A Speaker-Dependent Deep Learning Approach to Joint Speech Separation and Acoustic Modeling for Multi-Talker Automatic Speech Recognition

Yan-Hui Tu[1], *Jun Du*[1], *Li-Rong Dai*[1], *Chin-Hui Lee*[2]

[1]University of Science and Technology of China
[2]Georgia Institute of Technology

tuyanhui@mail.ustc.edu.cn, {jundu,lrdai}@ustc.edu.cn, chl@ece.gatech.edu

Abstract

We propose a novel speaker-dependent (SD) approach to joint training of deep neural networks (DNNs) with an explicit speech separation structure for multi-talker speech recognition in a single-channel setting. First, a multi-condition training strategy is designed for a SD-DNN recognizer in multi-talker scenarios, which can significantly reduce the decoding runtime and improve the recognition accuracy over the approaches that use speaker-independent DNN models with a complicated joint decoding framework. In addition, a SD regression DNN for mapping the acoustic features of mixed speech to the speech features of a target speaker is jointly trained with the SD recognition DNN for acoustic modeling. Our experiments on the Speech Separation Challenge (SSC) task show that the proposed SD recognition system under multi-condition training achieves an average word error rate (WER) of 3.8%, yielding a relative WER reduction of 65.1% from the proposed DNN pre-processing approach under clean-condition training [1]. Furthermore, the jointly trained DNN system generates a relative WER reduction of 13.2% from the state-of-the-art systems under multi-condition training.

Index Terms: multi-talker speech recognition, speaker-dependent model, single-channel speech separation, deep neural networks, joint training

1. Introduction

In the mobile internet era, automatic speech recognition (ASR) techniques are widely used in many speech-enabled applications. However, multi-talker ASR with a single microphone is still quite challenging because of the coupled problems of speech separation and ASR of poorly separated target speech. Even with the availability of dual-microphone setting in most of today's mobile devices, the speech separation performance is still unsatisfactory. As early as 2006, the monaural speech separation and recognition challenge (SSC) [2] was launched which aimed at recognizing speech of a target speaker given the single-channel mixed speech corrupted by an interfering talker. Historically, all the approaches to solving this problem could be mainly divided into three categories. First, an interaction between target and competing speech signals with their temporal dynamics were modeled using factorial hidden Markov model (FHMM) [3, 4, 5, 6] for separation, followed by a joint decoding strategy for ASR. Second, non-negative matrix factorization (NMF) [7, 8] was adopted for speech separation. Finally in [9, 10, 11], approaches based on computational auditory scene analysis (CASA) to estimate a time-frequency mask of each speaker have been proposed. Among all the submissions to SSC, the IBM superhuman system [3], belonging to the first category, performed the best and even exceeded human listeners on the challenge task.

Recent advances in deep learning [12, 13], especially deep neural networks (DNN), have led to a great success in a number of speech processing areas. For example, the hybrid DNN-HMM structure [14, 15, 16] was widely adopted in ASR systems for acoustic modeling in place of the traditional GMM-HMM. In source separation, a series of DNN based approaches [17, 18, 19] were also proposed to separate each target speaker from mixed speech. Furthermore, for single-channel multi-talk speech recognition, one remarkable work in [20] utilized a novel DNN architecture to jointly model the two mixing speakers with a weighted finite-state transducer (WFST) based decoder, which outperformed the IBM superhuman system.

However, both the state-of-the-art approaches in [3, 20] use a joint decoding framework which requires an additional computational complexity. Meanwhile, those methods cannot be easily extended to scenarios with more than two competing speakers. To alleviate these difficulties, we concentrate our attention on extracting information of the target speaker which is more relevant in source separation and ASR in multi-talker scenarios. In [1], speaker-dependent (SD) DNN models were designed for speech separation as a pre-processing module for the subsequent speech recognition task using clean-condition trained GMM-HMMs, yielding a better recognition accuracy than the IBM system with a more efficient decoder. In this study, we extend the speaker-dependent concept from speech separation to multi-talker speech recognition.

