SUPERVISED CONTRASTIVE EMBEDDINGS FOR BINAURAL SOURCE LOCALIZATION

Duowei Tang[*], *Maja Taseska,*[†] *and Toon van Waterschoot*

KU Leuven, Dept. of Electrical Engineering (ESAT-STADIUS/ETC), Leuven, Belgium
{duowei.tang, maja.taseska, toon.vanwaterschoot}@esat.kuleuven.be [‡]

ABSTRACT

Recent data-driven approaches for binaural source localization are able to learn the non-linear functions that map measured binaural cues to source locations. This is done either by learning a parametric map directly using training data, or by learning a low-dimensional representation (embedding) of the binaural cues that is consistent with the source locations. In this paper, we adopt the second approach and propose a parametric embedding to map the binaural cues to a low-dimensional space, where localization can be done with a nearest-neighbor regression. We implement the embedding using a neural network, optimized to map points that are close in the latent space (the space of source azimuths or elevations) to nearby points in the embedding. We show that the proposed embedding generalizes well in acoustic conditions different from those encountered during training, and provides better results than unsupervised embeddings previously used for localization.

***Index Terms*—** binaural source localization, manifold learning, supervised embedding.

1. INTRODUCTION

To localize sources, the human auditory system uses binaural features extracted from acoustic signals, such as the Interaural Phase Differences (IPDs) and Interaural Level Differences (ILDs) [1]. Computational localization algorithms in robot audition [2], hearing aids, virtual reality [3], etc., try to mimic this process and estimate the binaural cues from microphone signals. However, the acoustic channels introduce uncertainties in the binaural cues due to reverberation, making source localization challenging. Traditionally, robustness to reverberation has been tackled with statistical model-based approaches [4–6].

In contrast, data-driven approaches are able to learn the non-linear functions that map binaural cues to source locations, without an acoustic propagation model or a lookup table. A multilayer perceptron was used to model the nonlinear map already in the mid-nineties [7]. Recently, deep neural networks were used to learn the relationship between azimuth and binaural cues in [8], by exploiting head movements to resolve the front-back ambiguity. A different data-driven approach was used in [9, 10], where the relationship between source locations and binaural cues was modeled with a probabilistic piecewise linear function. By learning the function parameters, sources can be localized by probabilistic inversion. An implicit assumption of the piecewise linear model in [9, 10] is that similar source locations result in similar binaural cues. The same assumption is also used in non-parametric source localization algorithms based on manifold learning in [11, 12]. In this paper, we focus on data-driven source localization approaches, inspired by low-dimensional manifold learning [11, 12].

Manifold learning methods that rely on smoothness in measurement space with respect to the underlying source locations, might generalize poorly to varying acoustic conditions. The uncertainties in the binaural cue measurements introduced by reverberation, introduce variations in the measurement space neighborhoods that might not be consistent with source locations. In this paper, we propose a parametric embedding to map the binaural cues to a low-dimensional space, where localization can be done with a nearest-neighbor regression. This paradigm is often used in the machine learning community as a pretraining stage for classifiers [13]. We implement the embedding with a neural network, optimized with a contrastive loss function [14], such that binaural cues recorded from signals with similar source locations, have a small Euclidean distance in the embedding. This approach generalizes better to unseen acoustic conditions than unsupervised manifold learning used in [11, 12]. The paper is organized as follows. In Section 2, we revise the binaural cue extraction and we formulate the problem. In Section 3, we provide a brief overview of related work. The proposed method is presented in Section 4 and experimental results are shown in Section 5.

2. DATA MODEL AND PROBLEM FORMULATION

Let $s_1(\tau)$ and $s_2(\tau)$ denote the signals captured at the left and right microphones in a binaural recording setup in a reverberant environment. In this work, we extract the binaural cues in the Short-time Fourier Transform (STFT) domain, as in [10, 15]. Let $S_1(t, k)$ and $S_2(t, k)$ denote the STFT coefficients of $s_1(\tau)$ and $s_2(\tau)$, where t and k are the time and frequency index, respectively. At a time-frequency bin (t, k) an ILD α_{tk} and an IPD ϕ_{tk} are defined as

$$\alpha_{tk} = 20 \log_{10} \frac{|S_1(t,k)|}{|S_2(t,k)|}, \quad \phi_{tk} = \angle \frac{S_1(t,k)}{S_2(t,k)}. \tag{1}$$

