = \frac{\sum_{t=1}^T\sum_{m=1}^M\sum_{k=1}^K q^{(i)}_{W_{dt}}(m)\ \alpha^{(i)}_{km}\ \left(\mathrm{tr}\left(\boldsymbol{S}^{(i)}_{km}\boldsymbol{a}_{dk}\boldsymbol{a}_{dk}^\top\right) + (y_{dt}-\boldsymbol{a}^{\mathrm{t}}_{dk}\boldsymbol{\mu}^{(i)}_{km}-b_{dk})^2\right)}{\sum_{t=1}^T\sum_{m=1}^M\sum_{k=1}^K q^{(i)}_{W_{dt}}(m)\ \alpha^{(i)}_{km}} \tag{4}$$

(3). This can be seen as the *separation* step, as it provides the assignment probability of each observation to the sources.

M: $\boldsymbol{\psi}^{(i+1)} = \arg\max_{\boldsymbol{\psi}} \mathrm{E}_{q^{(i)}_W q^{(i)}_{X,Z}}\left[\log p(\boldsymbol{y},\boldsymbol{W},\boldsymbol{X},\boldsymbol{Z}\ ;\ \boldsymbol{\psi})\right]$. This corresponds to the update of sources' weights $\boldsymbol{\lambda}^{(i)}$ and ILD variances $\boldsymbol{\Sigma}^{(i)} = \mathrm{diag}(\sigma_1^{2(i)}...\sigma_D^{2(i)})$, as given in (4).

Missing frequencies: An important challenge in real-world sound source localization is that natural sounds such as speech have a sparse spectrum, and hence generate ILD spectrograms with only a few frequency-time points. Our probabilistic formulation straightforwardly generalizes to such missing observations. In (1) and (2) $q^{(i)}_{W_{dt}}(m)$ is set to 0 for all m if the ILD y_{dt} is missing, *i.e.* the recorded acoustic level is below a given threshold at this point.

Initialization strategies: Extensive experiments have shown that the VEM objective function, called the variational free energy, had a large number of local maxima using real world sound mixtures. This may be due to the combinatorial sizes of the set of all possible binary masks $\mathcal{W}$ and the set of all possible affine transformation assignments $\mathcal{Z}$. Indeed, the procedure has shown to be more sensitive to initialization and to get trapped in suboptimal solutions more often as the size of the spectrogram and the number of transformation K increased. On the other hand, too few local affine transformations K make the mapping very imprecise. We thus developed a novel efficient way to deal with the well established local maxima problem, referred to as *multi-scale initialization*. The idea is to train PPAM at different *scales*, *i.e.*, with a different number of transformation K each time, yielding to different sets of trained parameters $\widetilde{\theta}_K$ where, *e.g.*, $K = 1, 2, 4, 8 \ldots, 64$. When proceeding to the inverse mapping, we first run the VEM algorithm from a random initialization using $\widetilde{\theta}_1$. We then use the obtained masks and positions to initialize a new VEM algorithm using $\widetilde{\theta}_2$, then $\widetilde{\theta}_4$, and so forth until the desired value for K. To further improve the convergence of each *sub-scale* algorithm an additional constraint was added, referred to as *progressive masking*. During the first iteration, the mask of each source is constrained such that all the frequency bins of each time frame are assigned to the same source. This

method used	1 source		2 sources				3 sources			
	Az	El	Az	El	SDR	SIR	Az	El	SDR	SIR
VESSL T1	**2.1**±2.1	**1.1**±1.2	**4.7**±11	**2.9**±9.9	**3.8**±1.7	**6.1**±1.7	**17** ±34	**8.7**±19	**1.7**±1.5	2.1±1.5
VESSL T2	3.5±3.3	2.4±2.6	8.2±16	4.7±11	3.3±1.6	5.2±1.6	19 ±35	9.1±18	1.5±1.5	1.8±1.5
MESSL-G	5.6±9.4		14±21		2.3±1.6	6.0±4.3	18±28		1.3±1.2	**2.2**±4.4
mixture					0.0±2.5	0.2±2.5			-3.2±2.3	-3.0±2.3
oracle					12± 1.6	21 ±2.0			11± 1.7	20 ±2.1