Accordingly, a novel speaker-dependent DNN for joint modeling of the front-end separation and back-end ASR is proposed to separate and recognize the target speaker simultaneously. The main contributions are summarized as follows. First, a multi-condition training strategy by synthesizing a large-scale training set via very limited data from each speaker is adopted to boost the speaker-dependent speech recognition performance in the multi-talker scenario, achieving a significantly lower word error rate (WER) and smaller runtime latency in comparison to all the existing speaker-independent (SI) approaches on the SSC task. Second, a speaker-dependent regression DNN for mapping the log mel-filterbank (LMFB) features of mixed speech to speech of the target speaker is adopted as the front-end, which is different from the proposed pre-processing DNN using log-power spectra features [1]. Finally, the speaker-dependent front-end DNN can be seamlessly concatenated and jointly trained with the speaker-dependent back-end DNN for acoustic modeling as a hybrid DNN architecture, which explicitly normalizes the variability from other interfering speakers.

Experimental results on the SSC task show that our speaker-dependent approach is quite robust to the interference of a com-

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

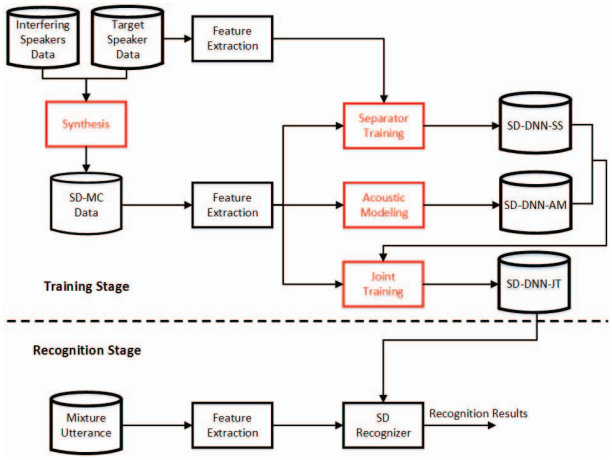

Figure 1: SD recognition system in multi-talker scenarios.

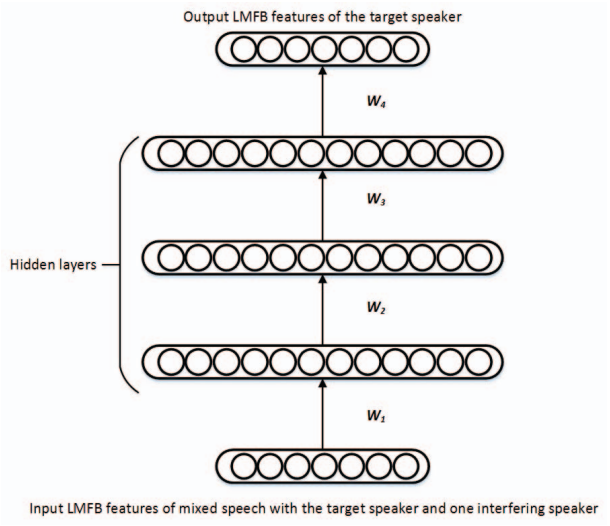

Figure 2: Illustration of SD-DNN-SS.

peting speaker even in low target-to-masker ratio (TMR) conditions. The best configured multi-condition system achieves an average WER of 3.8% across different TMRs, yielding a relative WER reduction of 65.1% from the proposed DNN preprocessing approach under clean-condition training [1]. Furthermore, the jointly trained DNN system can generate a relative WER reduction of 13.2% from the state-of-the-art high-performance systems under multi-condition traing.

