SINGLE-CHANNEL MIXED SPEECH RECOGNITION USING DEEP NEURAL NETWORKS

Chao Weng[1*] *, Dong Yu*[2]*, Michael L. Seltzer*[2]*, Jasha Droppo*[2]

[1] Georgia Institute of Technology, Atlanta, GA, USA
[2] Microsoft Research, One Microsoft Way, Redmond, WA, USA
[1]chao.weng@ece.gatech.edu, [2]{dongyu, mseltzer, jdroppo}@microsoft.com

ABSTRACT

In this work, we study the problem of single-channel mixed speech recognition using deep neural networks (DNNs). Using a multi-style training strategy on artificially mixed speech data, we investigate several different training setups that enable the DNN to generalize to corresponding similar patterns in the test data. We also introduce a WFST-based two-talker decoder to work with the trained DNNs. Experiments on the 2006 speech separation and recognition challenge task demonstrate that the proposed DNN-based system has remarkable noise robustness to the interference of a competing speaker. The best setup of our proposed systems achieves an overall WER of 19.7% which improves upon the results obtained by the state-of-the-art IBM superhuman system by 1.9% absolute, with fewer assumptions and lower computational complexity.

Index Terms— DNN, multi-talker ASR, WFST

1. INTRODUCTION

While significant progress has been made in improving the noise robustness of speech recognition systems, recognizing speech in the presence of a competing talker remains one of the most challenging unsolved problems in the field. To study the specific case of single-microphone speech recognition in the presence of competing talker, a monaural speech separation and recognition challenge [1] was issued in 2006. It enabled researchers to apply a variety of techniques on the same task and make comparisons between them. Several types of solutions were proposed. Model based approaches [2, 3, 4] use factorial GMM-HMM [5] to model the interaction between target and competing speech signals and their temporal dynamics, then the joint inference or decoding determined the two most likely speech signals or spoken sentences given the observed speech mixture. In computational auditory scene analysis (CASA) and missing feature approaches [6, 7, 8], certain segmentation rules operate on low-level features to estimate a time-frequency mask that isolates the signal components that belong to the each speaker. This mask is used either to reconstruct the signal or directly inform the decoding process. Some other approaches including [9] and [10] utilize the non-negative matrix factorization (NMF) for the separation and pitch-based enhancement. Among all the submissions to the challenge, the IBM superhuman system [2] performed the best and even exceeded what human listeners could do on the challenge task (see Table 2). Their system consists of three main components: a speaker recognizer, a separation system, and a speech recognizer. The separation system requires as input the speaker identities and signal gains that are output from the speaker recognition system. In practice, it is usually necessary to enumerate several of the most probable speaker combinations and run the whole system multiple times. This may be impractical when the number of speakers is large. The separation system uses factorial GMM-HMM generative models with 256 Gaussians to model the acoustic space for each speaker. While this was sufficient for the small vocabulary in the challenge task, it is a very primitive model for a large vocabulary task. However, with a larger number of Gaussians, performing inference on the factorial GMM-HMM becomes computationally impractical. Moreover, the system assumes the availability of speaker-dependent training data and a closed set of speakers between training and test.

Recently, acoustic models based on deep neural networks (DNNs) [11] have shown great success on large vocabulary tasks [12]. However, few, if any, previous work has explored how DNNs could be used in the multi-talker speech recognition scenario. High-resolution features are typically favored by speech separation system, while the fact that a conventional GMM-HMM ASR system is incapable of compactly modeling the high-resolution features usually forces researchers to perform speech separation and recognition separately. However, DNN-based systems have been shown to work significantly better on spectral-domain features than cepstral-domain features [13], and have shown outstanding robustness to speaker variation and environment distortions [14, 15]. In this work, we aim to build a unified DNN-based system, which can simultaneously separate and recognize two-talker speech in a manner that is more likely to scale up to a larger task. We propose several methods for co-channel speech recognition that combine multi-style training with different objective functions defined specifically for the multi-tasker task. The phonetic probabilities output by the DNNs will then be decoded by a WFST-based decoder modified to operate on multi-talker speech. Experiments on the 2006 speech separation and recognition challenge data demonstrate the proposed DNN based system has remarkable noise robustness to the interference of competing talker. The best setup of our systems achieves 19.7% overall WER, which is 1.9% absolute improvement over the state-of-the-art IBM system with less complexity and fewer assumptions.

