INTERSPEECH 2014

Binaural Deep Neural Network Classification for Reverberant Speech Segregation

Yi Jiang[1], *DeLiang Wang*[2], *RunSheng Liu*[1]

[1]Department of Electronic Engineering, Tsinghua University, Beijing 100084, P.R. China
[2]Department of Computer Science & Engineering and Center for Cognitive & Brain Sciences,
The Ohio State University, Columbus, Ohio 43210, USA

jiangyi09@mails.tsinghua.edu.cn, dwang@cse.ohio-state.edu,lrs-dee@tsinghua.edu.cn

Abstract

While human listening is robust in complex auditory scenes, current speech segregation algorithms do not perform well in noisy and reverberant environments. This paper addresses the robustness in binaural speech segregation by employing binary classification based on deep neural networks (DNNs). We systematically examine DNN generalization to untrained configurations. Evaluations and comparisons show that DNN based binaural classification produces superior segregation performance in a variety of multisource and reverberant conditions.

Index Terms: Computational auditory scene analysis (CASA), binaural speech segregation, deep neural networks (DNN), binary classification, room reverberation

1. Introduction

The performance gap between human listeners and speech segregation systems remains large in noisy and reverberant environments despite extensive research in signal processing. Inspired by human auditory processing, computational auditory scene analysis (CASA)[1] has shown considerable promise in the last decade. A commonly used computational goal in CASA is the ideal binary mask (IBM), which is a two-dimensional matrix of binary label where 1 indicates the target signal dominates the corresponding time-frequency (T-F) unit and 0 otherwise. The estimation of the IBM thus corresponds to binary classification of T-F units. So far, classification based speech separation primarily uses monaural features such as pitch, amplitude modulation spectrogram (AMS), and gammatone frequency cepstral coefficients (GFCC) [2, 3].

With two ears, human listening is robust under both noisy and reverberant conditions. Binaural cues contribute to auditory scene analysis[4]. Speech segregation by classifying binaural feature was investigated before using kernel density estimation [5]. In this study, we explore deep neural networks (DNNs) [6]for binaural classification in order to achieve robust speech segregation.

There are a number of binaural features, peak interaural time difference, interaural coherence, and interaural phase difference [7, 8, 9]. We focus on two principal binaural cues, interaural time difference (ITD) and interaural level difference (ILD) [5, 10], which have been widely used as sound localization and location based segregation [1, 11].

In the following section, we present an overview of our classification-based binaural speech segregation system. Section 3 describes how to extract binaural features and perform DNN classification. We present the evaluation results in Section 4, including on untrained source locations. Comparisons with several related systems are also presented in this section. We conclude the paper in Section 5.

2. System overview

The proposed classification-based binaural speech segregation system is shown in Figure 1. The same two auditory filterbanks are used to decompose the left-ear and right-ear input signals into T-F units. A T-F unit corresponds to a certain channel in a filterbank at a certain time frame. Binaural features are calculated for each corresponding pair of T-F units. As ITD and ILD features vary with frequency channels [5, 8], we train a DNN classifier for each frequency channel, and the training labels are provided by the IBM. In grouping, the T-F units with the target label (unity) comprise the segregated target stream.

The binaural input signals are produced by a set of binaural impulse responses (BIRs). The ROOMSIM package[12], which uses measured head related transfer functions (HRTFs) from the KEMAR dummy head in combination with the image method for simulating room acoustics, is used to generate the BIRs. The training and test speech signals are chosen from the TIMIT corpus. We use the babble noise in the NOISEX corpus [13] The signals were originally sampled at 16 kHz. We upsampled them to 44.1 kHz to match the sampling rate of the BIRs, and then downsampled to 16 kHz for periphery and subsequent processing.

3. Binaural feature extraction and classification

3.1. Auditory periphery

We use the gammatone filterbank[15] as the auditory filterbanks in Figure 1. The gammatone filterbank shows equivalent rectangular bandwidth (ERB), and each filter's impulse response is described below:

$$g(t, f_c) = t^{n-1} e^{-2\pi b(f_c)t} \cos(2\pi f_c t + \phi) \quad t \geq 0 \qquad (1)$$

where c denotes the filter channel, and we use a total of 64 channels for each ear model. The center frequency of the filter, f_c , varies from 50Hz to 8000Hz. $b(f_c)$ indicates the bandwidth. The filter order, n , is 4, and ϕ is the phase which is set to zero. This peripheral analysis is commonly used in CASA[1].

