Unsupervised feature learning on monaural DOA estimation using convolutional deep belief networks

Yan Chen[1], Mengyao Zhu[1], Nicolas Epain[2], Craig Jin[2]

[1]School of Communication & Information Engineering, Shanghai University, Shanghai, China

[2]CARLab, School of Electrical and Information Engineering, the University of Sydney, Sydney, Australia

Email: zhumengyao@shu.edu.cn

ABSTRACT

In recent years, deep learning approaches have gained significant interest as a way of building hierarchical representations from unlabeled data. Additionally, in the field of sound direction-of-arrival (DOA) estimation, the binaural features like interaural time or phase difference and interaural level difference, or monaural cues like spectral peaks and notches are often used to estimate sound DOA. Although these binaural or monaural cues successfully explained human sound DOA ability, its accuracy and extent are all limited to the human knowledge on feature extracting methods. In this paper, we are interested in applying a deep learning approach to monaural sound localization based on monaural spectral cues. A convolutional deep belief network is applied to monaural auditory spectrograms processed by a computational auditory model to learning monaural features automatically. The learned features are then regressed using a support vector regression (SVR) model for sound DOA estimation tasks. Moreover, our feature representations learned from unlabeled monaural auditory spectrograms are then compared to the well-known binaural features. The results indicate that our monaural model shows reasonable performance at the task of 3D DOA estimation.

Keywords: Unsupervised feature learning; Monaural DOA estimation; SVR

1. INTRODUCTION

Normal-hearing human listeners have a remarkable performance in estimating the direction of arrival (DOA) of specific sound sources even in the presence of a noisy background, in reverberation, or in the presence of other concurrent sound sources. Human sound DOA evaluation is generally acknowledged to be a binaural process that depends predominantly on the use of the interaural time or phase difference (ITD and IPD, respectively) and the interaural level difference (ILD). However, it has been clear for some time that there are essential features of human sound DOA estimation that can not be explained by ITD and ILD alone. For example, in the median plane, the binaural cues are virtually absent and, although they may still be perceptually relevant such as for externalization of sounds, they are probably not for sound localization [1]. As well, with respect to locations outside the median plane, the interaural differences are approximately constant around an entire cone of positions, the so-called “cone of confusion”. It is now well know that the folds of the pinnae perform a direction-dependent filtering of the incoming sound, providing additional (monaural spectral) cues essential for estimating sound DOA on these cones.

The studies of monaural sound DOA estimation have captured the interest of hearing scientists since early 21th century. In the past, Zakarauskas and Cynader [2] and Hofman and Opstal [3] were among the first proposing explicit functional models of sound localization based on monaural spectral cues. Later, Langendijk and Bronkhorst [4] and Bremen *et al*. [5] used a probabilistic approach to model their results from sound DOA estimation experiments. All those models roughly approximate peripheral auditory processing in order to obtain internal representations of the incoming sounds. Furthermore, they follow a template-based approach, assuming that listeners create an internal template of their specific head-related transfer function (HRTF) as a result of monaural learning process [6]. The more similar the representations of the incoming sound compared to a template entry, the larger the assumed probability of responding at polar angle corresponding to that entry. Langendijk and Bronkhorst [4] demonstrated a correspondence between their model

predictions and experimental outcomes for individual listeners by means of likelihood statistics.

However, these monaural spectrums localization models have some significant weaknesses. The first problem is that extraction of pinnae filtering characteristics from an incoming sound requires knowledge of the spectrum of the sound source. And it is clearly unreasonable to postulate that listeners know, in any precise sense, the spectra of all potential sounds. The second problem is that these monaural DOA estimation models are based on template comparison approaches which require an internal template of listener's specific HRTFs and estimate the final sound DOA through comparing the auditory spectrums with the specific subject's HRTFs. These processes all depend on the subject, so their extent and accuracy are restricted to human knowledge and computational complexity. Furthermore, the monaural spectrums are only used for some particular planes or distinguishing front-back directions.

