ACOUSTIC SPACE LEARNING FOR SOUND-SOURCE SEPARATION AND LOCALIZATION ON BINAURAL MANIFOLDS

ANTOINE DELEFORGE

INRIA Grenoble Rhône-Alpes, 655 Avenue de l'Europe
Saint-Ismier, 38334, France
E-mail: antoine.deleforge@inria.fr

FLORENCE FORBES

INRIA Grenoble Rhône-Alpes, 655 Avenue de l'Europe
Saint-Ismier, 38334, France
E-mail: florence.forbes@inria.fr

RADU HORAUD

INRIA Grenoble Rhône-Alpes, 655 Avenue de l'Europe
Saint-Ismier, 38334, France
E-mail: radu.horaud@inria.fr

In this paper we address the problems of modeling the acoustic space generated by a full-spectrum sound source and of using the learned model for the localization and separation of multiple sources that simultaneously emit sparse-spectrum sounds. We lay theoretical and methodological grounds in order to introduce the *binaural manifold* paradigm. We perform an in-depth study of the latent low-dimensional structure of the high-dimensional interaural spectral data, based on a corpus recorded with a human-like audiomotor robot head. A non-linear dimensionality reduction technique is used to show that these data lie on a two-dimensional (2D) smooth manifold parameterized by the motor states of the listener, or equivalently, the sound source directions. We propose a *probabilistic piecewise affine mapping* model (PPAM) specifically designed to deal with high-dimensional data exhibiting an intrinsic piecewise linear structure. We derive a closed-form expectation-maximization (EM) procedure for estimating the model parameters, followed by Bayes inversion for obtaining the full posterior density function of a sound source direction. We extend this solution to deal with missing data and redundancy in real world spectrograms, and hence for 2D localization of natural sound sources such as speech. We further generalize the model to the challenging case of multiple sound sources and we propose a variational EM framework. The associated algorithm, referred to as *variational EM for source separation and localization* (VESSL) yields a Bayesian estimation of the 2D locations and time-frequency masks of all the sources. Comparisons of the proposed approach with several existing methods reveal that the combination of acoustic-space learning with Bayesian inference enables our method to outperform state-of-the-art methods.

Keywords: binaural hearing; sound localization; sound-source separation; manifold learning; mixture of regressors; EM inference.

1. Introduction

The remarkable abilities of humans to localize one or several sound sources and to identify their content from the perceived acoustic signals have been intensively studied in psychophysics[1], computational auditory analysis[2], and more recently in the emerging field of *robot hearing*[3]. A classical example which illustrates the difficulty of understanding these human skills, is the well known *cocktail party effect* introduced by Cherry[4] that still challenges today's methods[5]: How are listeners able to decipher speech in the presence of other sound sources, including competing talkers? While human listeners solve this problem routinely and effortlessly, this is still a challenge in computational audition.

There is behavioral and physiological evidence that human listeners use *binaural cues* in order to estimate the direction of a sound source[6] and that sound localization plays an important role for solving the cocktail party problem[7,2]. Two binaural cues seem to play an essential role, namely the interaural level difference (ILD) and the interaural time difference (ITD), or its spectral equivalent the interaural phase difference (IPD). Both ILD and IPD are known to be subject-dependent and frequency-dependent cues, due to the so-called *head related transfer function* (HRTF) generated by the shape of the head, pinna and torso, which filters signals arriving at each eardrum. These complex shapes induce a non-linear dependency of the HRTFs on the sound source direction[1,2]. It is known that the spatial information provided by interaural-difference cues within a restricted band of frequency is spatially ambiguous, particularly along a roughly vertical and front/back dimension[7]. This suggests that humans and mammals make use of full spectrum information for 2D sound source localization[8,9]. This is confirmed by biological models of the auditory system hypothesizing the existence of neurons dedicated to the computation of interaural cues in specific frequency bands[2].

A lot of computational techniques exist to extract ITD, ILD and IPD from binaural recordings, either in the time domain using cross-correlation[10,11], or in the time-frequency domain using Fourier analysis[12,13] or gammatone filters[14]. However, the problem of localizing and/or separating several sound sources remains a challenge in computational auditory scene analysis, for several reasons. Firstly, the mapping from interaural cues to sound-source positions is usually unknown, complex and non-linear due to the transfer function of microphones which cannot be easily modeled. Secondly, auditory data are corrupted by noise and reverberations. Thirdly, an interaural value at a given frequency 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 to one of the sources is hard to estimate. The first problem, *i.e.*, mapping audio cues to source positions, is central: Yet, it has received little attention in computational audition. Most existing approaches approximate this mapping based on simplifying assumptions, such as direct-path source-to-microphone propagation[15], a sine interpolation of ILD data from a human HRTF dataset[16], or a spiral ear model[17]. These simplifying assumptions are often not valid in real world conditions. Following this view, accurately modeling a real world binaural system would require a prohibitively high number of parameters including the exact shape and acoustic properties of the recording device and of the room, which are unaccessible in practice. Due to this difficulty, the vast majority of current binaural sound localization approaches mainly focus on a rough estimation of a frontal azimuth angle, or *one-dimensional* (1D) localization[10,12,14,16], and very few perform 2D localization[17]. Alternatively, some approaches[18,19,13] bypass the explicit mapping model and perform 2D localization by exhaustive search in a large HRTF look-up table associating source directions to interaural spectral cues. However, this process is unstable and hardly scalable in practice as the number of required associations yields prohibitive memory and computational costs.

On the other hand, a number of psychophysical studies have suggested that the ability of localizing sounds is learned at early stages of development in humans and mammals[20,21,22]. That is, the link between auditory features and source locations might not be hardcoded in the brain but rather learned from experience. The sensorimotor theory, originally laid by Poincaré[23] and more recently investigated by O'Regan[24], suggests that experiencing the sensory

consequences of voluntary motor actions is necessary for an organism to learn the perception of space. For example, Held and Hein[25] showed that neo-natal kittens deprived of the ability of moving while seeing could not develop vision properly. Most notably, Aytekin et al.[22] proposed a sensorimotor model of sound source localization using HRTF datasets of bats and humans. In particular, they argue that biological mechanisms could learn sound localization based solely on acoustic inputs and their relation to motor states.

In this article, we get inspiration from these psychological studies to propose a *supervised learning* computational paradigm for multiple sound source separation and localization. In other words, the acoustic properties of a system are first learned during a *training phase*. A key element of this study is the existence of *binaural manifolds*, which are defined and detailed in this article. Their existence is asserted through an in-depth study of the intrinsic structure of high-dimensional interaural spectral cues, based on real world recordings with a human-like auditory robot head. A non-linear dimensionality reduction technique is used to show that these cues lie on a two-dimensional (2D) smooth manifold parameterized by the motor states of the system, or equivalently, the sound source directions. With this key observation in mind, we propose a *probabilistic piecewise affine mapping* model (PPAM) specifically designed to deal with high-dimensional data presenting such a locally linear structure. We explain how the model parameters can be learned using a closed-form expectation-maximization (EM) procedure, and how Bayes inversion can be used to obtain the full posterior density function of a sound source direction given a new auditory observation. We also show how to extend this inversion to deal with missing data and redundancy in real world spectrograms, and hence for two-dimensional localization of natural sound sources such as speech. We further extend this inversion to the challenging case of multiple sound sources emitting at the same time, by proposing a variational EM (VEM) framework. The associated algorithm, referred to as *variational EM for source separation and localization* (VESSL) yields a Bayesian estimation of the 2D locations and time-frequency masks of all the sources. We introduced some parts of this paradigm in previous work[26,27],

and we propose in this article a unified, more detailed and in-depth presentation of the global methodology.

The remainder of this article is organized as follows. Section 1.1 provides a brief literature overview of sound source separation and space mapping. Section 2 describes how interaural cues are computed in our framework. Section 3 presents the audio-motor dataset specifically recorded for this study. Section 4 describes the manifold learning method used to prove the existence of smooth binaural manifolds parameterized by source directions. Section 5 describes the PPAM model, its inference using EM, as well as the associated inverse mapping formula extended to missing and redundant observations in spectrograms. Section 6 further extends the inversion to the case of multiple sound sources with the VESSL algorithm. Section 7 shows detailed performance of PPAM and VESSL in terms of mapping, sound separation and sound localization under various conditions. Results are compared to state-of-the-art methods. Finally, Section 8 concludes and provides directions for future work.

1.1. *Related work*

Sound Source Separation. The problem of sound source separation has been thoroughly studied in the last decades and several interesting approaches have been proposed. Some methods[28,29] achieve separation with a single microphone, based on known acoustic properties of speech signals, and are therefore limited to a specific type of input. Other techniques such as independent component analysis (ICA)[30] require as many microphones as the number of sources. A recently explored approach is to use Gaussian complex models for spectrograms and estimate source parameters using an expectation-maximization algorithm[31]. However, these methods are known to be very sensitive to initialization due to the large dimensionality of the parameter space, and often rely on external knowledge or other sound source separation algorithms for initialization.

Another category of methods uses binaural localization cues combined with *time-frequency masking* for source separation[16,18,32]. Time-frequency masking, also called binary masking, allows the separation of more sources than microphones from a mixed signal, with the assumption that a single source is ac-

tive at every time-frequency point. This is referred to as the *W-disjoint orthogonality* assumption[15] and it has been shown to hold, in general, for simultaneous speech signals. Recently, Mandel et al.[32] proposed a probabilistic model for multiple sound source localization and separation based on interaural spatial cues and binary masking. For each sound source, a binary mask and a discrete distribution over interaural time delays is provided. This can be used to approximate the frontal azimuth angle of the sound source using a direct-path sound propagation model, if the distance between the microphones is known.

