A NOVEL LSTM-BASED SPEECH PREPROCESSOR FOR SPEAKER DIARIZATION IN REALISTIC MISMATCH CONDITIONS

Lei Sun[1]*, Jun Du*[1]*, Tian Gao*[1]*, Yu-Ding Lu*[2]*, Yu Tsao*[2]*, Chin-Hui Lee*[3]*, Neville Ryant*[4]

[1]University of Science and Technology of China, Hefei, Anhui, China
[2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[3]Georgia Institute of Technology, Atlanta, Georgia, USA
[4]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA

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

ABSTRACT

In this study, we investigate on the effects of deep learning based speech enhancement as a preprocessor to speaker diarization in quite challenging realistic environments involving the background noises, reverberations and overlapping speech. To improve the generalization capability, the advanced long short-term memory (LSTM) architecture with the novel design of hidden layers via densely connected progressive learning and output layer via multiple-target learning is proposed for preprocessing. We build the deep model using synthesized training data pairs generated from WSJ0 reading-style speech and more than 100 noise types. Surprisingly, this proposed preprocessor demonstrates a strong generalization capability to speaker diarization with the realistic noisy speech in highly mismatched conditions, in terms of the speaking style, interferences, and the interaction between them. Tested on three challenging tasks, namely AMI, ADOS, and SeedLings, the state-of-the-art diarization system with the novel LSTM-based speech preprocessor can yield consistent and significant reductions of diarization error rate (DER) over the systems using unprocessed noisy speech and traditional enhancement methods.

Index Terms— Speaker diarization, deep learning based speech enhancement, densely connected progressive learning, multiple-target learning, highly mismatch condition

1. INTRODUCTION

Speaker diarization, also referred as the *who speak when* task, is to segment an audio recording into speaker homogeneous regions [1], which is an essential part for many applications, such as speaker adaptation in large vocabulary continuous speech recognition (LVCSR) systems, speaker based indexing, navigation and retrieval [2]. A complete diarization system should contain the following stages: audio preprocessing, speech activity detection (SAD), acoustic feature extraction, segmentation/clustering algorithms. And segmentation/clustering, namely the core diarization algorithm, is one most important part. Historically, the diarization algorithms could be summarized into two categories. The first one is based on hidden Markov model (HMM) where each hidden state is a Gaussian mixture model (GMM) corresponding to a speaker [3]. The other one is the non-parametric method based on information theory [4]. Besides the diarization algorithms, the other parts of the pipeline can also affect the final diarization results. For instance, the long-term conversational features for better overlap detection [5], the i-vectors to enhance speaker specific information [6], and system fusion strategies [7, 8] were proposed and widely used.

Inevitably, a practical diarization system should address the environmental robustness problem in real applications, e.g., the background noises, reverberations and overlapping speech. However, for this challenging problem, few studies have focused on front-end speech enhancement or preprocessing for speaker diarization, especially in the single-channel case. For traditional enhancement approaches like Wiener filtering [3], LogMMSE [9], there are many limitations in real applications, e.g., the weakness of tracking non-stationary noises, due to the model assumptions made during the inference. Moreover, the resulting denoised speech often suffers from annoying artifact called musical noise [10], which can even degrade the performance of speaker diarization system in realistic adverse environments.

With the emergence of supervised deep learning techniques, the great efforts are also made for speech enhancement area recently. In [11, 12], ideal ratio masks (IRMs) were used to make binary classification on time-frequency (T-F) units for speech separation. Deep denoising autoencoder was also studied on speech enhancement in [13]. Previously, we proposed a deep neural network (DNN) framework to learn the direct mapping from noisy to clean speech in log-power spectral (LPS) domain, which demonstrated its superiority to the traditional enhancement methods [14, 15], especially for tracking the non-stationary noises. In [16], Gao $et, al.$ proposed a DNN-based progressive learning (PL) approach which aimed at decomposing the complicated regression problem into a series of subproblems. To increase the modeling capability, in [17], we adopted the advanced long short-term memory recurrent neural network (LSTM-RNN) with multiple-target learning of both LPS and IRM to capture the long-term contextual information and make a better prediction.

