HAL
open science

2D Sound-Source Localization on the Binaural Manifold

Antoine Deleforge, Radu Horaud

To cite this version:

Antoine Deleforge, Radu Horaud. 2D Sound-Source Localization on the Binaural Manifold. MLSP 2012 - IEEE Workshop on Machine Learning for Signal Processing, Sep 2012, Santander, Spain. pp.1-6, 10.1109/MLSP.2012.6349784 . hal-00768657

HAL Id: hal-00768657

https://inria.hal.science/hal-00768657

Submitted on 22 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

2D SOUND-SOURCE LOCALIZATION ON THE BINAURAL MANIFOLD

Antoine Deleforge and Radu Horaud

INRIA Grenoble Rhône Alpes

ABSTRACT

The problem of 2D sound-source localization based on a robotic binaural setup and audio-motor learning is addressed. We first introduce a methodology to experimentally verify the existence of a locally-linear bijective mapping between sound-source positions and high-dimensional interaural data, using manifold learning. Based on this local linearity assumption, we propose an novel method, namely probabilistic piecewise affine regression, that learns the localization-to-interaural mapping and its inverse. We show that our method outperforms two state-of-the art mapping methods, and allows to achieve accurate 2D localization of natural sounds from real world binaural recordings.

1. INTRODUCTION

The human extraordinary ability of localizing one or several sound sources from the perceived acoustic signals has been intensively studied in cognition [1], in computational audition [2], and more recently in the emerging field of *robot hearing* [3]. The most commonly used cues in binaural computational sound-source localization (SSL) are interaural cues, namely the ITD (interaural time difference), the IPD (interaural phase difference) and the ILD (interaural level difference). A lot of techniques exist to evaluate their values, either in the time domain using cross-correlation [4], or in the time-frequency domain using Fourier analysis [5, 6] or gammatone filters [7]. Once such cues are computed they need to be mapped to a source position. The vast majority of current SSL approaches mainly focus on frontal azimuth estimation, i.e., 1D localization [4, 5, 7, 8], while very few perform 2D localization [9]. All these approaches rely on a simplifying geometric or parametric model of sound propagation in the binaural system. Simple models assume a direct path propagation from source to microphones (single attenuation coefficient and delay) while more advanced ones use a spherical-head model for ITD propagation (Woodworths formula), approximate ILD data from a human *head related transfer function* (HRTF) dataset with a sine function [8] or rely on a spiral ear model [9]. In all these cases, extra parameters are needed such as the distance between microphones, the radius of the head, the sound speed, the shape of the ear or the sine coefficients. Following this view, the number of parameters required to accurately model a real world binaural system may become prohibitively high including the exact shape of the recording device, of the room, and all their acoustic properties. This is often unaccessible in practice, which is an important limitation for real-world SSL. On the other hand, in the particular case of a hearing robot, other sensory data may be available, such as the motor states (proprioception) or source position based on vision.

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 the latent space $\mathbb{R}^L$ and the observation space $\mathbb{R}^D$ 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 audio, it was notably used in text-to-speech synthesis [10], voice conversion [11] or articulatory-acoustic mapping systems [12], but never thoroughly examined for SSL. In this study, we focus on a particular class of mapping techniques where the relationship between the two spaces is approximately locally linear. This approximation is particularly relevant for data lying on a smooth Riemanian manifold, which by definition is locally homeomorphic to an Euclidean space. Given a set of K unknown linear transformations, locally linear mapping can be split in (i) assigning training couples to transformations and (ii) using linear regression to estimate the parameters of each transformation. These two tasks can be achieved within a probabilistic framework, where observations are seen as the realization of random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, while assignments to transformations are modeled as hidden variables $\boldsymbol{Z}$. Although the mixture of linear regression model (MLR) [13] represents data by a set of linear transformations, the transformations are not local since $\boldsymbol{Z}$ is independent of $\boldsymbol{X}$. This is unsuitable for manifold data. Mapping methods based on Gaussian mixture models (GMM) [10, 11, 12], widely used in audio applications, were recently unified by the mixture of probabilistic linear regression model (MPLR) [14]. In MPLR, posteriors $p(\boldsymbol{Z}|\boldsymbol{x})$ or $p(\boldsymbol{Z}|\boldsymbol{x}, \boldsymbol{y})$ are first obtained by estimating a GMM with the EM algorithm on latent data or joint data. Then, transformations parameters are estimated with weighted lin-

978-1-4673-1026-0/12/$31.00 ©2012 IEEE

ear regression. As the two tasks are achieved sequentially, the partitioning of the latent space is not optimal in the maximum likelihood sense. Conversely, the mixture of experts model (MoE) and its probabilistic view [15] allows to jointly optimize the partitioning of the latent space and local transformation parameters using EM. However, MoE is a generic model inspired by gating networks, and was not specifically designed for piecewise linear data.