Space Mapping. The task of learning a mapping between two spaces can be summarized as follows: if we are given a set of training couples $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N \subset \mathbb{R}^L \times \mathbb{R}^D$, how can we obtain a relationship between $\mathbb{R}^L$ and $\mathbb{R}^D$ such that given a new observation in one space, its associated point in the other space is deduced? This problem has been extensively studied in machine learning, and offers a broad range of applications. In this study, we are interested in mapping a high-dimensional space to a low dimensional space, where the relationship between the two spaces is approximately locally linear. Indeed, it is showed in Section 4 that high dimensional spectral interaural data lie on a smooth Riemanian manifold, which by definition is locally homeomorphic to an Euclidean space. High-to-low-dimensional mapping problems are often solved in two steps, *i.e.*, dimensionality reduction followed by regression. Methods falling into this category are partial least-squares[33], sliced inverse regression (SIR)[34] and kernel SIR[35], but they are not designed to model local linearity. An attractive probabilistic approach for local linearity is to use a mixture of locally linear sub-models. A number of methods have explored this paradigm in the Gaussian case, such as mixture of linear regressors[36], mixture of local experts[37] and methods based on joint Gaussian mixture models[38,39,40,41] (GMM), that were recently unified in a common supervised and unsupervised mapping framework[42]. Joint GMM has particularly been used in audio applications such as text-to-speech synthesis, voice conversion or articulatory-acoustic mapping systems. In Section 5, we propose a variant of the mixture of local experts model[37] specifically adapted to locally linear and high-to-low dimensional mapping with an attractive geometrical

interpretation. The concept of *acoustic space learning* introduced in this article, *i.e.*, the idea of learning a mapping from interaural cues to 2D source positions and exploiting this mapping for sound source separation and localization does not seem to have been explored in the past.

2. Extracting Spatial Cues from Sounds

To localize sound sources, one needs to find a representation of sounds that 1) is independent of the sound source *content*, *i.e.*, the emitted signal and 2) contains discriminative spatial information. These features can be computed in the time domain, but contain richer information when computed for different frequency channels, *i.e.*, in the time-frequency domain. This section presents the time-frequency representation and spatial features used in this study.

2.1. *Time-Frequency Representation*

A time-frequency representation can be obtained either using Gammatone filter banks[14], which are mimicking human auditory representation, or short-term Fourier transform (STFT) analysis[12,13]. The present study uses STFT as it is more directly applicable to source separation through binary-masking, as addressed in Section 6. First, the complex-valued spectrograms associated with the two microphones are computed with a 64 ms time-window and 8 ms hop, yielding $T = 126$ frames for a 1 s signal. These values proved to be a good compromise between computational time and time-frequency resolution. Since sounds are recorded at 16,000 Hz, each time window contains 1,024 samples which are transformed into $F = 512$ complex Fourier coefficients associated to positive frequency channels between 0 and 8,000 Hz. For a binaural recording made in the presence of a sound source located at $\boldsymbol{x}$ in a listener-centered coordinate frame, we denote with $\{s_{ft}^{(\mathrm{S})}\}_{f,t=1}^{F,T}$ the complex-valued spectrogram emitted by the source, and with $\{s_{ft}^{(\mathrm{L})}\}_{f,t=1}^{F,T}$ and $\{s_{ft}^{(\mathrm{R})}\}_{f,t=1}^{F,T}$ the left and right perceived spectrograms.

2.2. *Interaural Spectral Cues*

The HRTF model provides a relationships between the spectrogram emitted from source position

$\boldsymbol{x}$ and the perceived spectrograms:

$$\begin{cases} s_{ft}^{(\mathrm{L})} = h_f^{(\mathrm{L})}(\boldsymbol{x})\, s_{ft}^{(\mathrm{S})} \\ s_{ft}^{(\mathrm{R})} = h_f^{(\mathrm{R})}(\boldsymbol{x})\, s_{ft}^{(\mathrm{S})}. \end{cases} \quad (1)$$

where $\boldsymbol{h}^{(\mathrm{L})}$ and $\boldsymbol{h}^{(\mathrm{R})}$ denote the left and right nonlinear HRTFs. The *interaural transfer function* (ITF) is defined by the ratio between the two HRTFs, *i.e.*, $I_f(\boldsymbol{x}) = h_f^{(\mathrm{R})}(\boldsymbol{x})/h_f^{(\mathrm{L})}(\boldsymbol{x}) \in \mathbb{C}$. The interaural spectrogram is defined by $\hat{I}_{ft} = s_{ft}^{(\mathrm{R})}/s_{ft}^{(\mathrm{L})}$, so that $\hat{I}_{ft} = I_f(\boldsymbol{x})$. This way, $\hat{I}_{ft}$ does not depend on the emitted spectrogram value $s_{ft}^{(\mathrm{S})}$ but only on the emitting source position $\boldsymbol{x}$. However, this equality holds only if the source is emitting at (f,t) (*i.e.*, $s_{ft}^{(\mathrm{S})} \neq 0$). Since natural sounds have a null acoustic level at most time-frequency bins, associated interaural spectrograms will be very sparse, *i.e.*, they will have many *missing values*. To characterize missing interaural spectrogram values, we introduce the binary variables χ_{ft} so that $\chi_{ft} = 0$ if the value at (f,t) is missing and $\chi_{ft} = 1$ otherwise. They can be determined using a threshold on left and right spectral powers $|s_{ft}^{(\mathrm{L})}|^2$ and $|s_{ft}^{(\mathrm{R})}|^2$. We define the *ILD spectrogram* $\boldsymbol{\alpha}$ and the *IPD spectrogram* $\boldsymbol{\phi}$ as the log-amplitude and phase of the interaural spectrogram $\hat{I}_{f,t}$:

$$\begin{cases} \alpha_{ft} = 20\log|\hat{I}_{ft}| \in \mathbb{R}, \\ \phi_{ft} = \exp(j\arg(\hat{I}_{ft})) \in \mathbb{C}. \end{cases} \quad (2)$$

The phase difference is expressed in the complex domain, or equivalently $\mathbb{R}^2$, to avoid problems due to phase circularity. This way, two nearby phase values will be nearby in terms of Euclidean distance. In the particular case of a sound source emitting white noise from $\boldsymbol{x}$, we have $\chi_{ft} = 1$ for all (f,t), *i.e.*, the sound source is emitting at all (f,t) points. One can thus compute the temporal means $\bar{\boldsymbol{\alpha}} \in \mathbb{R}^F$ and $\bar{\boldsymbol{\phi}} \in \mathbb{R}^{2F}$ of ILD and IPD spectrograms. These vectors will be referred to as the *mean interaural vectors* associated with $\boldsymbol{x}$. The well established duplex theory[7] suggests that ILD cues are mostly used at high frequencies (above 2 kHz) while ITD (or IPD) cues are mostly used at low frequencies (below 2 kHz) in humans. Indeed, ILD values are similar at low frequencies because the HRTF can be neglected, and the phase difference becomes very unstable with respect to the source position at high frequencies. To account for these phenomena, the initial binaural cues are split into two distinct vectors, namely the mean *low*-ILD and *high*-ILD and the mean *low*-IPD and *high*-IPD vectors, where *low* corresponds to frequency channels between 0 and 2 kHz and *high* corresponds to frequency channels between 2 kHz and 8 kHz. Frequency channels below 300 Hz were also removed in *low*-IPD vectors as they did not prove to contain significant information, due to a high amount of noise in this frequency range.

3. The CAMIL Dataset: Audiomotor Sampling of the Acoustic Space

We developed a technique to gather a large number of interaural vectors associated with source positions, in an entirely unsupervised and automated way. Sound acquisition is performed with a Sennheiser MKE 2002 acoustic dummy-head linked to a computer via a Behringer ADA8000 Ultragain Pro-8 digital external sound card. The head is mounted onto a robotic system with two rotational degrees of freedom: a pan motion and a tilt motion (see Fig. 1). This device is specifically designed to achieve precise and reproducible movements[a]. The emitter – a loud-speaker – is placed at approximately 2.7 meters ahead of the robot, as shown on Fig. 1. The loud-speaker's input and the microphones' outputs are handled by two synchronized sound cards in order to simultaneously record and play. All the experiments are carried out in real-world conditions, *i.e.*, in a room with natural reverberations and background noise due to computer fans.

Rather than placing the emitter at known 3D locations around the robot, it is kept in a fixed reference position while the robot records emitted sounds from different motor states. Consequently, a sound source direction is directly associated with a pan-tilt motor state (ψ, θ). A great advantage of the audio-motor method is that it allows to obtain a very dense sampling of almost the entire acoustic space. The overall training is fully-automatic and takes around 15 hours, which could not be done manually. However, it also presents a limitation: A recording made at a given motor-state only approximates what would be perceived if the source was actually placed at the corresponding relative position in the room. This approximation holds only if the room has relatively few

[a]Full details on the setup at http://team.inria.fr/perception/popeye/

Fig. 1: A binaural dummy head is placed onto a motor system which can perform precise and reproducible pan and tilt motions (left). The emitter (a loud-speaker) is placed in front of the robot (right).

asymmetries and reverberations, which might not be the case in general. Note that when this is the case, a sound localization system trained with this dataset could be used to directly calculate the head movement pointing toward an emitting sound source. This could be done without needing inverse kinematics, distance between microphones or any other parameters. Alternatively, one could learn the room acoustics together with the system acoustics using, *e.g.*, visual markers related to an emitter successively placed in different positions around the static head. Interestingly, the remainder of this study does not depend on the specific way source-position-to-interaural-cues associations are gathered.

Recordings are made from $10,800$ uniformly spread motor states covering the entire reachable motor-state space of the robot: 180 pan rotations ψ in the range $[-180°, 180°]$ (left-right) and 60 tilt rotations θ in the range $[-60°, 60°]$ (top-down). Hence, the source location spans a $360°$ azimuth range and a $120°$ elevation range in the robot's frame, with $2°$ between each source direction. There is a one-to-one association between motor states and source directions and they are indifferently denoted by $\{\boldsymbol{x}_n\}_{n=1}^N \in \mathcal{X}$. Note that the space $\mathcal{X}$ has a cylindrical topology. The speaker is located in the far field of the head during the experiments (> 1.8 meters), and Otani

et al.[43] showed that HRTFs mainly depend on the sound source direction while the distance has less impact in that case. This is why sound source locations are expressed with two angles in this work.

For each $\boldsymbol{x}_n \in \mathbb{R}^2$, two binaural recordings are performed: (i) white noise which can be used to estimate $\bar{\boldsymbol{\alpha}}_n$ and $\bar{\boldsymbol{\phi}}_n$, and (ii) a randomly selected utterance amongst 362 samples from the TIMIT dataset[44]. These utterances are composed of 50% female, 50% male and they last 1 to 5 seconds. Sparse interaural spectrograms are computed from these recordings and are used to test our algorithms on natural sounds (Section 7). The resulting dataset is publicly available online and is referred to as the *Computational Audio-Motor Integration through Learning* (CAMIL) dataset[b].