Nevertheless, the recent progress in deep learning based speech enhancement are not widely investigated for speaker diarization. This might be partially explained as that the generalization problem of deep learning approaches in highly mismatched conditions can be well solved. In other words, it is unrealistic for speaker diarization to rebuild a matched speech preprocessor for a specific scenario. For example in [18], a feature mapping DNN is used to map the noisy features to clean features, which improve the diarization performance in relatively matched condition. But it can not well generalize to mismatch testing conditions due to the simple NN architecture and the binding feature design with the diarization algorithm. To address these issues, in this paper, we conduct a comprehensive study on well-designed deep learning based speech enhancement approach as a preprocessor to speaker diarization to

978-1-5386-4658-8/18/$31.00 ©2018 IEEE

test the generalization capability in highly mismatched conditions. Accordingly, the advanced LSTM architecture with the novel design of hidden layers via densely connected progressive learning and output layer via multiple-target learning is proposed for preprocessing. We build the deep model using synthesized training data pairs generated from WSJ0 reading-style speech and more than 100 noise types. Surprisingly, this proposed preprocessor demonstrates a strong generalization capability to speaker diarization with the realistic noisy speech in highly mismatched conditions, in terms of the speaking style, interferences, and the interaction between them. Tested on three challenging tasks, namely AMI, ADOS, and SeedLings, the state-of-the-art diarization system with the novel LSTM-based speech preprocessor can yield consistent and significant reductions of diarization error rate over the systems using unprocessed noisy speech and traditional enhancement methods. Furthermore, as a preprocessor, our approach is independent with the diarization system, e.g., the feature design, showing its superiority to the feature mapping method in [18].

2. DIARIZATION SYSTEM

In this section, we briefly introduce the state-of-the-art information bottleneck(IB) diarization algorithm proposed in [2] and the error metric of diarization system, which are used in our experiments.

2.1. Information Bottleneck Framework

Suppose we have a set of speech segment $\boldsymbol{X} = \{x_1, x_2, ..., x_m\}$ to be clustered, and $\boldsymbol{Y} = \{y_1, y_2, ...y_n\}$ denotes the set of relevance variables carrying meaningful information about the desired clutering output $\boldsymbol{C} = \{c_1, c_2, ...c_p\}$. Thus the optimal clustering can be found by maximmize the following function:

$$F = I(C, Y) - \frac{1}{\beta} I(C, X) \tag{1}$$

where $I(\cdot, \cdot)$ denotes the mutual information between two sets of random variables. β is a Lagrange multiplier representing the tradeoff between amount of information preserved $I(C, Y)$ and compression of the initial representation $I(C, X)$. The objection function can be optimized by agglomerative Information Bottleneck (aIB) method, which is a greedy approach. The algorithm is initialized with each input element $x_i \in X$ as an individual cluster. Subsequently, two clusters are merged to guarantee the decrease in the objective function is minimum at the current step. More optimization details can be found in [4].

2.2. Error Metric

To measure the performance of a diarzation system, we use the diarization error rate (DER) [3], which is defined by the evaluations campaigns organized by NIST. It compares the differences between the ground-truth reference segmentation and the generated diarization output. The final DER result is the sum of three types of errors:

$$\text{DER} = E_{\text{Miss}} + E_{\text{FA}} + E_{\text{Spkr}} \tag{2}$$

where E_{Miss} is the percentage of missed speech error (speaker not attributed when speech segment exists), E_{FA} is the percentage of false alarm error (speaker attributed in non-speech segment), E_{Spkr} is the percentage of speaker misclassification error (wrong speaker labeling according to reference segmentation). Lower DER indicates better diarization performance. Additionally, a non-scoring collar of 250 msec [1] is generally adopted in both sides of the ground-truth segment boundaries to eliminate the effects of inevitably inaccurate labeling.