In recent years, deep learning approaches have gained significant interest as a way of detecting features automatically from high dimensional data, such as image or speech. Many promising approaches have been proposed to process the steps of deep networks ([7], [8], [9], [10]). Deep learning algorithms try to learn simple features in the lower layers and more complex features in the higher layers from image or video data. They have been successfully used in a wide variety of domains, including handwritten digits and human motion capture data and face recognition etc. Deep learning approaches have also been applied to auditory data. In [11], the convolutional deep belief networks (CDBN) have been applied to auditory data and show good performance for multiple audio classification tasks. In [12], Hamel applied a deep belief network (DBN) to extract relevant features from audio for music information retrieval tasks through SVM and the learned features performed significantly better than conventional MFCCs. In [13], Nam applied the DBN to musical data and evaluated the detected features on classification-based polyphonic piano transcription. The results showed the learned features outperformed the baseline features.

In this paper, we apply convolutional deep belief networks to unlabeled monaural auditory spectrograms (data processed by an auditory model) and evaluate the learned feature representations for sound DOA estimation through a support vector regression (SVR) model. We compare the localization performance with the binaural models using ITD, ILD and interaural across correlation coefficient (IACC). The results are quite equivalent for sound DOA estimation.

The paper is organized as follows. In the section 2, we introduce the convolutional deep belief network and human auditory processing model respectively. The detail process of unsupervised feature learning is given in section 3. The DOA estimation experiment and the results analysis is given in section 4. Finally, conclusions are presented in section 5.

2. ALGORITHMS

2.1 Convolutional Deep Belief Network

First, we briefly review the Convolutional Restricted Boltzmann Machine (CRBM) ([10], [14], [15]) as building blocks for convolutional deep belief networks (CDBN). The CRBM is similar to the RBM [16], but the weights between the hidden and visible layers are shared among all locations in the hidden layer.

The energy function for RBM can be defined as:

$$E(\mathrm{v},\mathrm{h}) = -\sum_{m,n} v_m W_{mn} h_n - \sum_n b_n h_n - \sum_m c_m v_m \tag{1}$$

where v_m is *m-th* visible unit while h_n is *n-th* hidden unit, b_n are hidden unit biases and c_m are visible unit biases.

The joint probability distribution of (v, h) is defined as follows:

$$P(\mathrm{v},\mathrm{h}) = \frac{1}{Z}\exp(-E(\mathrm{v},\mathrm{h})) \tag{2}$$

$$Z = \sum_{v,h} e^{-E(v,h)} \tag{3}$$

Then the block Gibbs sampling of RBM can be performed using the following conditional distributions:

$$p(h_n = 1|\mathrm{v}) = \sigma(\sum_m W_{mn} v_m + b_n) \tag{4}$$

$$p(v_m = 1|\mathrm{h}) = \sigma(\sum_n W_{mn} h_n + c_m) \tag{5}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$.

Since all units in one layer are conditionally independent given the other layer, influence in the network is efficiently performed using block Gibbs sampling. Lee [17] further developed a convolutional RBM with "probabilistic max-pooling", which shrinks the representation of the detection layer by a constant factor. More specifically, each unit in a pooling layer computes the maximum activation of the units in a small region of the detection layer. Shrinking the representation with max-pooling allows higher-layer representations to be invariant to small translations of the input and reduces the computational burden.

As shown in figure 1, the basic CRBM consists of two layers: an input (visible) layer V and a hidden layer H. The input layer consists of an $N_V \times Nv$ array of binary units. The hidden layer consists of K group, where each group is an $N_H \times N_H$ array of binary units, resulting in $N_H{}^2K$ hidden units. Each of the K groups is associated with a $N_W \times N_W$ filter ($N_W = N_V - N_H + 1$). The filter weights are shared across all the hidden units within the group. In addition, each hidden group has a bias b_k and all visible units share a single bias c. The pooling layer also has K groups of units, and each group of the pooling layer has $N_P \times N_P$ binary units. For each $k \in \{1,...K\}$, the pooling layer P^k shrinks the representation of the detection layer H^k by a factor of C along each dimension, where C is a small integer. The detection layer H^k is partitioned into blocks of size $C \times C$, and each block α is connected to one binary unit p_α^k in the pooling layer (i.e., $N_p = N_H/C$). The energy function of this probabilistic max-pooling CRBM is then defined as follows.

$$E(\mathrm{v},\mathrm{h}) = -\sum_k \sum_{i,j} \left(h_{i,j}^k \left(\tilde{W}^k * v \right)_{i,j} + b_k h_{i,j}^k \right) - c \sum_{i,j} v_{i,j} \tag{6}$$

$$\text{subject to } \sum_{(i,j)\in B_\alpha} h_{i,j}^k \le 1, \forall k, \alpha \tag{7}$$

where i and j is respectively the index of a block α, k is the hidden group index.