The remainder of this paper is organized as follows. In Section 2, we describe our multi-style DNN training and the different multi-talker objective functions used to train the networks. The WFST-based joint decoder is introduced in Section 3. We report experimental results in Section 4 and summarize our work in Section 5.

2. DNN MULTI-STYLE TRAINING WITH MIXED SPEECH

Although a DNN-based acoustic model has proven to be more robust to environmental perturbations, it was also shown in [14] that the robustness holds well only for the input features with modest distortions beyond what was observed in the training data. When there exist severe distortions between training and test samples, it

*The work was performed while the first author was an intern at Microsoft Research, Redmond, WA, USA

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

System/Method	IBM superhuman	Human	Next best
WER	21.6%	22.3%	34.2%

Table 1. Overall keywords WERs of three systems/methods on the 2006 challenge task. IBM superhuman: Hershey et al. [2] ; Human: human listeners; Next best: the system by Viranen [3].

is essential for DNNs to see examples of representative variations during training in order to generalize to the severely corrupted test samples. Since that we are dealing with a challenging task where the speech signal from the target speaker is mixed with a competing one, a DNN-based model will generalize poorly if it is trained only on single-speaker speech, as will be shown in Section 4. One way to circumvent this issue is using a multi-style training strategy [16] in which training data is synthesized to be representative of the speech expected to be observed at test time. In our case, this means corrupting the clean single-talker speech database with samples of competing speech from other talkers at various levels and then training the DNNs with these created multi-condition waveforms. In the next sections, we describe how this multi-condition data can be used to create networks that can separate multi-talker speech.

2.1. High and Low Energy Signal Models

In each mixed-speech utterance, we assume that one signal is the target speech and one is the interference. The labeling is somewhat arbitrary as the system will decode both signals. The first approach assumes that one signal has higher average energy than the other. Under this assumption, we can identify the target speech as either the higher energy signal (positive SNR) or the lower energy signal (negative SNR). Thus in our first system, two DNNs are used: given a mixed-speech input, one network is trained to recognize the higher energy speech signal while the other one is trained to recognize the low energy speech signal. Suppose we are given a clean training dataset $\mathcal{X}$, we first perform energy normalization so that each speech utterance in the data set has the same power level. To simulate the acoustical environments where the target speech signal has higher average energy or lower average energy, we randomly choose another signal from the training set, scale its amplitude appropriately and mix it with the target speech. Denote by $\mathcal{X}_H$,$\mathcal{X}_L$ the two multi-condition datasets created as described. For the high energy target speaker, we train the DNN models with the loss function,

$$\mathcal{L}_{\mathrm{CE}}(\theta) = -\sum_{x_t \in \mathcal{X}_H} \log p(s_j^{\mathrm{H}}|x_t;\theta), \tag{1}$$

where s_j^{H} is the reference senone label at t^{th} frame. Note that the reference senone labels comes from the alignments on the uncorrupted data. This was critical to obtaining good performance in our experiments. Similarly, the DNN models for the low energy target speaker can be trained on the dataset $\mathcal{X}_L$. With the two created dataset $\mathcal{X}_L$ and $\mathcal{X}_H$, we can also train the DNNs as denoisers using the minimum square error (MSE) loss function,

$$\mathcal{L}_{\mathrm{MSE}}(\theta) = \sum_{x_t \in \mathcal{X}_H} |\hat{y}(x_t;\theta) - y_t|^2, \quad y_t \in \mathcal{X}, \tag{2}$$

where $y_t \in \mathcal{X}$ is the corresponding clean speech features and $\hat{y}(x_t;\theta)$ is the estimation of the uncorrupted inputs using the deep denoiser. Similarly, the denoiser for the low energy target speaker can be trained on the dataset $\mathcal{X}_L$.

2.2. High and Low Pitch Signal Models

One potential issue with the above training strategy based on high and low energy speech signals is that the trained models may perform poorly when mixed signals have similar average energy levels, *i.e.* near 0dB SNR. The reason is that the problem is ill-defined in this region since one cannot reliably label one signal as the higher or lower energy signal. Since it is far less likely that the two speakers will speak with the same pitch, we propose another approach in which DNNs are trained to recognize the speech with the higher or lower pitch. In this case, we only need to create a single training set $\mathcal{X}_P$ from original clean dataset $\mathcal{X}$ by randomly choosing an interfering speech signal and mixing it with the target speech signal. The training also requires a pitch estimate for both the target and interfering speech signals which will be used to select appropriate labels for DNN training. The loss function for training the DNN for the high pitch speech signals is thus,

$$\mathcal{L}_{\mathrm{CE}}(\theta) = -\sum_{x_t \in \mathcal{X}_P} \log p(s_j^{\mathrm{HP}}|x_t;\theta), \tag{3}$$

where s_j^{HP} is the reference senone label obtained from the alignments on the speech signal with the higher average pitch. Similarly, a DNN for the lower pitch speech signals can be trained with the senone alignments of the speech signal with the lower average pitch.