Our speaker diarization system is implemented using IB toolkits DiarTk [19]. 19 Mel-frequency cesptrum coefficients (MFCCs) are extracted as the acoustic features by HTK toolkit [20] with a 10-msec frame shift with a 30-msec analysis window. As reported in [21], the awareness of the number of speakers has little effect on final diarization performance. So we manually set the max cluster number. In our experiments, we use oracle SAD segments derived from human transcriptions and ignore the overlap segments. Accordingly, the first two parts, namely E_{Miss} and E_{FA}, have no effects on final DER. We focus on the speaker misclassification error to highlight the effects of enhancement methods.

3. THE NOVEL ARCHITECTURE FOR PREPROCESSOR

In the conventional DNN architecture for speech enhancement [15], the important temporal information is only considered via frame expansion. To model time sequences, recurrent neural networks (RNN) seem to have a congenital advantage by using recursive structures between the previous frames and the current frame to capture the long-term contextual information. However, the conventional RNN can not hold information for a long period and the optimization of RNN parameters via the back propagation through time (BPTT) faces the problem of the vanishing and exploding gradients [22]. The problems can be well alleviated by the invention of LSTM [23] which introduces the concepts of memory cell and a series of gates to dynamically control the information flow. As shown in Fig. 1, all LSTM layers consist of memory cells.

To further improve the generalization capability of LSTM architecture, the novel design of hidden layers via densely connected progressive learning and output layer via multiple-target learning is presented (denoted as LSTM-PL-MTL), as illustrated in Fig. 1, which is inspired by our previous work [17, 16]. The overall LSTM architecture aims to predict the clean LPS features given the input noisy LPS features with acoustic context. All the target layers are designed to learn intermediate speech with higher SNRs or clean speech. For the input and multiple 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 [16] with dense structures [24] in which the input and the estimations of intermediate target are spliced together to learn next target. Then, a weighted MMSE criterion in terms of MTL is designed to optimize all network parameters randomly initialized with K target layers as follows:

$$\begin{aligned}
E &= \sum_{k=1}^{K} \alpha_k E_k + E_{\text{IRM}} \\
E_k &= \frac{1}{N} \sum_{n=1}^{N} \|\mathcal{F}_k(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{k-1}, \mathbf{\Lambda}_k) - \mathbf{x}_n^k\|_2^2 \\
E_{\text{IRM}} &= \frac{1}{N} \sum_{n=1}^{N} \|\mathcal{F}_{\text{IRM}}(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{K-1}, \mathbf{\Lambda}_{\text{IRM}}) - \mathbf{x}_n^{\text{IRM}}\|_2^2
\end{aligned} \tag{3}$$

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

The LSTM-PL-MTL architecture in Fig. 1 is (257*7)-1024-(257+257)-1024-(257+257+257)-1024-(257:257), denoting the dimension of both LPS and IRM feature vectors is 257, 7-frame expansion is used for input, the number of LSTM memory cells in each layer is 1024, and K=3. α_K=1.0 and α_k=0.1 ($k < K$). As a comparison, a two-layer direct mapping LSTM network which only consists of conventional recurrent LSTM layers, is built as our baseline model. Note that the IRM output of LSTM-PL-MTL is used to reconstruct the enhanced speech waveform. By using new structure, the enhanced speech has better intelligibility and less distortions. More details of the preprocessor can refer to [25].