4. The Manifold Structure of Interaural Cues

In this section, we analyze the intrinsic structure of the mean high- and low-ILD and -IPD vectors previously described. While these vectors live in a high-dimensional space, they might be parameterized by motor states and hence, could lie on a lower L-dimensional manifold, with $L = 2$. We propose to experimentally verify the existence of a Riemannian manifold structure[c] using non-linear dimensionality

---

[b]http://perception.inrialpes.fr/~Deleforge/CAMIL_Dataset/.
[c]by definition, a Riemannian manifold is locally homeomorphic to a Euclidean space.
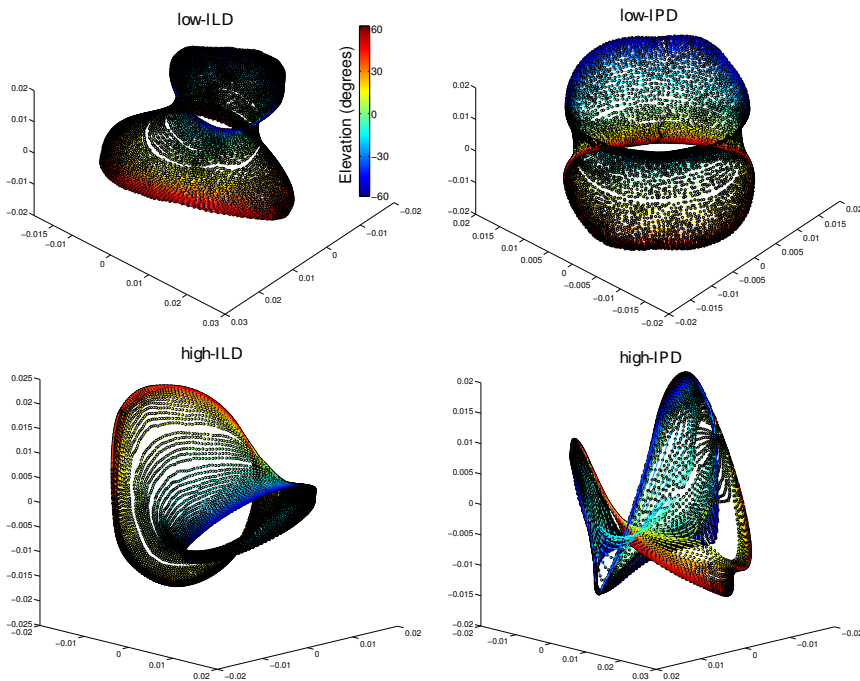
Fig. 2: Low-dimensional representations of mean interaural vectors using non-linear dimensionality reduction (LTSA). For visualization purpose, points with the same ground truth elevation are linked with a colored line in azimuth order. The three axes correspond to the first three eigenvectors of the global-alignment matrix. Obtained point clouds are zero-centered and arbitrary scaled.

Fig. 3: Low-dimensional representations of mean interaural vectors using linear dimensionality reduction (PCA) (azimuths in $[-90°, 90°]$). The two axes correspond to the first two eigenvectors of the correlation matrix.

reduction, and examine whether the obtained representations are homeomorphic to the motor-state space. Such a homeomorphism would allow us to validate or not the existence of a locally linear bijective mapping between motor states (or equivalently, source directions) and the interaural data gathered with our setup.

4.1. *Local Tangent Space Alignment*

If the interaural data lie in a linear subspace, linear dimensionality reduction methods such as principal component analysis (PCA), could be used. However, in the case of a non-linear subspace one should use a manifold learning technique, *e.g.*, diffusion kernels as done in[45]. We chose to use local tangent-space alignment (LTSA)[46] because it essentially relies on the assumption that the data are locally linear, which is our central hypothesis. LTSA starts by building a local neighborhood around each high-
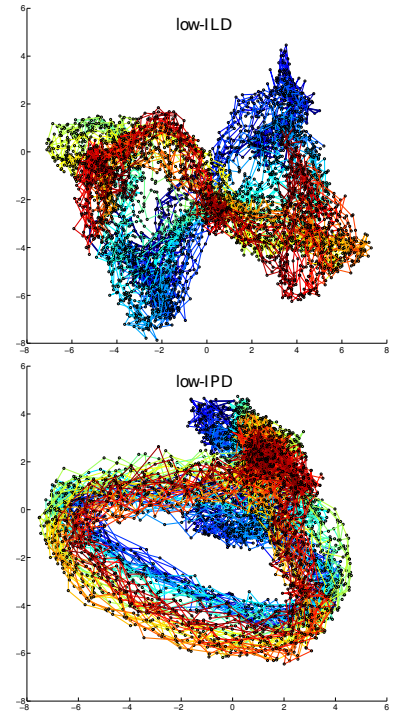
dimensional observation. Under the key assumption that each such neighborhood spans a linear space of low dimension L corresponding to the dimensionality of the tangent space, *i.e.*, a Riemannian manifold, PCA can be applied to each one of these neighborhoods thus yielding as many L-dimensional data representations as points in the dataset. Finally, a global map is built by optimal alignment of these local representations. This global alignment is done in the L-dimensional space by computing the L largest eigenvalue-eigenvector pairs of a *global alignment matrix* **B** (see[46] for details).

Two extensions are proposed to adapt LTSA to our data. Firstly, LTSA uses the *k-nearest neighbors* (*k*NN) algorithm to determine neighboring relationships between points, yielding neighborhoods of identical size k over the data. This has the advantage of always providing graphs with a single connected component, but it can easily lead to inappropriate edges between points, especially at boundaries or in

the presence of outliers. A simple way to overcome these artifacts is to implement a *symmetric* version of kNN, by considering that two points are connected if and only if each of them belongs to the neighborhood of the other one. Comparisons between the outputs of standard and symmetric kNN are shown in Fig. 4. Although symmetric kNN solves connection issues at boundaries, it creates neighborhoods of variable size, and in particular some points might get disconnected from the graph. Nevertheless, it turns out that ignoring such isolated points is an advantage since it may well be viewed as a way to remove outliers from the data. In our case the value of k is set manually; in practice any value in the range $[15, 25]$ yields satisfying results.
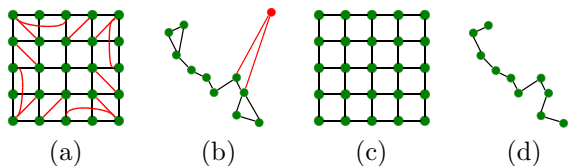

Fig. 4: Differences between standard kNN (a,b) and symmetric kNN (c,d) on a grid of points with boundaries ($k = 4$) and in the presence of an outlier ($k = 2$).

Secondly, LTSA is extended to represent manifolds that are homeomorphic to the 2D surface of a cylinder. The best way to visualize such a 2D curved surface is to represent it in the 3D Euclidean space and to visualize the 3D points lying on that surface. For this reason, we retain the $L + 1 = 3$ largest eigenvalue-eigenvector pairs of the global alignment matrix **B** such that the extracted manifolds can be easily visualized.

4.2. *Binaural Manifold Visualization*

Low-dimensional representations obtained using LTSA are shown in Fig. 2. Mean *low*-ILD, *low*-IPD, and *high*-ILD surfaces are all smooth and homeomorphic to the motor-state space (a cylinder), thus confirming that these cues can be used for 2D binaural sound source localization based on locally linear mapping. However, this is not the case for the mean *high*-IPD surface which features several distortions, elevation ambiguities, and crossings. While these computational experiments confirm the duplex theory for IPD cues, they surprisingly suggest that

ILD cues at low frequencies still contain rich 2D directional information. This was also hypothesized but not thoroughly analyzed in[32]. One can therefore concatenate full-spectrum ILD and low-frequency IPD vectors to form an observation space of dimension 730, referred to as the *ILPD space*. Similarly, one can define sparse *ILPD spectrograms* for general sounds. These results experimentally confirm the existence of *binaural manifolds*, *i.e.*, a locally-linear, low-dimensional structure hidden behind the complexity of interaural spectral cues obtained from real world recordings in a reverberant room.

For comparison, Fig. 3 shows the result of applying PCA to $5,400$ mean *low*-ILD vectors and *low*-IPD vectors corresponding to frontal sources (azimuths in $[-90°, 90°]$). The resulting representations are extremely distorted, due to the non-linear nature of binaural manifolds. This rules out the use of a linear regression method to estimate the interaural-to-localization mapping and justifies the development of an appropriate piecewise-linear mapping method, detailed in the next section.

5. Probabilistic Acoustic Space Learning and Single Sound Source Localization

The manifold learning technique described above allows retrieving intrinsic two-dimensional spatial coordinates from auditory cues and shows the existence of a smooth, locally-linear relationship between the two spaces. We now want to use these results to address the more concrete problem of localizing a sound source. To do this, we need a technique that maps any high-dimensional interaural vector onto the $L = 2$ dimensional space of sound source direction: azimuth and elevation. We will refer to the process of establishing such a mapping as *acoustic space learning*. To be applicable to real world sound source localization, the mapping technique should feature a number of properties. First, it should deal with the sparsity of natural sounds mentioned in Section 2, and hence handle missing data. Second, it should deal with the high amount of noise and redundancy present in the interaural spectrograms of natural sounds, as opposed to the clean vectors obtained by averaging white noise interaural spectrograms during training (Section 2). Finally, it should allow further extension to the more complex

case of mixtures of sound sources that will be addressed in Section 6. An attractive approach embracing all these properties is to use a Bayesian framework. We hence view the sound source localization problem as a probabilistic space mapping problem. This strongly contrasts with traditional approaches in sound source localization, which usually assume the mapping to be known, based on simplified sound propagation models. In the following section we present the proposed model, which may be viewed as a variant of the mixture of local experts model[37] with a geometrical interpretation.

5.1. *Probabilistic Piecewise Affine Mapping*

Notations Capital letters indicate random variables, and lower case their realizations. We both use the following equivalent notations: $p(V = v)$ and $p(v)$ for the probability or density of variable V at value v. Subscripts $n \in \{1 \ldots N\}$, $d \in \{1 \ldots D\}$, $k \in \{1 \ldots K\}$, $t \in \{1 \ldots T\}$ and $m \in \{1 \ldots M\}$ are respectively indexes over training observations, interaural vector components, affine transformations, spectrogram time frames and sources.

As mentioned in Section 3, the complete space of sound source positions (or motor states) used in the last sections has a cylindrical topology. However, to make the inference of the mapping model described below analytically tractable, it is preferable for the low-dimensional space to have a linear (Euclidean) topology. Indeed, this notably allows standard Gaussian distributions to be defined over that space which are easy to deal with in practice. To do so, we will simply use a subset of the complete space of sound source positions, corresponding to azimuths between $-160°$ and $160°$ degrees. This subset will be used throughout the article. To generalize the present work to a cylindrical space of source positions, one could imagine learning two different mapping models: one for frontal sources and one for rearward sources. Localization could then be done based on a mixture of these two models.