Assuming that a single sound source is active, we follow the binaural feature extraction approach from [10], and compute time-averaged ILDs and IPDs across T frames as follows

$$a_k = T^{-1} \sum_{t=1}^{T} \alpha_{tk}, \quad p_k = T^{-1} \sum_{t=1}^{T} \exp(j\phi_{tk}). \tag{2}$$

[*] Duowei Tang is sponsored by the Chinese Scholarship Council (CSC) (no. 201707650021)

[†] Maja Taseska is a Postdoctoral Fellow of the Research Foundation Flanders - FWO-Vlaanderen (no. 12X6719N).

[‡] This research work was carried out at the ESAT Laboratory of KU Leuven. The research leading to these results has received funding from the KU Leuven Internal Funds C2-16-00449 and VES/19/004, and the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

By concatenating the ILDs, and the real and imaginary parts of the IPDs in selected frequency ranges $[k_1, k_2]$ and $[k_3, k_4]$, the binaural information is summarized in a measurement vector $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^D$,

$$\boldsymbol{x} = [a_{k_1}, \ldots, a_{k_2}, \mathcal{R}\{p_{k_3}\}, \mathcal{I}\{p_{k_3}\}, \ldots, \mathcal{R}\{p_{k_4}\}, \mathcal{I}\{p_{k_4}\}]^{\mathrm{T}}. \tag{3}$$

It is known that IPDs carry reliable location cues below 2 kHz [1], while ILDs contribute to localization at higher frequencies as well [10]. Hence, we used the ranges $[k_1, k_2] = [200, 7000]$ Hz and $[k_3, k_4] = [200, 2500]$ Hz. For an STFT window of 1024 samples at 16 kHz, this results in a 729-dimensional vector $\boldsymbol{x}$.

Hence, a pair of signals $s_1(\tau)$ and $s_2(\tau)$ is associated to a vector $\boldsymbol{x} \in \mathcal{X}$. We refer to $\mathcal{X}$ as the *measurement* space. Let the unknown source location be denoted by $u \in \mathcal{U}$. We refer to $\mathcal{U}$ as the *latent space*. $\mathcal{U}$ is one-dimensional if one considers azimuth or elevation separately, and two-dimensional if the angles are considered simultaneously. Given a training set of N pairs $\mathcal{T} = \{(\boldsymbol{x}_i, u_i)\}_{i=1}^N$, the localization problem consists of finding a function h

$$\hat{u} = h(\boldsymbol{x}), \quad h : \mathcal{X} \to \mathcal{U}. \tag{4}$$

that accurately maps measurements to latent variables. In this work, we implement h in a non-parametric fashion, using Nearest-Neighbor (NN) regression in a suitable low-dimensional space. Therefore, our main objective is to learn an embedding function f that maps the vectors $\boldsymbol{x}$ to a low-dimensional space which preserves latent space neighborhoods, i.e.,

$$\boldsymbol{z} = f(\boldsymbol{x}), \quad f : \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^d, \quad d << D. \tag{5}$$

We propose a supervised framework to learn a parametric function f that satisfies these properties, when the source azimuth, or elevation are the latent variables. Distance estimation is not considered. A NN regression function $h : \mathcal{Z} \to \mathcal{U}$ is then used for localization.

3. BACKGROUND AND PRIOR WORK

If the microphone location in a given room is fixed, the authors in [12] showed that features extracted from binaural signals can be embedded in a low-dimensional space $\mathcal{Z}$, in a way that recovers source locations. The framework in [12] is based on unsupervised manifold learning, in particular, *Laplacian eigenmaps* (LEM) [16].

Unsupervised manifold learning approaches often start by computing a similarity matrix $\boldsymbol{K} \in \mathbb{R}^{N \times N}$, with entries $\boldsymbol{K}[i, j]$ related to the Euclidean distances $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$. One way to compute $\boldsymbol{K}$ is using nearest-neighbors, i.e., $\boldsymbol{K}[i, j] = \boldsymbol{K}[j, i] = 1$ if $\boldsymbol{x}_i$ is among the M nearest neighbors of $\boldsymbol{x}_j$, or if $\boldsymbol{x}_j$ is among the M nearest neighbors of $\boldsymbol{x}_i$ (in Euclidean distance). A second way is using an exponentially decaying kernel function, such as the Gaussian

$$\boldsymbol{K}[i, j] = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{\varepsilon}\right), \tag{6}$$

where ε is the kernel bandwidth. Such kernel was used for source localization in [12]. Given the similarity matrix $\boldsymbol{K}$, the neighborhood-preserving cost function of LEM is given by [16]

$$\underset{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N}{\arg\min} \sum_{i,j=1}^{N} \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2 \, \boldsymbol{K}[i, j], \tag{7}$$

which enforces that points with large affinity $\boldsymbol{K}[i, j]$, are to be mapped to points with a small Euclidean distance $\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2$.