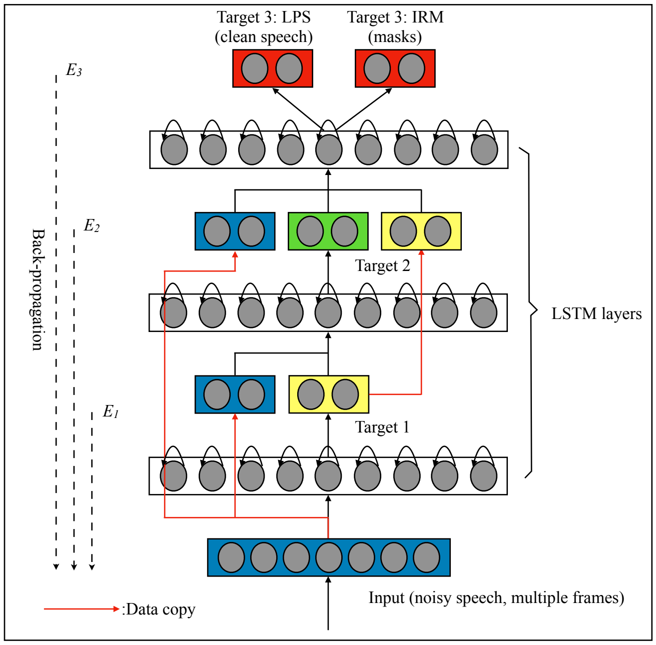

Fig. 1. An illustration of the novel architecture for preprocessor.

4. EXPERIMENTS AND RESULTS

4.1. Database

To train deep models for enhancement, 115 noise types are adopted and clean speech of WSJ0 corpus is employed as in [16]. 7138 utterances (about 12 hours of reading style speech) from 83 speakers, denoted as SI-84 training set, are corrupted with the above mentioned 115 noise types at three SNR levels (-5dB, 0dB and 5dB) to build a 36-hour training set, consisting of pairs of clean and noisy utterances.

A variety of datasets were used as the testing sets to explore the generalization capability of enhancement methods in adverse conditions, including the widely-used AMI meeting corpus [26], the soon-to-be-published ADOS (Autism Diagnostic Observation Schedule) corpus [27], and SeedLings corpus of day-long child language acquisition recordings [28].

Table 1. High mismatches between training and testing.

	Training	Testing		
Corpus	WSJ0	ADOS	SeedLings	AMI
Distance	Near	Far	Near	Far
Style	Reading	Conversation		
Interferences	Additive noise	Background noises, reverberations		
Interaction	Simulation	Unknown, real noisy speech		
Child?	None	Kids	6-month baby	None

AMI corpus consists more than one hundred of recorded meetings of 4-5 participants. The conversations are recorded in parallel on multiple devices, including a tabletop array of 8 microphones and head-mounted microphones on each meeting participant. We conduct our experiments on two conditions: ‘SDM’, a single distance microphone from the far-field array; ‘MDM’, a beamformed version of multiple distant microphones in the array by using BeamformIt tool [29]. Both the development and evaluation sets of AMI are adopted for testing, including 34 meetings totally.

The conversations in ADOS are recorded between children and trained clinicians in a semi-structured interview mode. The recordings are extracted from the ceiling microphone as described in [27]. The recordings of SeedLings are extracted via a device worn in a breast pocket by the child, so the sounds of the child and those surroundings are all included. Thus, the acoustic environment for the conversations varies throughout the day. More interestingly, all these children in SeedLings are only 6 months old who can seldom speak clear sentences.

Please note that all the test sets are recorded in realistic noisy conditions. Therefore, there are high mismatches between the training and testing, as shown in Table 1. For example, the reading style speech data are used in training while all testing data are conversational speech. The far-field speech data are recorded in AMI and ADOS. For the interferences, only additive noises are considered in training while both background noises and reverberations with unknown interaction with speech exist in testing. Both ADOS and SeedLings have child data. All these mismatches make those tasks quite challenging for our proposed enhancement approach.

4.2. Results

Several notations are first defined. ‘Noisy’ denotes unprocessed noisy speech. ‘Log-mmse’ and ‘PSNR-wiener’ represent two traditional methods [9, 30] for the comparison with deep learning based approaches. ‘DNN’ is the regression-based deep learning approach in [15]. ‘LSTM’ and ‘LSTM-PL-MTL’ are two LSTM-based approaches introduced in Section 3.