2. System Overview

In Figure 1 we illustrate the proposed SD recognition system in multi-talker scenarios. In the training stage, a speaker-dependent multi-condition (SD-MC) training set is first designed for a target speaker. Then the LMFB feature pairs of mixed speakers and the target speaker are extracted from the training data samples, which are used for joint training of the speaker-dependent DNNs for speech separation and acoustic modeling, denoted as SD-DNN-SS and SD-DNN-AM, respectively, and finally a hybrid DNN (SD-DNN-JT) is generated. In the recognition stage, as in the conventional procedure, the LMFB features of the mixed speech are directly fed to the hybrid DNN, which internally extracts speech of the target speaker using SD-DNN-SS and then recognizes its speech content accordingly. In the next section, the highlighted modules in Figure 1, namely the design of the SD-MC training set and the proposed joint training procedure, will be elaborated.

3. Training of Speaker-Dependent DNNs

3.1. Design of a SD-MC Training Set

In the conventional SI ASR system, the multi-condition training strategy (e.g., [21]) is widely used to improve the robustness in noisy environments. But for multi-talker scenarios, this concept cannot be directly applicable because it is difficult to differentiate the competing speakers. So in the IBM superhuman system, the clean-condition trained GMM-HMMs were adopted with two streams of each speaker from the separation module for subsequent joint decoding. Only in a recent work [20], a DNN architecture to simultaneously model two speakers at the output layer could accommodate the multi-condition training strategy for SI recognition system. However, its flexibility to more mixed speakers and runtime latency will still need to be addressed in real applications. In our proposed SD recognition framework, as the ambiguity between speakers has been reduced by focusing on the target speaker. The needed SD-MC training set can be inherently designed with the following procedure: (i) In the time domain, the waveform of each target speaker utterance is added with a time-synchronized segment of different interfering speakers normalized by a specified TMR to form a mixture utterance; (ii) By randomizing both the interfering segments and the TMR levels, a large-scale SD-MC data set can be synthesized even if only a very limited target speaker data set is provided, e.g., about 15-minute training data for each speaker for the SSC task.

For training of SD-DNN-AM with the synthesized SD-MC training set, the labels of the mixture utterances are corresponding to those of the underlying target speaker utterance. In this way, the HMMs of the speech units are guided to learn the phonetic information of the target speaker while the speech segments of other interfering speakers are forced to be aligned to the "non-speech" units, like the filler segments in keyword spotting [22, 23]. With this SD recognizer, it can perform a "selective" recognition of speech segments corresponding to the target speaker and ignore the segments of other competing speakers.

3.2. Training of Speech Separation DNNs

Although the SD-DNN-AM built with the SD-MC training data can achieve a quite competitive recognition performance, the interferences from other speakers as the irrelevant variabilities are not explicitly addressed. Motivated by the pre-processing approach to extract speech of the target speaker [1, 24], here we adopt the DNN as a regression model to directly map the features from the mixed speakers to the target. It can be considered as an irrelevant variability normalization step [25, 26] for the SD recognizer. As shown in Figure 2, both the input and output layers consist of multiple frames as the acoustic context. And the estimated target LMFB features can be used to retrain the SD-DNN-AM models.

For training SD-DNN-SS, the stereo set consisting of the same SD-MC mixed data described in Section 3.1 with the underlying target speaker data is adopted. We follow the proposed approach in [24]. An unsupervised pre-training step via the re-

stricted Boltzmann machine (RBM) [27] is first conducted in a layer-by-layer manner. Then with the pre-trained parameters, supervised fine-tuning is performed with a minimum mean squared error (MMSE) criterion between the DNN output and the reference LMFB features of the target speaker:

$$E = \frac{1}{N}\sum_{n=1}^{N} \|\hat{\boldsymbol{x}}_{n-\tau}^{n+\tau}(\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}, \boldsymbol{b}) - \boldsymbol{x}_{n-\tau}^{n+\tau}\|_2^2 + \kappa\|\boldsymbol{W}\|_2^2 \quad (1)$$