Table 1. Comparing the average and standard deviation (Avg±Std) of azimuth (Az) and elevation (El) angular errors in degrees, as well as Signal to Distortion Ratio (SDR) and Signal to Inteferer Ratio (SIR) for 600 test sounds with 1 to 3 sources using different methods.

is done by adding a product over t in $q_{W_{dt}}(m)$ (3). Similarly to what is done in [14], this constraint is then progressively released at each iteration by dividing time frames in 2,4,8... frequency *blocks* until the total number of frequency bins is reached. These two strategies dramatically increased both the algorithm's performance and speed[1].

Algorithm termination: MAP estimates for missing data can be easily obtained at convergence of the algorithm by maximizing respectively the final $q^{(\infty)}_{X,Z}(\boldsymbol{x},\boldsymbol{z})$ and $q^{(\infty)}_{W}(\boldsymbol{w})$ probability distributions. We have $(\boldsymbol{X}^{MAP}_m, Z^{MAP}_m) = (\boldsymbol{\mu}^{(\infty)}_{\hat{k}m}, \hat{k})$ where $\hat{k} = \arg\max_{k=1:K} \alpha^{(\infty)}_{km} \ |\boldsymbol{\Sigma}^{(\infty)}_{km}|^{-1/2}$ and $W^{MAP}_{dt} = \arg\max_{m=1:M} q^{(\infty)}_{W_{dt}}(m)$. Note that as shown in Fig. 1, the algorithm not only provides MAP estimates, but also complete posterior distributions over both the 2D space of sound source positions and the space of binary masks.

4. EXPERIMENTS

The proposed algorithm (VESSL) was tested using the CAMIL dataset[2] [15] which consists of binaural recordings made in the presence of sound sources emitting white noise and random utterances from the TIMIT speech dataset. Recordings are all made in a reverberant room and are associated to the ground truth emitter's direction in the microphones' frame, *i.e.*, azimuth and elevation. $N = 9,600$ directions are available in the dataset, corresponding to 160 azimuths in the range $[-160°, 160°]$, 60 elevations in the range $[-60°, 60°]$ and an average angular distance between points (*density*) of $2°$. ILD spectrograms were obtained from the log-ratio between the left and right power spectrograms. Spectrograms were computed using short-time Fourier transform with a 64ms time-window and 8ms overlap, yielding $T = 126$ D-dimensional *ILD vectors* per second, where $D = 512$ corresponds to the number of frequencies in the range 0-8000Hz. The mapping parameters $\boldsymbol{\theta}$ were trained with PPAM using *mean ILD vectors*, *i.e*, the temporal mean of the ILD spectrograms, associated to the ground truth (azimuth and elevation) of the emitter. The training was done on recordings corresponding to white-noise emitters such that all the frequencies are present. In order to test the algorithm, we used both single source recordings and mixtures of 2 to 3 sources obtained by summing utterances emitted from different positions, so that at least two sources were emitting at the same time in 60% of the test sounds.

We evaluated VESSL using two sets of PPAM parameters. The first parameters were estimated from the training set *T1* with $N_1 = 9,600$ positions, density $\delta = 2°$ and using $K = 128$. The second parameters were estimated from the decimated set *T2* with $N_2 = 1,530$ positions (density $\delta = 5°$) and using $K = 64$. Localization and separation results are compared to the state-of-the-art EM-based sound source separation and localization algorithm MESSL [14] in table 1. 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 in terms of separation [3, 4, 16, 17]. Note that this algorithm, as well as the vast majority of existing source localization methods [3–5, 7, 9, 10], do not make use a training set 2D source locations and hence they only provide time difference of arrival for each source, *i.e.*, frontal azimuth and no elevation. For the comparison to be fair, results given for MESSL correspond to test with only frontal sources (azimuth in $[-90°, 90°]$). We evaluated separation performance using the standard metrics Signal to Distortion Ratio (SDR) and Signal to Interferer Ratio (SIR) introduced in [18]. SDR and SIR results of both methods were also compared to those obtained with the ground truth binary masks or *oracle masks* [3] and to those of the original mixture. Oracle masks provide an upper bound that cannot be reached in practice as it requires to know the original signals. Conversely, the mixture scores provide a lower bound, as no mask is applied.