The cost function has a closed-form solution, given by the largest eigenvectors of $\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{K}$, where $\boldsymbol{D}$ is a diagonal matrix with entries $\boldsymbol{D}[i, i] = \sum_{j=1}^N \boldsymbol{K}[i, j]$. If $\{\boldsymbol{\psi}_i\}_{i=1}^N$ denote the eigenvectors of $\boldsymbol{P}$, with eigenvalues $1 = \lambda_1 > \lambda_2 \geq \ldots, \geq \lambda_N$, the d-dimensional LEM embedding is given by [16]

$$\boldsymbol{z}_i = f(\boldsymbol{x}_i) = [\, \boldsymbol{\psi}_2[i], \ \boldsymbol{\psi}_3[i], \ \ldots, \ \boldsymbol{\psi}_{d+1}[i]]^{\mathrm{T}}, \tag{8}$$

where the constant eigenvector $\boldsymbol{\psi}_1$ is not included [16, 17]. The LEM embedding f is non-parametric, and the low-dimensional representation $\boldsymbol{z}$ of a new measurement $\boldsymbol{x}$ is obtained as a linear combination of the training points $\{\boldsymbol{z}_i\}_{i=1}^N$ [18]. However, this procedure is often insufficiently accurate and represents a disadvantage of LEM and of spectral embeddings in general.

Besides the promising performance of spectral embeddings for localization [11, 12, 19], their major drawback is the assumption that neighborhoods in the measurement space are consistent with the source locations. Although the assumption was shown to hold when all signals are recorded in one room, for fixed microphone locations [9, 12, 19], this is not the case when the signals are filtered by various acoustic channels in different enclosures.

4. SUPERVISED EMBEDDING FOR LOCALIZATION

We propose a parametric embedding, designed to preserve neighborhoods in terms of source locations. The framework consists of defining the neighborhoods and a suitable cost function (Section 4.1), and training a neural network to implement the embedding (Section 4.2). Note that a similar supervised approach is used for various classification tasks in machine learning [14, 20–22].

4.1. Supervised neighborhoods and contrastive loss

Consider two labeled measurements $(\boldsymbol{x}_i, u_i)$ and $(\boldsymbol{x}_j, u_j)$. Let $d_u(u_i, u_j) = |u_i - u_j|$ denote the distance in the one-dimensional latent space $\mathcal{U}$, where u_i, u_j corresponds to the source azimuth or elevation. A neighborhood indicator $y_{ij} \in \{0, 1\}$ is defined as

$$y_{ij} = \begin{cases} 0, & \text{if} \quad d_u(u_i, u_j) > \epsilon_u \\ 1, & \text{if} \quad d_u(u_i, u_j) \leq \epsilon_u, \end{cases} \tag{9}$$

for a neighborhood size ϵ_u. We seek to learn a parametric function $f_W : \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^d$, with parameters W, that maps $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ to their low-dimensional images $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$. If $y_{ij} = 1$, the Euclidean distance $\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2$ should be small, and if $y_{ij} = 0$, then $\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2$ should be large. For a given embedding function f_W, we have

$$\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2 = \|f_W(\boldsymbol{x}_i) - f_W(\boldsymbol{x}_j)\|_2. \tag{10}$$

A *contrastive loss function* over the parameters W, tailored for neighborhood preservation has been proposed in [14] for non-linear dimensionality reduction, and is given by

$$L(W) = \sum_{i=1}^{N} \sum_{j=1}^{N} \Big(y_{ij} \, \|f_W(\boldsymbol{x}_i) - f_W(\boldsymbol{x}_j)\|_2^2 + (1 - y_{ij}) \max(0, \mu_{ij} - \|f_W(\boldsymbol{x}_i) - f_W(\boldsymbol{x}_j)\|_2)^2 \Big). \tag{11}$$

