A REGRESSION APPROACH TO BINAURAL SPEECH SEGREGATION VIA DEEP NEURAL NETWORK

Nana Fan, Jun Du, Li-Rong Dai

National Engineering Laboratory for Speech and Lauguage Information Processing
University of Science and Technology of China, Hefei, P.R.China
nnf8185@mail.ustc.edu.cn, {jundu,lrdai}@ustc.edu.cn

Abstract

This paper proposes a novel regression approach to binaural speech segregation based on deep neural network (DNN). In contrast to the conventional ideal binary mask (IBM) method using DNN with the interaural time difference (ITD) and interaural level difference (ILD) as the auditory features, the log-power spectra (LPS) features of target speech are directly predicted via a regression DNN model by concatenating the monaural LPS features and the binaural features as the input. As for the binaural features, the sub-band ILDs based on LPS features are designed which are verified to be more effective than the full-band ILDs. Our experiments show that our proposed approach can significantly outperform IBM-based speech segregation in terms of both objective measures of speech quality and speech intelligibility for noisy and reverberant environments.

Index Terms: binaural speech segregation, interaural level difference, sub-band binaural features, deep neural network

1. Introduction

The acoustic signals coming from the real world are often corrupted by environmental noises, reflections, interferences of other speakers, leading to the degradation of the speech quality and intelligibility. Techniques to segregate clean speech from mixed noisy signals are essential in many speech-enabled applications. The conventional speech segregation algorithms have been developed for several decades such as spectral subtraction [1] and Wiener filter method [2]. In multi-channel conditions using the microphone array, the beamforming approaches [3] and multi-channel Wiener filter [4, 5] have been widely explored. However, various assumptions such as the stationarity of noises and the space independence between target speech and interference are made in most of those methods , which can not be satisfied in realistic acoustic conditions. Thus the separated speeches are often affected by musical noise artifacts [6].

Inspired by the human auditory processing system, more and more researchers focus on the auditory scene analysis (ASA). In the study of computational theory of ASA, Wang [7] has proposed that the goal of CASA based speech segregation is to estimate an ideal time-frequency (T-F) mask. The ideal binary mask (IBM) estimation problem has motivated a plenty of investigations based on classification methods [8]. In [9] it shows that Gaussian-kernel support vector machines (SVMs) achieve better performances than Gaussian mixture models (GMMs). However, to further improve the performance, the constructed model must be able to learn complex relations of statistics and be more generalizable and adaptable.

In recent years, deep learning based approaches are adopted for speech signal processing due to its powerful ability to construct statistical models using the large data set. Monaural segregation is formulated as a binary classification problem in [10, 11, 12], which are based on the T-F mask estimation using deep neural networks (DNNs). IBM method has also driven the interests on feature study to improve the signal to noise ratio (SNR) of the segregated speech. Monaural features extracted from T-F unit level have been comprehensively studied in [13], such as gammatone frequency cepstral coefficients (GFCCs), mel-frequency cepstral coefficients (MFCCs). In binaural separation cases, the binaural features such as interaural time difference (ITD) and interaural level difference (ILD) are employed as binaural cues for better T-F units classification [14, 15, 16].

In this study, we propose a DNN-based regression approach to binaural speech segregation. Although the IBM classification approach can improve the SNR, the speech quality after segregation is not quite good. The segregated speech suffers from the discontinuity in time and frequency bands, which is caused by the direct setting of T-F units to 0/1 in IBM. While a modified ideal ratio mask might improve the situation [17], here we consider a regression based method [18] where the clean speech features are directly predicted from the noisy input using the non-linear mapping function by DNN. Since the regression approach is based on the feature generation and can fully utilize the neighbouring frames and full frequency band information, the speech continuity can be better preserved. The log-power spectra (LPS) features are used as monaural features for both input and output layers. Furthermore, the binaural sub-band ILD features are designed based on the LPS features from the two channels, which are also adopted as input features and have been verified to be effective in segregation.

The rest of this paper is organized as follows. In Section 2, the IBM classification based approach is briefly reviewed. Section 3 introduces our proposed regression approach including the monaural/binaural features and DNN architecture. Section 4 presents the experimental results. Finally we draw the conclusion in Section 5.