The gammatone filter responses are further decomposed into time frames. Here we use 20-ms frame length with 10-ms frame shift. The resulting T-F representation is called a cochleagram, with a two-dimensional matrix of T-F units.

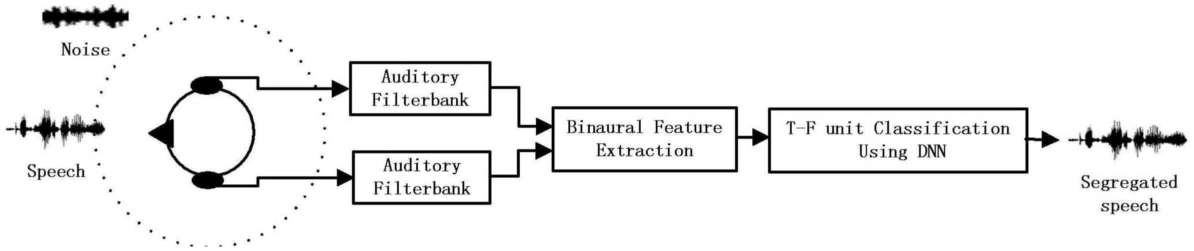

Figure 1: *Schematic diagram of the proposed binaural DNN classification system.*

3.2. Feature extraction

With the available binaural signals, we extract the two primary binaural features of ITD and ILD. The ITD is calculated from the normalized cross-correlation between the two ear signals (denoted as l, r , for left and right ear respectively) in frequency channel c and time frame m . The cross-correlation function, indexed by time lag τ , is described in the following equation:

$$CCF(c,m,\tau) = \frac{\sum\limits_{n}(x_{cm,l}(n) - \bar{x}_{cm,l})(x_{cm,r}(n-\tau) - \bar{x}_{cm,r})}{\sqrt{\sum\limits_{n}(x_{cm,l}(n) - \bar{x}_{cm,l})^2}\sqrt{\sum\limits_{n}(x_{cm,r}(n-\tau) - \bar{x}_{cm,r})^2}} \quad (2)$$

In the above equation, n indexes the signal sample in the T-F unit. The time lag τ varies between -1 ms and 1 ms. The overbar indicates averaging. For the 16 kHz sampling rate, we obtain 32-D features for each pair of T-F units.

The ILD corresponds to the energy ratio in dB between two ears:

$$ILD(c,m,i) = 10\log\frac{x_l(c,m,i)^2}{x_r(c,m,i)^2} \quad (3)$$

where i indexes the ILD(c,m) feature, and is set to 2 in this study. With 20-ms frame length, each feature index corresponds to an ILD value over a 10-ms duration. Overall, we obtain a 34-D binaural feature vector for each unit pair.

3.3. DNN classification

Each subband DNN classifier consists of an input layer, two hidden layers and an output layer. The 34-D binaural features are the inputs. Each layer contains 200 hidden neurons. We follow the approach in [3] where restricted Boltzmann machines (RBMs) are used for pre-training. The learning rate of the pre-training is set to 0.001 for the first hidden layer, and 0.1 for the other hidden layers. The batch size is 256 and the momentum rate is set to 0.5. Finally, the standard back-propagation optimization is applied for supervised fine-tuning. The learning rate decreases linearly from 1 to 0.001 in 50 epochs. The output layer labels a T-F unit as 1 if target speech dominates or 0 otherwise.

4. Evaluation and comparison

In this section, we present two sets of experiments to evaluate the speech segregation performance of the proposed system. With the ROOMSIM package and KEMAR dummy head, we create a library of BIRs. The simulated room has the dimension of 6m×4m×3m. The position of the listener is fixed at 2.5m×2.5m×2m. Reflection and absorption coefficients of the wall surfaces are uniform. The reflection paths of a particular sound source are obtained using the image model for a small rectangular room. The reverberation times (T_{60}) are approximately set to 0.3s and 0.7s. We also use the anechoic setting as a baseline. All sound sources are presented at the same distance of 1.5 m from the listener (in the available space of each room configuration). We generate BIRs for azimuth angles between zero and 360°, spaced by 5°. All elevation angles are set to zero. Speech utterances and babble noise are drawn randomly from the database and are convolved with selected BIRs to generate the mixtures with defined SNRs. We generate 600 mixtures(about 120000 T-F units) to train the DNN classifiers, and use 50 sentences to evaluate the performance of the proposed algorithm in each condition. We also use 2000 mixtures to train DNN classifiers, and the HIT-FA results have about two percent improvements.

