INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Binaural reverberant Speech separation based on deep neural networks

Xueliang Zhang [1], *DeLiang Wang* [2,3]

[1] Department of Computer Science, Inner Mongolia University, China
[2] Department of Computer Science and Engineering, The Ohio State University, USA
[3] Center for Cognitive and Brain Sciences, The Ohio State University, USA

cszxl@imu.edu.cn, dwang@cse.ohio-state.edu

Abstract

Supervised learning has exhibited great potential for speech separation in recent years. In this paper, we focus on separating target speech in reverberant conditions from binaural inputs using supervised learning. Specifically, deep neural network (DNN) is constructed to map from both spectral and spatial features to a training target. For spectral features extraction, we first convert binaural inputs into a single signal by applying a fixed beamformer. A new spatial feature is proposed and extracted to complement spectral features. The training target is the recently suggested ideal ratio mask (IRM). Systematic evaluations and comparisons show that the proposed system achieves good separation performance and substantially outperforms existing algorithms under challenging multi-source and reverberant environments.

Index Terms: binaural speech separation, room reverberation, deep neural network (DNN), beamforming.

1. Introduction

In real-world environments, speech signal is usually degraded by concurrent sound sources and their reflections from the surfaces in physical space. Separating the target speech in such an environment is important for many applications such as hearing aid design, robust automatic speech recognition (ASR) and mobile communication. However, speech separation remains a considerable challenge despite extensive research over decades.

Since target speech and background noise usually overlap in time and frequency, it is hard to remove the noise without speech distortion in monaural separation. However, speech and interfering sources are often located at different positions of the physical space, and one can exploit the spatial information for speech separation by using two or more microphones. Many algorithms are proposed in the literature. Fixed and adaptive beamformers are common signal processing techniques for multi-microphone speech separation [18]. The delay-and-sum beamformer is the simplest and most widely used fixed beamformer, and it can be steered to a specific direction by adjusting the phase for each microphone and adds the signals from different microphones. One limitation of a fixed beamformer is that it needs a large array to achieve high-fidelity separation. The minimized variance distortionless response (MVDR) [5] beamformer is a representative adaptive beamformer, which minimizes the output energy while imposing linear constraints to maintain energies from the direction of the target speech. Compared with fixed beamformers, adaptive beamformers provide better performance in certain conditions, like strong and relatively few interfering sources. However, adaptive beamformers are more sensitive than fixed beamformers to microphone array errors such as sensor mismatch and mis-steering, and to correlated reflections arriving from nontarget directions [1]. The performance of both fixed and adaptive beamformers diminishes in the presence of room reverberation.

Localization-based clustering [12][23] is a popular method for unsupervised binaural separation. In general, two steps are taken. The localization step is to build the relationship between source locations and interaural parameters, such as interaural time difference (ITD) and interaural level difference (ILD), in individual time-frequency (T-F) units. The separation step is to assign each T-F unit into a different sound source by clustering or histogram picking. In [12], these two steps are jointly estimated by using an expectation-maximization algorithm.

Treating speech separation as a supervised learning problem has become popular in recent years, particularly since deep neural networks (DNNs) were introduced for supervised monaural speech separation [20]. Pertilä and Nikunen [14] used spatial feature and DNN for multichannel speech separation. Recently, Jiang *et al*. [9] extract binaural and monaural features and train a DNN for each frequency band to perform binary classification. Their results show that even a single monaural feature can improve separation performance in reverberant conditions when interference and target are very close to each other.

In this study, we address the problem of binaural speech separation in reverberant environments. The proposed system is supervised in nature, and employs DNN. Both spatial and spectral features are extracted to provide complementary information for speech separation. Motivated by recent analysis of training targets, our DNN training aims to estimate the IRM. In addition, we conduct feature extraction on full-band signals and train only one DNN to predict the IRM across all frequencies. In the following section, we present our DNN-based binaural speech separation system. The evaluation, including a description of comparison methods, is provided in Section 3. We present the experimental results and comparison in Section 4, and conclude the paper in Section 5.