where $\hat{\boldsymbol{x}}_{n-\tau}^{n+\tau}$ and $\boldsymbol{x}_{n-\tau}^{n+\tau}$ are the n^{th} $D(2\tau+1)$-dimensional vectors of estimated and reference LMFB features of the target speaker, respectively. $\boldsymbol{y}_{n-\tau}^{n+\tau}$ is a $D(2\tau+1)$-dimensional vector of input mixed speech features with neighbouring left and right τ frames as the acoustic context. $\boldsymbol{W}$ and $\boldsymbol{b}$ denote all the weight and bias parameters. κ is the regularization weighting coefficient to avoid over-fitting. The objective function is optimized using back-propagation with a stochastic gradient descent method in mini-batch mode of N sample frames.

3.3. Acoustic Modeling and Joint Training

To build the acoustic model SD-DNN-AM, we follow the recipe in [14, 15, 28]. First, GMM-HMMs trained on clean utterances of the target speaker are used to generate the frame-level senone labels for the SD-MC data set. Then the layer-by-layer generative pre-training [15] followed by discriminative pre-training [28] is conducted. Finally, the cross-entropy (CE) criterion is adopted to update all the parameters.

So far, the SD-DNN-SS and SD-DNN-AM are separately learned using different objective functions. However, the learning objective of SD-DNN-AM is closer to the recognition performance. Meanwhile, the output layer of SD-DNN-SS can be completely overlapped with the input layer of SD-DNN-AM. It is straightforward to concatenate two DNNs to form a hybrid DNN (SD-DNN-JT). And a joint training procedure as illustrated in Figure 3 can be described as follows.

Step 1: Train a SD-DNN-SS to eliminate the interferences of other speakers. Meanwhile, the speech distortions of the target speakers might be also generated.

Step 2: Train a SD-DNN-AM with the SD-MC training set as an initial model. Then fine-tune all the parameters with the SD-DNN-SS generated features.

Step 3: Concatenate SD-DNN-SS and SD-DNN-AM as one SD-DNN-JT and fine-tune all the parameters of SD-DNN-JT via the CE criterion. And the speech distortions in **Step 1** might be alleviated via this joint training step.

4. Experiments and Result Analysis

Our experiments were conducted on the SSC corpus [29]. The challenge task was to recognize the keywords from simple *target* sentences when presented with a simultaneous *masker* sentence with a very similar structure [2]. All the training and test materials were drawn from the GRID corpus [30]. There were 34 speakers for both the training and test sets, including 18 males and 16 females. For the training set, 500 clean utterances were randomly selected from the GRID corpus for each speaker. The test set of the SSC corpus consisted of two-speaker mixtures at a range of TMRs from -9dB to 6dB with an increment of 3dB. The fixed grammar contains six parts: command, color, preposition, letter (with W excluded), number, and adverb. During the test phase, the speaker who uttered the color

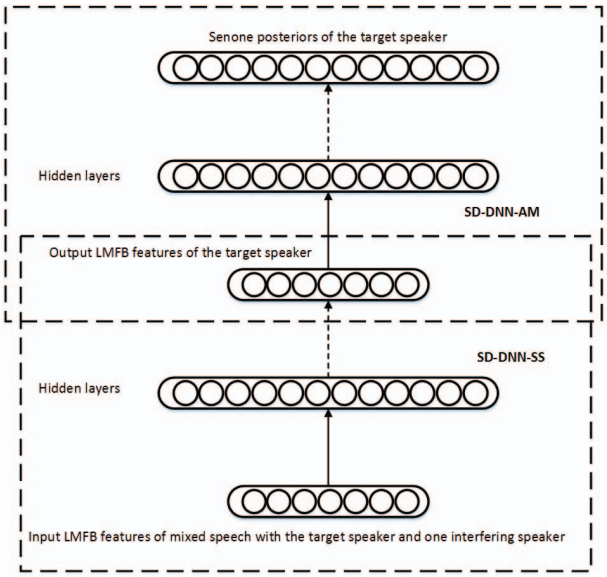

Figure 3: Illustration of the joint training procedure.