In this paper we show that high-dimensional binaural observations, parameterized by low-dimensional motor-state parameters, enjoy an intrinsic manifold structure, and hence are locally linear. With this property in mind, we propose a novel SSL method based on a *probabilistic piecewise affine regression* model (PPAR) specifically designed to deal with high-dimensional acoustic data that have an intrinsic low-dimensional manifold structure. PPAR is based on a generative model that is used to learn the mapping between a set of motor states and associated binaural cues. We show that this mapping can be inverted in closed-form to obtain the full posterior density function $p(\boldsymbol{X}|\boldsymbol{y})$ in the latent space given a new observation $\boldsymbol{y}$. We further generalize this inversion to the case of missing and/or redundant observations in order to solve for *two-dimensional* SSL of natural, sparse sounds. The proposed method uses an audio-motor learning stage with white noise and does not require any a priori knowledge on the system's geometry and parameters.

The remainder of this paper is organized as follows. Section 2 describes the binaural model and the audio-motor paradigm. Section 3 validates the manifoldness of the binaural data. Section 4 describes the PPAR method and its associated Bayesian inverse mapping extended to missing and redundant observations. Section 5, compares PPAR with MPLR [14], and shows SSL results obtained with PPAR. Section 6 concludes and provides directions for future work.

2. THE AUDIO-MOTOR MODEL AND DATA

Any SSL method needs a content-independent sound representation containing as much spatial information as possible. Rather than single ILD and ITD coefficients, we use frequency-dependent ILD and IPD values calculated from a short-term Fourier transform (STFT) analysis, as done in e.g. [5, 6]. First, the complex-valued spectrograms associated with the two microphones are computed with a 64ms time-window and 8ms overlap, yielding $T = 126$ frames for a 1s signal. Since sounds are recorded at 16,000Hz, each time window contains 1,024 samples which are transformed into $F = 512$ complex Fourier coefficients associated to frequency channels between 0 and 8,000Hz. For a binaural recording made in the presence of a single 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. The HRTF model provides a relationships between the emitted and the perceived spectrograms points:

$$s_{ft}^{(\mathrm{L})} = h_f^{(\mathrm{L})}(\boldsymbol{x})\, s_{ft}^{(\mathrm{S})} \;\text{ and }\; s_{ft}^{(\mathrm{R})} = h_f^{(\mathrm{R})}(\boldsymbol{x})\, s_{ft}^{(\mathrm{S})} \qquad (1)$$

where $\boldsymbol{h}^{(\mathrm{L})}$ and $\boldsymbol{h}^{(\mathrm{R})}$ denote the left and right non-linear HRTFs. The *interaural transfer function* (ITF) is defined by the ratio between the two HRTFs, i.e., $I_f(\boldsymbol{x}) = h_f^{(\mathrm{R})}/h_f^{(\mathrm{L})} \in \mathbb{C}$. The interaural spectrogram is defined by $\hat{I}_{ft} = s_{ft}^{(\mathrm{R})}/s_{ft}^{(\mathrm{L})}$, so that $\hat{I}_{ft} \approx I_f(\boldsymbol{x})$. This way, $\hat{I}_{ft}$ does not depend on the emitted spectrogram value $s_{ft}^{(S)}$ but only on the emitting source position $\boldsymbol{x}$. However, this approximation holds only if the source is emitting at (f,t) (i.e. $s_{ft}^{(\mathrm{S})} \neq 0$). Therefore, interaural spectrograms of natural sounds are sparse. Missing interaural spectrograms values will be characterized by $\chi_{ft} = 0$ 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* α and the *IPD spectrogram* ϕ as the log-amplitude and phase of the interaural spectrogram $\hat{I}_{f,t}$:

$$\alpha_{ft} = 20 \log |\hat{I}_{ft}| \in \mathbb{R}, \;\; \phi_{ft} = \exp(j \arg(\hat{I}_{ft})) \in \mathbb{C} \qquad (2)$$