2.3. Instantaneous High and Low Energy Signal Models

Finally, we can also train the DNNs based on the instantaneous energy in each frame rather than the average energy of the utterance. Even an utterance with an average energy of 0 dB will have non-zero instantaneous SNR values in each frame, this means there is no ambiguity in the labeling. We only need to create one training set $\mathcal{X}_I$ by mixing speech signals and computing the instantaneous frame energies in the target and interfering signal. The loss function for the instantaneous high energy signal is given by,

$$\mathcal{L}_{\mathrm{CE}}(\theta) = -\sum_{x_t \in \mathcal{X}_I} \log p(s_j^{\mathrm{IH}}|x_t;\theta), \tag{4}$$

where s_j^{IH} corresponds to the senone label from the signal source which contains higher energy at frame t. In this scenario, since we are using a frame-based energy rather than an utterance-based energy as the criterion for separation, there is uncertainty as to which output corresponds to the target or interferer from frame to frame. For example, the target speaker can have higher energy in one frame and lower energy in the next frame. We will address this in the decoder described in the next section.

3. JOINT DECODING WITH DNN MODELS

For the DNNs based on instantaneous energy, we need to determine which of the two DNN outputs belongs to which speaker at each frame. To do so, we introduce a joint decoder that can take the posterior probability estimates from the instantaneous high-energy and low-energy DNNs to jointly find best two state sequences, one for each speaker. The standard recipe for creating the decoding graph in the WFST framework can be written as,

$$\mathrm{HCLG} = \min(\det(H \circ C \circ L \circ G)), \tag{5}$$

where H, C, L and G represent the HMM structure, phonetic context-dependency, lexicon and grammar respectively, and $\circ$ is WFST composition. The input labels of the HCLG are the identifiers of context-dependent HMM states (senone labels), and the output labels represent words. Denote by θ^{H} and θ^{L} instantaneous high and low energy signal DNN models trained as described in Section 2.3. The task of the joint decoder is to find best two state

sequence in the 2-D joint state space such that the sum of each state-sequence log-likelihood is maximized,

$$(\mathbf{s}^{1*}, \mathbf{s}^{2*}) = \underset{(\mathbf{s}^1, \mathbf{s}^2) \in \{\mathbf{s}^1 \times \mathbf{s}^2\}}{\operatorname{argmax}} p(x_{1:T}|\mathbf{s}^1; \theta^{\mathrm{H}}, \theta^{\mathrm{L}}) \cdot p(x_{1:T}|\mathbf{s}^2; \theta^{\mathrm{H}}, \theta^{\mathrm{L}}). \tag{6}$$

The key part of the proposed decoding algorithm is joint token passing on the two HCLG decoding graphs. The main difference in token passing between joint decoding and conventional decoding is that now each token is associated with two states rather than one in the decoding graph. Figure 1 shows a toy example to illustrate the joint token passing process: suppose the token for the first speaker is at state 1, and the token associated with the second speaker is at state 2. For the outgoing arcs with non-ϵ input labels (those arcs that consume acoustic frames), the expanded arcs will be the Cartesian product between the two outgoing arc sets. The graph cost of each expanded arc will be the semiring multiplication of the two. The acoustic cost of each expanded arc is computed using the senone hypotheses from the two DNNs for the instantaneous high and low energy. Because we need to consider both cases where either one of the two sources has the higher energy, the acoustic cost is given by the combination with higher likelihood,

$$\begin{aligned} \mathcal{C} = \max\{&p(x_t|s^1; \theta^{\mathrm{H}}) \cdot p(x_t|s^2; \theta^{\mathrm{L}}), \\ &p(x_t|s^1; \theta^{\mathrm{L}}) \cdot p(x_t|s^2; \theta^{\mathrm{H}})\}. \end{aligned} \tag{7}$$