2. System description

The proposed system is illustrated in Figure 1. Binaural input signals are generated by placing the target speaker in a reverberant space with many other simultaneously interfering talkers forming a spatially diffuse, speech babble. To separate the target speech from the background noise, the left-ear and right-ear signals are first fed into two modules to extract the spectral and spatial features separately. In the upper module, a beamformer is employed to preprocess the two-ear signals to produce a single signal for spectral feature extraction. In the lower module, the left-ear and right-ear signals are first decomposed into T-F units independently. Then, the spectral and spatial features are combined to form the final features. Our computational goal is to estimate the IRM. We train a DNN to map the final features to the IRM. After obtaining a ratio mask

http://dx.doi.org/10.21437/Interspeech.2017-297

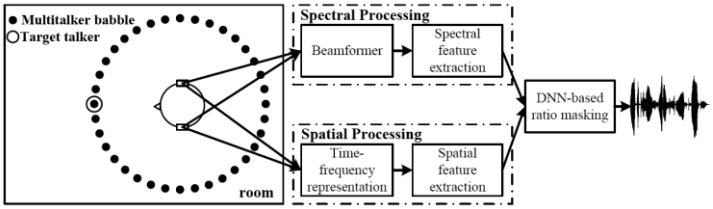


Figure 1: *Schematic diagram of the proposed binaural separation system.*

from the trained DNN, the waveform signal of the target speech is synthesized from the sound mixture and the mask [19].

2.1. Features extraction

2.1.1. Spectral features

We employ the delay-and-sum (DAS) beamformer to process the left-ear and right-ear signals into a single signal before extracting monaural spectral features. The rationale for proposing beamforming before spectral feature extraction is twofold. First, beamforming enhances the target signal, and second, it avoids an adhoc decision of having to choose one side for monaural feature extraction, as done in [9] for instance.

After beamforming, we extract amplitude modulation spectrum (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP) and mel-frequency cepstral coefficients (MFCC). In [21], these features are shown to be complementary and have been successfully used in DNN-based monaural separation. It should be mentioned that the complementary feature set originally proposed in [21] is extracted at the unit level. We extract the complementary feature set at the frame level as done in [22].

2.1.2. Spatial features

We first decompose both the left-ear and right-ear signals into cochleagrams [19]. Specifically, the input mixture is decomposed by the 64-channel gammatone filterbank with center frequencies ranging from 50 Hz to 8000 Hz on the equivalent rectangular bandwidth rate scale. The output of each channel is divided into 20-ms frame length with a 10-ms frame shift and half-wave rectified. With a 16 kHz sampling rate, the signal in a T-F unit has 320 samples.

With binaural input signals, we extract two primary binaural features of ITD and ILD. The ITD is calculated from the normalized Cross Correlation Function (CCF) between the left- and right-ear signals, denoted by subscript *l*,*r* respectively. The CCF of a T-F unit pair, indexed by time lag τ, is defined as

$$CCF(c,m,\tau) = \frac{\sum_k x_{cm,l}(k)x_{cm,r}(k-\tau)}{\sqrt{\sum_k x_{cm,l}^2(k)}\sqrt{\sum_k x_{cm,r}^2(k-\tau)}} \quad (1)$$

In the above formula, τ varies between -1 ms and 1 ms, $x_{cm,l}$ and $x_{cm,r}$ represent the left- and right-ear signals of the unit at channel c and frame m, respectively, and k indexes a signal sample of a T-F unit. For the 16 kHz sampling rate, the dimension of CCF is 33. In [9], CCF values are directly used as a feature vector to distinguish the signals coming from different locations.