The phase difference is expressed in the complex domain, or equivalently $\mathbb{R}^2$, to avoid problems due to phase circularity. In the particular case of a sound source emitting white noise from $\boldsymbol{x}$, we have $\boldsymbol{\chi}_t = \mathbf{1}$ for all t, i.e., the sound source is emitting at all (f,t) points. One can thus compute the temporal means $\bar{\boldsymbol{\alpha}}(\boldsymbol{x}) \in \mathbb{R}^F$ and $\bar{\boldsymbol{\phi}}(\boldsymbol{x}) \in \mathbb{R}^{2F}$ of ILD and IPD spectrograms. These mean vectors will be referred to as the mean ILD and the mean IPD vectors associated to $\boldsymbol{x}$. The well established duplex theory suggests that ILD cues are mostly used at high frequencies (above 1.5 kHz) while ITD (or IPD) cues are mostly used at low frequencies (below 1.5kHz) in humans. Indeed, ILD values are similar because the HRTF can be neglected at low frequencies, 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 96 frequency channels between 0 and 1.5kHz and *high* corresponds to 416 frequency channels between 1.5kHz and 8kHz.

To automatically gather a large number of such vectors associated with different source positions, we used the same technique and robot setup as in [3]. A binaural acoustic dummy-head is mounted onto a pan-tilt (ψ, θ) motor system. The emitter (a loud speaker) is fixed in a reference position at 2.5 meters in front of the robot, while the robot is placed in 160 pan angles $\psi \in [-160°, 160°]$ (left-right) and 60 tilt angles $\theta \in [-60°, 60°]$ (up-down), with 2° steps,

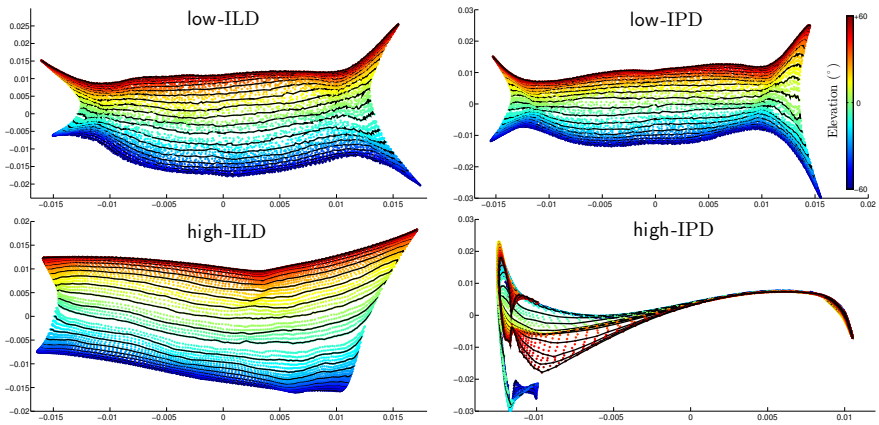

Fig. 1: Two-dimensional representations of mean interaural vectors using LTSA. For visualization purposes, points corresponding to the same tilt angle (elevation) have the same color and they are linked in pan angle (azimuth) order with a black line.

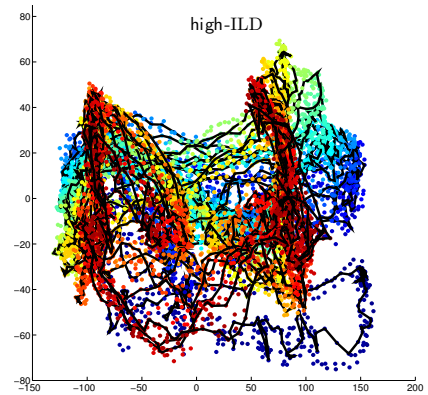

Fig. 2: Low-dimensional representations of mean *high*-ILD vectors using PCA.