5. CONCLUSION AND FUTURE WORK

With a similar computational time, VESSL outperforms state-of-the art separation scores from MESSL and performs accurate 2D localization in the challenging case of noisy real-world recordings of multiple sparse sound sources emitting from a wide range of directions, using spectral ILD only. This pushes VESSL forward, as a promising method for robustly addressing SSL using a training stage (calibration). Future work will include adding spectral interaural phase differences in the model, testing the robustness to changes in the rever-

[1] About $15\times$ real time speed for $M = 2$, $K = 64$ and MATLAB code.

[2] http://perception.inrialpes.fr/ Deleforge/CAMIL_Dataset/

berating properties of the room where the training has been performed, or using audiovisual training procedures [19, 20].

6. REFERENCES

[1] M. Aytekin, C. F. Moss, and J. Z. Simon, "A sensorimotor approach to sound localization," *Neural Computation*, vol. 20, no. 3, pp. 603–635, 2008.

[2] A. Deleforge and R. P. Horaud, "2D sound-source localization on the binaural manifold," in *IEEE International Workshop on Machine Learning for Signal Processing*, September 2012.

[3] O. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1830–1847, 2004.

[4] J. Mouba and S. Marchand, "A source localization/separation/respatialization system based on unsupervised classification of interaural cues," in *Int. Conf. on Digital Audio Effects*, 2006.

[5] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Proc. NIPS*, 2007, pp. 953–960.

[6] A. Deleforge and R. P. Horaud, "A latently constrained mixture model for audio source separation and localization," in *LVA/ICA*, Tel Aviv, Israel, March 2012, pp. 372–379.

[7] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *proc. DAFx*, 2003, pp. 209–213.

[8] A. R. Kullaib, M. Al-Mualla, and D. Vernon, "2d binaural sound localization: for urban search and rescue robotics," in *proc. Mobile Robotics*, Istanbul, Turkey, September 2009, pp. 423–435.

[9] R. Liu and Y. Wang, "Azimuthal source localization using interaural coherence in a robotic dog: modeling and application," *Robotica*, vol. 28, no. 7, pp. 1013–1020, 2010.

[10] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 20, no. 5, pp. 1503–1512, 2012.

[11] J. Hörnstein, M. Lopes, J. Santos-Victor, and F. Lacerda, "Sound localization for humanoid robots – building audio-motor maps based on the HRTF," in *IEEE/RSJ IROS*, 2006, pp. 1170–1176.

[12] F. Keyrouz, W. Maier, and K. Diepold, "Robotic localization and separation of concurrent sound sources using self-splitting competitive learning," in *Proc. of IEEE CI-ISP*, Hawaii, Apr. 2007, pp. 340–345.

[13] M. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," *Bayesian Statistics*, pp. 453–464, 2003.

[14] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE TASLP*, vol. 18, pp. 382–394, 2010.

[15] A. Deleforge and R. P. Horaud, "The cocktail party robot: Sound source separation and localisation with an active binaural head," in *ACM/IEEE HRI*, Boston, MA, March 2012.

[16] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech, Audio, Lang. Proc.*, 2005.

[17] H. Sawada, S. Araki, and S. Makino, "A Two-Stage Frequency-Domain Blind Source Separation Method for Underdetermined Convolutive Mixtures," in *Work. App. of Sig. Proc. to Audio and Acoustics*, 2007.

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

[19] V. Khalidov, F. Forbes, and R. P. Horaud, "Conjugate mixture models for clustering multimodal data," *Neural Computation*, vol. 23, no. 2, pp. 587–602, February 2011.

[20] V. Khalidov, F. Forbes, and R. P. Horaud, "Calibration of a binocular-binaural sensor using a moving audio-visual target," Tech. Rep. 7865, INRIA Grenoble Rhone-Alpes, January 2012.