With the equation above, we can also tell which speaker has higher

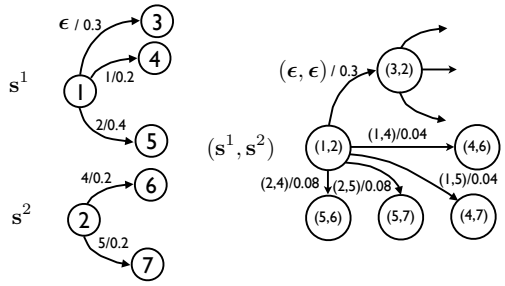

Fig. 1. A toy example illustrating the joint token passing on the two WFST graph: $\mathbf{s}^1$, $\mathbf{s}^2$ denote state space corresponds to one of two speakers; $(\mathbf{s}^1, \mathbf{s}^2)$ represent the joint state space.

energy in the corresponding signal at certain frame t along this search path. For the arcs with ϵ input labels, the expansion process is bit tricky. As the ϵ arcs are not consuming acoustic frames, to guarantee the synchronization of the tokens on two decoding graphs, a new joint state for current frame has to be created (see the state $(3, 2)$ in the Fig.1).

One potential issue of our joint decoder is that we allow free energy switching frame by frame while decoding the whole utterance. Yet, we know that in practice, the energy switching should not typically occur too frequently. This issue can be overcome by introduce a constant penalty in certain searching path when the louder signal has changed from last frame. Alternatively, we can estimate the probability that a certain frame is the energy switching point and let the value of the penalty adaptively changed with it. Since we created the training set by mixing the speech signals, the energy of each original speech frame is available. We can use it to train a DNN to predict whether the energy switch point occurs at certain frame. If we let θ^S represent the models we trained to detect the energy switching point, the adaptive penalty on energy switching is given by,

$$\mathcal{P} = -\alpha \cdot \log p(y_t|x_t; \theta^S). \tag{8}$$

Systems	Conditions						
	Clean	6dB	3dB	0dB	-3dB	-6dB	-9dB
GMM	4.0	38.5	54.7	70.5	82.3	89.3	94.2
DNN	0.7	32.5	48.8	66.3	78.4	86.3	91.8

Table 2. WERs (%) of baseline GMM-HMM and DNN-HMM systems

4. EXPERIMENTS

4.1. The Challenge Task and Scoring Procedure

The main task of 2006 monaural speech separation and recognition challenge is to recognize the keywords (numbers and letters) from the speech of a target speaker in the presence of another competing speaker using a single microphone. The speech data of the challenge task is drawn from GRID corpus [17]. The training set contains 17,000 clean speech utterances from 34 difference speakers (500 utterances for each speaker). The evaluation set includes 4,200 mixed speech utterances in 7 conditions, clean, 6dB, 3dB, 0dB, -3dB, -6dB, -9dB target-to-mask ratio (TMR) and the development set contains 1,800 mixed speech utterances in 6 conditions (no clean condition). The fixed grammar contains six parts: command, color, preposition, letter (with W excluded), number, and adverb, *e.g.* "place white at L 3 now". During the test phase, the speaker who utters the color 'white' is treated as the target speaker. The evaluation metric is the WER on letters and numbers spoken by the target speaker. Note that the WER on all words will be much lower, and unless otherwise specified, all reported WERs in the following experiments are the ones evaluated only on letters and numbers.

4.2. Baseline System

The baseline system is built using a DNN trained on the original training set consisting of 17,000 clean speech utterances. We first train a GMM-HMM system using 39-dimension MFCCs features with 271 distinct senones. Then we use 64 dimension log mel-filterbank as features and context window of 9 frames to train the DNN. The DNN has 7 hidden layers with 1024 hidden units at each layer and the 271-dimensional softmax output layer, corresponding to the senones of the GMM-HMM system. The following training scheme will be used through all the DNN experiments: the parameter initialization is done using layer by layer using generative pre-training [18] following by discriminative pre-training [19]. Then the network is discriminatively trained using backpropagation. The mini-batch size is set to 256 and the initial learning rate is set to 0.008. After each training epoch, we validate the frame accuracy on the development set, if the improvement is less than 0.5%, we shrink the learning rate by the factor of 0.5. The training process is stopped after the frame accuracy improvement is less than 0.1%. The WERs of the baseline GMM-HMM and DNN-HMM system are shown in Table 2. As can be seen, the DNN-HMM system trained only on clean data performs poorly in all SNR conditions except the clean condition, confirming the necessity of DNN multi-style training.