Let $\mathcal{X} \subset \mathbb{R}^L$ denote the low-dimensional space of source positions and $\mathbb{R}^D$ the high-dimensional observation space, *i.e.*, the space of interaural cues. The computational experiments of Section 4 suggest that there exists a smooth, locally linear bijection $g : \mathcal{X} \rightarrow$ $\mathcal{Y} \subset \mathbb{R}^D$ such that the set $\mathcal{Y} = \{g(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}\}$ forms an L−dimensional manifold embedded in $\mathbb{R}^D$. Based on this assumption, the proposed idea is to compute a piecewise-affine probabilistic approximation of g from a training data set $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N \subset \mathcal{X} \times \mathcal{Y}$ and to estimate the inverse of g using a Bayesian formulation. The local linearity of g suggests that each point $\boldsymbol{y}_n$ is the image of a point $\boldsymbol{x}_n \in \mathcal{R}_k \subset \mathcal{X}$ by an affine transformation τ_k, plus an error term. Assuming that there is a finite number K of such affine transformations τ_k and an equal number of associated regions $\mathcal{R}_k$, we obtain a piecewise-affine approximation of g. An assignment variable $z_n \in \{1 \ldots K\}$ is associated with each training pair $(\boldsymbol{x}_n, \boldsymbol{y}_n)$ such that $z_n = k$ if $\boldsymbol{y}_n$ is the image of $\boldsymbol{x}_n \in \mathcal{R}_k$ by τ_k. This allows us to write ($\mathbb{I}_{\{z_n=k\}} = 1$ if $z_n = k$, and 0 otherwise):

$$\boldsymbol{y}_n = \sum_{k=1}^K \mathbb{I}_{\{z_n=k\}}(\boldsymbol{A}_k\boldsymbol{x}_n + \boldsymbol{b}_k) + \boldsymbol{e}_n \qquad (3)$$

where the $D \times L$ matrix $\mathbf{A}_k$ and the vector $\boldsymbol{b}_k \in \mathbb{R}^D$ define the transformation τ_k, and $\boldsymbol{e}_n \in \mathbb{R}^D$ is an error term capturing both the observation noise and the reconstruction error of affine transformations. If we make the assumption that the error terms $\boldsymbol{e}_n$ do not depend on $\boldsymbol{x}_n$, $\boldsymbol{y}_n$ or $\boldsymbol{z}_n$, and are independent identically distributed realizations of a Gaussian variable with $\mathbf{0}$ mean and diagonal covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\sigma_{1:D}^2)$, we obtain:

$$p(\boldsymbol{y}_n|\boldsymbol{x}_n, Z_n = k; \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}_n; \mathbf{A}_k\boldsymbol{x}_n + \boldsymbol{b}_k, \boldsymbol{\Sigma}) \quad (4)$$

where $\boldsymbol{\theta}$ designates all the model parameters (see (9)). To make the affine transformations local, we set z_n to the realization of a hidden multinomial random variable Z_n conditioned by $\boldsymbol{x}_n$:

$$p(Z_n = k|\boldsymbol{x}_n; \boldsymbol{\theta}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)}{\sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)}, \qquad (5)$$

where $\boldsymbol{c}_k \in \mathbb{R}^L$, $\boldsymbol{\Gamma}_k \in \mathbb{R}^{L \times L}$ and $\sum_k \pi_k = 1$. We can give a geometrical interpretation of this distribution by adding the following *volume equality* constraints to the model:

$$|\boldsymbol{\Gamma}_1| = \cdots = |\boldsymbol{\Gamma}_K| \text{ and } \pi_1 = \cdots = \pi_K = 1/K \quad (6)$$

One can verify that under these constraints, the set of K regions of $\mathcal{X}$ maximizing (5) for each k defines a Voronoi diagram of centroids $\{\boldsymbol{c}_k\}_{k=1}^K$, where the Mahalanobis distance $||.||_{\boldsymbol{\Gamma}_k}$ is used instead of the Euclidean one. This corresponds to a compact probabilistic way of representing a general partitioning of the low-dimensional space into convex regions of

equal volume. Extensive tests on simulated and audio data showed that these constraints yield lower reconstruction errors, on top of providing a meaningful interpretation of (5). To make our generative model complete, we define the following Gaussian mixture prior on $\boldsymbol{X}$:

$$p(\boldsymbol{X}_n = \boldsymbol{x}_n; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k) \quad (7)$$

which allows (5) and (7) to be rewritten in a simpler form:

$$p(\boldsymbol{X}_n = \boldsymbol{x}_n, Z_n = k; \boldsymbol{\theta}) = \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k). \quad (8)$$

The graphical model of PPAM is given in Fig. 6(a) and Fig. 5 shows a partitioning example obtained with PPAM using a toy data set.
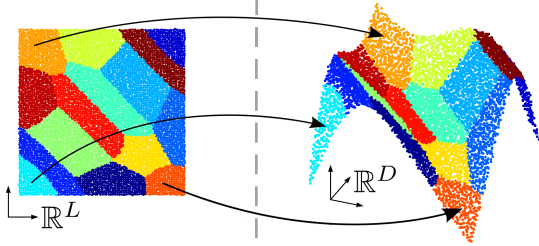

Fig. 5: Space partitioning and locally affine mapping on a toy data set ($N = 9600, K = 15, L = 2, D = 3$). Colors encode regions in $\mathbb{R}^D$ maximizing (5). Observe how these regions (associated with affine transformations) are adjusted to the geometry of the observation manifold.

The inference of PPAM can be achieved with a closed-form and efficient EM algorithm maximizing the observed-data log-likelihood $\log p(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta})$ with respect to the model parameters:

$$\boldsymbol{\theta} = \left\{ \{\boldsymbol{\Gamma}_k, \boldsymbol{c}_k, \mathbf{A}_k, \boldsymbol{b}_k\}_{k=1}^K, \boldsymbol{\Sigma} \right\}. \quad (9)$$

The EM algorithm is made of two steps: E (Expectation) and M (Maximization). In the E-step posteriors $r_{kn}^{(i)} = p(Z_n = k | \boldsymbol{x}_n, \boldsymbol{y}_n; \boldsymbol{\theta}^{(i-1)})$ at iteration i are computed from (4), (5) and Bayes inversion. In the M-step we maximize the expected complete-data log-likelihood $\mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta}^{(i)})}[\log p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} | \boldsymbol{\theta})]$ with respect to parameters $\boldsymbol{\theta}$. We obtain the following closed-form expressions for the parameters updates

under the volume equality constraints (6):

$$\boldsymbol{c}_k^{(i)} = \sum_{n=1}^N \frac{r_{kn}^{(i)}}{\bar{r}_k^{(i)}} \boldsymbol{x}_n, \quad \boldsymbol{\Gamma}_k^{(i)} = \frac{\mathbf{S}_k^{(i)}}{|\mathbf{S}_k^{(i)}|^{\frac{1}{L}}} \sum_{j=1}^K \frac{\bar{r}_j^{(i)}}{N} |\mathbf{S}_j^{(i)}|^{\frac{1}{L}}$$

$$\mathbf{A}_k^{(i)} = \overline{\mathbf{Y}}_k^{(i)} \overline{\mathbf{X}}_k^{(i)\dagger}, \quad \boldsymbol{b}_k^{(i)} = \sum_{n=1}^N \frac{r_{kn}^{(i)}}{\bar{r}_k^{(i)}} (\boldsymbol{y}_n - \mathbf{A}_k^{(i)} \boldsymbol{x}_n),$$

$${\sigma^2}_d^{(i)} = \frac{1}{K} \sum_{k=1}^K \sum_{n=1}^N \frac{r_{kn}^{(i)}}{\bar{r}_k^{(i)}} (y_{dn} - \boldsymbol{a}_{dk}^{(i)\top} \boldsymbol{x}_n - b_{dk}^{(i)})^2 \quad (10)$$

where † is the Moore-Penrose pseudo inverse operator, $(.,.)$ denotes horizontal concatenation and:

$$\mathbf{S}_k^{(i)} = \sum_{n=1}^N r_{kn}^{(i)} / \bar{r}_k^{(i)} (\boldsymbol{x}_n - \boldsymbol{c}_k^{(i)})(\boldsymbol{x}_n - \boldsymbol{c}_k^{(i)})^\top$$

$$\bar{r}_k^{(i)} = \sum_{k=1}^K r_{kn}^{(i)}, \quad \mathbf{A}_k^{(i)} = (\boldsymbol{a}_{1k}^{(i)}, \dots, \boldsymbol{a}_{Dk}^{(i)})^\top$$

$$\overline{\mathbf{X}}_k^{(i)} = (r_{k1}^{(i)\frac{1}{2}}(\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_k^{(i)}) \dots r_{kN}^{(i)\frac{1}{2}}(\boldsymbol{x}_N - \bar{\boldsymbol{x}}_k^{(i)}))$$

$$\overline{\mathbf{Y}}_k^{(i)} = (r_{k1}^{(i)\frac{1}{2}}(\boldsymbol{y}_1 - \bar{\boldsymbol{y}}_k^{(i)}) \dots r_{kN}^{(i)\frac{1}{2}}(\boldsymbol{y}_N - \bar{\boldsymbol{y}}_k^{(i)}))$$

$$\bar{\boldsymbol{x}}_k^{(i)} = \sum_{n=1}^N r_{kn}^{(i)} / \bar{r}_k^{(i)} \boldsymbol{x}_n, \quad \bar{\boldsymbol{y}}_k^{(i)} = \sum_{n=1}^N r_{kn}^{(i)} / \bar{r}_k^{(i)} \boldsymbol{y}_n.$$

Initial posteriors $r_{kn}^{(0)}$ can be obtained either by estimating a K-GMM solely on $\boldsymbol{X}$ or on joint data $[\boldsymbol{X}; \boldsymbol{Y}]$ ($[.;.]$ denotes vertical concatenation) and then go on with the M-step (10). The latter strategy generally provides a better initialization and hence a faster convergence, at the cost of being more computationally demanding.