4.2.1. Evaluations on AMI

On SDM data of AMI corpus, we conduct a comprehensive comparison between different enhancement methods. Several observations could be made in Fig. 2. First, the traditional methods like Log-mmse and PSNR-wiener do not work well and even lead to diarization performance degradation. Second, the generalization capability of DNN is poor due to its simple architecture. Finally, both LSTM and LSTM-PL-MTL achieve better DER results than unprocessed system for all settings of maximum cluster number. The well-designed LSTM-PL-MTL yields the best results and significantly

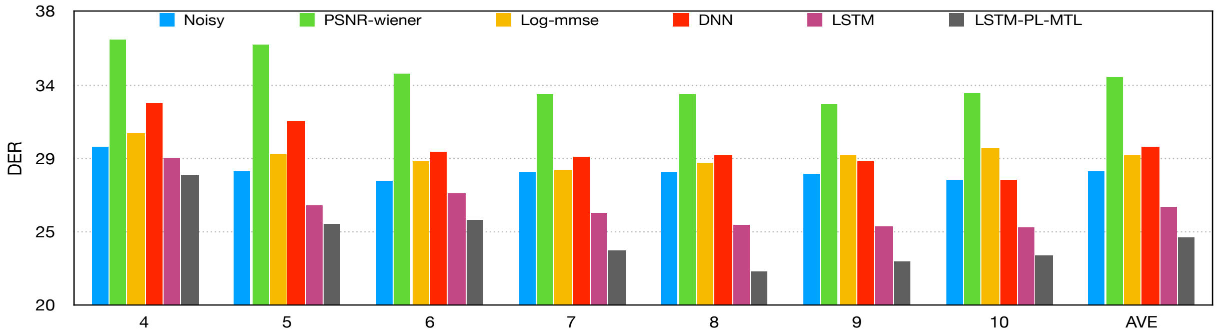

Fig. 2. Diarization results across different maximum cluster numbers on SDM data of AMI corpus.

outperforms LSTM for all cases, demonstrating the strong generalization capability to mismatched conditions.

For the MDM data, we can also use the deep learning based approaches as a post-processing to eliminate the nonlinear distortions by the beamforming algorithm. As shown in Table 2, the similar trend to SDM case can be observed. The proposed LSTM-PL-MTL achieves a relative DER reduction of 16.7% over MDM beamformed data, indicating that deep learning based post-processing is strongly complementary with the conventional beamforming techniques.

Table 2. Average DER results (in %) across different settings of maximum cluster number (4-10) on MDM data of AMI corpus.

Noisy	DNN	LSTM	LSTM-PL-MTL
25.9	26.4	22.5	21.6

Table 3. Average DER results (in %) across different settings of maximum cluster number (2-10) on ADOS and SeedLings.

	Noisy	Log-mmse	PSNR-wiener	LSTM-PL-MTL
ADOS	36.3	40.0	36.0	29.2
SeedLings	45.3	47.0	46.7	39.2

4.2.2. Evaluations on ADOS and SeedLings

We give the DER comparison among the traditional methods and the best LSTM-PL-MTL architecture on ADOS and SeedLings, as listed in Table 3. Similar to the results on AMI corpus, the traditional methods can not improve the performance over the unprocessed case while LSTM-PL-MTL yields relative DER reductions of 19.6% and 13.5% for ADOS and SeedLings, respectively.

Finally, to illustrate why LSTM-PL-MTL based speech preprocessing can help diarization, we choose a conversation from SeedLings which contains a female adult and a male child, and the max cluster number is set to 2. The overall DERs for each recording are 44.5% before preprocessing and 13.2% after preprocessing. A 15-second segment is shown in Fig. 3. Before denoising, many adult segments are misclassified as child due to the presence of huge grating noises from carry-on microphone. After denoising, all the adult segments are correctly classified. However, a short child segment in the ending is largely eliminated as our deep model mistakes it as noises, which can be explained as the child data is not used as the speech data to train our enhancement model. To address this problem, we need to include child data for training of deep models in the future.