i.e., $N = 9,600$ uniformly distributed *motor states*. Hence, the source location spans a $320°$ azimuth range and a $120°$ elevation range in the robot's frame. Notice that there is a one-to-one association between motor states and source locations. They will be denoted by $\{\boldsymbol{x}_n\}_{n=1}^N$. For each $\boldsymbol{x}_n \in \mathbb{R}^2$, two binaural recordings are made: (i) white noise which can be used to estimate $\bar{\boldsymbol{\alpha}}(\boldsymbol{x}_n)$ and $\bar{\boldsymbol{\phi}}(\boldsymbol{x}_n)$, and (ii) a randomly picked utterance amongst 362 samples from the TIMIT dataset [16]. These are 50% female, 50% male and they last 1-5s. Sparse interaural spectrograms can be computed from these recordings and are used to test our sound localization algorithm on natural sounds in section 5. All the experiments were carried out in real-world conditions, i.e., a room with natural reverberations and background noise.

3. THE MANIFOLD 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 should be parameterized by motor states and hence, lie on a lower L-dimensional manifold with $L = 2$. We propose to experimentally verify the existence of this manifold structure using non-linear dimensionality reduction, and examine whether obtained representations are homeomorphic to the motor-state space. Such a homeomorphism would allow us to confirm (or invalidate) the existence of a locally linear bijective mapping between motor states (or equivalently, source positions) and the interaural data gathered with our setup.

If the interaural data lie in a linear subspace, a standard dimensionality reduction method such as PCA could be used. However, in the case of a non-linear subspace one should use a manifold learning technique, e.g., diffusion kernels [17]. Alternatively, we chose to use local tangent-space alignment (LTSA) [18] 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 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 low-dimensional data representations as points in the data set. Finally a global low-dimensional *map* is built by optimal alignment of these local representations (see [18] for details). Two-dimensional[1] maps obtained using LTSA are shown in Fig. 1. Mean *low*-ILD, *low*-IPD, and *high*-ILD maps are smooth and homeomorphic to the motor-state space, thus confirming that these cues can be used for 2D binaural SSL based on locally linear mapping. However, this is not the case for the mean *high*-IPD map 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 enough 2D sound-source position information. We can therefore concatenate full-spectrum ILD and low-frequency IPD vectors to form an observation space in $\mathbb{R}^D$ (with $D = 704$), which will be referred to as the *ILPD space*. Similarly, we can define sparse *ILPD spectrograms* for general sounds.

Fig. 2 shows the result of applying PCA *globally* to the mean *high*-ILD data with $L = 2$. One may observe that the resulting map is extremely distorted, due to the non-linear nature of the *high*-ILD manifold. 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.

[1] When applying PCA locally, we observed a significant eigengap between the 2nd and 3rd eigeinvalues, thus validating the choice $L = 2$.

4. PROBABILISTIC PIECEWISE AFFINE MAPPING

Let $\mathcal{X} \subset \mathbb{R}^L$ be a subset of the low-dimensional latent space, i.e., the space of sound-source positions (or motor states), and $\mathbb{R}^D$ be the high-dimensional observation space, i.e., the space of ILPD cues. The computational experiments of section 3 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 t_k, plus an error term. Assuming that there is a finite number K of such affine transformations t_k and an equal number of associated regions $\mathcal{R}_k$ we obtain a piecewise-affine approximation of g. With each training-couple $(\boldsymbol{x}_n, \boldsymbol{y}_n)$ we associate an assignment variable $\boldsymbol{z}_n = (z_{1n} \dots z_{Kn})^\top$ such that $z_{kn} = 1$ if $\boldsymbol{y}_n$ is the image of $\boldsymbol{x}_n \in \mathcal{R}_k$ by t_k and 0 otherwise. This allows us to write:

$$\boldsymbol{y}_n = \sum_{k=1}^K z_{kn}(\mathbf{A}_k \boldsymbol{x}_n + \boldsymbol{b}_k) + \boldsymbol{e}_n \tag{3}$$

where the $D \times L$ matrix $\mathbf{A}_k$ and the vector $\boldsymbol{b}_k \in \mathbb{R}^D$ define the transformation t_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 iid realizations of a Gaussian variable with $\mathbf{0}$ mean and diagonal covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\sigma_{1:D}^2)$ we obtain:

$$p(\boldsymbol{y}_n | \boldsymbol{x}_n, z_{kn} = 1; \Theta) = \mathcal{N}(\boldsymbol{y}_n; \mathbf{A}_k \boldsymbol{x}_n + \boldsymbol{b}_k, \boldsymbol{\Sigma}) \tag{4}$$

where Θ designates all the model parameters (8). To make the affine transformations local, we model $\boldsymbol{z}_n$ by the realization of a hidden multinomial random variable conditioned by $\boldsymbol{x}_n$:

$$p(z_{kn} = 1 | \boldsymbol{x}_n; \Theta) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)}{\sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)}, \tag{5}$$