The parameter μ_{ij} is a positive real-valued margin, such that $\mu_{ij}/2$ can be interpreted as the radius of circles centered on $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$. If the circles intersect and $y_{ij} = 0$, the two dissimilar pairs are too close in the embedding space, thus increasing the *contrastive loss* in (11). On the other hand, if $y_{ij} = 1$, large distances are penalized, enforcing f_W to preserve neighborhoods. It is important to note that in [14], where the *contrastive loss* was first proposed for classification, $\mu_{ij} \equiv \mu$ is a constant margin. In our application, the latent space of azimuths or elevations is continuous. To accurately preserve its geometry, we propose an adaptive margin, based on the distance in the latent space, as follows

$$\mu_{ij} = \frac{\exp\left(d_u(u_i, u_j)\right)}{\exp\left(d_u(u_i, u_j)\right) + 1}. \qquad (12)$$

Thus, as $d_u(u_i, u_j)$ decreases, the margin μ_{ij} decreases as well.

4.2. Learning the embedding and NN localization

We implement f_W with a neural network with two fully connected hidden layers with $4D$ and $2D$ neurons, respectively. The output layer has 3 neurons, corresponding to a 3-dimensional embedding space i.e., $d = 3$. The hidden neurons have a ReLU, and the output neurons have a linear activation. Training scheme for minimizing (11), called *siamese* architecture, was proposed in [21] and used for various tasks in [14,20]. It consists of two identical branches that implement f_W, taking a pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ as an input. The measurements $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are passed through the branches (one per branch), and the cost is evaluated in (11) using y_{ij} and the outputs $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ of the branches. To avoid overfitting, we used dropout layers. The dropout rate after the input layer was in the range $[0.1, 0.2]$, and after the hidden layers it was in the range $[0.2, 0.3]$. The dropout rates were fine-tuned using a line search, based on the performance on a validation set.

A key aspect of the Siamese scheme is the selection of pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for training. For small datasets, one could consider all pairs and proceed with training on randomized batches of data. However the polynomial growth of the number of pairs results in memory problems even for moderately large datasets. To solve this problem, we implemented pair selection during training as follows. First, we create B pairs from each batch. Then we randomly select $L \leq B/2$ similar pairs (i.e. $y_{ij} = 1$) and L dissimilar pairs (i.e. $y_{ij} = 0$), ensuring that the number of similar and dissimilar pairs is balanced in each batch. By training a large number of epochs, all pairs will be eventually considered in the optimization process with a large probability. An important direction for future research is to explore and understand different strategies of pair selection, and their effect on the embedding properties.

Once the weights of f_W are optimized, we compute the embedding of a new $\boldsymbol{x}$ by a forward-pass through the network. Let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K$ denote the K nearest neighbors of $\boldsymbol{z}$ in the training set. The latent variable (azimuth or elevation) is then estimated as

$$\hat{u} = \sum_{i=1}^{K} w_i u_i, \quad \text{with} \quad w_i = \frac{\exp\left(-\frac{\|\boldsymbol{z}-\boldsymbol{z}_i\|_2^2}{\varepsilon}\right)}{\sum_{j=1}^{K} \exp\left(-\frac{\|\boldsymbol{z}-\boldsymbol{z}_j\|_2^2}{\varepsilon}\right)}. \qquad (13)$$

The bandwidth ε of the exponential kernel is obtained as the median of the squared distances from the K neighbors, i.e.,

$$\varepsilon = \text{median}\left(\|\boldsymbol{z} - \boldsymbol{z}_1\|_2^2, \ldots, \|\boldsymbol{z} - \boldsymbol{z}_K\|_2^2\right). \qquad (14)$$

Note that if the embedding is accurately preserving neighborhoods, the choice of regression weights is not critical. For instance w_i can be inversely proportional to $\|\boldsymbol{z} - \boldsymbol{z}_i\|_2^2$. However, in our experiments, the latter generally leads to less accurate location estimates than exponentially decaying weights.