Here, we propose a new 2-dimensional (2D) ITD feature. The first dimension is the CCF value at an estimated time lag $\tilde{\tau}$, corresponding to the direction of the target speech, which can be estimated by DOA method. The second dimension is the maximum value of CCF, which reflects the coherence of the left and right ear signals, and has been used for selecting binaural cues for sound localization [4]. The reasons for proposing these two features are as follows. The maximum CCF value is used to distinguish directional sources from diffuse sounds. For a directional source, the maximum CCF value should be close to 1, whereas for a diffuse sound it is close to 0. The CCF value at the estimated target direction is to differentiate the target speech and the interfering sounds that come from different directions. Specifically, we have

$$ITD(c,m) = \begin{bmatrix} CCF(c,m,\tilde{\tau}) \\ \max_{\tau} CCF(c,m,\tau) \end{bmatrix} \quad (2)$$

ILD corresponds to the energy ratio in dB, and it is calculated for each unit pair as below

$$ILD(c,m) = 10\log_{10}\frac{\sum_k x_{cm,l}^2(k)}{\sum_k x_{cm,r}^2(k)} \quad (3)$$

To sum up, the spatial features in each T-F unit pair are composed of 2D ITD and 1D ILD. We concatenate all the unit-level features at a frame to form the frame-level spatial feature vector. For 64-channel cochleagrams, the total dimension is 192 for each time frame.

2.2. Training target

We employ the ideal ratio mask (IRM) [22] as the training target,

$$IRM(c,m) = \sqrt{\frac{S^2(c,m)}{S^2(c,m)+N^2(c,m)}} \quad (4)$$

where $S^2(c,m)$ and $N^2(c,m)$ denote the speech and noise energy, respectively, in a given T-F unit. This mask is essentially the square-root of the classical Wiener filter, which is the optimal estimator in the power spectrum [10]. The IRM is obtained using a 64-channel gammatone filterbank.

The IRM has been shown to be preferable to the Ideal Binary Mask (IBM) [22]. We employ the IRM in individual frames as the training target, which provides the desired signal at the frame level for supervised training.

2.3. DNN training

A DNN is trained to estimate the IRM using the frame-level features described earlier. The DNN includes 2 hidden layers, each with 1000 units. We find that this relatively simple DNN architecture is effective for our task. The rectified linear unit (ReLU) activation function [13] is used for the hidden layers and the sigmoid activation function is used for the output layer. The cost function is mean square error (MSE). Weights of the DNN are randomly initialized. The adaptive gradient algorithm (AdaGrad) [3] is utilized for back propagation. We also employ the dropout technique [16] on hidden units to avoid overfitting. The dropout rate is 0.5. The total number of training epochs is 100. The batch size is 512. To incorporate temporal context, we use an input window that spans 9 frames (4 before and 4 after) to predict one frame of the IRM.

3. Experimental setup

3.1. Dataset

For both training and test datasets, we generate binaural mixtures by placing the target speaker in a reverberant space

with many interfering speech sources simultaneously. A reverberant signal is generated by convolving a speech signal with a binaural room impulse response (BRIR). In this study, we use two sets of BRIRs. One is simulated by software, called *BRIR Sim Set*. The other is measured in real rooms, called *BRIR Real Set*. These sets were generated or recorded at the University of Surrey.

The *BRIR Sim Set* is obtained from a room simulated using CATT-Acoustics modeling software. The simulated room is shoebox-shaped with dimensions 6m × 4m × 3m (length, width, height). The reverberation time (T60) was varied in the range [0, 1] second with 0.1s increments by changing the absorption coefficient of all six surfaces. The impulse responses are calculated with the receiver located at the center of the room at a height of 2 m and the source at a distance of 1.5 m from the receiver. The sound source was placed at the head height with azimuth between -90° and 90° spaced by 5°.

The *BRIR Real Set* is recorded in four rooms with different sizes and reflective characteristics, and their reverberation times are 0.32s, 0.47s, 0.68s and 0.89s. The responses are captured using a head and torso simulator (HATS) and a loudspeaker. The loudspeaker was placed around the HATS on an arc in the median plane with a 1.5 m radius between ±90° and measured at 5° intervals.