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:

$$|\boldsymbol{\Gamma}_1| = \dots = |\boldsymbol{\Gamma}_K| \text{ and } \pi_1 = \dots = \pi_K = 1/K \tag{6}$$

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 latent space into convex regions of equal volume. Extensive tests on simulated and audio data showed that these constraints yielded 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 the latent variables:

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

This model yields a closed-form and efficient EM algorithm maximizing the observed-data log-likelihood $\log p(\mathbf{X}, \mathbf{Y}; \Theta)$ with respect to the model's parameters:

$$\Theta = \left\{ \{\boldsymbol{\Gamma}_k, \boldsymbol{c}_k, \mathbf{A}_k, \boldsymbol{b}_k, \pi_k\}_{k=1}^K, \boldsymbol{\Sigma} \right\} \tag{8}$$

Posterior probabilities $r_{kn}^{(i)} = p(z_{kn} = 1 | \boldsymbol{x}_n, \boldsymbol{y}_n; \Theta^{(i-1)})$ are computed in the E-step from (4), (5) and Bayes inversion. The M-step maximizes the expected complete-data log-likelihood $\mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{Y},\Theta^{(i)})}[\log p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} | \Theta)]$. We obtain the following closed-form expressions for the parameters updates under the volume equality constraints (6):

$$\boldsymbol{c}_k^{(i)} = \sum_{n=1}^N \frac{r_{kn}^{(i)}}{\bar{r}_k^{(i)}} \boldsymbol{x}_n, \ \boldsymbol{\Gamma}_k^{(i)} = \left(\sum_{k=1}^K \frac{\bar{r}_k^{(i)}}{N} |\mathbf{S}_k^{(i)}|^{\frac{1}{L}} \right) \frac{\mathbf{S}_k^{(i)}}{|\mathbf{S}_k^{(i)}|^{\frac{1}{L}}} \tag{9}$$

$$\mathbf{A}_k^{(i)} = \overline{\mathbf{Y}}_k^{(i)} \overline{\mathbf{X}}_k^{(i)\dagger}, \ \ \boldsymbol{b}_k^{(i)} = \sum_{n=1}^N \frac{r_{kn}^{(i)}}{\bar{r}_k^{(i)}} (\boldsymbol{y}_n - \mathbf{A}_k^{(i)} \boldsymbol{x}_n), \tag{10}$$

$$\sigma_d^{2(i)} = \frac{1}{K} \sum_{k=1}^K \sum_{n=1}^N \frac{r_{kn}^{(i)}}{\bar{r}_k^{(i)}} (y_{dn} - \boldsymbol{a}_{dk}^{(i)\top} \boldsymbol{x}_n - b_{dk}^{(i)})^2, \tag{11}$$

where † is the Moore-Penrose pseudo inverse operator and:

$$\begin{aligned}
\mathbf{S}_k^{(i)} &= \textstyle\sum_{n=1}^N r_{kn}^{(i)} / \bar{r}_k^{(i)} (\boldsymbol{x}_n - \boldsymbol{c}_k^{(i)})(\boldsymbol{x}_n - \boldsymbol{c}_k^{(i)})^\top \\
\bar{r}_k^{(i)} &= \textstyle\sum_{k=1}^K r_{kn}^{(i)}, \ \mathbf{A}_k^{(i)} = (\boldsymbol{a}_{1k}^{(i)} \dots \boldsymbol{a}_{Dk}^{(i)})^\top, \boldsymbol{a}_{dk}^{(i)} \in \mathbb{R}^L \\
\overline{\mathbf{X}}_k^{(i)} &= (r_{k1}^{(i)\frac{1}{2}} (\boldsymbol{x}_1 - \bar{\boldsymbol{x}}_k^{(i)}) \dots r_{kN}^{(i)\frac{1}{2}} (\boldsymbol{x}_N - \bar{\boldsymbol{x}}_k^{(i)})) \\
\overline{\mathbf{Y}}_k^{(i)} &= (r_{k1}^{(i)\frac{1}{2}} (\boldsymbol{y}_1 - \bar{\boldsymbol{y}}_k^{(i)}) \dots r_{kN}^{(i)\frac{1}{2}} (\boldsymbol{y}_N - \bar{\boldsymbol{y}}_k^{(i)})) \\
\bar{\boldsymbol{x}}_k^{(i)} &= \textstyle\sum_{n=1}^N r_{kn}^{(i)} / \bar{r}_k^{(i)} \boldsymbol{x}_n \text{ and } \bar{\boldsymbol{y}}_k^{(i)} = \sum_{n=1}^N r_{kn}^{(i)} / \bar{r}_k^{(i)} \boldsymbol{y}_n
\end{aligned}$$