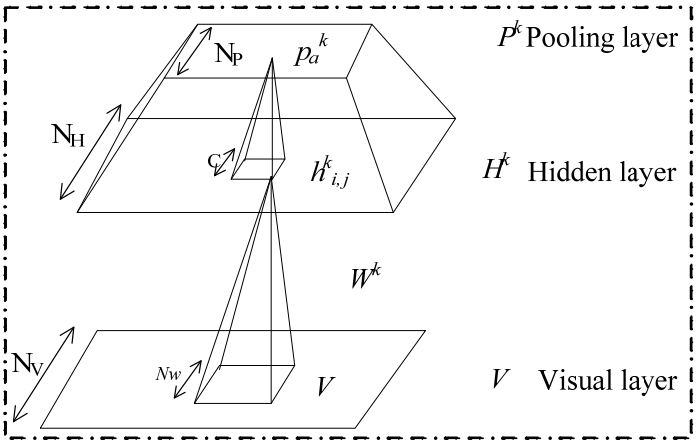

Figure 1 – Convolutional RBM with probabilistic max-pooling. For simplicity, only group k of the hidden layer and the pooling layer are shown.

Therefore, when given the visible layer V, the detection layer H and pooling layer P can be calculated. Group k receives the following bottom-up signal from layer V:

$$I(h_{i,j}^k) = b_k + (\tilde{W}^k * v)_{ij} \tag{8}$$

The conditional probability is given by:

$$P(h_{i,j}^{k}=1|\mathrm{v})=\frac{\exp(I(h_{i,j}^{k}))}{1+\sum_{(i',j')\in B_{\alpha}}\exp(I(h_{i',j'}^{k}))} \tag{9}$$

$$P(p_{\alpha}^{k}=1|\mathrm{v})=1-\frac{1}{1+\sum_{(i',j')\in B_{\alpha}}\exp(I(h_{i',j'}^{k}))} \tag{10}$$

For training the CRBM, computing the exact gradient for the log-likelihood term is intractable, but contrastive divergence (CD) [16] can be used to approximate the gradient effectively. We can obtain the CDBN by stacking several max-pooling CRBMs on top of one another. Training of CDBN is accomplished with the greedy, layer-wise procedure described in [17]. Once a given layer is trained, its weights are frozen, and its activations are used as input to the next layer.

2.2 Auditory Processing

For applying the CDBN to monaural auditory spectrograms, we first apply an auditory model to the audio signal. The effective auditory model we take in the peripheral auditory part of CASP (Computational auditory Signal Processing and Perception) [18] model. But in this paper, we remove the modulation filter-bank and optimal detector of CASP model because its output is three dimensional which is complicated for CDBN. The successive processing stages follow the same order as in the human ear and are briefly described below.

(1) The input is scaled to be represented in pascal and filtered with two linear phase FIR filters, simulating the outer and middle ear transfer functions.

(2) The frequency analysis performed by the basilar membrane is modeled with a DRNL (Dual-Resonance Nonlinear) filter bank. In DRNL filter bank, linear and compressively nonlinear processing units operate in parallel, simulating differences in input/output functions among three different sites along the cochlear partition. At low signal levels (below 30-40 dB SPL), the nonlinear part behaves linearly. At medium signal levels (40-70 dB SPL), the nonlinear part is compressive. At high signal levels (above 70-80 dB SPL), the output of the linear path dominates the sum. The filtered audio signal is spilt into 38 frequency sub-bands spaced on an ERB (equivalent rectangular bandwidth) scale ranging from 80Hz to 18000Hz after DRNL filtering.

(3) The transformation of elastic waves into electric neural signals by the hair cells is simulated by a half-wave rectification followed by a low pass filter. This allows to keeping the fine structure of the signal at low frequencies while extracting the envelope of the signal at high frequencies. Each sub-band signals is processed by half-wave rectification and low pass filter, obtaining a sub-band signal envelope.

(4) A squaring expansion is introduced to reflect the square-law behavior of rate-versus-level functions of the neural response in the AN. After this step, we can get audio signal energy for each sub-band.

