A Progressive Deep Learning Approach to Child Speech Separation

Xin Wang[1], *Jun Du*[1], *Lei Sun*[1], *Qing Wang*[1], *Chin-Hui Lee*[2]

[1]University of Science and Technology of China, Hefei, Anhui, China
[2]Georgia Institute of Technology, Atlanta, Georgia, USA

wx0304@mail.ustc.edu.cn, jundu@ustc.edu.cn, sunlei17@mail.ustc.edu.cn,
xiaosong@mail.ustc.edu.cn, chl@ece.gatech.edu

Abstract

We propose a progressive learning approach to separating child speech from signals with mixed adult speech in a speaker-independent manner based on a densely connected long short-term memory (LSTM) architecture to deal with limited training data issue in child speech. First, by measuring the speech dissimilarities between children and adults using i-vectors, we demonstrate that distances between child and adult speech are large enough to warrant a possible separation through establishing child and adult speech groups. Accordingly, we present a novel LSTM design with densely connected hidden layers and stacked inputs containing progressively obtained intermediate targets that are learnt via multiple-target learning for speech separation between child and adult groups. Experimental results on a simulation corpus show that the proposed framework can yield consistent and significant gains of objective measures over the LSTM baseline for child speech separation. Furthermore, our preliminary results on the SeedLing corpus with realistic recordings for child language acquisition show that our approach can achieve better overall separation performances than LSTM baseline when comparing spectrograms of separated speech, implying a potential for speaker diarization involving child speech.

Index Terms: speaker dissimilarity measure, long short-term memory, densely connected nextworks, progressive learning, multi-task learning, source separation, child speech separation

1. Introduction

Speech separation [1] aims at segregating mixed speech into voices of individual speakers. Child speech separation, referring to separating child speech from the utterances mixed up with adult speech, potentially has a variety of important applications, such as child speech recognition [2] and child-involved speaker diarization [3]. Acoustic and linguistic characteristics of child speech are quite different from those of adults. For instance, child speech is characterized by higher pitch and formant frequencies with respect to adult speech [4]. Several techniques, e.g., Gaussian mixture models (GMMs) [5] and i-vectors [6], have been proposed to classify or identify speech of children and adults in the past [7], [8]. However, it is difficult for these techniques to deal with overlapped speech that children and adults speak at the same time.

Many techniques focusing on speech separation between adults have been proposed in the past. For example, Roweis [9] employs factorial hidden Markov models (FHMMs) to learn the information of a speaker and then separates the speech mixture through computing a mask function. Another popular approach is non-negative matrix factorization (NMF) [10] which decomposes the signal into sets of bases and weight matrices. A non-negative back-propagation algorithm was proposed in [11] to build a deep network with non-negative parameters. However, these aforementioned supervised methods are not always applicable to practical scenarios due to a lack of prior knowledge of speakers. Therefore in an unsupervised methods, computational auditory scene analysis (CASA) [12] , inspired by the ability of human auditory perception to recover signals of interest from background distractions, is widely adopted without assuming any knowledge about mixing speakers. Unsupervised clustering for sequential grouping is adopted to convert simultaneous streams to two clusters in [13] by maximizing the ratio of between-cluster and within-cluster distances.

Recently, deep neural networks (DNNs) [14], [15] have been utilized in many speech processing areas, such as speech enhancement [16], [17], and speech dereverberation [18], [19], which proves a new direction for speech separation. In [20], long short-term memory recurrent neural network (LSTM-RNN) was used. In [21, 22], ideal ratio masks (IRMs) were used to make binary classification on time-frequency (T-F) units. In [23], Gao et al. proposed DNN-based progressive learning (PL) which aimed at decomposing complicated regression into a series of subproblems. To increase the modeling capability, in [24], Sun et al. adopted LSTM-RNN with multiple-target learning [25] of both log-power spectra (LPS) and IRM to capture the long-term contextual information. DNN-based semi-supervised speech separation had also been proposed in [26]. DNN-based unsupervised speech separation had also been studied in [27]. Both studies dealt only with adult speech separation. However research in child speech separation was still quite limited in the literature.