"white" was treated as the target speaker. The evaluation metric was the WER on letters and numbers spoken by the target speaker. Note that the recognition performances were evaluated on the test mixture utterances, including combinations of the same gender and different genders.

As for front-end and back-end processing, we follow most of the configurations in [20]. First, 64-dimensional LMFB features with a context window of 9 frames were adopted to train both the SD-DNN-SS and SD-DNN-AM components. The architecture of SD-DNN-SS was 576-2048-2048-2048-576, which denote that the size was 576(64*9, τ=4) at the input layer, 2048 for the 3 sigmoidal hidden layers, and 576 for the output layer. Meanwhile, the SD-DNN-AM had 7 sigmoidal hidden layers with 2048 hidden units in each layer and the final soft-max output layer with 534 units corresponding to the tied states of HMM. The mini-batch size was set to 256. Other parameter settings can be referred to [24, 31, 32].

4.1. Experiments under Clean-condition Training

In the first set of experiments, both the performances of SI and SD DNN-HMM systems on the test set of all 34 target speakers under clean-condition training are compared in Table 1 as the baselines. For the SI system, one set of DNN acoustic model was trained using all 17,000 clean utterances from 34 target speakers. And for the SD system, 34 sets of DNNs were separately trained using 500 clean utterances from each target speaker. Obviously, it was a mismatch testing scenario under clean-condition training. Although the SD system slightly outperformed the SI system, both systems yielded very poor performance, especially under low TMRs, which implied the necessity of multi-condition training.

4.2. Experiments under Multi-condition Training

In the following, 6 target speakers, including 3 males (IDs: $\{9, 15, 32\}$) and 3 females (IDs: $\{11, 23, 24\}$), were randomly selected for training and testing, because training both SD-DNN-SS and SD-DNN-AM with the SD-MC training set (typically more than 100-hour speech data) was time-consuming.

Table 1: WER comparison of SI and SD DNN-HMM systems under clean-condition training on the test set of all 34 target speakers with different TMRs.

System	6dB	3dB	0dB	-3dB	-6dB	-9dB
SI	32.8	47.1	63.3	76.9	84.2	90.9
SD	31.5	45.6	59.1	72.8	82.3	89.8

Table 2: WER comparison of SD DNN-HMM systems under clean-condition (Clean) and multi-condition (Multi) training on the test set of 6 selected target speakers with different TMRs.

System	6dB	3dB	0dB	-3dB	-6dB	-9dB	Avg.
Clean	32.3	47.2	61.9	78.3	85.2	92.3	66.2
Multi	19.7	23.9	25.4	28.2	31.7	39.4	28.1

We intend to complete experiments with all 34 speakers later.

Table 2 lists a WER comparison of the SD DNN-HMM systems under clean-condition and multi-condition training on the test set of the 6 selected target speakers with different TMR levels. For clean-condition training, 500 clean utterances of each target speaker were used. Then each clean utterance was corrupted with speech segments randomly selected from utterances of other 33 interfering speakers normalized by a specified TMR level. To cover the 6 TMR levels, ranging from -9 dB to 6 dB with an increment of 3 dB, in the test set, 3000 (500×6) mixture utterances in total were adopted in multi-condition training for each target speaker. First, the average WERs of the 6 target speakers in different TMRs under clean-condition training were similar to those of the SD system in Table 1, which show the 6 randomly selected speakers have a good representation of the set of 34 speakers, i.e., the results in the next three subsections would represent the typical performances for the entire 34 speakers. Second, multi-condition training significantly reduced the average WER from 66.2% in clean-condition training to 28.1%, yielding a relative WER reduction of 57.6%.