Initial posteriors $r_{kn}^{(0)}$ can be obtained either by estimating a K-GMM solely on $\boldsymbol{X}$ (GMM) or jointly on $(\boldsymbol{X}, \boldsymbol{Y})$ (GMM-J). One may see that the MPLR-GMM and MPLR-GMM-J methods proposed in [14] are strictly equivalent to these initializations followed by a single parameter estimation using (10). Unlike [14], we use a consistent probabilistic model enforcing partitioning within parameter estimation, a component-dependent reconstruction error $\boldsymbol{\Sigma}$, and an E-step allowing to iterate until convergence to a maximum of the log-likelihood. This significantly reduces the global reconstruction error (section 5). Fig. 3 shows a partitioning example obtained with our method using a toy data set.

Let $\widetilde{\Theta}$ denote the parameters estimated with our algorithm; we describe now a method based on Bayesian inversion to estimate the unknown position $\boldsymbol{x}$ of a sound

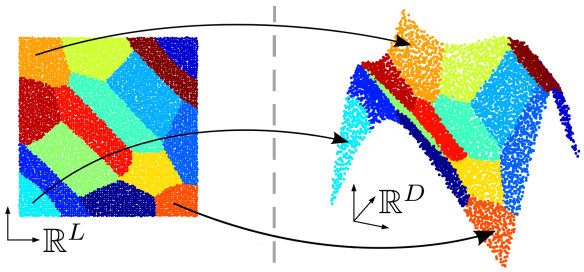

Fig. 3: Latent space partitioning and locally affine mapping on a toy data set ($N = 9600, K = 15, L = 2, D = 3$). Colors encode regions in $\mathbb{R}^D$ maximizing (5). Observe how well these regions (associated with affine transformations) are adjusted to the geometry of both the latent space and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

source given its observed (possibly sparse) ILPD spectrogram $\boldsymbol{Y}_{\boldsymbol{\chi}} = \{y_{dt}; \chi_{dt}\}_{t,d=1}^{T,D}$ as defined in section 2 and 3. If we suppose that the observations are assigned to the same position $\boldsymbol{x}$ and transformation $\boldsymbol{z}$, it follows from our model (4), (5), (7) that the posterior distribution $p(\boldsymbol{x}|\boldsymbol{Y}_{\boldsymbol{\chi}}; \widetilde{\Theta})$ is a GMM $\sum_{k=1}^{K} \alpha_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \mathbf{V}_k)$ in $\mathbb{R}^L$ with parameters:

$$\boldsymbol{\mu}_k = \mathbf{V}_k \left(\widetilde{\boldsymbol{\Gamma}}_k^{-1} \widetilde{\boldsymbol{c}}_k + \sum_{d,t=1}^{D,T} \frac{\chi_{dt}}{\widetilde{\sigma}_d^2} \widetilde{\boldsymbol{a}}_{dk} (y_{dt} - \widetilde{b}_{dk}) \right), \quad (12)$$