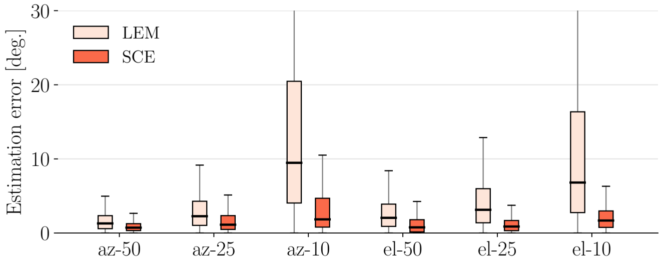

Figure 1: Localization accuracy for azimuth (az) and elevation (el) on the CAMIL dataset, for different sizes of the training set.

5. EXPERIMENTS

The proposed *supervised contrastive embedding* (SCE) is compared to the *Laplacian eigenmaps* (LEM) in a NN localization framework. As the neighborhoods for LEM are defined in the input space, a single embedding is used to estimate both azimuth and elevation. Although a single SCE embedding can be trained to estimate azimuth and elevation simultaneously as well, a system with two separately trained embeddings provided better results for the same amount of data. This result is not unexpected, as each system is trained to learn a simpler 1-dimensional latent space. For the NN-regression in (13), 50 neighbors are used in all experiments. The margin ϵ_u in (9) is set to 3° both for azimuth and elevation. We implemented the LEM using a nearest neighbor matrix $\boldsymbol{K}$, which in our experiments, provided better results than the Gaussian kernel used in [12, 19].

5.1. Fixed acoustic conditions

In this experiment, we used the CAMIL dataset [10] of binaural recordings, made with a dummy head in a reverberant room. The source is at a fixed position, 2.7 m from the head, while sounds are recorded for 10800 pan-tilt states of the head. This results in source azimuth and elevation in the range $[-180^\circ, 180^\circ]$ and $[-60^\circ, 60^\circ]$, respectively, (with 2° resolution). Training was done using white noise (1 s per recording) in three experiments, by using 50%, 25%, and 10% randomly selected pan-tilt states. We used STFT window length of 1024 samples at 16 kHz. The test set contains recordings of 1-5 s speech samples from the TIMIT corpus [23]. In addition, 15 dB spatially uncorrelated white noise was added. The angle estimation error statistics are summarized in Figure 1. The proposed SCE outperforms the LEM in all cases, achieving lower median estimation error, and significantly lower variance. Notably, the LEM performance deteriorates for small training sets, with a median error of 9.5° in azimuth and 6.8° in elevation, for the 10% training set. The proposed SCE maintains low median errors of $0.7^\circ, 1.1^\circ$, and 1.8° in azimuth and $0.8^\circ, 0.9^\circ$, and 1.7° in elevation, for the three training sets, respectively. The embeddings for the 50% training set are shown in Figure 2, where the consistency with source locations is visible both in terms of azimuth and elevation.

5.2. Varying acoustic conditions

To evaluate the embeddings in varying acoustic conditions, we used the VAST dataset [24] of simulated binaural room impulse responses of a KEMAR dummy head [25, 26]. The training set consists of 15 rooms with reverberation time 0.1-0.4 s. Each room contains spherical grids of positions with radii 1, 1.5, and 2 meters, centered at 9 positions. Two test sets are provided: in the first set,

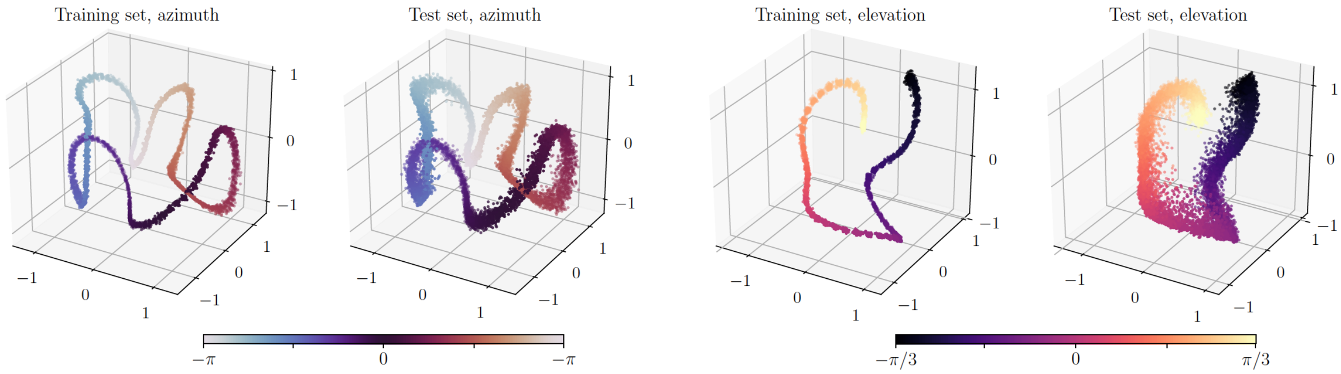

Figure 2: Scatter plot of the embeddings of the CAMIL training and test sets. The latent azimuth and elevation are coded in color.