2. Review of the IBM approach

A classification based approach has been proposed for binaural speech segregation [15] where DNN is employed to estimate the IBM of the input noisy speech. The overall architecture is shown in Figure 1. First, the input speech is processed by gammatone filters with a set of frequency bands, which is followed by the framing with 20ms frame length and 10ms frame shift. Then to utilize the information in binaural channels, both the monaural GFCC features and the binaural features including cross-correlation function (CCF) , ILD, and ITD are extracted from the T-F representation. A classification DNN with two hid-

978-1-5090-4294-4/16/$31.00 ©2016 IEEE

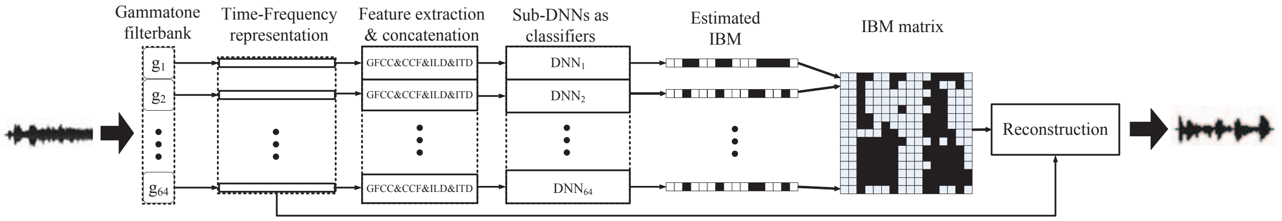

Figure 1: *The overall procedure of IBM based binaural speech segregation system.*

den layers is built to model the relationship of input features and the corresponding IBM labels. A pre-training stage [19, 20] is performed to obtain an initial set of DNN parameters, which are fed to the supervised fine-tuning stage [21]. To reconstruct the separated speech, the estimated binary mask is directly applied to the noisy speech T-F representation. And the target can be restored by adding up the values from each frequency channel. Note that the phase correction and windowing are conducted at the reconstruction stage [22].

3. The proposed DNN regression approach

The system flowchart of our proposed approach is illustrated in Figure 2. In the training stage, noisy LPS monaural features of the left channel and ILD binaural features are extracted from the training utterances and concatenated with the frame expansion as the input of the regression DNN. And the clean LPS features of the left channel are adopted as the learning targets of DNN. In the testing stage, the estimated clean LPS features from the output of the well-trained DNN are used to reconstruct the segregated speech with the noisy phases [18]. In the following subsections, the feature design and DNN architecture are elaborated.

3.1. The monaural and binaural feature design

To make a full use of the information from binaural channels in the proposed DNN regression model, we employ LPS features of the left channel as the monaural features and the ILD between the two channels as the binaural features, which are concatenated as the input of DNN. LPS features can maintain the full information (excluding the phase information) of the original speech signals and provide the perceptually relevant parameters [23, 24]. The binaural ILD features can be calculated in a similar way as the IBM approach reviewed in section 2. The basic idea is to utilize the energy differences between two channels in one frame. In this study we investigate several types of ILDs based on LPS features:

- Global ILD: one-dimensional ILD defined as the ratio between the sum of full-band LPS features from two channels.
- Full-band ILD: an ILD vector with the same dimension as the LPS features, where each element is defined on each linear frequency bin.
- Sub-band ILD: an ILD vector with each element defined on the sub-band of LPS features.

The first two ILDs can be formulated as:

$$ILD_{\text{global}} = \frac{\sum_{d=1}^{D} LPS_{\text{left}}^{d}}{\sum_{d=1}^{D} LPS_{\text{right}}^{d}} \qquad (1)$$