In this paper, we present a novel LSTM design with densely connected hidden layers and stacked inputs containing progressively obtained intermediate targets that are learnt via multiple-target learning for speech separation between child and adult groups in a speaker-independent manner, extending from [27] in which separation of adult speaker groups was considered. By measuring the dissimilarities, we found that segregating child speech mixed with adult speech should be the easiest among separation tasks because i-vector based distance between child and adult speech is the largest among all mixed speaker groups to guarantee a possible separation. However, it is not easy to collect a large set of good-quality training utterances from young children aged between 2 and 5 because they don't follow specific recording instructions as well as adults. Based on such restrictions, we propose a progressive learning framework to generate intermediate target outputs and stack them together with the original limited-sized mixed input feature vectors to increase the amount of effective training samples. Multiple-target learning is also employed to improve modeling effectiveness for speech separation [25]. Experimental results on simulation data using adult speech from WSJ0 [28] and child speech from PhonBank [29] demonstrate that the proposed approach can yield

Table 1: *The average distances across all speaker pairs for different combinations.*

Combination	C-A	C-C				A-A			
		All	M-F	M-M	F-F	All	M-F	M-M	F-F
Distance	224.54	215.87	223.15	211.92	191.54	172.82	184.76	184.73	129.98

consistent gains in common objective measures over our LSTM baseline for child speech separation. Our model is also robust to different ages of children. Furthermore, our preliminary results on SeedLing corpus [3] with realistic recordings for child language acquisition show that the proposed approach outperforms the LSTM baseline based on observing spectrograms of separated speech, implying a potential for speaker diarization applications, involving child speech.

2. Child-Adult Speech Dissimilarity

Based on our previous studies, speaker separability can be tied to distances between speaker groups [30] by adopting i-vectors [6] represent each speare and the Euclidean distance is then utilized to measure a speaker dissimilarity between the i-th and j-th speakers as follows:

$$D(i,j) = \|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2 \tag{1}$$

where $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are 100-dimensional i-vectors of two speakers in the training dataset described in Section 4.

In terms of the combination of child and adult, a mixture of two speakers generally belongs to three cases, namely mixing of child-adult (C-A), child-child (C-C), and adult-adult (A-A) speaker groups. For the C-C and A-A cases, we can further divide each case into three subclasses, namely mixing of male-female (M-F), male-male (M-M), and female-female (F-F) speaker groups. Table 1 gives the averaged distances across all speaker pairs for each of the nine combinations corresponding to the above mentioned nine input mixture cases. We use 81 adult speakers and 128 child speakers in the training set. First, the C-A combination yields the largest distance among C-A, C-C and A-A cases, which implies that the mixtures of child and adult should have a better separability than the mixtures of C-C and A-A. By considering that many studies for speech separation in A-A case have been explored, speech separation in C-A should be also feasible. Second, for the C-C and A-A mixtures, the case of mixtures from different genders yields larger distance than that of the same-gender mixtures, which implies that the different-gender mixtures should have a better separability than the same-gender mixtures. Moreover, the F-F combination seems to be more challenging. In this study, we just focus on the separation of child speech from the speech mixed up with adult speech (C-A case), ignoring the gender influence.

To visualize the similarity between two individual objects in a low-dimensional space, each object to be studied can be represented by a point and the points are elaborately arranged in order to approximate the distances between pairs of objects. We adopted multidimensional scaling (MDS) [31] to graphically describe the relationship conveyed by aforementioned distance measurements. The MDS graphs of i-vector based distance matrices for the 81 adult speakers and 129 child speakers are shown in Figure 1. In this figure, the blue and red points represent the adult and child speakers, respectively. Figure 1 confirms that the child and the adult groups could be well separated in two clusters for most cases, which motivates our proposed LSTM-based approach in the next section.