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

$$\alpha_k \propto (|\mathbf{V}_k| / |\widetilde{\boldsymbol{\Gamma}}_k|)^{\frac{1}{2}} \exp\Big(-\tfrac{1}{2} \big(\sum_{d,t=1}^{D,T} \tfrac{\chi_{dt}}{\widetilde{\sigma}_d^2} (y_{dt} - \widetilde{b}_{dk})^2 + \widetilde{\boldsymbol{c}}_k^\top \widetilde{\boldsymbol{\Gamma}}_k^{-1} \widetilde{\boldsymbol{c}}_k + \boldsymbol{\mu}_k^\top \mathbf{V}_k^{-1} \boldsymbol{\mu}_k \big) \Big) \quad (14)$$

where the weights $\{\alpha_k\}_{k=1}^{K}$ are normalized to sum to 1. This formulation is more general than the unique, complete observation case ($T = 1, \boldsymbol{\chi}_1 = \mathbf{1}$). Several estimates can be inferred from this posterior distribution. We used the posterior expectation $\mathbb{E}[\boldsymbol{x}|\boldsymbol{Y}_{\boldsymbol{\chi}}] = \sum_{k=1}^{K} \alpha_k \boldsymbol{\mu}_k$ which yields the lowest average localization errors. However, the posterior distribution may have several high modes, in which case the posterior expectation becomes a misleading estimator. This notably happens if the sound spectrum is extremely sparse. In this case, one may preserve the full posterior distribution and, for instance, combine it with other external probabilistic knowledge.

5. EXPERIMENTS AND RESULTS

We compared the proposed algorithm with two different initialization strategies, PPAR-GMM and PPAR-GMM-J, to the two algorithms proposed in [14], MPLR-GMM and MPLR-GMM-J. We used these four algorithms for training with the complete set of $N = 9,600$ ILPD white-noise samples available from the audio-motor recordings. The algorithm performance was evaluated using the mean reconstruction error,

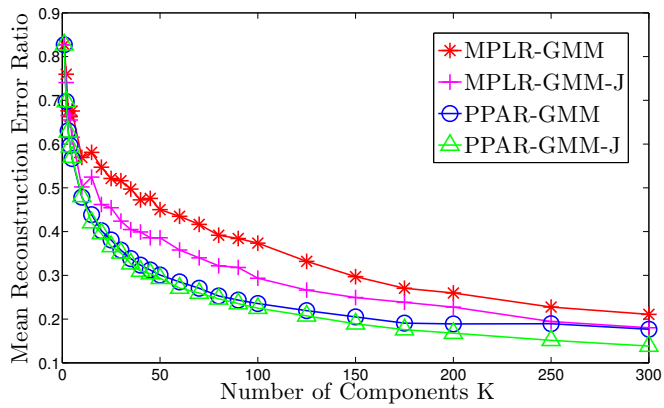

Fig. 4: Comparison of the proposed method (PPAR) with MPLR [14] as a function of the number of components.

namely the ratio $\text{mean}_d[\sigma_d^2]/\text{var}_n[y_{dn}]$, where σ_d^2 is defined by (11). As shown in Fig. 4, the proposed method significantly outperforms the two others at a minor additional computational cost. This is not surprising since PPAR yields maximum likelihood estimators of the affine transformations and space partitioning obtained after several EM iterations[2], while MPLR-GMM and MPLR-GMM-J perform just one optimization step (10). GMM-J initialization strategy will be used from now on since it provides better results than GMM initialization. Interestingly, the only free parameter of our method, namely K, can be chosen based on a compromise between computational cost and precision, since the higher the value of K the lower the error. Next we evaluate the proposed inverse

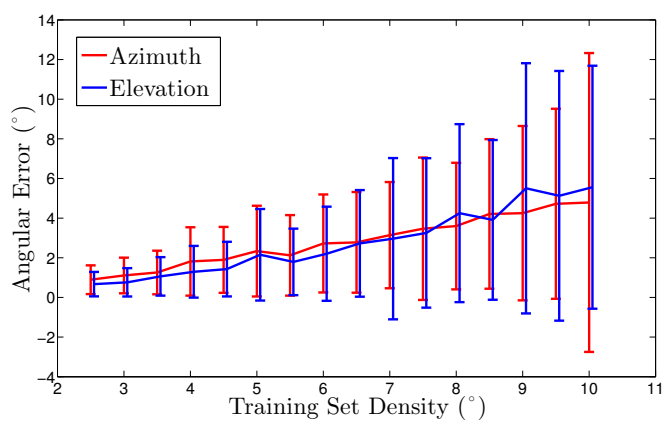

Fig. 5: Average and standard deviation of angular errors in white noise localization as a function of the training set's density.

mapping method in the case of a complete, unique observation, i.e., sound source localization using a mean ILPD vectors computed from white noise. Localization error is evaluated as a function of the *density* of the training set. We represent the density by the angle δ, corresponding to the average pan and tilt absolute difference between a point and its nearest