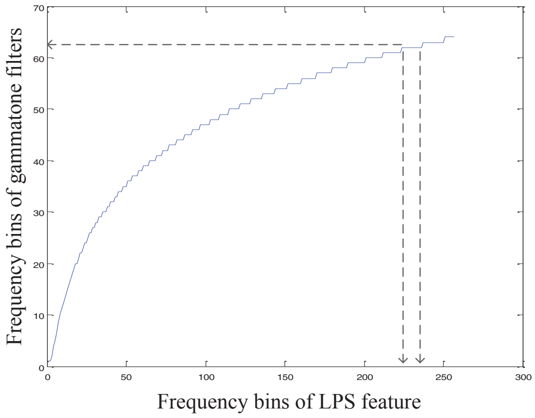

Figure 3: *Illustration of mapping linear frequency bins of LPS features into the sub-bands of gammatone filters.*

$$ILD_{\text{full}}^{d} = \frac{LPS_{\text{left}}^{d}}{LPS_{\text{right}}^{d}}, d = 1, 2, \ldots, D \qquad (2)$$

where LPS_{left}^{d} and LPS_{right}^{d} are the d^{th} elements of LPS feature vectors for the left and right channel, respectively. D is the dimension of LPS feature vector. Obviously, the information provided by the global ILD is limited. Meanwhile, the high-dimensional full-band ILD features do not necessarily improve the performance when concatenated with the high-dimensional monaural features. Accordingly, the sub-band ILD features are designed as a tradeoff. To determine the number of sub-bands, the D ($D = 257$) linear frequency bins for LPS features are mapped to D_{sub} frequency sub-bands for 64-gammatone filter banks, which is illustrated in Figure 3. As there are no linear frequency bins corresponding to a few low frequency gammetone bands, we finally get 61 sub-bands ($D_{\text{sub}} = 61$). Then in each sub-band, the ILD can be calculated as

$$ILD_{\text{sub}}^{i} = \frac{\sum_{d_i \le d < d_{i+1}} LPS_{\text{left}}^{d}}{\sum_{d_i \le d < d_{i+1}} LPS_{\text{right}}^{d}}, i = 1, 2, \ldots, D_{\text{sub}} \qquad (3)$$

where d_i is the starting index of the i^{th} sub-band. Please note that this design of sub-band ILDs is just one type of implementation using gammatone frequency bands to simulate the frequency selectivity of human ears described in CASA [22].

3.2. The DNN architecture

The DNN architecture adopted here, as shown in Figure 4, is similar to that in [21] as a nonlinear regression model and has shown a powerful modeling capability [18]. The multi-layer structure, including input layer, output layer and several hidden

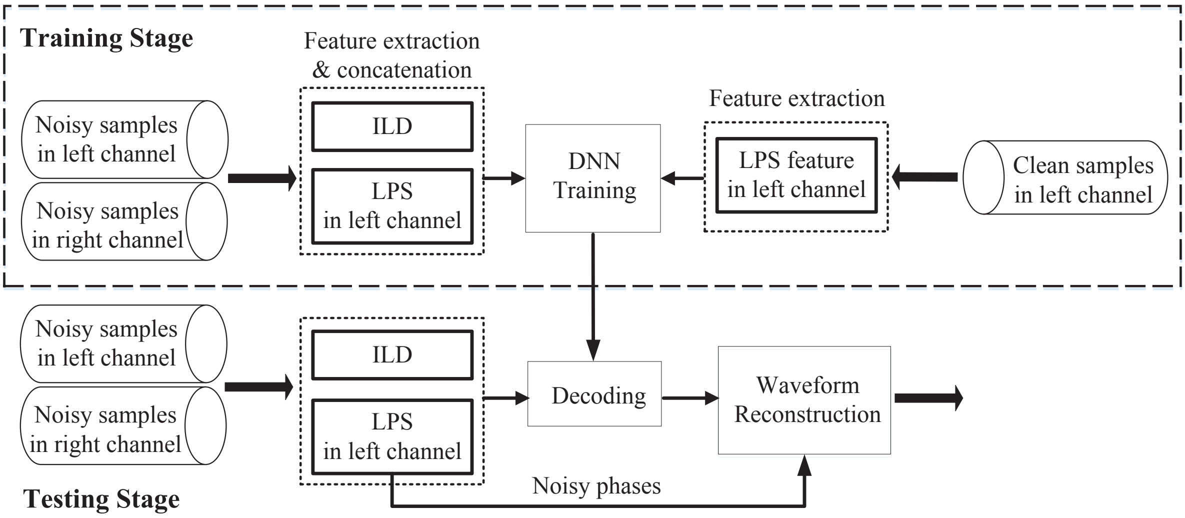
Training Stage
Feature extraction & concatenation
Noisy samples in left channel
Noisy samples in right channel
ILD
LPS in left channel
DNN Training
Feature extraction
LPS feature in left channel
Clean samples in left channel
Noisy samples in left channel
Noisy samples in right channel
ILD
LPS in left channel
Decoding
Waveform Reconstruction
Noisy phases
Testing Stage