As described in Section 3.1, the design of the SD-MC training set can be scalable by using a huge amount of synthesized mixture data. Table 3 shows a WER comparison of SD DNN-HMM systems on the test set of the 6 selected target speakers under multi-condition training with different amounts of training sets. Three multi-condition trained SD systems, S1, S2, and S3, using different amounts of training data, respectively, were compared. S1 was exactly the same as the Multi system in Table 2. S2 was a modified version of S1, where each clean utterance of the target speaker was repeatedly 34 times corresponding to all 34 speakers giving a total of 102000 ($500\times34\times6$) training utterances for training S2. In obtaining S3 we adopted a different TMR setting from S2, namely ranging from -10 dB to 10 dB with an increment of 1 dB, generating a set of 357000 ($500\times34\times21$) training utterances approximately equal to about 150 hours of speech data. To our surprise, WERs for all TMRs were significantly reduced with the increase of training data amounts in terms of the resolutions for interfering speakers (from S1 to S2) and the TMR levels (from S2 to S3). the S3 system achieved an average WER of 3.8%, representing a relative WER reduction of 86.5% and most likely the best published results so far in literature, from S1 with an WER of 28.1%.

4.3. Experiments with Jointly Trained DNN Models

Finally, on top of the high-performance S3 system, we examine the effectiveness of our proposed jointly trained SD-DNN-JT system as shown in Table 4. In most TMR levels, significant performance gains could be observed from the SD-DNN-JT system with an average WER of 3.3%, or a relative WER reduction of 13.2% from the multi-condition trained SD-DNN-AM system. One more interesting observation was that the WERs of the SD-DNN-JT system among the TMR range from -6 dB to 3 dB were exactly corresponding to the WERs of the SD-DNN-AM system from -3 dB to 6 dB, with an increment of 3 dB in TMR, which indicated that the SD-DNN-JT could play the role of improving the TMR of the input mixture utterances via the SD-DNN-SS structure to reduce the impact of the interferences. In comparison to a WER of 10.9% obtained with the proposed pre-processing DNN approach in [1], a relative WER reduction of 69.7% could be observed. Even the worst recognition performance of SD-DNN-JT at -9 dB (a WER of 6.3%) was much better than the best performance of the pre-processing DNN approach at 6 dB (a WER of 7%).

Table 3: WER comparisons of SD DNN-HMM systems on the test set of 6 selected target speakers under multi-condition training with different amounts of training data (3000, 102000, and 357000 training utterances for S1, S2 and S3, respectively).

System	6dB	3dB	0dB	-3dB	-6dB	-9dB	Avg.
S1	19.7	23.9	25.4	28.2	31.7	39.4	28.1
S2	6.3	7.1	9.1	9.8	10.6	11.2	9.1
S3	2.1	2.8	3.5	3.5	4.3	6.3	3.8

Table 4: WER comparison of the multi-condition trained SD-DNN-AM system (Multi) and the jointly trained SD-DNN-JT system (Joint) on the test set of 6 selected target speakers.

System	6dB	3dB	0dB	-3dB	-6dB	-9dB	Avg.
Multi	2.1	2.8	3.5	3.5	4.3	6.3	3.8
Joint	2.1	2.1	2.8	3.5	3.5	5.6	3.3
[1]	7	8.5	9.2	11.3	12.7	16.9	10.9

5. Conclusion and Future Work

In this paper, we have proposed a novel speaker-dependent approach to jointly performing speech separation and acoustic modeling in one hybrid DNN architecture for single-channel automatic speech recognition of mixture speech in a multi-talker scenario. Coupling with the multi-condition training strategy, very promising speech recognition results on the SSC task were achieved. As for future work, we will further verify the effectiveness of the proposed framework on large vocabulary speech recognition. Moreover, we will investigate more adverse environments, including speaker interferences and background and convolutional noises. Finally the feasibility of designing a SD recognizer on portable devices will also be explored as one customization example in the mobile internet era.

6. Acknowledgment

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] Y.-H. Tu, J. Du, L.-R. Dai, and C.-H. Lee, "Speech Separation based on signal-noise-dependent deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2015.

