Deep Neural Networks for Single-Channel Multi-Talker Speech Recognition

Chao Weng, *Student Member, IEEE,* Dong Yu, *Senior Member, IEEE,* Michael L. Seltzer, *Senior Member, IEEE,* Jasha Droppo, *Senior Member, IEEE*

***Abstract*—We investigate techniques based on deep neural networks (DNNs) for attacking the single-channel multi-talker speech recognition problem. Our proposed approach contains five key ingredients: a multi-style training strategy on artificially mixed speech data, a separate DNN to estimate senone posterior probabilities of the louder and softer speakers at each frame, a WFST-based two-talker decoder to jointly estimate and correlate the speaker and speech, a speaker switching penalty estimated from the energy pattern change in the mixed-speech, and a confidence based system combination strategy. Experiments on the 2006 speech separation and recognition challenge task demonstrate that our proposed DNN-based system has remarkable noise robustness to the interference of a competing speaker. The best setup of our proposed systems achieves an average word error rate (WER) of 18.8% across different SNRs and outperforms the state-of-the-art IBM superhuman system by 2.8% absolute with fewer assumptions.**

***Index Terms*—DNN, noise robustness, multi-talker ASR, single-channel, WFST, joint decoding**

I. Introduction

WHILE significant progress has been made in improving the noise robustness of automatic speech recognition (ASR) 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 [2] was carried out 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 use factorial GMM-HMM [3] to model the interaction between the target and competing speech signals and their temporal dynamics [4], [5], [6]. A joint inference or decoding strategy is used to determine the two most likely speech signals or spoken sentences given the observed mixed speech. In computational auditory scene analysis (CASA) and missing feature approaches, certain segmentation rules operate on low-level features to estimate a time-frequency mask that isolates the signal components belonging to different speakers [7], [8], [9]. This mask is used either to reconstruct the signal or to inform the decoder directly. Some other approaches utilize the non-negative matrix factorization [10] or pitch-based enhancement [11] techniques. Among all the submissions to the challenge, the IBM superhuman system [4] performed the best and even exceeded what human listeners could do on the challenge task.

The IBM superhuman system consists of three main components: a speaker recognizer, a separation system, and a speech recognizer. The speaker recognizer estimates speaker identities and amplitudes for the separation system. For maximum performance, the entire system is run multiple passes for several different combinations of the most probable speaker identities. The separation system uses factorial GMM-HMM generative models with 256 Gaussians to model the acoustic space for each speaker. For a large vocabulary task, performing inference on the factorial GMM-HMM can be expensive. Although the techniques elaborated in [12] can be used to control the complexity of inference for larger tasks, the assumptions on the availability of speaker-dependent training data and a closed set of speakers between training and test could still be issues for the system being applied in a real large system.

Recently, acoustic models based on deep neural networks (DNNs) [13] have shown great successes on both LVCSR [14] and robust ASR tasks [15]. However, few, if any, previous work has explored how DNNs could be used in the multi-talker speech recognition scenario. For speech separation, DNNs have an advantage over conventional GMM-HMM ASR systems that are incapable of compactly modeling the high-resolution feaures typically favored by speech separation systems. This deficiency usually forces researchers to perform speech separation and recognition separately. However, DNN-based systems have been shown to perform as well or better on spectral-domain features than cepstral-domain features [16], and have shown outstanding robustness to speaker variation and environmental distortions [17].

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-speaker 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 that the

C. Weng is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. He conducted this work while working as an intern at Microsoft Research (e-mail: chao.weng@ece.gatech.edu)

Dong Yu, corresponding author, is with Microsoft Research, One Microsoft Way, Redmond, WA, USA (e-mail: dongyu@microsoft.com, phone: +1 425-707-9282)

Michael L. Seltzer and Jasha Droppo are with Microsoft Research, One Microsoft Way, Redmond, WA, USA (e-mail: {mseltzer, jdroppo}@microsoft.com)

Manuscript received ; revised ;

A short version with earlier results has been presented at ICASSP 2014 [1]

proposed DNN based system has remarkable noise robustness to the interference of a competing talker. Our best system achieves 18.8% overall word error rate (WER), which is 2.8% absolute better than that obtained with the state-of-the-art IBM system with fewer assumptions.

The remainder of this paper is organized as follows. We first review DNN based approaches to robust speech recognition in Section II. In Section III, we describe our multi-style DNN training strategy and the different multi-talker objective functions used to train the networks. The WFST-based joint decoder is introduced in Section IV. We report experimental results in Section V and summarize our work in Section VI.