To generate a diffuse multitalker babble (see [11]), we use the TIMIT corpus [6] which contains 6300 sentences, with 10 sentences spoken by each of 630 speakers. Specifically, 10 sentences of each speaker in the TIMIT corpus are first concatenated. Then, we randomly choose 37 speakers, one for each source location as depicted in Fig. 1. A random slice of each speaker is cut and convolved with the BRIR corresponding to its location. Finally, we sum the convolved signals to form the diffuse babble, which is also non-stationary. The IEEE corpus [8] is employed to generate reverberant binaural target utterances, and it contains 720 utterances spoken by a female speaker. The target source is fixed at azimuth 0°, in front of the dummy head (see Fig. 1). To generate a reverberant target signal, we convolve an IEEE utterance with the BRIR at 0°. Finally, the reverberant target speech and background noise are summed to yield two binaural mixtures.

For the training and development sets, we respectively select 500 and 70 sentences from the IEEE corpus and generate binaural mixtures using *BRIR Sim Set* with 4 T60 values of 0s, 0.3s, 0.6s and 0.9s; T60 = 0s corresponds to the anechoic condition. So, the training set includes 2000 mixtures. The remaining 150 IEEE sentences are used to generate the test set. To evaluate the proposed method, we use three sets of BRIRs to build test sets called *simulated matched room*, *simulated unmatched room* and *real room*. For the simulated matched-room test set, we use the same simulated BRIRs as the ones in the training stage. For the simulated unmatched-room test set, the *BRIR Sim Set* with T60's of 0.2s, 0.4s, 0.8s and 1.0s are used. The real-room test set is generated by using *BRIR Real Set*. The SNR of the mixtures for training and test is set to -5 dB, which is the average at the two ears. It means that the SNR at a given ear may vary around -5 dB due to the randomly generated background noise and different reverberation times. In SNR calculations, the reverberant target speech, not its anechoic version, is used as the signal.

3.2. Evaluation criteria

We quantitatively evaluate the performance of speech separation by short-time objective intelligibility (STOI) [17]. This metric measures objective intelligibility by computing the correlation of short-time temporal envelopes between target and separated speech, resulting in a score in the range of [0, 1], which can be roughly interpreted as the percent-correct predicted intelligibility.

3.3. Comparison methods

We compare the performance of the proposed method with several other prominent and related methods for binaural speech separation. The first kind is beamforming and we choose DAS and MVDR beamformers for comparison. As described earlier, the DAS beamformer is employed as a preprocessor in our system. The MVDR beamformer minimizes the output energy while imposing linear constraints to maintain the energy from the direction of the target speech. Both the DAS and MVDR beamformer need the DOA (direction of arrival) estimation. Because the location of the target speaker is fixed in our evaluation, we provide the target direction to the beamformers, which facilitates the implementation.

The next one is a joint localization and segregation approach [12], dubbed as MESSL, which uses spatial clustering for source localization. Given the number of sources, MESSL iteratively modifies Gaussian mixture models (GMMs) of interaural phase difference and ILD to fit the observed data. Across frequency integration is handled by linking the GMMs models in individual frequency bands to a principal ITD.

The third comparison method employs DNN to estimate the IBM [9]. First, input binaural mixtures are decomposed into 64-channel subband signals. At each frequency channel, CCF, ILD and monaural GFCC (gammatone frequency cepstral coefficient) features are extracted and used to train a DNN for subband classification. Each DNN has two hidden layers each containing 200 sigmoidal units. Weights of DNNs are pre-trained with restricted Boltzmann machines. The subband binaural classification algorithm is referred as SBC in the following.

4. Evaluation and comparison

4.1. Simulated-room conditions

In this test condition, we evaluate the performance of the proposed algorithm in the simulated rooms, which are divided into two parts: matched and unmatched conditions. As mentioned earlier, for matched-room conditions, test reverberated mixtures are generated by using the same BRIRs as in the training stage, where the T60s are 0.3s, 0.6s and 0.9s. For the unmatched-room conditions, the BRIRs for generating reverberated mixtures are still simulated ones, but the T60s are different from those in training conditions and take the values of 0.2s, 0.4s, 0.8s and 1.0s. The results of STOI are shown in Table I. Since left-ear and right-ear signals are very similar, we just list the STOI scores of the unprocessed mixtures at the left ear, referred to "MIX_L".