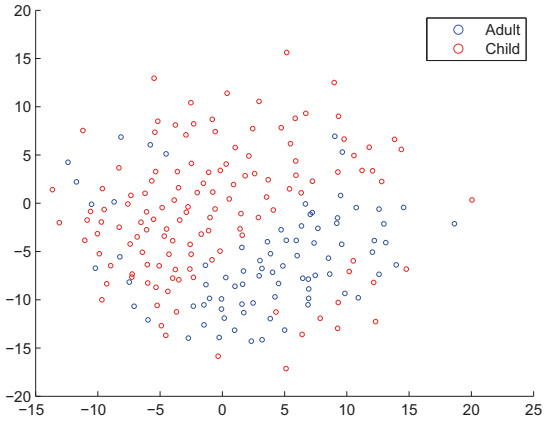

Figure 1: *Multidimensional scaling graph of the i-vector distances among all the speakers in the training set.*

3. Proposed Densely Connected LSTM Architecture for Chile Speech Separation

In training DNNs, temporal information is only utilized via frame expansion. To model time sequences, RNN seems to have an advantage with recursive structures between the previous frames and the current frame to capture the long-term contextual information. However, the conventional RNN cannot hold information for a long period and optimization of RNN parameters via back propagation through time (BPTT) faces the problem of the vanishing and exploding gradients [32]. The problems can be well alleviated by LSTM [33] which introduces memory cells and a series of gates to dynamically control the information flow.

To further improve the generalization capability of LSTM architecture, the design of hidden layers via densely connected progressive learning and output layer via multiple-target learning (MTL) is presented (denoted as LSTM_PL_MTL), as illustrated in Figure 2. This architecture is motivated by the previous work [23,24] in speech enhancement. The overall LSTM architecture aims to predict the child LPS features given the input mixed LPS features of child and adult progressively. All the target layers are designed to learn intermediate speech with higher signal-to-interference ratios (SIRs) or target child speech. For example, the input SIR of mixed speech is 0dB, then two intermediate learning targets are 10dB and 20dB speech while the final target is the child speech (infinity dB). For the input and multiple learning targets, LSTM layers are used to link between each other. This stacking style network can learn multiple targets progressively and efficiently. In order to make full use of the rich set of information from the multiple learning targets, we update the progressive learning in [23] with dense structures [34] in which the input and the estimations of intermediate target are spliced together to learn next target. In the output layer, besides the target child LPS features, the child IRM features are

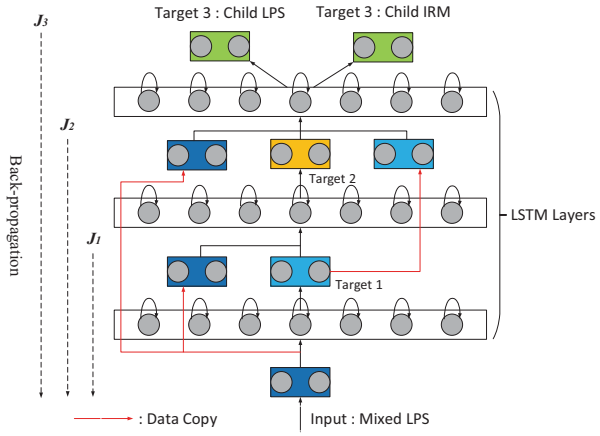

Figure 2: *The proposed LSTM_PL_MTL architecture for child speech separation.*

also adopted as another learning target and defined as:

$$x^{\text{IRM}}(t,f) = \frac{C(t,f)}{C(t,f)+A(t,f)} \tag{2}$$

where $C(t,f)$ and $A(t,f)$ denote the energy of the child speech and adult speech at time frame t and frequency bin f, respectively. Then, a weighted minimum mean squared error (MMSE) criterion in terms of multitask learning is designed to optimize all network parameters randomly initialized with K target layers as follows:

$$J = \sum_{k=1}^{K} \lambda_k J_k + J_{\text{IRM}} \tag{3}$$

$$J_k = \frac{1}{N}\sum_{n=1}^{N} \left\| \mathcal{F}_k(\hat{\boldsymbol{x}}_n^0, \hat{\boldsymbol{x}}_n^1, ..., \hat{\boldsymbol{x}}_n^{k-1}, \boldsymbol{\Lambda}_k) - \boldsymbol{x}_n^k \right\|_2^2 \tag{4}$$

$$J_{\text{IRM}} = \frac{1}{N}\sum_{n=1}^{N} \left\| \mathcal{F}_{\text{IRM}}(\hat{\boldsymbol{x}}_n^0, \hat{\boldsymbol{x}}_n^1, ..., \hat{\boldsymbol{x}}_n^{K-1}, \boldsymbol{\Lambda}_{\text{IRM}}) - \boldsymbol{x}_n^{\text{IRM}} \right\|_2^2 \tag{5}$$

where J_k is the mean square error (MSE) corresponding to k^{th} target layer while J_{IRM} is the MSE for MTL with IRM in the final output layer. $\hat{\boldsymbol{x}}_n^k$ and $\boldsymbol{x}_n^k$ are the n^{th} D-dimentional vectors of estimated and reference target LPS feature vectors for k^{th} ($k > 0$) target layer, respectively, with N representing the mini-batch size. $\hat{\boldsymbol{x}}_n^0$ denotes the n^{th} vector of input mixed LPS features with acoustic context. $\mathcal{F}_k(\hat{\boldsymbol{x}}_n^0, \hat{\boldsymbol{x}}_n^1, ..., \hat{\boldsymbol{x}}_n^{k-1}, \boldsymbol{\Lambda}_k)$ is the neural network function for k^{th} target with the dense structure using the previously learned intermediate targets from $\hat{\boldsymbol{x}}_n^0$ to $\hat{\boldsymbol{x}}_n^{k-1}$, and $\boldsymbol{\Lambda}_k$ represents the parameter set of the weights and bias vectors before k^{th} target layer, which are optimized with gradient descent. $\boldsymbol{x}_n^{\text{IRM}}$, $\mathcal{F}_k(\hat{\boldsymbol{x}}_n^0, \hat{\boldsymbol{x}}_n^1, ..., \hat{\boldsymbol{x}}_n^{K-1}, \boldsymbol{\Lambda}_{\text{IRM}})$ and $\boldsymbol{\Lambda}_{\text{IRM}}$ are corresponding versions to IRM targets. λ_k is the weighting factor for the k^{th} target layer.

4. Experiments and Result Analysis

In our experiments, the adult speech data was derived from the WSJ0 corpus [28]. 1000 utterances for 129 children from 2 years old to 5 years old were adopted as child speech derived from PhonBank project [29]. The whole 7138 utterances (about 12 hours speech) from 83 speakers, denoted as SI-84 training set, were mixed with the above mentioned 1000 child utterances at three SNR levels (-5dB, 0dB and 5dB) to build a 36-hour training set, consisting of pairs of child and mixed utterances. The 330 utterances from 12 other adult speakers, namely the Nov92 WSJ evaluation set, were used to construct the test set for each combination of SNR levels (-10dB, -5dB, 0dB, 5dB) with 235 utterances from 32 unseen children.

Table 2: *Average performance comparison of different SNRs on the test set between LSTM_Baseline and LSTM_PL_MTL.*

SNRs	Systems	SSNR	PESQ	STOI
-10dB	LSTM_Baseline	0.47	1.76	0.55
	LSTM_PL_MTL	0.50	2.02	0.61
-5dB	LSTM_Baseline	1.16	1.93	0.60
	LSTM_PL_MTL	1.57	2.29	0.67
0dB	LSTM_Baseline	1.93	2.09	0.65
	LSTM_PL_MTL	3.71	2.54	0.73
5dB	LSTM_Baseline	2.65	2.24	0.69
	LSTM_PL_MTL	6.20	2.76	0.77