Figure 2: *A block diagram of proposed regression DNN based binaural speech segregation system.*

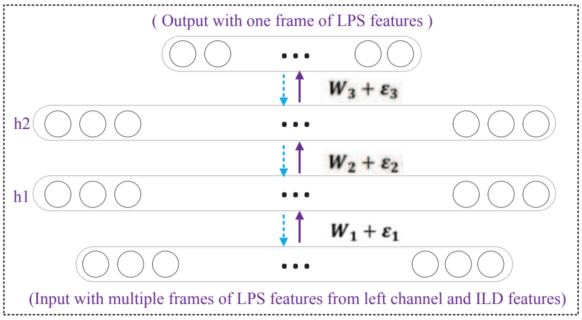
(Output with one frame of LPS features)
$W_3 + \varepsilon_3$
h2
$W_2 + \varepsilon_2$
h1
$W_1 + \varepsilon_1$
(Input with multiple frames of LPS features from left channel and ILD features)

Figure 4: *Illustration of regression DNN architecture*

layers, performs as a mapping function from input noisy LPS features and binaural features to clean target features. The utilization of acoustic contexts (multiple neighbouring frames) in the input layer can further improve the continuity of the recovered speech. Sigmoidal activation functions are used. For the DNN training, a random initialization is first conducted, which is followed by the fine-tuning stage using back-propagation algorithm [25] with the minimum mean squared error (MMSE) objective function between the target and estimated LPS features:

$$E = \frac{1}{N}\sum_{n=1}^{N} \parallel \hat{\boldsymbol{X}}_n(\boldsymbol{Y}_{n-\tau}^{n+\tau}, \boldsymbol{W}, \boldsymbol{b}) - \boldsymbol{X}_n \parallel_2^2, \qquad (4)$$

where E is the mean squared error, $\hat{\boldsymbol{X}}_n(\boldsymbol{Y}_{n-\tau}^{n+\tau}, \boldsymbol{W}, \boldsymbol{b})$ and $\boldsymbol{X}_n$ represent the estimated and target LPS feature vectors at the n^{th} frame, respectively. N is the mini-batch size. $\boldsymbol{Y}_{n-\tau}^{n+\tau}$ is the input monaural and binaural feature vector with neighbouring $2 * \tau + 1$ frames. $(\boldsymbol{W}, \boldsymbol{b})$ denote the weight and bias parameters to be learned. For the l^{th} layer, the $(\boldsymbol{W}^l, \boldsymbol{b}^l)$ parameters can be updated as follows,

$$\Delta(\boldsymbol{W}_{\text{new}}^l, \boldsymbol{b}_{\text{new}}^l) = -\lambda \frac{\partial E}{\partial(\boldsymbol{W}^l, \boldsymbol{b}^l)} + \omega \Delta(\boldsymbol{W}^l, \boldsymbol{b}^l) \quad (5)$$

$$(\boldsymbol{W}_{\text{new}}^l, \boldsymbol{b}_{\text{new}}^l) = (\boldsymbol{W}^l, \boldsymbol{b}^l) + \Delta(\boldsymbol{W}_{\text{new}}^l, \boldsymbol{b}_{\text{new}}^l) \qquad (6)$$

where ω is the momentum and λ is the learning rate.

4. Experiments

To compare with the IBM-based classification method for the binaural speech segregation [15], we employed the similar settings for training and testing. TIMIT database [26] was used to create the simulated noisy and reverberant speech. 500 utterances were used for training while 50 utterances were for testing. The simulated binaural data was created by using the Matlab toolkit “Roomsim” [27] with the head related transformation function (HTRF). All noisy data was obtained in anechoic or reverberant conditions by corrupting the clean speech with the babble noise at the azimuth of 45° and 0dB SNR while the clean speech source was placed in the front of the receivers. The speech waveforms are all 16kHz sampled. As for the short-time Fourier analysis, the frame length is 20ms and the frame shift is 10ms. Then the dimension of the LPS feature vector is 257.