Table 1: *Average STOI scores (%) of different methods in simulated matched-room and unmatched-room conditions*

	T_{60}	**MIX_L**	**DAS**	**MVDR**	**MESSL**	**SBC**	**Pro.**
Matched-room	0.0s	58.00	63.56	63.75	65.92	63.65	74.66
	0.3s	53.13	58.61	58.78	58.66	62.79	74.88
	0.6s	44.08	50.82	50.84	51.89	55.08	68.53
	0.9s	44.58	48.20	48.15	48.46	53.37	65.39
	Avg.	49.05	55.30	55.38	56.23	58.72	70.87
Unmatched-room	0.2s	55.28	61.80	61.91	60.52	64.68	74.95
	0.4s	47.98	54.46	54.64	55.91	59.02	70.40
	0.8s	39.99	47.06	47.01	47.12	54.27	65.92
	1.0s	39.05	45.21	45.01	46.05	51.64	62.82
	Avg.	45.58	52.13	52.14	52.40	57.40	68.52

Compared the unprocessed mixtures, the proposed system obtains the absolute STOI gain about 22% on average (i.e. from 49% to 71%) in the simulated matched-room conditions and 23% in the simulated unmatched-room conditions. From Table 1, we can see that the proposed system outperforms the other comparison methods in anechoic and all reverberation conditions. DAS and MVDR have similar results, because the background noise is quite diffuse; it can be proven that MVDR and DAS become identical when noise is truly diffuse. For the supervised learning algorithms, both SBC and the proposed algorithm exhibit good generalization in the unmatched-room conditions.

4.2. Real-room conditions

In this test condition, we use the *BRIR Real Set* to evaluate the proposed separation system and compare it with other methods. The STOI results are given in Table 2. The proposed system achieves the best results in all four room conditions. Compared with unprocessed mixtures, the average STOI gain is about 20% (i.e. from 43% to 63%), which is consistent with that in simulated room conditions.

Table 2: *Average STOI scores (%) of different methods in real room conditions.*

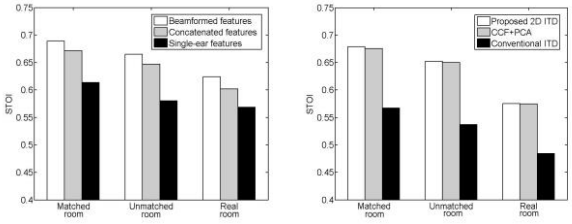
Room	MIX_L	DAS	MVDR	MESSL	SBC	Pro.
A (0.32s)	47.49	53.71	53.84	54.39	53.37	66.70
B (0.47s)	41.29	48.10	48.08	48.61	42.95	61.96
C (0.68s)	44.33	51.31	50.86	52.11	54.13	64.78
D (0.89s)	39.61	45.48	45.58	45.35	48.52	60.57
Avg.	43.18	49.65	49.59	50.12	49.74	63.50

From the above experimental results, we can see that the proposed algorithm outperforms SBC which is also a DNN-based separation algorithm. One of the differences is that the proposed algorithm employs ratio masking for separation, while SBC utilizes binary masking. A simple way to turn a binary mask to a ratio mask in the context of DNN is to directly use the outputs of the subband DNNs, which can be interpreted as posterior probabilities with values ranging from 0 to 1. With such soft masks, SBC's average STOI scores are 63.25% for matched-room conditions, 61.96% for unmatched-room conditions and 55.80% for real-room conditions. These results represent significant improvements over binary masks, but they are still not as high as those of the proposed algorithm.