For signal analysis, speech was sampled at 16 kHz. A 512-point discrete Fourier transform (DFT) of each overlapping windowed frame was computed. Then 257-dimensional LPS vectors normalized by global mean and variance were used to train LSTMs. The phase required to reconstruct waveform was directly extracted from the mixed speech [35] and the child speech waveform was reconstructed from the estimated spectral magnitude and the mixed speech phase with an overlap-add method. The Microsoft Computational Network Toolkit (CNTK) [36] was used for training. For progressive learning systems, one LSTM layer was used to connect the input layer and target layers. Each target SIR gain was 10dB. The 1-frame input and the estimations of intermediate target are spliced together to learn next target. The number of LSTM memory cells in each layer was 1024, and the parameter λ_k in Equation 3 was all set to 0.1. The IRM output of LSTM_PL_MTL was used to reconstruct the separated speech waveform. As a comparison, a direct mapping LSTM network in [24] with the architecture 257-1024-1024-1024-257, consisting of three LSTM layers and 1024 memory cells for each LSTM layer, was built as our baseline model (denoted as LSTM_Baseline). The learning rate of the two models for the fine-tuning was set to 0.001 for the first 20 epochs and 0.0001 for the next 30 epochs. Segmental signal-to-noise ratio (SSNR in dB), perceptual evaluation of speech quality (PESQ) [37], and short-time objective intelligibility (STOI) [38] were adopted to evaluate the performances of separated speech.

Table 2 shows the average SSNR, PESQ and STOI on the whole test set between LSTM_Baseline and LSTM_PL_MTL. Clearly, the proposed LSTM_PL_MTL approach yielded consistent and significant improvements over the LSTM_Baseline approach for all different SNRs of the mixed speech, e.g., a SSNR gain of 1.78 dB, a PESQ gain of 0.45, a STOI gain of 0.08 at 0dB. Figure 3 shows the spectrograms of an utterance example at -5dB. The LSTM_Baseline achieved a good interference reduction but with severe child speech distortion and child speech loss. Meanwhile LSTM_PL_MTL generated the child speech with less speech distortion as shown in the blue rectangles, especially preserving the structure in the high frequency band. All these results illustrated the superiority of the proposed LSTM_PL_MTL approach over the conventional LSTM approach in terms of separation capability.

Table 3: *Average performance comparison of different ages on the test set between LSTM_Baseline and LSTM_PL_MTL.*

Age	Systems	SSNR	PESQ	STOI
2	LSTM_Baseline	1.30	1.93	0.59
	LSTM_PL_MTL	2.92	2.38	0.65
3	LSTM_Baseline	1.82	2.15	0.64
	LSTM_PL_MTL	3.75	2.52	0.72
4	LSTM_Baseline	1.88	2.06	0.66
	LSTM_PL_MTL	3.60	2.52	0.74
5	LSTM_Baseline	1.88	1.97	0.62
	LSTM_PL_MTL	3.69	2.40	0.70

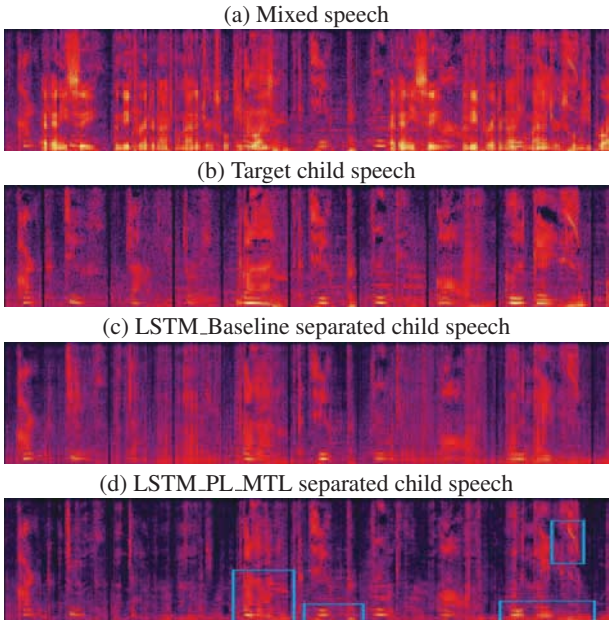

Figure 3: *Spectrograms of an utterance example at -5dB from the simulated test set.*