Given a set of parameters $\boldsymbol{\theta}$, a mapping from a test position $\boldsymbol{x}_t \in \mathbb{R}^L$ to its corresponding interaural cue $\boldsymbol{y}_t \in \mathbb{R}^D$ is obtained using the *forward conditional density* of PPAM, *i.e.*, $p(\boldsymbol{y}_t | \boldsymbol{x}_t; \boldsymbol{\theta}) =$

$$\frac{\sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k) \mathcal{N}(\boldsymbol{y}_t; \mathbf{A}_k \boldsymbol{x}_t + \boldsymbol{b}_k, \boldsymbol{\Sigma})}{\sum_{j=1}^K \pi_j \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{c}_j, \boldsymbol{\Gamma}_j)} \quad (11)$$

while a mapping from an interaural cue to its corresponding source position is obtained using the *inverse conditional density*, *i.e.*, $p(\boldsymbol{x}_t | \boldsymbol{y}_t; \boldsymbol{\theta}) =$

$$\frac{\sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*) \mathcal{N}(\boldsymbol{x}_t; \mathbf{A}_k^* \boldsymbol{y}_t + \boldsymbol{b}_k^*, \boldsymbol{\Sigma}_k^*)}{\sum_{j=1}^K \pi_j \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{c}_j^*, \boldsymbol{\Gamma}_j^*)} \quad (12)$$

where $\boldsymbol{c}_k^* = \mathbf{A}_k \boldsymbol{c}_k + \boldsymbol{b}_k$, $\boldsymbol{\Gamma}_k^* = \boldsymbol{\Sigma} + \mathbf{A}_k \boldsymbol{\Gamma}_k \mathbf{A}_k^\top$, $\mathbf{A}_k^* = \boldsymbol{\Sigma}^* \mathbf{A}_k^\top \boldsymbol{\Sigma}^{-1}$, $\boldsymbol{b}_k^* = \boldsymbol{\Sigma}_k^* (\boldsymbol{\Gamma}_k^{-1} \boldsymbol{c}_k - \mathbf{A}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{b}_k)$, and $\boldsymbol{\Sigma}_k^* = (\boldsymbol{\Gamma}_k^{-1} + \mathbf{A}_k^\top \boldsymbol{\Sigma}^{-1} \mathbf{A}_k)^{-1}$. Note that both (11) and (12) take the form of a Gaussian mixture distribution in one space given an observation in the other space. These Gaussian mixtures are parameterized in two different ways by the observed data
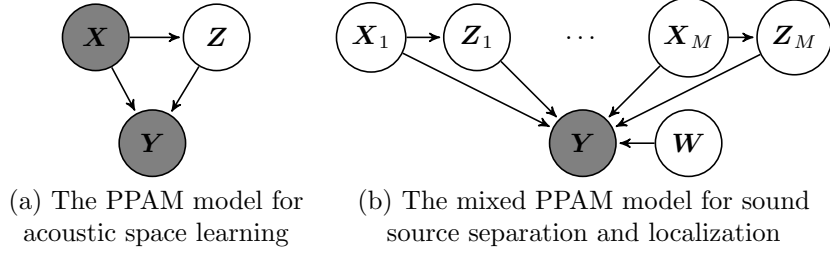
(a) The PPAM model for acoustic space learning

(b) The mixed PPAM model for sound source separation and localization

Fig. 6: Graphical models of PPAM and mixed PPAM. White means unobserved, gray means observed.

and PPAM's parameters. One can use their expectations $\mathbb{E}[\boldsymbol{Y}_t|\boldsymbol{x}_t;\boldsymbol{\theta}]$ and $\mathbb{E}[\boldsymbol{X}_t|\boldsymbol{y}_t;\boldsymbol{\theta}]$ to obtain *forward* and *inverse* mapping functions. The idea of learning the mapping from the low-dimensional space of source position to the high-dimensional space of auditory observations and then inverting the mapping for sound source localization using (12) is crucial. It implies that the number of parameters $\boldsymbol{\theta}$ to estimate is $K(D(L+2)+L+L^2+1)$ while it would be $K(L(D+2)+D+D^2+1)$ if the mapping was learned the other way around. The latter number is 130 times the former for $D = 512$ and $L = 2$, making the EM estimation impossible in practice due to over-parameterization.

5.2. *Localization From Sparse Spectrograms*

So far, we have considered mapping from a vector space to another vector space. However, as detailed in Section 2, spectrograms of natural sounds are time series of possibly noisy sparse vectors. The Bayesian framework presented above allows to deal with this situation straightforwardly. Let $\widetilde{\boldsymbol{\theta}}$ denote a set of parameters estimated with PPAM and let $\boldsymbol{Y}_{\boldsymbol{\chi}} = \{y_{dt}; \chi_{dt}\}_{t,d=1}^{T,D}$ be an observed sparse interaural spectrogram (χ_{dt} denotes missing values as defined in Section 2.2). Note that d may represent the index of an ILD or an IPD value at a given frequency. If we suppose that all the observations are assigned to the same source position $\boldsymbol{x}$ and transformation z, it follows from the model (4), (5), (7) that the posterior distribution $p(\boldsymbol{x}|\boldsymbol{Y}_{\boldsymbol{\chi}};\widetilde{\boldsymbol{\theta}})$ is a GMM $\sum_{k=1}^{K} \rho_k \mathcal{N}(\boldsymbol{x};\boldsymbol{m}_k,\mathbf{V}_k)$ in $\mathbb{R}^L$ with parameters:

$$\boldsymbol{m}_k = \mathbf{V}_k\left(\widetilde{\boldsymbol{\Gamma}}_k^{-1}\widetilde{\boldsymbol{c}}_k + \sum_{d,t=1}^{D,T} \frac{\chi_{dt}}{\widetilde{\sigma}_d^2}\widetilde{\boldsymbol{a}}_{dk}(y_{dt}-\widetilde{b}_{dk})\right), \quad (13)$$

$$\mathbf{V}_k = \left(\widetilde{\boldsymbol{\Gamma}}_k^{-1} + \sum_{d,t=1}^{D,T}\frac{\chi_{dt}}{\widetilde{\sigma}_d^2}\widetilde{\boldsymbol{a}}_{dk}\widetilde{\boldsymbol{a}}_{dk}^\top\right)^{-1} \text{and} \quad (14)$$

$$\rho_k \propto \frac{|\mathbf{V}_k|^{\frac{1}{2}}}{|\widetilde{\boldsymbol{\Gamma}}_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\Big(\textstyle\sum_{d,t=1}^{D,T} \frac{\chi_{dt}}{\widetilde{\sigma}_d^2}(y_{dt}-\widetilde{b}_{dk})^2 \right.$$
$$\left. + \widetilde{\boldsymbol{c}}_k^\top\widetilde{\boldsymbol{\Gamma}}_k^{-1}\widetilde{\boldsymbol{c}}_k + \boldsymbol{m}_k^\top\mathbf{V}_k^{-1}\boldsymbol{m}_k\Big)\right) \quad (15)$$

where the weights $\{\rho_k\}_{k=1}^K$ are normalized to sum to 1. This formulation is more general than the unique, complete observation case ($T = 1, \boldsymbol{\chi} = \mathbf{1}$) provided by (12). The posterior expectation $\mathbb{E}[\boldsymbol{x}|\boldsymbol{Y}_{\boldsymbol{\chi}}] = \sum_{k=1}^K \rho_k\boldsymbol{m}_k$ can be used to obtain an estimate of the sound source position given an observed spectrogram. Alternatively, one may use the full posterior distribution and, for instance, combine it with other external probabilistic knowledge to increase the localization accuracy or extract higher order information.

6. Extension to Multiple Sound Sources

We now extend the single sound source localization method described in Section 5 to multiple sound source separation and localization. With our model, this problem can 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 extend the PPAM model presented in the previous section to this more general case and propose a variational expectation-maximization (VEM) approach[47] to solve for the model inference. The VEM algorithm described below will be referred to as *variational EM for sound separation and localization* (VESSL). Typical examples of the algorithm's inputs and outputs are shown in Fig. 7.

6.1. *The Mixed PPAM Model*

Given a time series of T interaural cues $\boldsymbol{Y} =$

$\{\boldsymbol{Y}_t\}_{t=1}^T \subset \mathbb{R}^D$, we seek the M emitting sound source positions, denoted by $\boldsymbol{X} = \{\boldsymbol{X}_m\}_{m=1}^M \subset \mathbb{R}^L$ (M is assumed to be known). To deal with mixed data, we introduce a *source assignment* variable $\boldsymbol{W} = \{W_{dt}\}_{d=1,t=1}^{D,T}$ such that $W_{dt} = m$ when Y_{dt} is generated from source m. The only observed data are the interaural cues $\boldsymbol{Y}$ while all the other variables $\boldsymbol{W} \in \mathcal{W}$, $\boldsymbol{X} \in \mathcal{X}$ and $\boldsymbol{Z} \in \mathcal{Z}$ are hidden. To account for $\boldsymbol{W}$, the observation model rewrites $p(\boldsymbol{y}_t | \boldsymbol{w}_t, \boldsymbol{x}, \boldsymbol{z}) = \prod_d p(y_{dt} | w_{dt}, \boldsymbol{x}_{w_{dt}}, z_{w_{dt}})$ where $p(y_{dt} | W_{dt} = m, \boldsymbol{X}_m = \boldsymbol{x}_m, Z_m = k) = \mathcal{N}(y_{dt}; \boldsymbol{a}_{dk}^\top \boldsymbol{x}_m + \boldsymbol{b}_{dk}, \sigma_d^2)$. 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 and frequency, so that $p(\boldsymbol{w}) = \prod_{d,t} p(w_{dt})$. We define the prior on source assignments by $p(W_{dt} = m) = \lambda_{dm}$, where λ_{dm} are positive numbers representing the relative presence of each source in each frequency channel (source weights), so that $\sum_{m=1}^M \lambda_{dm} = 1$ for all d. We will write $\boldsymbol{\Lambda} = \{\lambda_{dm}\}_{d=1,m=1}^{D,M}$. Finally, source assignments and positions are assumed independent, so that we get the following hierarchical decomposition of the full model: $p(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\psi}) =$

$$p(\boldsymbol{Y} | \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\psi}) p(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\psi}) p(\boldsymbol{W}; \boldsymbol{\psi}) \qquad (16)$$

where

$$\boldsymbol{\psi} = \{\{\boldsymbol{\Gamma}_k, \boldsymbol{c}_k, \mathbf{A}_k, \boldsymbol{b}_k\}_{k=1}^K, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}\} \qquad (17)$$

denotes the complete set of model parameters. This extended PPAM model for multiple sound sources will be referred to as the *mixed PPAM* model and is represented with a graphical model in Fig. 6(b).