In the experiments below, we compare the performance of the proposed method with three representative binaural separation methods from the literature. Roman et al.' method[16] performs binaural segregation in multi-source reverberant environments. The target attenuation correlates with the relative strength of the target to the mixture. The second comparison method is the joint localization and segregation approach presented in[9], dubbed MESSL, which uses spatial clustering for source localization. The system requires the specification of the number of sources and iteratively fits GMM models of interaural phase difference and ILD to the observed data using an EM procedure. Across frequency integration is handled by tying the GMM models in individual frequency bands to a principal ITD. DUET [17] is a popular blind source separation method capable of separating an arbitrary number of sources using only two mixtures (microphones). For Roman et al. and MESSL, we use the implementations provided by their respective authors. The DUET implementation comes from its author's book [17]. All of the comparison systems' parameters are adjusted to get the optimal results.

To measure classification-based separation performance, we use HIT-FA as our main evaluation criterion, which has been shown to be well correlated to human intelligibility [18].The HIT rate is the percent of correctly classified target-dominant T-F units in the IBM. The FA (false-alarm) rate is the percent of wrongly classified interference-dominant T-F units. In addition to this measure of classification accuracy, we adopt the IBM-modulated SNR measure to give another indication of the segregation performance, where the resynthesized speech from the ideal binary mask is used as the ground truth [1].

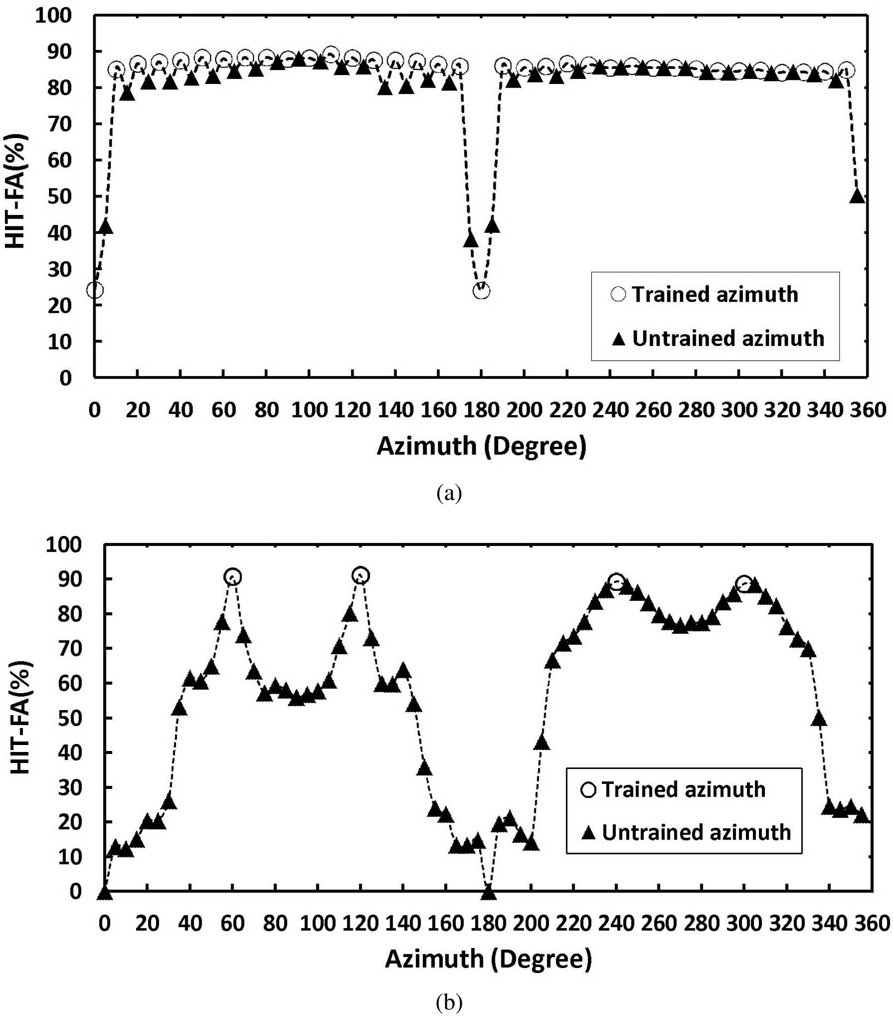

Figure 2: HIT-FA performance for two-source segregation at various interference training azimuths.