(5) To account for the non-linear adaptation stage, a chain of five feedback loops circuits are adopted to make the stationary inputs more compressed while the non-stationary inputs with less compression.

(6) To imitate the frequency masking effects of the auditory system, we use the method presented in our previous work ([19], [20]), which consists in adding frequency spreading and masking effects model to solve the problem that the signal energy is centered at low frequency and simulate frequency masking effects of human ear. We then can obtain 38 sets of time data, namely an auditory spectrogram.

3. UNSUPERVISED FEATURE LEARNING

In this section, we illustrate what the CDBN network “learns” from a set of monaural auditory spectrograms of different sound DOAs and show the specific process of how to extract the monaural features through visualization.

We trained the first and second-layer CDBN representations using large unlabeled monaural auditory spectrograms. First, we convolved a single-channel anechoic sound (a saxophone music is used) with left-ear head related impulse responses (HRIRs) of different directions and to obtain a set of left ear sound signals. Then we input these sound signals into the CASP model imitating the auditory process of human ear and obtained auditory images of different directions. Because each auditory spectrogram is not square and its size is considerable large, the input image size of original CDBN was adjusted to match the left-ear auditory spectrogram size. Similarly, the max-pooling ratio

(local neighborhood size) was adjusted for non-square. We then trained the first-layer features with filter size (N_W) of 7, and the local neighborhood size was 2×5. We further trained the second-layer features using the max-pooled first-layer activations as input with a filter length of 5 and a max-pooling ratio of 2.

In Figure 2, we show the auditory spectrogram of original dry sound signal and two randomly selected wet acoustic signal auditory spectrograms which correspond to two different sound directions. The length of sound is about 23ms. Different colors represent different energies in the figure. For each spectrogram, the abscissa indicates time samples which is 87 samples because a frame length is 1024 and in the process of backward masking, there is a down-sampling by a factor of 6. The ordinate represents 38 frequency sub-bands and 1-38 respectively corresponds to 80-18000Hz. We will describe in detail the methods to learn monaural features using a CDBN.

When we input the first left-ear auditory spectrogram into the CDBN model, it randomly initializes a group of weight *W*, visual layer biases *c* and hidden layer biases *b*. According to the formula (8), we can calculate the bottom-up signals *I*, therefore, the conditional probability of hidden layer can be obtained by equation (9). Based on the independence between the input layer and the hidden layer, the reconstituted input layer probability can be computed by equation (5) and then the reconstituted hidden layer probability can be obtained by formula (9). A Gibbs sampling process is accomplished. After that, CDBN model begins to compute increments of its weight *W*, bias *b*, *c* and update the above three parameters according the following equations.

$$\Delta W_{ij} = \varepsilon\left(\left\langle v_i h_j \right\rangle_{\text{data}} - \left\langle v_i h_j \right\rangle_{\text{recon}}\right) \tag{11}$$

$$W = W + \Delta W \tag{12}$$

$$\Delta c_i = \varepsilon\left(\left\langle v_i \right\rangle_{\text{data}} - \left\langle v_i \right\rangle_{\text{recon}}\right) \tag{13}$$

$$c = c + \Delta c \tag{14}$$

$$\Delta b_j = \varepsilon\left(\left\langle h_j \right\rangle_{\text{data}} - \left\langle h_j \right\rangle_{\text{recon}}\right) \tag{15}$$

$$b = b + \Delta b \tag{16}$$

where ε is learning rate, $\langle\cdot\rangle_{\text{data}}$ represents original model distribution, $\langle\cdot\rangle_{\text{recon}}$ represents reconstituted model distribution.

After updating the above three parameters, the CDBN model has a new weight *W*, bias *b* and *c* which correspond to the minimum error between original input image and its reconstituted image. Then, the reduced information is input to the CDBN model, convolves with the new weight *W*, adds new hidden layer bias *b* and also gets a bottom-up signals *I* , the following steps are identical to the steps of the input spectrogram, so we won't repeat. Similarly, the other left-ear auditory spectrograms are also exactly the same as the first two inputs until all auditory images are processed, the training process is finished. When all the parameters of CDBN have been trained, we compute the pooling-layer's representations of auditory spectrograms through shrinking the representations of the hidden layers by a constant factor. Specifically, the first auditory spectrogram convolves with the trained weight *W* and adds bias, obtaining the bottom-up signals *I* by the equation (8). According to the formula (10), we can calculate the pooling-layer's representations of the first spectrogram. So do the other auditory spectrograms.