Notice that the training stage, where position-to-interaural-cue couples $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N \subset \mathcal{X} \times \mathcal{Y}$ are given (Section 5), may be viewed as a particular instance of mixed PPAM where $T = M = N$ and with $\boldsymbol{X}$ and $\boldsymbol{W}$ being completely known. Amongst the parameters $\boldsymbol{\psi}$ of mixed PPAM, the values of $\{\boldsymbol{\Gamma}_k, \boldsymbol{c}_k, \mathbf{A}_k, \boldsymbol{b}_k\}_{k=1}^K$ have thus already been estimated during this training stage, and only the parameters $\{\boldsymbol{\Sigma}, \boldsymbol{\Lambda}\}$ remain to be estimated, while $\boldsymbol{X}$ and $\boldsymbol{W}$ are now hidden variables. $\boldsymbol{\Sigma}$ is re-estimated to account for possibly higher noise levels in the mixed observed signals compared to training.

6.2. Inference with Variational EM

From now on, $\mathbb{E}q$ denotes 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

$$q^{(i)}(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{z}) = q_W^{(i)}(\boldsymbol{w}) \, q_{X,Z}^{(i)}(\boldsymbol{x}, \boldsymbol{z}) \qquad (18)$$

where $q_W^{(i)}$ and $q_{X,Z}^{(i)}$ are probability distributions on $\mathcal{W}$ and $\mathcal{X} \times \mathcal{Z}$ respectively. Such a factorization may seem drastic but its main beneficial effect is to replace potentially complex 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 goals are to update $q_{X,Z}$ and q_W in turn. Closed-form expressions for these sub-steps at iteration (i), initialization strategies, and the algorithm termination are detailed below.

E-XZ step: The update of $q_{X,Z}$ is given by:

$$q_{X,Z}^{(i)}(\boldsymbol{x}, \boldsymbol{z}) \propto \exp \mathbb{E}q_W^{(i-1)}[\log p(\boldsymbol{x}, \boldsymbol{z} | \boldsymbol{y}, \boldsymbol{W}; \boldsymbol{\psi}^{(i)})].$$

It follows from standard algebra that

$$q_{X,Z}^{(i)}(\boldsymbol{x}, \boldsymbol{z}) = \prod_{m=1}^M \alpha_{km}^{(i)} \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{km}^{(i)}, \mathbf{S}_{km}^{(i)})$$

where $\boldsymbol{\mu}_{km}^{(i)}, \mathbf{S}_{km}^{(i)}, \alpha_{km}^{(i)}$ are given in (19),(20) and the weights $\{\alpha_{km}^{(i)}\}_{k=1}^K$ are normalized to sum to 1 over k for all m. One can see this step as the *localization step*, since it corresponds to estimating a mixture of Gaussians over the latent space $\mathcal{X}$ of positions for each source. When $M = 1$, $\boldsymbol{W}$ is entirely determined and $q_{W_{dt}} = \chi_{dt}$. Thus, we can directly obtain the probability density $q_{X,Z}$ of the sound source position using (19),(20), and we recover exactly the single-source formula (13), (14), (15).

E-W step: The update of q_W is given by:

$$q_W^{(i)}(\boldsymbol{w}) \propto \exp \mathbb{E}q_{X,Z}^{(i)}[\log p(\boldsymbol{w} | \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\psi}^{(i)})].$$

It follows that $q_W^{(i)}(\boldsymbol{w}) = \prod_{d,t} q_{W_{dt}}^{(i)}(w_{dt})$ where $q_{W_{dt}}^{(i)}$ is given in (21) and is normalized to sum to 1 over m. This can be seen as the *sound source separation step*, as it provides the probability of assigning each observation to each source.

M step: We need to solve for:

$$\boldsymbol{\psi}^{(i+1)} = \text{argmax}_{\boldsymbol{\psi}} \, \mathbb{E}q_W^{(i)} q_{X,Z}^{(i)}[\log p(\boldsymbol{y}, \boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\psi})].$$
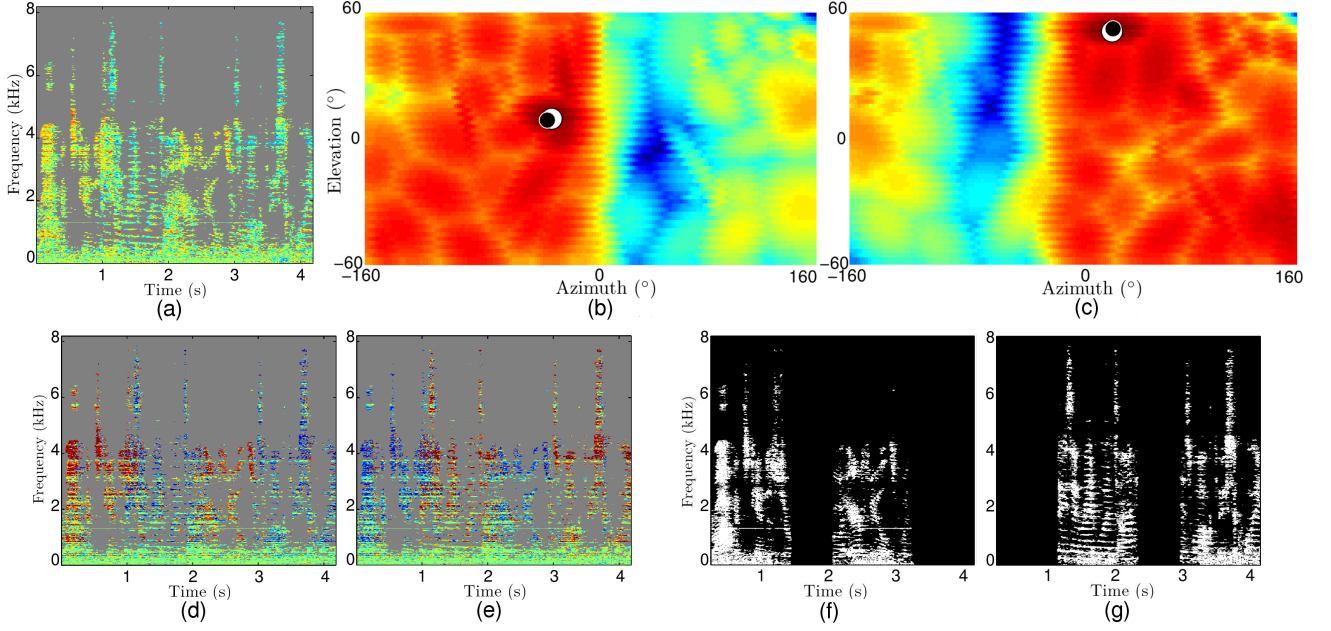
Fig. 7: (a) Input ILD spectrogram. (b,c) Output log-density of each source position as determined by $q_{X,Z}^{(\infty)}$. Ground-truth source positions are noted with a black dot, and the peak of the log-density with a white circle. (d,e) Output source assignment probabilities $q_W^{(\infty)}$. (f,g) Ground truth binary masks. Red color denotes high values, blue color low values, and grey colors missing observations.

$$\boldsymbol{\mu}_{km}^{(i)} = \mathbf{S}_{km}^{(i)}\left(\boldsymbol{\Gamma}_k^{-1}\boldsymbol{c}_k + \sum_{d,t}\sigma_d^{-2}q_{W_{dt}}^{(i-1)}(m)(y_{dt} - b_{dk})\boldsymbol{a}_{dk}\right), \quad \mathbf{S}_{km}^{(i)} = \left(\boldsymbol{\Gamma}_k^{-1} + \sum_{d,t}\sigma_d^{-2}q_{W_{dt}}^{(i-1)}(m)\boldsymbol{a}_{dk}\boldsymbol{a}_{dk}^{\top}\right)^{-1}, \quad (19)$$

$$\alpha_{km}^{(i)} \propto \frac{|\mathbf{S}_{km}^{(i)}|^{\frac{1}{2}}}{|\boldsymbol{\Gamma}_k|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\sum_{d,t}\frac{\chi_{dt}}{\sigma_d^2}(y_{dt} - \widetilde{b}_{dk})^2 + \boldsymbol{c}_k^{\top}\boldsymbol{\Gamma}_k^{-1}\boldsymbol{c}_k + \boldsymbol{\mu}_{km}^{(i)\top}\mathbf{S}_{km}^{(i)-1}\boldsymbol{\mu}_{km}^{(i)}\right)\right\}, \quad (20)$$

$$q_{W_{dt}}^{(i)}(m) \propto \chi_{dt}\lambda_{dm}^{(i)}\prod_{k=1}^{K}\exp\left\{-\frac{\alpha_{km}^{(i)}}{2\sigma_d^2}\left(\text{tr}(\mathbf{S}_{km}^{(i)}\boldsymbol{a}_{dk}\boldsymbol{a}_{dk}^{\top}) + (y_{dt} - \boldsymbol{a}_{dk}^{\top}\boldsymbol{\mu}_{km}^{(i)} - b_{dk})^2\right)\right\}, \quad (21)$$

$$\lambda_{dm}^{(i)} = \frac{1}{T}\sum_t q_{W_{dt}}^{(i)}(m), \ \sigma_d^{2(i)} = \frac{\sum_{t,m,k}q_{W_{dt}}^{(i)}(m)\,\alpha_{km}^{(i)}\left(\text{tr}(\mathbf{S}_{km}^{(i)}\boldsymbol{a}_{dk}\boldsymbol{a}_{dk}^{\top}) + (y_{dt} - \boldsymbol{a}_{dk}^{\top}\boldsymbol{\mu}_{km}^{(i)} - b_{dk})^2\right)}{\sum_{t,m,k}q_{W_{dt}}^{(i)}(m)\,\alpha_{km}^{(i)}}. \quad (22)$$

This reduces to the update of the source weights $\boldsymbol{\Lambda}^{(i)}$ and noise variances $\boldsymbol{\Sigma}^{(i)} = \text{diag}(\sigma_1^{2(i)}...\sigma_D^{2(i)})$ as given by (22).