4.1. Two-source segregation

In this set of experiments, we fix the target source at azimuth 0°, i.e. just in front of the dummy head. Then we train the DNNs classifiers at various noise locations. All training mixtures have 0 dB SNR.

As shown in Figure 2(a), the interference source used in training systematically varies between 0° and 350°, spaced by 10°. In testing, we place the interference source at the azimuths between 0° and 355° in 5° steps. In this way, in half of the test configurations the interference azimuths are not used in training whereas the other half of the test configurations used the trained interference angles. As shown in Figure 2(a), the HIT-FA rates are above 80% for most interference azimuths and are close to 90% for some azimuths. When the interference locations are close to the target sound, at azimuths of 0°, 5°, 175°, 180°, 185° and 355°, the HIT-FA rates are down to 25%. As expected, the proposed system cannot separate the target speech from a nearby interference source on the basis of binaural cues. Note that, even though the target azimuth is at 0°, the binaural cues at the azimuth of 180° (i.e. right in the back) are almost the same as those at the target azimuth. The trained locations yield the higher HIT-FA rates than the nearby untrained locations. At the better ear, for the interference source located between 185° and 355°, the performance differences between trained locations and untrained locations are small. In Figure 2(b), we train the system at 4 interference locations, at azimuth 60°, 120°, 240° and 300°. These trained locations produce the four peak points of HIT-FA rates. The HIT-FA rates decrease as the test interference locations move away the trained locations. Comparing the results in Figure 2(a) and Figure 2(b), it is clear that the more the trained angles cover the azimuth space, the better the trained system performs at untrained angles.

As another evaluation, we use the babble noise located between 0° and 350° spaced by 10° to train the DNNs. Then an untrained interference angle located at 45° is used to test the system. In this test condition, we now vary the input SNR and the classification and SNR results are shown in Table 1. As shown in the table, the proposed system produces strong performance in terms of HIT-FA rate and SNR. With the input SNR decreases, the HIT-FA rate decreases gradually. Even at the input SNR of -15 dB, the HIT-FA rate of 78.91% is still high in comparison with the monaural separation method in [2]. Our informal listening shows that we can recognize segregated speech in this very low SNR condition.

Table 1: Two-source segregation results with respect to input.

Input SNR (dB)	HIT (%)	FA (%)	HIT-FA (%)	Output SNR (dB)
-15	80.96	2.06	78.91	1.79
-10	86.46	4.16	82.30	6.46
-5	89.10	4.40	84.70	10.72
0	92.68	7.23	85.45	14.34
5	94.30	9.41	84.89	17.29
10	94.64	10.89	83.75	18.45

4.2. Multi-source segregation with reverberation

In the second set of experiments, the SNRs of the training and test mixtures are set to 0dB. The target speech is again located right in front of the KEMAR dummy head, at 0°. Azimuths of the interfering sources are selected between 0° and 350°, spaced by 10°, to train the DNN classifiers. The four interfering sources are located at the azimuths of 45°, -45°, 135° and -135°. Note that test (evaluation) results are obtained from untrained interference locations.

4.2.1. Results without reverberation

We first present test results in anechoic conditions. The SNR results from our system and the three comparison systems are given in Table 2. The proposed system produces the best results in all test conditions. The MESSL results are better than those of the other two comparison systems, both of which also produce improved SNR in all test conditions.

Table 2: SNR performance in multi-source environments with no reverberation.

Sources	Proposed	Roman et al.	MESSL	DUET
2	14.34	5.06	11.16	3.22
3	8.76	4.55	7.43	4.53
5	8.55	4.10	7.54	5.34

4.2.2. Results with reverberation

In these test conditions, we use two simulated rooms with T_{60} set to 0.3s and 0.7s. For comparison, we also include the condition with $T_{60} = 0$s (i.e. the anechoic case). The evaluation is performed in the 5-source conditions. The SNR results from our algorithm and the comparison methods are plotted in Figure 3.