4.3. Multi-style Trained DNN Systems

To investigate the use of multi-style training for the high and low energy signal models, we generated two mixed-speech training datasets. The high energy training set, which we refer to as Set I, was created as follows: for each clean utterance, we randomly choose three other utterances and mixed them with the target clean utterance under 4 conditions, clean, 6dB, 3dB, 0dB. (17,000 $\times$ 12); II. The low energy training set, referred to as Set II, was created in a similar manner but the mixing was done under 5 conditions, clean, and TMRs of 0dB, -3dB, -6dB, -9dB. (17,000 $\times$ 15). Then we use

Systems	Conditions						AVG
	6dB	3dB	0dB	-3dB	-6dB	-9dB	
DNN	32.5	48.8	66.3	78.4	86.3	91.8	67.4
DNN I	**4.5**	**16.8**	56.8	-	-	-	-
DNN II	-	-	52.6	33.6	**18.4**	**17.4**	-
IBM [2]	15.4	17.8	**22.7**	**20.8**	22.1	30.9	**21.6**
DNN I+II	**4.5**	16.9	49.8	39.8	21.7	19.6	25.4

Table 3. WERs (%) of the DNN systems for high and low energy signals

Systems	Conditions		
	6dB	3dB	0dB
Denoiser I + DNN	16.8	32.2	65.9
Denoiser I + DNN (retrained)	6.3	17.3	**56.3**
DNN I	**4.5**	**16.8**	56.8

Table 4. WERs (%) of deep denoisers for high and low energy signals

these two training sets to train two DNN models, DNN I and II, for high and low energy signals respectively, and listed the results in Table 3. From the table, we can see the results are surprisingly good, especially in the cases where two mixing signals have large energy level difference, *i.e.* 6dB, -6dB, -9dB. By combining the results from DNN I and II systems using the rule that the target speaker always utters the color white, the combined DNN I+II system achieves 25.4% WER compared to 67.4% which obtained with the DNN trained only on clean data. Then we experimented with the multi-style trained deep denoiser. With the same training set I, we train a DNN as a front-end denoiser as described in Section 2.1. With trained deep denoiser, we try two different setups: the first one we directly feed denoised features to the DNN trained on the clean data; in the second setup, we retrained another DNN on the denoised data and conduct the experiments. We list the results of both setups in the Table 4. From the above experiments, there are two noteworthy points. First, the system with the DNN trained to predict senone labels seems slightly better than the one with a trained deep denoiser followed by another retrained DNN. This implies that DNN is capable learning robust representations automatically, there may be no need to extract hand-crafted features in the front-end. The combined system DNN I+II is still not good as the state-of-the-art IBM superhuman system. The main reason is that the system performs very poorly in the cases where two mixing signals have very close energy level, *i.e.* 0dB, -3dB. This coincides with our concerns discussed earlier. Specifically, the multi-style training strategy for the high and low energy signals has the potential issue of assigning conflicting labels during training.

For the high and low pitch signals models, we first estimate the pitch for each speaker from the clean training set. Then we combine the Train Set I and Train Set II to form Set III ($17{,}000 \times 24$) to train two DNNs for high and low pitch signals respectively. When training the DNNs for the high pitch signals, we assign the label from the alignments on clean speech utterances corresponding to the high pitch talker; When training the DNNs for the low pitch signals, we assign the label from the alignments corresponding to the low pitch talker. With the two trained DNN models, we do the decoding independently as before and combine the decoding results using the rules that the target speaker always utters the color white. We list the WERs in Table 5. As can be seen, the system with the high and low pitch signal models performs better than the one with the high and low energy models in the 0dB case, but worse in the other cases.

4.4. DNN System with Joint Decoder

Finally, we use training set III to train two DNN models for instantaneous high and low energy signals as described in Section 2.3. With these two trained models, we perform a joint decoding as described

Systems	Conditions						AVG
	6dB	3dB	0dB	-3dB	-6dB	-9dB	
DNN I+II	**4.5**	**16.9**	49.8	39.8	**21.7**	**19.6**	**25.4**
DNN III	14.5	22.1	**30.8**	41.9	52.8	59.6	36.9