II. DNN-based Approaches to Noise Robust ASR

DNNs can be exploited in either the feature or the model space to improve noise robustness in ASR systems. When used in the feature space as a front-end feature denoiser, DNNs or recurrent neural networks (RNNs) try to capture the mapping from the noisy speech to the clean speech [18], [19]. In these methods, DNNs are usually trained on clean and noisy stereo pairs of observations to minimize the mean squared error (MSE) between the estimated clean speech and the actual clean speech [17]. Alternatively, the mismatch between the noisy and clean speech can be treated as a mismatch between the training and testing conditions and feature-space DNN adaptation techniques such as the linear input network (LIN) [20] or the feature discriminative linear regression (fLDR) [21] can be applied to learn a linear feature transformation to compensate for the mismatch, similar to the linear discriminative analysis (LDA) [22] and feature-space maximum likelihood linear regression (fMLLR) [23] often used in the Gaussian mixture model (GMM) based ASR systems. Since DNNs and RNNs can extract more invariant features than the raw acoustic feature at the output and hidden layers [17], they can also be used to generate the so called Tandem [24], [25] feature if it is extracted from the output layer, or bottleneck feature [26], [27] if it is extracted from a hidden layer with smaller number of neurons, to improve the noise robustness. These features can be concatenated with the original MFCCs features and used in the conventional GMM-hidden Markov model (HMM) system [24]. The advantage of all these feature space approaches is the flexibility in selecting the back-end systems.

When used as a model space method, DNNs can generate the posterior probability of each HMM state, in place of GMMs, in the DNN-HMM hybrid setup. Unlike a conventional GMM-HMM LVCSR system which uses GMMs to generate the state emission log-likelihood of the observation feature vector x_t for certain tied state or senone s_t at frame t, a DNN-HMM hybrid system [13] uses pseudo log-likelihood as the state emissions,

$$\log p(x_t|s_t) \propto \log p(s_t|x_t) - \log p(s_t), \tag{1}$$

where $p(s_t|x_t)$ are the state posteriors output from DNNs and the state priors $\log p(s_t)$ can be estimated using the state alignments on the training speech data. The input feature vectors x_t to the first layer of DNNs usually use a context of l frames [13], *e.g.* $l = 9$ or $l = 11$. The most straightforward way of DNN-based approaches to noise robustness in the model space is multi-style training [28]: either collecting or artificially creating (e.g., by corrupting the clean database with noise samples of various levels and types) acoustic samples under different acoustical environments and training DNNs with all these data. Recently, multiple DNN model training methods, e.g, noise-aware training and dropout, have been introduced in [15] to lead more accurate senone prediction under various noise conditions for DNN-HMM hybrid systems. In [29], the recurrent architecture is introduced into the DNN-HMM hybrid system and the authors can achieve state-of-the-art performances on both the 2nd CHiME challenge (track 2) [30] and Aurora-4 tasks without front-end preprocessing, speaker adaptive training or multiple decoding passes. Recently long short-term memory (LSTM) recurrent architectures are also introduced in the hybrid system [31], [32] to further improve the robustness of the acoustic models.

III. DNN Multi-style Training with Co-channel Speech

Although DNN-based acoustic models have been proven to be more robust to environmental perturbations, it was also shown in [17] that the robustness holds well only for input features with modest distortions beyond what was observed in the training data. When there exist severe distortions between training and test samples, it is essential for DNNs to see examples of representative variations during training in order to generalize to the severely corrupted test samples. Since 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 V. As mentioned in Section II, one way to circumvent this issue is using a multi-style training strategy [28] in which training data are 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 synthesized multi-condition waveforms. In this section, we describe how this multi-condition data can be used to train networks that can separate multi-talker speech.

A. Speech Separation Based on Average Energy

In each mixed-speech utterance, the target speech is mixed with an interference speech. Since the decoder does not know which is the target and which is the interference beforehand it needs to decode both signals. To separate two speakers, some information is needed. Our first approach assumes that one signal has higher average energy than the other. Under this assumption, the target speech either has higher average energy (positive SNR) or lower average energy (negative SNR). We recognize two speech streams at the same time using two DNNs: 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, which we will refer to as the high and low energy models respectively. Suppose we are given a single-speaker speech training set $\mathcal{X}$. To synthesize the mixed-speech dataset we first perform energy normalization so that each speech utterance in the dataset 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. More specifically, for the multi-condition dataset used to train the high energy signal models, we need to decrease the amplitude of those speech waveforms mixed with the target speech to various levels while for the multi-condition dataset used to train the low energy signal models we need to increase the amplitude level accordingly. Denote the high-energy and low-energy datasets created as described by $\mathcal{X}_H$,$\mathcal{X}_L$, respectively. For the high energy target speaker, we train a DNN model with the loss function,

$$\mathcal{L}_{\mathrm{CE}}(\theta) = -\sum_{x_t \in \mathcal{X}_H} \log p(s_t^{\mathrm{H}} | x_t; \theta), \quad (2)$$

where s_t^{H} is the reference senone label at t^{th} frame from the high-energy speaker. Note that the reference senone labels come from the alignments on the uncorrupted data. This was critical to obtain good performance in our experiments. A DNN model for the low energy target speaker can be similarly trained using the dataset $\mathcal{X}_L$.