As shown in the Figure 3, the proposed system gives the best results in all reverberant conditions. As reverberation increases, the performance of the proposed system decreases rather gradually. The performance gap between our system and MESSL becomes larger in reverberant conditions.

5. Concluding remarks

In this study, we have proposed a DNN-based classification algorithm for binaural speech segregation. To our knowledge, this is the first study that introduces deep neural networks to location-based separation. The evaluation results show that the proposed system achieves better segregation than representative binaural separation algorithms. Even at very low input SNRs, the proposed system still yields good segregation performance. In addition, the performance decreases only gradually with increased room reverberation. The results from this initial evaluation also indicate encouraging generalization to untrained spatial configurations.

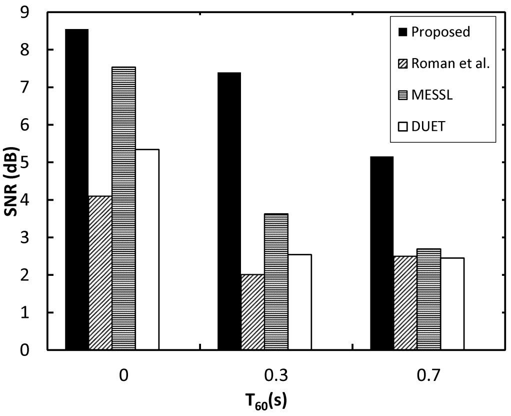

Figure 3: *SNR results in 5-source, reverberant conditions.*

We believe that the classification framework is a very promising direction for future development[19]. In this framework, for example, it is straightforward to include monaural features to complement binaural features to further improve segregation performance, especially when the target signal and interfering sources are either co-located or close to one another. We are currently developing this framework and expanding evaluation scenarios.

6. Acknowledgements

This work was performed while the first author was a visiting scholar at the Ohio State University. The authors wish to thank the Ohio Supercomputing Center for providing computing resources. The authors thank Nicoleta Roman, Michael Mandel and Yuxuan Wang for making implementations of their algorithms available to us.

7. References

[1] D.L. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley/IEEE Press, Hoboken, NJ, 2006.

[2] G. Kim, Y. Lu, Y. Hu, and P.C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 126, pp. 1486–1494, 2009.

[3] Y.X. Wang and D.L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 21, pp. 1381–1390, 2013.

[4] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, Cambridge, MA, 1990.

[5] N. Roman, D.L. Wang, and G.J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, pp. 2236–2252, 2003.

[6] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, pp. 1527–1554, 2006.

[7] T. May, S. van de Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 20, pp. 2016–2030, 2012.

[8] S. Keronen, H. Kallasjoki, U. Remes, and G. J. Brown, "Mask estimation and imputation methods for missing data speech recognition in a multisource reverberant environment," *Comput Speech Lang. Proc.*, vol. 27, pp. 798–819, 2013.

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

[10] 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, and Lang. Proc.*, vol. 19, pp. 1–13, 2011.

[11] J. Blauert, *Spatial Hearing The Psychophysics of Human Sound Localization*, MIT Press, Cambridge, MA, 1997.

[12] D. R. Campbell, *The ROOMSIM User Guide (v3.3)*, 2004, available: http://media.paisley.ac.uk/ campbell/Roomsim/.

[13] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database andan experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.

[14] Christopher Hummersone, Russell Mason, and Tim Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 18, pp. 1867–1871, 2010.

[15] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditoryfilterbank based on the gammatone function," Tech. Rep., APU, 1988.

[16] N. Roman, S. Srinivasan, and D. L. Wang, "Binaural segregation in multisource reverberant environments," *J. Acoust. Soc. Am.*, vol. 120, pp. 4040–4051, 2006.

[17] S. Makino, T. Lee, and Eds. H. Sawada, *Blind Speech Separation*, Springer, New York, 2007.

[18] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, pp. 1673–1682, 2008.

[19] N. Roman, D.L. Wang, and G.J. Brown, "A classification-based cocktail party processor," in *NIPS*, 2003, pp. 1425–1432.