Initialization strategies: Extensive real world experiments show that the VEM objective function, called the variational free energy, have a large number of local maxima. This may be due to the combinatorial size of the set of all possible binary masks $\mathcal{W}$ and the set of all possible affine transformation assignments $\mathcal{Z}$. Indeed, the procedure turns out 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 transformations K are increased. On the other hand, too few local affine transformations K make the mapping very impre-

cise. We thus developed a new 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 transformations K at each scale, yielding different sets of trained parameters $\widetilde{\theta}_K$ where, *e.g.*, $K = 1, 2, 4, 8\ldots, 64$. When proceeding with 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 the algorithm at each scale, an additional constraint is added, namely *progressive masking*. During the first iteration, the mask

Method	ILD only		IPD only		ILPD	
used	Azimuth	Elevation	Azimuth	Elevation	Azimuth	Elevation
PPAM	**2.2±1.9**	**1.6±1.6**	**1.5±1.3**	**1.5±1.4**	**1.8 ±1.6**	**1.6±1.5**
MPLR	2.4±2.2	2.2±2.1	1.8±1.7	1.9±1.7	2.2 ±2.0	2.1±2.0
SIR-1	41±34	17±13	36±25	24±17	41±34	16±13
SIR-2	27±28	14±13	32±23	11±11	26±28	14±13

Table 1: Comparing the average and standard deviation (Avg±Std) of azimuth and elevation angular errors in degrees using different types of interaural vectors obtained from white noise recordings and different mapping techniques.

of each source is constrained such that all the frequency bins of each time frame are assigned to the same source. This is done by adding a product over t in the expression of $q_{W_{dt}}^{(1)}(m)$ (21). Similarly to what is done in[32], this constraint is then progressively released at each iteration by dividing time frames in $2, 4, 8 \ldots, F$ frequency blocks until the total number of frequency bins F is reached. Combining these two strategies dramatically increases both the performance and speed[d] of the proposed VESSL algorithm.

Algorithm termination: Maximum a posteriori (MAP) estimates for missing data can be easily obtained at convergence of the algorithm by maximizing respectively the final probability distributions $q_{X,Z}^{(\infty)}(\boldsymbol{x}, \boldsymbol{z})$ and $q_W^{(\infty)}(\boldsymbol{w})$. We have $(\boldsymbol{X}_m^{MAP}, Z_m^{MAP}) = (\boldsymbol{\mu}_{km}^{(\infty)}, \hat{k})$ where $\hat{k} = \text{argmax}_k \alpha_{km}^{(\infty)} |\boldsymbol{\Sigma}_{km}^{(\infty)}|^{-1/2}$ and $W_{dt}^{MAP} = \text{argmax}_m q_{W_{dt}}^{(\infty)}(m)$. Note that as shown in Fig. 7, the algorithm not only provides MAP estimates, but also complete posterior distributions over both the 2D space of sound source positions $\mathcal{X}$ and the space of source assignments $\mathcal{W}$. Using the final assignment probabilities $q_{W_{dt}}^{(\infty)}(m)$ of source m, one can multiply recorded spectrogram values $s_{ft}^{(L)}$ and $s_{ft}^{(R)}$ by 1 if m has the highest emission probability at (f, t), and 0 otherwise (binary masking). This approximates the spectrogram $\{s_{ft}^{(m)}\}_{f,t=1}^{F,T}$ emitted by source m, from which the original temporal signal can be recovered using inverse Fourier transform, hence achieving sound source separation.

7.   Results on Acoustic Space Mapping, Source Localization and Sound Separation

We first compare PPAM to two other existing mapping methods, namely mixture of probabilistic linear regression (MPLR[41]), and sliced inverse regression (SIR[34]). MPLR may be viewed as a unifying view of joint GMM mapping techniques, which consists of estimating a standard Gaussian mixture model on joint variables $[\boldsymbol{X}; \boldsymbol{Y}]$ and uses conditional expectations to infer the mapping. Joint GMM has been widely used in audio applications[38,39,40,41]. SIR quantizes the low-dimensional data $\boldsymbol{X}$ into *slices* or clusters which in turn induces a quantization of the $\boldsymbol{Y}$-space. Each $\boldsymbol{Y}$-slice (all points $\boldsymbol{y}_n$ that map to the same $\boldsymbol{X}$-slice) is then replaced with its mean and PCA is carried out on these means. The resulting low-dimensional representation is then informed by $\boldsymbol{X}$ values through the preliminary slicing. We selected one (SIR-1) or two (SIR-2) principal axes for dimensionality reduction, 20 slices (the number of slices is known to have little influence on the results), and polynomial regression of order three (higher orders did not show significant improvements). We use $K = 64$ for both PPAM and MPLR. All techniques are trained on three types of cues: full spectrum ILD only, low-IPD only and ILPD cues (see Section 2 and 4). For each type of cue, the training is done on 10 random subsets of $4,800$ mean interaural vectors obtained from white noise recordings emitted by sources spanning a $[-160°, 160°]$ azimuth range and $[-60°, 60°]$ elevation range (section 3), and tested on the mean interaural vectors of the remaining $4,800$ source positions ($96,000$ tests in total). The resulting angular errors are shown in Table 1. As it may be seen, PPAM always performs best, validating the choice of piecewise-affine probabilistic technique adapted to the manifold structure of interaural data. The poor performance of SIR can be explained by the important non-linear nature of the data. Although pure IPD cues show slightly better localization results than pure ILD cues, it is better

---

[d]About 15 times real time speed for a mixture of 2 sources and $K = 64$ using a MATLAB implementation on a standard PC.
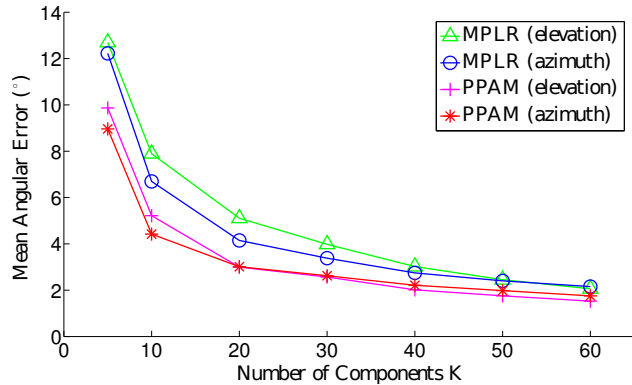
Fig. 8: Comparison of the proposed method (PPAM) with MPLR (Qiao and Minematsu 2009) as a function of the number of components.

Fig. 9: Average and standard deviation of angular errors in white noise localization as a function of the training set's sparsity.

to use both to maximize the available information in the case of noisy and/or missing data. Other experiments in this section are therefore carried out using ILPD cues.

Fig. 8 shows the influence of the number of affine components K on PPAM's and MPLR's localization results using similar test and training sets. Better performance of PPAM can be explained by the fact that the number of parameters to estimate in joint GMM methods increases quadratically with the dimensionality, and thus becomes prohibitively high using ILPD data ($D = 730$). Unsurprisingly, the localization error decreases when K increases. In practice, too high values of K (less than 20 samples per affine component) may lead to some degenerate covariance matrices in components where there are too few samples. Such components are simply removed along the execution of the algorithm, thus reducing K. In other words, the only free parameter K of PPAM can be chosen based on a compromise between computational cost and precision.

Fig. 9 shows the influence of the training set's *sparsity* on PPAM's localization results. We represent the sparsity by an angle δ, corresponding to the average pan and tilt absolute difference between a point and its nearest neighbor in the training set. The complete training set of sparsity 2° was uniformly decimated at random to obtain irregularly spaced and smaller training sets, while the test positions were randomly chosen from the remaining ones. For a given sparsity δ, 10 different decimated sets were used for training, and 100 source positions
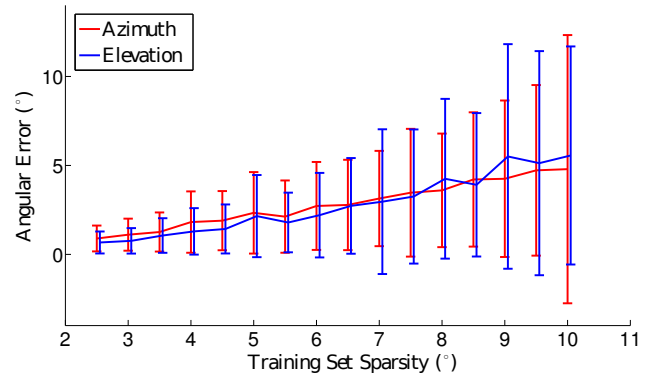
were estimated for each one, *i.e.*, 1000 localization tasks. K was chosen to have approximately 30 training samples per affine component. The mean and standard deviation of the errors in azimuth and in elevation are shown in Fig. 9. The mean localization errors are always smaller than half the training set sparsity, which illustrates the interpolation ability of our method. In addition, the error's standard deviation remains reasonable even for heavily decimated training sets (384 points only for $\delta = 10°$), thus showing that the overall performance is not much affected by the distribution and size of the training set being used. With the automatic audio-motor technique described in Section 3, recording a training set of sparsity 2° (9, 600 points) requires 2 hours and 30 minutes while it takes 25 minutes with a sparsity of 5° (1, 530 points).

We also tested the ability of PPAM to localize a single sound source using real world recordings of randomly located sources emitting random utterances (see Section 3), using the technique presented in Section 5.2. We use both the complete set of ILPD cues *T1* (sparsity $\delta = 2°$) and a partial set *T2* of sparsity $\delta = 5°$ for training. We set $K = 128$ for *T1* and $K = 64$ for *T2*. Test sounds are chosen to be outside of the training set *T2*. The spectral power threshold is manually set quite high, so that test spectrograms has 89.6% of missing data, on average. Results are compared to a baseline sound localization method, namely PHAT-histogram[48]. As the vast majority of existing source localization methods[15,49,12,16,10,14], PHAT-histogram provides a

Method	2 sources				3 sources			
used	Az	El	SDR	SIR	Az	El	SDR	SIR
VESSL($T1$)	**4.7±11**	**2.9±9.9**	**3.8±1.7**	**6.1±1.7**	**12 ±21**	**8.7±19**	**1.7±1.5**	**2.2±1.5**
VESSL($T2$)	8.2±16	4.7±11	3.3±1.6	5.2±1.6	16 ±24	9.1±18	1.5±1.5	1.8±1.5
MESSL-G	14±21		2.3±1.6	6.0±4.3	18±28		1.3±1.2	2.1±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 2: 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 mixtures of 2 to 3 sources using different methods.

time difference of arrival (TDOA) which can only be used to estimate frontal azimuth. The link between TDOAs and azimuths is estimated using linear regression on our dataset. For the comparison to be fair, PHAT is only tested on frontal sources. Comparison with the few existing binaural 2D sound source localization methods[17,18] is not possible because they rely on additional knowledge of the system being used that are not available. Means and standard deviations of azimuth (Az) and elevation (El) errors (Avg±Std) are shown in Table 3.

Method	1 source	
used	Az ($°$)	El ($°$)
PPAM($T1$)	**2.1 ± 2.1**	**1.1 ± 1.2**
PPAM($T2$)	3.5 ± 3.3	2.4 ± 2.6
PHAT	7.6 ± 9.4	

Table 3: Average single sound source localization error in degrees using PHAT and PPAM with two different training sets.