[2]Convergence was reached after 10 to 20 iterations using mean ILPD data.

neighbor in the training set. The complete training set of density 2° was uniformly decimated at random to obtain irregularly spaced and smaller training sets, while the test positions were randomly chosen from the complete set so that most of the test positions were outside the training set. For a given density δ, 10 different decimated sets were used for training, and 20 source positions were estimated for each one, i.e., 200 localization tasks. K was chosen such that there are approximately 30 training samples per affine transformation. The mean and standard deviation of the errors in azimuth and in elevation are shown in Fig. 5. The mean localization errors are always smaller than half the training set density, which illustrates the interpolation ability of our method. In addition, the error's standard deviation remains reasonable even for heavily decimated training sets, thus showing that the overall performance is not much affected by the distribution and size of the training set being used. No front-back azimuth or elevation confusions were observed, thanks to the asymmetry of the dummy head and to the spatial richness of ILPD cues. This is in contrast with most of the current binaural SSL approaches that focus only on frontal azimuth estimation.

Finally, we tested our SSL method with missing and redundant observations using randomly located sound sources emitting random utterances. The spectral power threshold was manually set quite high, so that the test ILPD spectrograms had 89.6% of missing data in average. Mean azimuth and elevation errors (Avg), standard deviations (Std), and percentage higher than $4\times$Avg (Out) over 500 tests are shown in the table below. The low Stds and low amount of high errors show the reliability and robustness of our method, even using real world recordings of very sparse sounds emitted from a wide range of azimuths and elevations.

Training density	Azimuth			Elevation		
	Avg	Std	Out	Avg	Std	Out
$\delta = 2^\circ$	2.0°	1.8°	1.2%	1.3°	1.2°	1.2%
$\delta = 6^\circ$	5.6°	5.2°	1.2%	4.5°	5.3°	1.4%

6. CONCLUSION

In this paper, we emphasized the manifold structure of real-world audio-motor data using ILD/IPD cues, and we set the basis of a novel probabilistic framework for accurate 2D sound source localization relying on this structure. More generally, we presented a new methodology for examining whether a dataset is intrinsically locally linear based on manifold learning, and proposed an adequate probabilistic mapping method relying on this property. The advantage of our Bayesian formulation is that it could serve as a basis for a wide variety of extensions and other applications. For example, a mixture of PPAR could be used for multiple sound sources localization. In the future, we plan to thoroughly examine how changes in the recording environment affect the acoustic manifold structure. Such a study and a proper adaptation of our mapping model would allow to make our sound source localization method robust to changes in reverberations and positions in the room.

7. 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 *ACM/IEEE HRI*, 2012.

[4] R. Liu and Y. Wang, "Azimuthal source localization using interaural coherence in a robotic dog: modeling and application," *Robotica*, 2010.

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

[6] A. Deleforge and R. P. Horaud, "A latently constrained mixture model for audio source separation and localization," in *LVA/ICA*, 2012.

[7] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Acoust., Speech, Signal Process.*, 2012.

[8] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *DAFx*, 2003.

[9] A. R. Kullaib, M. Al-Mualla, and D. Vernon, "2d binaural sound localization: for urban search and rescue robotics," in *Mobile Robotics*, 2009.

[10] A. Kain and MW Macon, "Spectral voice conversion for text-to-speech synthesis," in *ICASSP*, 1998.

[11] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Acoust., Speech, Signal Process.*, 1998.

[12] T. Toda, A. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, 2008.

[13] R. D. de Veaux, "Mixtures of linear regressions," *Comput. Stat. Data Anal.*, 1989.

[14] Y. Qiao and N. Minematsu, "Mixture of probabilistic linear regressions: A unified view of GMM-based mapping techiques," in *ICASSP*, 2009.

[15] L. Xu, M. I. Jordan, and G. E. Hinton, "An alternative model for mixtures of experts," in *NIPS*, 1995.

[16] J. S. Garofolo, L. F. Lamel, and W. M. Fisher, "The DARPA TIMIT acoustic-phonetic continuous speech corpus," 1993.

[17] R. Talmon, I. Cohen, and S. Gannot, "Supervised source localization using diffusion kernels," in *WASPAA*, 2011.

[18] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM Journal on Scientific Computing*, 2004.