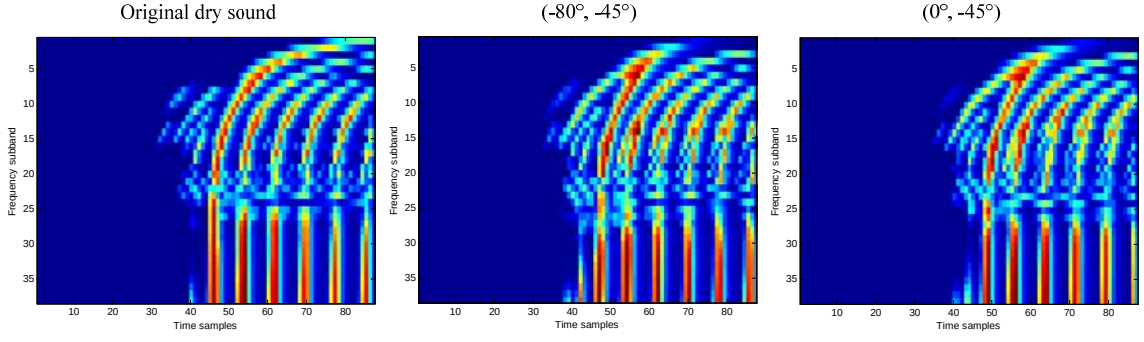

Figure 2 – (Left) Auditory spectrogram of original dry sound. (Middle) Auditory spectrogram of (-85°, -45°) wet acoustic signal. (Right) Auditory spectrogram of (0°, -45°) wet acoustic signal. The horizontal axis represents 87 time samples while the vertical axis represents 38 frequency sub-bands.

Once representations of pooling layers in the first-layer CDBN have all been computed, we use these activations as the input of the second-layer CDBN and the inference process is similar with the first-layer. The features shown in this paper refer activations in pooling layers.

In Figure 3, we show an analysis of two layers CDBN feature representations with respect to the DOA estimation tasks (Section 4). Note that the network was trained on unlabeled data, therefore, no information about original sound directions was given during training.

For comparison with the monaural CDBN features, randomly selected spectrograms of two different directions are shown. For the first-layer, size of feature maps is 16×16 which corresponds to 16×16 small patches. Each small patch corresponds to a unit in pooling layer and different colors in the image represent different activations. In the two feature maps, there are always some patches that their colors are different and these differences are used to distinguish the different directions of a sound. For the second-layer feature images of the CDBN, the size of them is 6×6, and each direction has a region which differs from other directions. Although we only display two learned features of sound directions, the features of other directions are also the similar.

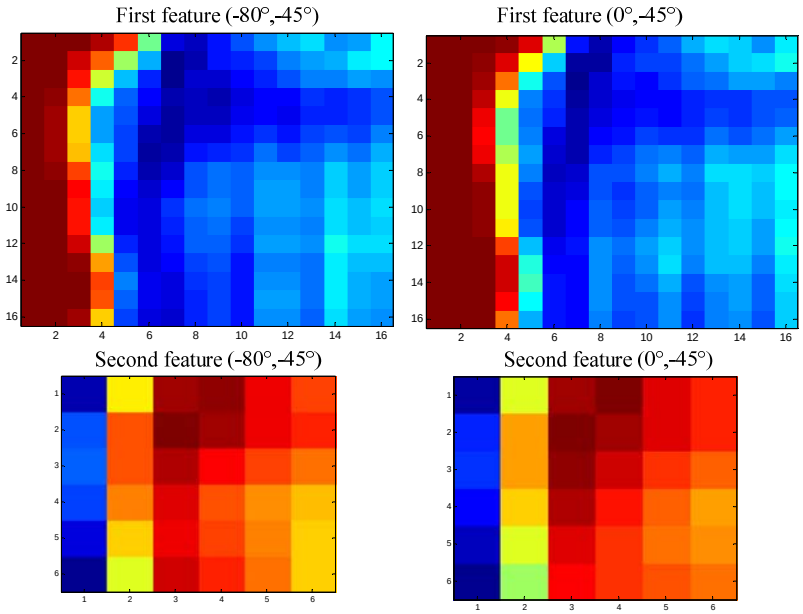

Figure 3– (Top) Visualization of two first-layer feature representations that differentially activate for two sounds directions of arrival (-80°,-45°)/ (0°,-45°). (Bottom) Visualization of the two second-layer feature representations that differentially activate for the same direction of arrival.