In addition, in order to show the robustness of the proposed approach to different ages of children, we make a comparison of average SSNR, PESQ and STOI across all SNR levels, as shown in Table 3. For the ages from 2 years old to 5 years old, the LSTM_PL_MTL consistently performed better than the conventional LSTM in the child-adult speech separation task.

The above mentioned experiments well demonstrated the effectiveness of the proposed child speech separation model on the simulation data. However, the realistic audio recordings in adverse acoustic environments should be much more challenging. Accordingly, we test our model on the real data from SeedLing corpus [3] which is designed for child language acquisition. As shown in Figure 4, the original utterance includes not only the mixed speech of child and adult but also the background noises and reverberations. We marked the child speech segments with red lines and adult speech segments with black lines. Obviously, the LSTM_Baseline could not well remove both the adult speech and background noises due to its limited generalization capability. But our proposed LSTM_PL_MTL model could well separate child speech from the quite noisy mixed speech compared with LSTM_Baseline, yielding much better inference removal and child speech preservation especially in high frequency bands as shown in blue rectangles. The

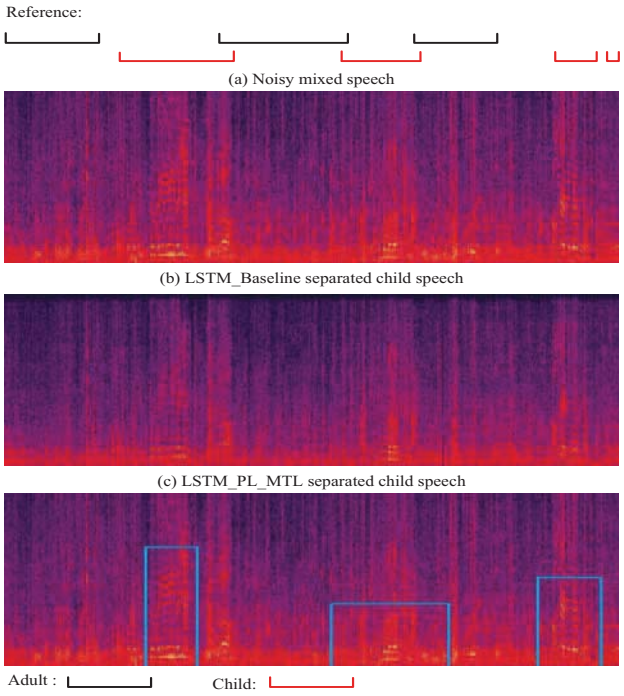

Figure 4: *Spectrograms of an utterance example from the realistic recordings of SeedLing corpus.*

LSTM_PL_MTL result is amazing because both child and adult speakers are unseen while the background noises and reverberations are not considered in the current framework. So the only reason should be the strong generalization capability from the novel design of LSTM architecture.

5. Conclusions

In this study, we propose a novel LSTM design with densely connected hidden layers and stacked inputs containing progressively obtained intermediate targets that are learnt via a multiple-target learning for child speech separation in a speaker-independent manner in order to reduce the impact of training data limitation because collecting a large set of training speech utterances from children aged 2 to 5 is not an easy task. In a preliminary set of experiments, our approach could yield a better performance and less speech distortions in child speech when compared with the conventional LSTM. Even in the quite noisy and challenging realistic conditions, our proposed approach can achieve satisfying performances. Inspired by the encouraging results, we will investigate to use the proposed approach to address the issue of detecting overlapped speech in speaker diarization involving child speech in future studies.