[2] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech and Language*, vol. 24, no. 1, pp. 1-15, 2010.

[3] T. T. Kristjansson, J. R. Hershey, P. A. Olsen, S. J. Rennie, and R. A. Gopinath, "Super-human multi-talker speech recognition: the IBM 2006 speech separation challenge system," in *INTERSPEECH*, 2006.

[4] T. Virtanen, "Speech recognition using factorial hidden markov models for separation in the feature space," in *INTERSPEECH*, 2006.

[5] R. J. Weiss and D. P. W. Ellis, "Monaural speech separation using source-adapted models," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007, pp. 114-117.

[6] Z. Ghahramani and M. I. Jordan, "Factorial hidden markov models," *Mach. Learn.*, vol. 29, no. 2-3, pp. 245-273, 1997.

[7] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *INTERSPEECH*, 2006.

[8] M. R. Every and P. J. B. Jackson, "Enhancement of harmonic content of speech based on a dynamic programming pitch tracking algorithm.," in *INTERSPEECH*, 2006.

[9] J. Barker, N. Ma, A. Coy, and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Computer Speech and Language*, vol. 24, no. 1, pp. 94-111, 2010.

[10] J. Ming, T. J. Hazen, and J. R. Glass, "Combining missing-feature theory, speech enhancement and speaker-dependent/-independent modeling for speech separation," in *INTERSPEECH*, 2006.

[11] Y. Shao, S. Srinivasan, Z.-Z. Jin, and D.-L. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Computer Speech and Language*, vol. 24, no. 1, pp. 77-93, 2010.

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

[13] G. 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.

[14] G. E Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30-42, 2012.

[15] A. Mohamed, G. E Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14-22, 2012.

[16] G. Hinton, L. Deng, D. Yu, G. E Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N Sainath, et al. "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2012.

[17] J. Du, Y.-H. Tu, Y. Xu, L.-R. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Proc. ICSP*, 2014, pp. 473-477.

[18] P.-S. Huang, M. Kim, M. H. Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *Audio, Speech and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 12, pp. 2136-2147, 2015.

[19] M. Zöhrer, R. Peharz, and F. Pernkopf, "Representation learning for single-channel source separation and bandwidth extension," *Audio, Speech and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 12, pp. 2398-2409, 2015.

[20] C. Weng, D. Yu, M. L. Seltzer, J. Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *Audio, Speech and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 10, pp. 1670-1679, 2015.

[21] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks - a study on speech recognition tasks," in *Proc. CoRR*, 2013, vol. abs/1301.3605.

[22] Y.-D. Zhang, J. R Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *ASRU*, 2009.

[23] J. G. Wilpon,C. H. Lee,L. R. Rabiner, "Application of hidden Markov models for recognition of a limited set of words in unconstrained speech," in *Proc. ICASSP*, 1989.

[24] Y.-H. Tu, J. Du, Y. Xu, L.-R. Dai, and C.-H. Lee, "Deep neural network based speech separation for robust speech recognition," in *Proc. ICSP*, 2014, pp. 532-536.

[25] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," in *Computer Speech and Language*, vol. 12, no. 2, pp. 75-98, 1998.

[26] Y. Hu, Q. Huo, "Irrelevant variability normalization based HMM training using VTS approximation of an explicit model of environmental distortions," in *INTERSPEECH*, 2007.

[27] Y. Bengio, "Learning deep architectures for AI," *Foundat. and Trends Mach. Learn.*, vol. 2, no. 1, pp. 1-127, 2009.

[28] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU*, 2011.

[29] M. Cooke and T.-W. Lee, Speech Separation Challenge, 2006. [http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm]

[30] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421-2424, 2006.

[31] 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.

[32] G. Hinton, "A practical guide to training restricted Boltzmann machines," UTML TR 2010-003, University of Toronto, 2010.