4. APPLICATION TO DOA ESTIMATION

In this section, we demonstrate that the monaural CDBN feature representations learned from the unlabeled auditory spectrograms of different sound directions can be useful for DOA estimation tasks. One constant sound source is considered in the following experiments.

We implemented the experiment on CIPIC HRTF database which was recorded in the CIPIC interface Laboratory at the UCDavis [21]. In the database, there are 1250 HRIRs (Head Related Impulse Response) represented in interaural polar coordination. Each direction corresponds to one of 50 different elevations form -45° to 230.625° in step of 5.625°, and one of 25 azimuths from -80° to 80° in uneven steps. A mono anechoic sound is convoluted with the left-ear HRIRs described above. After auditory model processing, we get 1250 spectrograms of different directions. We performed azimuths on this set and elevations evaluation on the other set in which elevation is between -45° and 90° in step of 5.625° because 90° to 237.625° is identical with -45° to 90° in elevation. For all directions, we randomly selected 70% data for training and 30% for testing. We measured the

evaluation accuracy by calculating Root Mean Square Error (RMSE) and the correlation between the measured angles in HRTF database and the evaluated angles. We computed the first and second-layer monaural CDBN features using the auditory spectrograms as input. We also computed ITD, ILD and IACC binaural features, widely-used and the most important features for generic DOA evaluation task. As a result, we evaluated these two types of features using standard supervised regression method, such as SVR. We report the estimation correlation and RMSE in the following tables and figures.

Table 1 –Azimuth estimation results of different features

Features / Results	ITD&ILD&IACC	CDBN L1	CDBN L2
Correlation(%)	98.1	98.7	98.4
RMSE (°)	6.446	4.937	5.546

Table 2 –Elevation estimation results of different features

Features / Results	ITD&ILD&IACC	CDBN L1	CDBN L2
Correlation (%)	98.1	98.5	98.3
RMSE (°)	6.390	5.174	5.643

Table 1 shows the azimuths estimation accuracy for each feature representations while Table 2 shows the precision of elevations estimation. In the tables, CDBN L1 represents the first-layer monaural features of CDBN and CDBN L2 means the second-layer activations. Comparing the second column with the third column, we can see that the correlation between original directions and estimated directions using first-layer CDBN representations is only improved by 0.6% for azimuths and 0.4% for elevations, but the RMSE dropped by 23% and 19%, respectively. It is proved that the learned monaural features can decrease the estimation RMSE value, so it even outperforms the most advanced binaural features (ITD, ILD and IACC) for DOA estimation. Comparing the second, third and the forth columns, the performance of three methods is quite equivalent.

Please note that, both elevation and azimuth estimation performance of CDBN methods with 1 layer and 2 layers are a bit better than binaural method in terms of correlation and RMSE. Consequently, our monaural CDBN representations are reasonable good comparing with state-of-the-art binaural features in the task of 3D DOA evaluation. After training process, CDBN's calculated is far less than traditional binaural feature extracting method.

Additionally, according to the traditional source localization theory, there is no interaural time difference and interaural level difference in the median plane. In fact, except the median plane, the interaural differences are approximately constant within an entire cone of positions called "cone of confusion". However, it's important to note that in our proposed method, the binaural cues (ITD, ILD and IACC) can achieve quite accurately elevation estimation. The reason for this is that in our experiment, we estimate elevations not only in the median plane but also in other vertical plane in which binaural cues are not absent.

5. CONCLUSIONS

In this paper, we applied convolutional deep belief network to monaural auditory spectrograms and evaluated these learned features on sound DOA estimation tasks. By leveraging a large amount of unlabeled data, our learned features often equaled or surpassed the binaural features like ITD, ILD and IACC, which are the most important features for human sound DOA estimation. Our experiment results show monaural features can also achieve good accuracy in sound DOA estimation by some

certain methods, which is important for the application of humanoid robotic sound localization. However, we have to emphasize that the work is presented in this paper is preliminary. In future work, we will vary the sound source, add reverberation and noise, and also multiple sounds.