4.3. Feature analysis

Our binaural speech separation system uses both spectral and spatial features. For spectral features, the DAS beamformer is employed as a preprocessor. The spatial features are formed by combining the proposed 2D ITD and ILD. To evaluate the effectiveness of the features, we further compare several alternatives with the same DNN configuration and training procedure (see Section 2.3).

One simple way to combine spectral and spatial analyses is to directly concatenate the left- and right-ear complementary features. We compare this feature vector with the proposed beamformed features and also single-ear monaural features (left-ear as in [9]). The interaural features are excluded here. Average STOI results are shown in Fig. 2(a). From the figure, we can see that extracting the spectral features on the output signal of the beamformer is better than concatenating the spectral features of the left- and right-ear signals. For spatial features, we compare the conventional ITD [15], CCF and the proposed 2D ITD. Since concatenating unit-level CCF vectors directly leads to a very high dimension, we perform principal component analysis (PCA) to reduce the dimension to 128, equal to the size of 2D ITD frame-level feature. The STOI results are shown in Fig. 2(b). We can see that the results with the conventional ITD are much worse than CCF plus PCA and the proposed 2D ITD. While proposed 2D ITD yields essentially the same results as CCF, it has an advantage of relative invariance to different target directions in addition to computational efficiency.



(a) (b)

Figure 2: *Comparison of the different features. (a) Spectral features. (b) Spatial features.*

5. Conclusions

In this work, we have proposed a DNN-based binaural speech separation algorithm which combines spectral and spatial features. DNN-based separation has shown its ability to improve speech intelligibility [7] even with just monaural spectral features. As demonstrated in previous work [9], binaural speech separation by incorporating monaural features represents a promising direction to further elevate separation performance.

For supervised speech separation, input features and training targets are both important. In this study, we make a novel use of beamforming to combine left-ear and right-ear monaural signals before extracting spectral features. In addition, we have proposed a new 2D ITD feature. With the IRM as the training target, the proposed system outperforms representative binaural separation algorithms in non-stationary background noise and reverberant environments, including a DNN-based subband classification algorithm [9]. Another issue is generalization to untrained environments. Our algorithm shows consistent results in unseen reverberant noisy conditions. This strong generalization ability is partly due to the use of effective features. Although only one noisy situation is considered, the noise problem can be addressed through large-scale training [2].

In the present study, the target speaker is fixed to the front direction and sound localization is not addressed. For the proposed algorithm, two parts need the target direction. One is DAS beamforming and the other is calculation of 2D ITD. Sound localization is a well-studied problem [19]. Recently, DNN is also used for sound localization [11], although only spatial features are considered. We believe that incorporating monaural separation is a good direction to improve the robustness of sound localization in adverse environments. One way to incorporate monaural separation is to employ spectral features for initial separation, from which reliable T-F units are selected for sound localization. Moreover, separation and localization could be done iteratively, analogous to [24].

6. Acknowledgements

This work was performed while the first author was a visiting scholar at the Ohio State University. This research was supported in part by a NSFC grant (No. 61365006), an NSF grant (IIS-1409431) and an AFOSR grant (No. FA9550-12-1-0130).

7. References

[1] Brandstein, M., and Ward, D., Eds. *Microphone arrays*: *signal processing techniques and applications*. New York, NY: Springer, 2001.

[2] Chen, J., Wang, Y., Yoho, S.E., Wang, D.L., and Healy, E.W. "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Amer.*, vol. 139, pp. 2604-2612, 2016.

[3] Duchi, J., Hazan, E., and Singer, Y. "Adaptive subgradient methods for online learning and stochastic optimization," *The Journal of Machine Learning Research*, vol. 12, pp. 2121-2159, 2011.

[4] Faller, C., and Merimaa, J. "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, pp. 3075-3089, 2004.

[5] Frost III, O.L. "An algorithm for linearly constrained adaptive array processing," in *Proceedings of the IEEE*, vol.60, pp.926-935, 1972.

[6] Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., and Dahlgren, N.L. "DARPA TIMIT acoustic phonetic continuous speech corpus," 1993, http://www.ldc.upenn.edu/Catalog/LDC93S1.html.