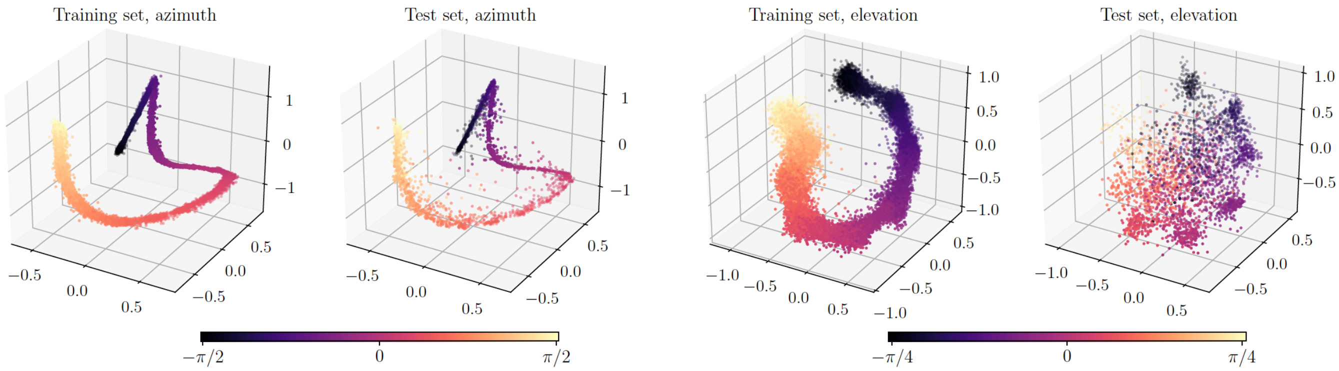

Figure 3: Scatter plot of the embeddings of the VAST training and test sets. The latent azimuth and elevation are coded in color.

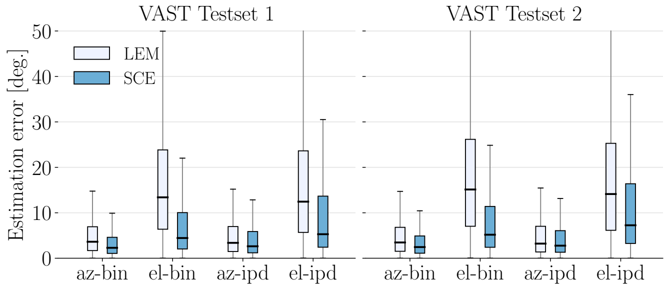

Figure 4: Localization accuracy for azimuth (az) and elevation (el) on the VAST dataset, for binaural cues consisting of both ILD and IPD (denoted by *bin*), and only IPD.

the source and receiver are placed at random positions in the same 15 rooms. In the second set, the source and receiver are placed in rooms of random width and length between 3×2 m and 10×4 m, with absorption profiles randomly picked from those of the training rooms. The receiver's height is fixed to 1.7 m. As done in the sample experiments published with the VAST dataset [24], we limited the azimuth to frontal angles $[-90°, 90°]$, and the elevation to $[-45°, 45°]$, resulting in 23523 training recordings. To focus on the influence of the varying room acoustics while exciting all frequencies, only white noise source signals were considered in this experiment. Due to the longer acoustic channels, compared to those in Section 5.1, we used an STFT window of 2048 samples.

The angle estimation error statistics are shown in Figure 4. The SCE outperforms the LEM embedding for both test sets. We also evaluated the embeddings of measurement vectors that only consist of IPDs. The results are similar for both types of measurements, with a median error difference of $0.3°$-$0.4°$ for azimuth and $0.8°$-$2°$ for elevation. Although this might indicate that the ILDs are inconsistent location cues across different acoustic channels, the claim is to be further investigated in more experiments. The embeddings for the first test set are shown in Figure 3. It can be seen that the elevation embedding generalizes somewhat poorly to the test set. Nonetheless, there are visible correctly embedded clusters which enable us to reach median errors of only $2°$-$2.5°$ worse than for azimuth (Figure 4). We believe that in future work, the elevation embedding can be improved with a better training strategy.