ACKNOWLEDGEMENTS

This work was supported in part by Innovation Program of Shanghai Municipal Education Commission No.12YZ024 and China Scholarship Council.

REFERENCES

1. M. Morimoto, K. Nomachi, “Binaural disparity cues in median-plane localization,” Journal of the Acoustical Society of Japan (E), 3(2), pp. 99-103, 1982.
2. P. Zakarauskas, M. S. Cynader, “A computational theory of spectral cue localization,” The Journal of the Acoustical Society of America, 94(3), pp. 1323-1331, 1993.
3. P. M. Hofman, A. J. Opstal, “Spectro-temporal factors in two-dimensional human sound localization,” The Journal of the Acoustical Society of America, 103(5), pp. 2634-2648, 1998.
4. E. H. A. Langendijk, A. W. Bronkhorst, “Contribution of spectral cues to human sound localization,” The Journal of the Acoustical Society of America, 112(4), pp. 1583-1596, 2002.
5. P. Bremen, M. M. Wanrooij, A. J. Opstal, “Pinna cues determine orienting response modes to synchronous sounds in elevation,” The Journal of Neuroscience, 30(1), pp. 194-204, 2010.
6. W. W. Wanrooij, A. J. Opstal, “Relearning sound localization with a new ear,” The Journal of neuroscience, 25(22), pp. 5413-5424, 2005.
7. M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun. “Efficient learning of sparse representations with an energy-based model,” Advances in neural information processing system, pp.1137-1144, 2006.
8. Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. “Greedy layer-wise training of deep networks,” Advances in neural information processing system, 2007.
9. H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. “An empirical evaluation of deep architectures on problems with many factors of variation,” Proceedings of the 24th International Conference on Machine learning. ACM, pp.473-480, 2007.
10. H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations,” Proceedings of the 26th Annual International Conference on Machine learning. ACM, pp.609-616, 2009.
11. H. Lee, P.T. Pham, Y. Largman, et al. “Unsupervised feature learning for audio classification using convolutional deep belief networks,” *NIPS*. vol. 9, pp.1096-1104, 2009.
12. P. Hamel, D. Eck, “Learning Features from Music Audio with Deep Belief Networks,” ISMIR, pp.339-344, 2010.
13. N. Juhan, N. Jiquan, L. Honglak, et al. “A Classification-Based Polyphonic Piano Transcription Approach Using Learned Feature Representations,” ISMIR, pp. 175-180, 2011.
14. G. Desjardins, Y. Bengio, “Empirical evaluation of convolutional RBMs for vision,” Technical report, pp.1-13, 2008.
15. M. Norouzi, M. Ranjbar, G. Mori, “Stacks of convolutional restricted Boltzmann machines for shift-invariant feature learning,” Computer Vision and Pattern Recognition, 2009, CVPR 2009, IEEE Conference on. IEEE, pp.2735-2742, 2009.
16. G. E. Hinton, S. Osindero, Y. W. Teh. “A fast learning algorithm for deep belief nets,” Neural computation, 2006, 18(7), pp.1527-1554, 2006.
17. H. Lee, R. Grosse, R. Ranganath, et al, “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations,” Proceedings of the 26th Annual International Conference on Machine Learning. ACM, pp.609-616, 2009.
18. L.J. Morten, D.E. Stephan, D. Torsten, “A Computational Model of Human Auditory Signal Processing and Perception,” *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 422-438, 2008.
19. J. Zheng, M.Y. Zhu, J.W. He, X.Q. Yu, “PEAQ Compatible Audio Quality Estimation Using Computational Auditory Model,” *ICONIP2012*, Nov.12-15, Doha, Qatar, 2012, LNCS 7666, pp. 83-90.
20. Zhu M, Zheng J, Jin C, et al. “Structural Similarity Analysis of Modulation for audio quality assessment,” Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, pp. 403-407, 2013.
21. V.R. Algazi, R.O. Duda, D.M. Thompson, and C. Avendano, “The CIPIC HRTF database,” In Proc. 2001 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.99–102,2001.