To distinguish different DNN systems in the subsequent experiments, we define the corresponding notations as follows.

- **IBM-DNN**: the IBM-based classification approach as the baseline [15].
- **R-DNN**: the regression DNN approach using only monaural LPS features.
- **R-DNN-Global**: the regression DNN approach using monaural LPS features and binaural global ILD features.
- **R-DNN-Full**: the regression DNN approach using monaural LPS features and binaural full-band ILD features.
- **R-DNN-Sub**: the regression DNN approach using monaural LPS features and binaural sub-band ILD features.

In each R-DNN system, two hidden layers with 2048 nodes in each were adopted and the input layer size was determined by the dimensions of both monaural and binaural features. Each model was trained by stochastic gradient descent (SGD) algorithm with a mini-batch size of 128 and totally 50 epochs were

Table 1: Average PESQs of different systems on the test set.

System	Anechoic	T60=0.3s
Noisy	1.45	1.75
IBM-DNN	1.68	1.61
R-DNN	1.56	2.16
R-DNN-Global	1.80	2.16
R-DNN-Full	2.36	2.21
R-DNN-Sub	**2.46**	**2.27**

Table 2: Average STOIs of different systems on the test set.

System	Anechoic	T60=0.3s
Noisy	0.6274	0.5487
IBM-DNN	0.8283	0.6843
R-DNN	0.5581	0.6338
R-DNN-Global	0.6666	0.6463
R-DNN-Full	0.8328	0.6707
R-DNN-Sub	**0.8333**	**0.6868**

set. The momentum was set as 0.5 and the learning rate was 0.1.

4.1. Objective Evaluation

Perceptual evaluation of speech quality (PESQ) was used to measure the segregated speech quality due to its correlation with subjective score [28]. Tabel 1 lists the average PESQ results of different systems in anechoic and reverberant conditions (T60=0.3s). The frame expansion was not used. For the reverberant condition, the PESQ of IBM-DNN system was even worse than the unprocessed noisy speech. Clearly, most of the regression DNN based systems outperformed the IBM-DNN system. And the ILD features concatenated with LPS features could significantly improve the performance due to the utilization of the binaural information. By the comparison of three binaural R-DNN systems, R-DNN-Sub achieved the best results which implied that the sub-band ILDs made a good tradeoff between the information insufficiency (R-DNN-Global) and high-dimensional problem (R-DNN-Full). Overall, the PESQ gains of 0.78 and 0.66 were yielded from IBM-DNN to R-DNN-Sub for the anechoic and reverberant conditions, respectively.

Short-time objective intelligibility (STOI) [29], which is highly relevant to the human speech intelligibility, is measured in Table 2. For the regression DNN based systems, similar observations could be made as in Tabel 1. The R-DNN-Sub achieved the best STOI results for all conditions. However, the performance gaps between R-DNN-Sub and IBM-DNN were not significant, especially for the reverberant condition. This is reasonable as the IBM concept is inherently proposed to improve the speech intelligibility rather than the speech quality.

Different settings of the frame expansion for the input layer of DNN are compared in Figure 5. With more neighbouring frames, better PESQ results could be obtained for all DNN systems. There was only one exception that the performance R-DNN-Full was not improved from the 5-frame to 7-frame setting, which might be due to the high-dimensional problem of full-band ILDs.

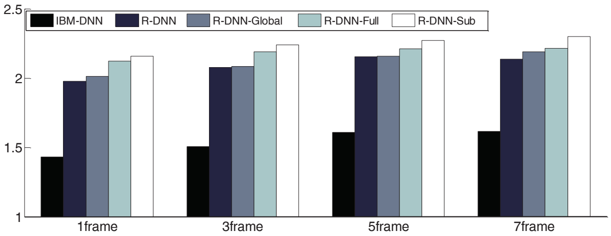

Figure 5: *Average PESQ results on test set using different input acoustic contexts. Test SNR = 0dB. T60 = 0.3s*