6. Acknowledgements

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005, MOE-Microsoft Key Laboratory of USTC, and Huawei Noah's Ark Lab.

7. References

[1] V. C. Shields, "Separation of added speech signals by digital comb filtering," *S.M. Thesis, Dept. of Electrical Engineering, MIT*, 1970

[2] D. Giuliani and M. Gerosa. "Investigating recognition of children's speech." In *Proc. ICASSP*, pp. 37-40, 2003.

[3] E. Bergelson, "Bergelson Seedlings HomeBank Corpus," doi:10.21415/T5PK6D, Accessed: 2017-08-22.

[4] S. Lee, A. Potamianos, and S. Narayanan, "Acoustic of childrens speech developmental changes of temporal and spectral parameters," *J. Acoust. Soc. Am.*, vol. 105, No.3, pp. 1455-1468, March 1999.

[5] G. J. McLachlan and K. E. Basford, "Mixture Models: Inference and Applications to Clustering." New York, NY, USA: Marcel Dekker, 1988.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, And Language Processing.*, vol. 19, pp. 788-798, 2011.

[7] Y. Zeng and Y. Zhang, "Robust children and adults speech classification," In *Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference,* vol. 4, pp. 721-725, 2007

[8] J. Grzybowska and S. Kacprzak, "Speaker age classification and regression using i-vectors," In *Proc. INTERSPEECH,* pp. 1402-1406, 2016

[9] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. EUROSPEECH*, pp. 1009-1012, 2003.

[10] M. Schmidt and R. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization, in *Proc. INTERSPEECH*, pp. 2614-2617, 2006.

[11] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 66-70, 2015.

[12] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications.* Hoboken, NJ, USA: Wiley, 2006

[13] K. Hu and D. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 21, no. 1, pp. 122-131, 2013.

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

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

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

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

[18] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, pp. 4628-4632, 2014.

[19] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A Reverberation-Time-Aware Approach to Speech Dereverberation Based on Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Lang. Process.*, vol. 25, no. 1, pp. 102-111, 2017.

[20] F. Weninger, J. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proceedings of GlobalSIP*, pp. 740-744, 2014.

[21] A. Narayana n 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*, pp. 7092-7096, 2013.

[22] Y. Wang, A. Narayanan, D.L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.* 22(12), 1849-1858 (2014)

[23] T. Gao, J. Du, Li.-R Dai, and C.-H Lee, "Snr-based progressive learning of deep neural network for speech enhancement," in *INTERSPEECH,* pp. 3713-3717, 2016.

[24] L. Sun, J. Du, Li.-R Dai, and C.-H Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *Hands-free Speech Communications and Microphone Arrays (HSCMA).2017 IEEE,* pp. 136-140, 2017.

[25] Xu, Yong, et al. "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement." *arXiv preprint arXiv:1703.07172*, 2017.

[26] Tu, Yanhui, et al. "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers." *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on. IEEE*, 2014.

[27] Wang, Yannan, et al. "Unsupervised single-channel speech separation via deep neural network for different gender mixtures." *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific. IEEE,* 2016.

[28] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete,"*Linguistic Data Consortium, Philadelphia*, 2007.

[29] http://media.talkbank.org/PhonBank/Eng-NA/PaidoEnglish/

[30] Wang, Yannan, et al. "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.7 (2017): 1535-1546.

[31] F. Young and R. Hamer, "Theory and applications of multidimensional scaling", *Hillsdale, NJ: Eribaum Associates*, 1994.

[32] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, pp. 1310-1318, 2013.

[33] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," 1999.

[34] G. Huang, Z . Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," *arXiv preprint arXiv:1608.06993,* 2016

[35] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 679-681, Aug. 1982.

[36] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang, et al., "An introduction to computational networks and the computational network toolkit," Microsoft Technical Report MSR-TR-2014-112, 2014

[37] A.W. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Recommendation, pp. 862, 2001.

[38] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125-2136, Sep. 2011.