The proposed acoustic space learning approach dramatically outperforms the baseline. No front-back azimuth or elevation confusions is observed using our method, thanks to the asymmetry of the dummy head and to the spatial richness of interaural spectral cues.

We finally tested the ability of VESSL (section 6) to localize and separate multiple sound sources emitting at the same time. We use both training sets *T1* and *T2* for training and mixtures of 2 to 3 sources for testing[e]. Mixtures are obtained by summing binaural recordings of sources emitting random utterances from random positions (see Section 3) so that at least two sources were emitting at the same time in 60% of the test sounds. Localization

and separation results are compared to state-of-the-art EM-based sound source separation and localization algorithm MESSL[32]. MESSL does not rely on acoustic space learning, and estimates a time difference of arrival for each source. As for PHAT-histogram, results given for MESSL correspond to tests with frontal sources only. The version MESSL-G that is 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[15,50,49,51]. We evaluate separation performance using the standard metrics signal to distortion ratio (SDR) and signal to interferer ratio (SIR) introduced in[52]. SDR and SIR scores of both methods are also compared to those obtained with the ground truth binary masks or *oracle masks*[15] and to those of the original mixture. Oracle masks provide an upper bound for binary masking methods that cannot be reached in practice because it requires knowledge of the original signals. Conversely, the mixture scores provide a lower bound, as no mask is applied. Table 2 shows that VESSL significantly outperforms MESSL-G both in terms of separation and localization, putting forward acoustic space learning as a promising tool for accurate sound source localization and separation.

8. Conclusion and Future Work

We showed the existence of binaural manifolds, *i.e.* an intrinsic locally-linear, low-dimensional structure hidden behind the complexity of interaural data obtained from real world recordings. Based on this key property, we developed a probabilistic framework to learn a mapping from sound source positions to interaural cues. We showed how this framework

---

[e]The binaural algorithms considered performed equally poorly on mixtures of 4 sources or more.

could be used to accurately localize in both azimuth and elevation and separate mixtures of natural sound sources. Results show the superiority of acoustic space learning compared to other sound source localization techniques relying on simplified mapping models.

In this work, auditory data are mapped to the motor-states of a robotic system, but the same framework could be exploited with mappings of different nature. Typically, for *unsupervised* sound source localization where one only has access to non-annotated auditory data from different locations, one may map auditory cues to their intrinsic coordinates obtained using manifold learning, as explained in Section 4. Such coordinates are spatially consistent, *i.e.*, two sources that are near will yield near coordinates, but are not linked to a physical quantity and may therefore be hard to interpret or evaluate. Alternatively, auditory data could be annotated with the pixel position of the sound source in an image, yielding audio-visual mapping instead of audio-motor mapping.

A direction for future work is to study the influence of changes in the reverberating properties of the room where the training is performed. The PPAM model could be extended by adding latent components modeling such changes[42], thus becoming more robust to the recording environment. Another open problem is determining the number of sound sources M in the mixture, which is generally challenging in source separation. In our framework, it corresponds to a model selection problem, which could be addressed using, *e.g.*, a Bayesian information theoretic analysis.

References

1. J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, (1997).
2. D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, IEEE Press, (2006).
3. A. Deleforge and R. P. Horaud, "The cocktail party robot: Sound source separation and localisation with an active binaural head," In *Proc. of the 7th ACM/IEEE International Conference on Human Robot Interaction (HRI)*, (2012).
4. E. C. Cherry, "Some experiment on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, **25**(5), 975–979, September (1953).
5. S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, **17**, 1875–1902, (2005).
6. L. Rayleigh, "On our perception of sound direction," *Philos. Mag.*, **13**, 214–232, April (1907).
7. J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annual Review of Psychology*, **42**, 135–159, (1991).
8. R. S. Woodworth and H. Schlosberg, *Experimental Psychology*, Holt, (1965).
9. P. M. Hofman and A. J. Van Opstal, "Spectro-temporal factors in two-dimensional human sound localization," *The Journal of the Acoustical Society of America*, **103**(5), 2634–2648, (1998).
10. R. Liu and Y. Wang, "Azimuthal source localization using interaural coherence in a robotic dog: modeling and application," *Robotica*, **28**(7), 1013–1020, (2010).
11. X. Alameda-Pineda and R. P. Horaud, "Geometrically constrained robust time delay estimation using non-coplanar microphone arrays," In *Proceeding of the 20th European Signal Processing Conference (EUSIPCO)*, 1309–1313, (2012).
12. M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," In *Proc. of the Neural Information Processing Systems (NIPS) conference*, 953–960, (2007).
13. A. Deleforge and R. P. Horaud, "A latently constrained mixture model for audio source separation and localization," In *Proc. of the 10th International Conference, LVA/ICA*, 372–379, (2012).
14. J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Transactions on Acoustics, Speech, Signal Processing*, **20**(5), 1503–1512, (2012).
15. O. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, **52**, 1830–1847, (2004).
16. H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," In *Proc. of the International Conference on Digital Audio Effects (DAFX)*, 209–213, (2003).
17. A. R. Kullaib, M. Al-Mualla, and D. Vernon, "2d binaural sound localization: for urban search and rescue robotics," In *Proc. Mobile Robotics*, 423–435, (2009).
18. F. Keyrouz, W. Maier, and K. Diepold, "Robotic localization and separation of concurrent sound sources using self-splitting competitive learning," In *Proc. of the IEEE Symposium on Computational Intelligence in Image and Signal Processing (CIISP)*, 340–345, (2007).
19. J. Hörnstein, M. Lopes, J. Santos-Victor, and F. Lacerda, "Sound localization for humanoid robots – building audio-motor maps based on the HRTF," In *Proc. of the IEEE/RSJ IROS International Conference on Intelligent Robots and Systems*, 1170–1176,

(2006).

20. P. M. Hofman, J. G. Van Riswick, A. J. Van Opstal, et al., "Relearning sound localization with new ears," *Nature neuroscience*, **1**(5), 417–421, (1998).

21. B. A. Wright and Y. Zhang, "A review of learning with normal and altered sound-localization cues in human adults," *International journal of audiology*, **45**(S1), 92–98, (2006).

22. M. Aytekin, C. F. Moss, and J. Z. Simon, "A sensorimotor approach to sound localization," *Neural Computation*, **20**(3), 603–635, (2008).

23. H. Poincaré, *The foundations of science; Science and hypothesis, the value of science, science and method*, New York: Science Press, (1929), Halsted, G. B. trans. of La valeur de la science, 1905.

24. J. K. O'Regan and A. Noe, "A sensorimotor account of vision and visual consciousness," *Behavioral and Brain Sciences*, **24**, 939–1031, (2001).

25. R. Held and A. Hein, "Movement-produced stimulation in the development of visually guided behavior," *J. Comp. Physiol. Psych.*, **56**(5), 872–876, (1963).

26. A. Deleforge and R. P. Horaud, "2D Sound-Source Localization on the Binaural Manifold," In *IEEE Workshop on Machine Learning for Signal Processing* (IEEE, Santander, Spain, 2012).

27. A. Deleforge, F. Forbes, and R. P. Horaud, "Variational EM for Binaural Sound-Source Separation and Localization," In *ICASSP 2013 - 38th International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Vancouver, Canada, 2013).

28. S. T. Roweis, "One microphone source separation," In *Advances in Neural Information Processing Systems*, volume 13, 793–799. MIT Press, (2000).

29. S. Bensaid, A. Schutz, and D. T. M. Slock, "Single microphone blind audio source separation using EM-Kalman filter and short+long term AR modeling," In *Latent Variable Analysis and Signal Separation*, 106–113, (2010).

30. P. Comon and C. Jutten, *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, Academic Press (Elsevier), (2010).

31. N.Q.K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio Signal and Language Processing*, **18**(7), 1830–1840, (2010).

32. M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech and Language Processing*, **18**(2), 382–394, (2010).

33. R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," *Subspace, Latent Structure and Feature Selection*, **3940**, 34–51, (2006).

34. K. C. Li, "Sliced inverse regression for dimension reduction," *The Journal of the American Statistical Association*, **86**(414), 316–327, (1991).

35. H. M. Wu, "Kernel sliced inverse regression with applications to classification," *Journal of Computational and Graphical Statistics*, **17**(3), 590–610, (2008).

36. R. D. de Veaux, "Mixtures of linear regressions," *Comput. Stat. Data Anal.*, **8**(3), 227–245, (1989).

37. L. Xu, M. I. Jordan, and G. E. Hinton, "An alternative model for mixtures of experts," In *Proc. of the Neural Information Processing Systems (NIPS) conference*, volume 7, 633–640, (1995).

38. A. Kain and MW Macon, "Spectral voice conversion for text-to-speech synthesis," In *proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, (1998).

39. Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Acoustics, Speech, Signal Processing.*, **6**, 131–142, (1998).

40. T. Toda, A. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, **50**(3), 215–227, (2008).

41. Y. Qiao and N. Minematsu, "Mixture of probabilistic linear regressions: A unified view of GMM-based mapping techniques," In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 3913–3916, (2009).

42. Antoine Deleforge, Florence Forbes, and Radu Horaud, "Mapping Learning with Partially Latent Output, Research report, August (2013).

43. M. Otani, T. Hirahara, and S. Ise, "Numerical study on source-distance dependency of head-related transfer functions," *The Journal of the Acoustical Society of America*, **125**(5), 3253–61, (2009).

44. J. S. Garofolo, L. F. Lamel, and W. M. Fisher, "The darpa timit acoustic-phonetic continuous speech corpus cdrom", (1993).

45. R. Talmon, I. Cohen, and S. Gannot, "Supervised source localization using diffusion kernels," In *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 245 –248, (2011).

46. Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM Journal on Scientific Computing*, **26**(1), (2004).

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

48. P. Aarabi, "Self-localizing dynamic microphone arrays," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, **32**(4), 474–484, (2002).

49. J. Mouba and S. Marchand, "A source localization/separation/respatialization system based on unsupervised classification of interaural cues," In *Proc. of the International Conference on Digital Audio Ef-*

fects, (2006).

50. H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Transactions on Audio, Speech and Language Processing*, **13**(1), 120–134, (2005).

51. H. Sawada, S. Araki, and S. Makino, "A Two-Stage Frequency-Domain Blind Source Separation Method for Underdetermined Convolutive Mixtures," In *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (2007).

52. E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, **14**(4), 1462–1469, (2006).