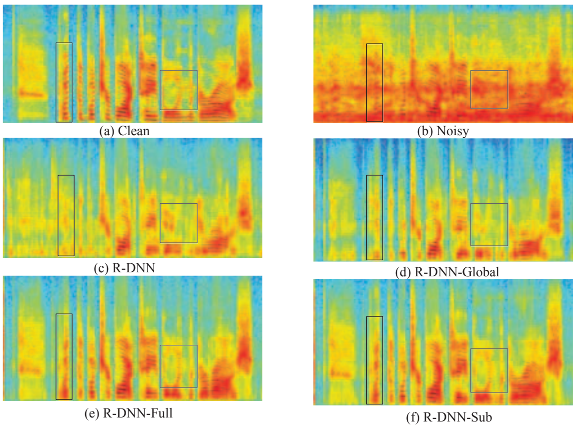

Figure 6: *Spectrograms of an anechoic 0dB utterance example segregated from different systems: (a) clean (upper left), (b) noisy (upper right), (c) R-DNN (middle left), (d) R-DNN-Global (middle right), (e) R-DNN-Full (bottom left), (f) R-DNN-Sub (bottom right).*

Figure 6 gives the spectrograms of an example utterance. The noisy spectrogram (upper right) was generated by corrupting the clean speech (upper left) with the babble noise at 0dB SNR. The areas of the clean spectrogram (upper left) in the black rectangles were severely degraded which could be well recovered by the R-DNN and R-DNN-Global systems. However, the fine structures could be better reconstructed in the R-DNN-Full and R-DNN-Sub systems which used more binaural information cues.

5. Conclusions

In this paper, we propose several DNN-based regression systems to segregate the binaural speech using monaural and binaural features. Compared with the IBM-based classification method, the direct mapping using the regression DNN can yield a better reconstructed speech in terms of speech quality and intelligibility. In addition, the sub-band ILD features deduced from LPS features can better utilize the binaural information and significantly improve the performance.

6. Acknowledgments

This work was partially funded by the National Nature Science Foundation of China under Grant No. 61305002, National Key Technology Support Program under Grants No. 2014BAK15B05, the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDB02070006, and MOE-Microsoft Key Laboratory of USTC.

7. References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[3] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2013.

[4] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone arrays*. Springer, 2001, pp. 39–60.

[5] A. Spriet, M. Moonen, and J. Wouters, "The impact of speech detection errors on the noise reduction performance of multi-channel wiener filtering and generalized sidelobe cancellation," *Signal Processing*, vol. 85, no. 6, pp. 1073–1088, 2005.

[6] A. Hussain, M. Chetouani, S. Squartini, A. Bastari, and F. Piazza, "Nonlinear speech enhancement: An overview," *Lecture Notes in Computer Science*, vol. 4391, pp. 217–248, 2007.

[7] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.

[8] M. L. Seltzer, B. Raj, and R. M. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

[9] K. Han and D. Wang, "A classification based approach to speech segregation," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, 2012.

[10] Y. Wang and D. Wang, "Cocktail party processing via structured prediction," in *Advances in Neural Information Processing Systems*, 2012, pp. 224–232.

[11] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 7, pp. 1381–1390, 2013.

[12] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1562–1566.

[13] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 2, pp. 270–279, 2013.

[14] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[15] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 12, pp. 2112–2121, 2014.

[16] 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," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 2016–2030, 2012.

[17] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7092–7096.

[18] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 1, pp. 7–19, 2015.

[19] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[20] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[21] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *Signal Processing Letters, IEEE*, vol. 21, no. 1, pp. 65–68, 2014.

[22] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[23] F. Xie and D. C. Van, "A family of mlp based nonlinear spectral estimators for noise reduction," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. 2. IEEE, 1994, pp. II–53.

[24] E. A. Wan and A. T. Nelson, "Networks for speech enhancement," *Handbook of neural networks for speech processing. Artech House, Boston, USA*, vol. 139, 1999.

[25] D. Rummelhart, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 9, pp. 533–536, 1986.

[26] 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," *Linguistic data consortium, Philadelphia*, vol. 33, 1993.

[27] D. Campbell, K. Palomaki, and G. Brown, "A matlab simulation of "shoebox" room acoustics for use in research and teaching," *Computing and Information Systems*, vol. 9, no. 3, p. 48, 2005.

[28] P. Recommendation, "862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Feb*, vol. 14, pp. 14–0, 2001.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2125–2136, 2011.