6. CONCLUSIONS

We proposed a framework for supervised dimensionality reduction of binaural cue measurements, followed by a nearest-neighbor source localization. Our work is based on recent results that apply manifold learning to extract source locations from binaural recordings, and the power of supervised learning to parametrize these manifolds. We used a contrastive training approach of a siamese architecture, to learn a parametric embedding that preserves local structure in terms of azimuth or elevation. We demonstrated promising results that show better generalization to varying acoustic conditions than unsupervised approaches. Problems to address in future work include neighborhood selection strategies, incorporating noise robustness during training, and investigating influence of different dummy heads.

7. REFERENCES

[1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997.

[2] S. Argentieri, P. Danès, and P. Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Comput. Speech Lang.*, vol. 34, no. 1, pp. 87–112, nov 2015.

[3] F. Keyrouz and K. Diepold, "Binaural Source Localization and Spatial Audio Reproduction for Telepresence Applications," *Presence Teleoperators Virtual Environ.*, vol. 16, no. 5, pp. 509–522, 2007.

[4] T. May, S. Van De Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 1, pp. 1–13, 2011.

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

[6] M. Mandel, R. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, 2010.

[7] M. S. Datum, F. Palmieri, and A. Moiseff, "An artificial neural network for sound localization using binaural cues," *J. Acoust. Soc. Am.*, vol. 100, no. 1, pp. 372–383, 1996.

[8] N. Ma, T. May, and G. J. Brown, "Exploiting Deep Neural Networks and Head Movements for Robust Binaural Localization of Multiple Sources in Reverberant Environments," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 12, pp. 2444–2453, 2017.

[9] A. Deleforge and R. Horaud, "2D sound-source localization on the binaural manifold," in *IEEE Int. Work. Mach. Learn. Signal Process. MLSP*, 2012.

[10] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic space learning for sound source separation and localization on binaural manifolds," *Int. J. Neural Syst.*, vol. 25, no. 1, 2015.

[11] B. Laufer, R. Talmon, and S. Gannot, "Relative transfer function modeling for supervised source localization," in *IEEE Work. Appl. Signal Process. to Audio Acoust.*, 2013, pp. 1–4.

[12] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "A Study on Manifolds of Acoustic Responses," in *Proc. Int. Conf. Latent Var. Anal. Signal Sep.*, 2015, pp. 203–210.

[13] R. Salakhutdinov and G. Hinton, "Learning a nonlinear embedding by preserving class neighbourhood structure," *Int. Conf. Artif. Intell. Stat.*, pp. 412–419, 2007.

[14] R. Hadsell, S. Chopra, and Y. Lecun, "Dimensionality Reduction by Learning an Invariant Mapping," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006.

[15] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 1, pp. 68–77, 2010.

[16] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 6, no. 15, pp. 1373–1396, 2003.

[17] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI: American Mathematical Society, 1997.

[18] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, "Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps and Spectral Clustering," *Adv. Neural Inf. Process. Syst.*, vol. 16, pp. 177–184, 2004.

[19] M. Taseska and T. van Waterschoot, "On spectral embeddings for supervised binaural source localization," in *Proc. 27th Eur. Signal Process. Conf.*

[20] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a Similarity Metric Discriminatively, with Application to Face Verification," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. IEEE, 2005, pp. 539–546.

[21] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," in *Adv. Neural Inf. Process. Syst.*, 1994, pp. 737–744.

[22] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1701–1708.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," 1993.

[24] C. Gaultier, S. Kataria, and A. Deleforge, "VAST: The Virtual Acoustic Space Traveler Dataset," in *Proc. Int. Conf. Latent Var. Anal. Signal Sep.*, 2017, pp. 68–79.

[25] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.*, vol. 97, no. 6, pp. 3907–3908, 1995.

[26] S. M. Schimmel, M. F. Muller, and N. Dillier, "A fast and accurate shoebox room acoustics simulator," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2009, pp. 241–244.