Table 5. WERs (%) of the DNN systems for high and low pitch signals

Systems	Conditions						AVG
	6dB	3dB	0dB	-3dB	-6dB	-9dB	
DNN	32.5	48.8	66.3	78.4	86.3	91.8	67.4
IBM [2]	15.4	17.8	22.7	20.8	22.1	30.9	21.6
DNN I+II	**4.5**	16.9	49.8	39.8	**21.7**	**19.6**	25.4
Joint Decoder	18.3	19.8	**19.3**	21.3	23.2	27.4	21.5
Joint Decoder I	16.1	18.7	20.5	19.6	23.6	26.8	20.9
Joint Decoder II	16.5	17.1	19.9	18.8	22.5	25.3	20.0
Combined	16.0	**16.6**	19.7	**18.8**	23.0	24.1	**19.7**

Table 6. WERs (%) of the DNN systems with the joint decoders.

in Section 3. The results of this Joint Decoder approach are shown in Table 6. The last two systems correspond to the cases where we introduce the energy switching penalties. The Joint Decoder I is the system with the constant energy switching penalty and Joint Decoder II is the system with adaptive switching penalty. To get the value of the energy switching penalties as defined in (8), we trained a DNN to estimate an energy switching probability for each frame.

4.5. System Combination

From Table 6, we can see that the DNN I+II system performs well in the cases where two mixing speech signals have large energy level difference, *i.e.* 6dB, -6dB, -9dB, while the Joint Decoder II system performs well in the cases where two mixing signals have similar energy level. This motivates us to do the system combination according to the energy difference between the two signals. To get energy level difference between two mixing signals, we use the deep denoisers for the high and low energy signals. The mixed signal is input to the two deep denoisers and the two resultant output signals will be used to estimate the high and low energy signals. Using these separated signals, we can calculate their energy ratio to approximate the energy difference of two original signals. We first tune and obtain a optimal threshold for the energy ratio on the development set, and use it for the system combination, *i.e.* if the energy ratio of two separated signals from the denoisers is higher than the threshold, we use system DNN I+II to decode the test utterance, otherwise the system Joint Decoder II will be used. The results are listed in Table 6.

5. CONCLUSIONS

In this work, we investigate DNN-based systems for single-channel mixed speech recognition by using multi-style training strategy. We also introduce a WFST-based joint decoder to work with the trained DNNs. Experiments on the 2006 speech separation and recognition challenge data demonstrate that the proposed DNN based system has remarkable noise robustness to the interference of competing speaker. The best setup of our proposed systems achieves 19.7% overall WER which improves upon the results obtained by the IBM superhuman system by 1.9% absolute, with making fewer assumptions and lower computational complexity.

6. ACKNOWLEDGEMENTS

We would like to thank Geoffrey Zweig, Frank Seide for their valuable suggestions and Kun Han (OSU) for the valuable discussions.

7. REFERENCES

[1] Martin Cooke, John R. Hershey, and Steven J. Rennie, "Monaural speech separation and recognition challenge.," *Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

[2] Trausti T. Kristjansson, John R. Hershey, Peder A. Olsen, Steven J. Rennie, and Ramesh A. Gopinath, "Super-human multi-talker speech recognition: the ibm 2006 speech separation challenge system.," in *INTERSPEECH*. 2006, ISCA.

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

[4] 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.

[5] Zoubin Ghahramani and Michael I. Jordan, "Factorial hidden markov models," *Mach. Learn.*, vol. 29, no. 2-3, pp. 245–273, Nov. 1997.

[6] Jon Barker, Ning Ma, André Coy, and Martin Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 94–111, Jan. 2010.

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

[8] Yang Shao, Soundararajan Srinivasan, Zhaozhang Jin, and DeLiang 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.

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

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

[11] Geoffrey E. Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[12] G.E. Dahl, Dong Yu, Li 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, jan. 2012.

[13] Jinyu Li, Dong Yu, Jui-Ting Huang, and Yifan Gong, "Improving wideband speech recognition using mixed-bandwidth training data in cd-dnn-hmm.," in *SLT*. 2012, pp. 131–136, IEEE.

[14] Dong Yu, Michael L. Seltzer, Jinyu Li, Jui-Ting Huang, and Frank Seide, "Feature learning in deep neural networks - a study on speech recognition tasks," *CoRR*, vol. abs/1301.3605, 2013.

[15] M. L. Seltzer, D. Yu, and Y.-Q. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP2013*, 2013.

[16] R. Lippmann, E. Martin, and D.B. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. ICASSP1987*, 1987.

[17] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, November 2006.

[18] 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, jan. 2012.

[19] Frank Seide, Gang Li, Xie Chen, and Dong Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *ASRU*, 2011, pp. 24–29.