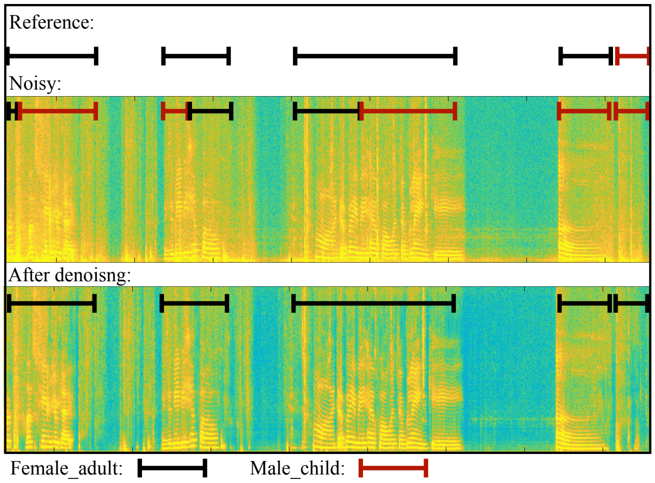

Fig. 3. Visualization of a diarization example from SeedLings.

5. CONCLUSION

In this paper, we investigate the effects of deep learning based speech enhancement as a preprocessor for speaker diarization. The proposed preprocessor based on well-designed LSTM architecture demonstrates a strong generalization capability with the realistic noisy speech in highly mismatched conditions and yields consistent and significant reductions of DER over the systems using unprocessed noisy speech and traditional enhancement methods.

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, and MOE-Microsoft Key Laboratory of USTC. Tencent has also funded this work. Note that this work was started at JSALT 2017, which was supported by JHU with gifts from Amazon, Apple, Facebook, Google and Microsoft. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), specifically the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

7. REFERENCES

[1] C. Wooters and M. Huijbregts, "The icsi rt07s speaker diarization system," *Multimodal Technologies for Perception of Humans*, pp. 509–519, 2008.

[2] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1382–1393, 2009.

[3] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[4] S. H. Yella and H. Bourlard, "Information bottleneck based speaker diarization of meetings using non-speech as side information," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 96–100.

[5] S. H. Yella and H. Bourlard, "Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1688–1700, 2014.

[6] S. Madikeri, I. Himawan, P. Motlicek, and M. Ferras, "Integrating online i-vector extractor with information bottleneck based speaker diarization system," Tech. Rep., Idiap, 2015.

[7] S.E. Tranter, "Two-way cluster voting to improve speaker diarisation performance," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*. IEEE, 2005, vol. 1, pp. I–753.

[8] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J.F. Bonastre, "The elisa consortium approaches in broadcast news speaker segmentation during the nist 2003 rich transcription evaluation," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004, vol. 1, pp. I–373.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[10] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, pp. 873–902. Springer, 2008.

[11] 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.

[12] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 4, pp. 826–835, 2014.

[13] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder.," in *Interspeech*, 2013, pp. 436–440.

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

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

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

[17] 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, 2017, pp. 136–140.

[18] W. Zhu, W. Guo, and G. Hu, "Feature mapping for speaker diarization in noisy conditions," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5445–5449.

[19] D. Vijayasenan and F. Valente, "Diartk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[20] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, et al., "The htk book," *Cambridge university engineering department*, vol. 3, pp. 175, 2002.

[21] M. Sinclair and S. King, "Where are the challenges in speaker diarization?," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7741–7745.

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

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

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

[25] G. Tian, D. Jun, et al., "Densely connected progressive learning for lstm-based speech enhancement.," in *Submitted to ICASSP*, 2018.

[26] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[27] J. Parish-Morris, M. Liberman, N. Ryant, C. Cieri, L. Bateman, E. Ferguson, and R. T. Schultz, "Exploring Autism Spectrum Disorders using HLT.," in *CLPsych*, 2016, pp. 74–84.

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

[29] X. Anguera, "Beamformit, the fast and robust acoustic beamformer," 2006.

[30] P. Scalart, "Speech enhancement based on a priori signal to noise estimation," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. IEEE, 1996, vol. 2, pp. 629–632.