[7] Healy, E.W., Yoho, S.E., Wang, Y., and Wang, D.L. "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Amer.*, vol. 134, pp. 3029-3038, 2013.

[8] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.

[9] Jiang, Y., Wang, D.L., Liu, R.S., *et al.* "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 22, pp. 2112-2121, 2014.

[10] Loizou, P. C., *Speech Enhancement: Theory and Practice*. BocaRaton, FL: CRC, 2007.

[11] Ma, N., Brown, G., and May, T. "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," in *Proc. Interspeech 2015*, pp. 3302-3306.

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

[13] Nair, V., and Hinton, G. "Rectified linear units improve restricted boltzmann machines," In *Proc. ICML 2010* pp. 807-814.

[14] Pertilä, P., and Nikunen, J. "Distant Speech Separation Using Predicted Time–frequency Masks from Spatial Features," *Speech Comm.*, vol. 68, pp. 97–106, 2015.

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

[16] Srivastava, N., Hinton, G., Krizhevsky, A., *et al.* "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, pp. 1929-1958, 2014.

[17] Taal, C.H., Hendriks, R.C., Heusdens, R., and Jensen, J. "An algorithm for intelligibility prediction of time frequency weighted noisy speech," *IEEE Trans. on Audio, Speech, Lang. Process.*, vol. 19, pp. 2125–2136, 2011.

[18] Van Veen, B.D., and Buckley, K.M. "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol.5, pp. 4-24, 1988.

[19] Wang, D.L., and Brown, G.J., Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hobboken, NJ: Wiley-IEEE Press, 2006.

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

[21] Wang, Y., Han, K., and Wang, D.L. "Exploring monaural features for classification-based speech segregation," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 21, pp. 270–279, 2013.

[22] Wang, Y., Narayanan, A., and Wang, D.L. "On training targets for supervised speech separation," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 22, pp. 1849-1858, 2014.

[23] Yilmaz, O., and Rickard, S. "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Process.*, vol. 52, pp. 1830-1847, 2004.

[24] Zhang, X., Zhang, H., Nie, S., *et al.* "A Pairwise Algorithm Using the Deep Stacking Network for Speech Separation and Pitch Estimation," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 24, pp. 1066-1078, 2016.

Index Termsâ€” Deep neural network, binaural blind speech separation, spectral and spatial, iterative DNN. 1. INTRODUCTION.Â [3] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, â€œAn experimen-tal study on speech enhancement based on deep neural networks,â€ IEEE Signal Processing Letters, vol. 21, no. 1, pp. 65â€“68, January 2014. [4] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Roux, J. R. Hershey, and B. Schuller, â€œSpeech en-hancement with LSTM recurrent neural networks and its application to noise-robust ASR,â€ in International Conference on Latent Variable Analysis and Signal Sep-aration, 2015, pp. 91â€“99. Speech signal degradation in real environments mainly results from room reverberation and concurrent noise. While human listening is robust in complex auditory scenes, current speech segregation algorithms do not perform well in noisy and reverberant environments. We treat the binaural segregation problem as binary classification, and employ deep neural networks (DNNs) for the classification task. The binaural features of the interaural time difference and interaural level difference are used as the main auditory features for classification. The monaural feature of gammatone frequency cepstral coefficients is also used to improve classification performance, especially when interference and target speech are collocated or very close to one another. Experiments based on real binaural RIRs and TIMIT dataset are provided to show the performance of the proposed system for reverberant speech mixtures, as compared with a model-based T-F masking technique proposed recently. Introduction.Â The success of deep neural networks (DNNs) in these applications inspires us to investigate its potential for improving the performance of stereo speech source separation algorithms.Â Compared with the monaural segregation of reverberant speech in [29], the stereo speech separation in [31] tends to be more robust due to the use of spatial information. In [7], GMM is used to model the MV and IPD/ILD cues that contain spatial information and the EM algorithm is used to estimate the model parameters and to derive the T-F mask.