The decoder process is simple with this approach. Each DNN operates and generates results independently and the result of the target speaker is selected from the two DNN outputs based on cues such as keywords recognized.

B. Speech Separation with Energy-Dependent Denoisers

As mentioned in Section II, in the feature enhancement approaches to robust ASR, DNNs are treated as front-end denoisers. With the same two synthesized datasets $\mathcal{X}_L$ and $\mathcal{X}_H$, the front-end deep denoiser for the high energy speaker can be trained to minimize the mean squared error (MSE) loss function,

$$\mathcal{L}_{\mathrm{MSE}}(\theta) = \sum_{x_t \in \mathcal{X}_H} |f(x_t; \theta) - x_t^{\mathrm{clean}}|^2, \quad x_t^{\mathrm{clean}} \in \mathcal{X}, \quad (3)$$

where $x_t^{\mathrm{clean}} \in \mathcal{X}$ is the corresponding clean speech features, *i.e.* the features generated on the original uncorrupted target speech, and $f(x_t; \theta)$ is the estimation of the uncorrupted inputs using the deep denoiser. Similarly, the denoiser for the low energy speaker can be trained on the dataset $\mathcal{X}_L$.

C. Speech Separation Based on Average Pitch

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 two speakers speak in similar average energy levels, *i.e.* near 0 dB SNR. This is because the contradictory labels generated from both speech signals may be used as labels to train the DNN under this condition. Since it is far less likely that the two speakers will speak with the same pitch, we propose another approach where DNNs are trained to recognize the speech with the higher or lower pitch which we will refer to as the high and low pitch signal models. 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, *i.e.* the senone labels always come from the alignments on the speech utterances with the higher average pitch when the high pitch signal models are being trained. The loss function for training the DNN for the high pitch speech signals is thus,

$$\mathcal{L}_{\mathrm{CE}}(\theta) = -\sum_{x_t \in \mathcal{X}_P} \log p(s_t^{\mathrm{HP}} | x_t; \theta), \quad (4)$$

where s_t^{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. In this approach, the two DNNs also operate independently during the decoding time just like in the high/low energy DNN case.

D. Speech Separation Based on Instantaneous Energy

Both the high/low energy models and the high/low pitch models have a common weakness. There is inherent ambiguity when two speakers talk with similar average energy or pitch. This motivates us to train the models based on the instantaneous characteristics (e.g., energy) of each individual frame rather than the whole utterance. Since even in utterances with an average SNR of 0 dB the instantaneous SNR at each frame will not likely to be zero, the label ambiguity problem can be significantly alleviated. To build such a model 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

$$\mathcal{L}_{\mathrm{CE}}(\theta) = -\sum_{x_t \in \mathcal{X}_I} \log p(s_t^{\mathrm{IH}} | x_t; \theta), \quad (5)$$

where s_t^{IH} corresponds to the senone label from the signal source which has higher energy at frame t,

$$s_t^{\mathrm{IH}} = \begin{cases} s_t^1 & \mathrm{Energy}(x_t^1) \geq \mathrm{Energy}(x_t^2) \\ s_t^2 & \mathrm{Energy}(x_t^1) < \mathrm{Energy}(x_t^2) \end{cases} . \quad (6)$$

Note that there still exist the cases in which the two signals have the same instantaneous energy. However, this happens very rarely and we simply use the labels of the first signal to train the DNN. Similarly, the instantaneous low energy signal models can be trained with the senone labels assigned using (6) but with reversed signs in the inequality conditions.

Different from the previous two approaches, the output of each DNN in this case is no longer associated with a single speaker since the relative instantaneous energy can change from frame to frame. For example, the target speaker can speak louder in one frame and softer in the next frame. To address this problem we introduce the joint decoding strategy in the next section.

IV. 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 WFST based 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 [33] can be written as,

$$\mathrm{HCLG} = \min(\det(H \circ C \circ L \circ G)), \qquad (7)$$

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. The HCLG graph has encoded all the information needed to decode an utterance, including HMM topologies and transitions probabilities, lexicon and pronunciation model scores and languages model scores except the acoustic scores which only can be evaluated when the speech frames to be decoded become available. Suppose now we want to decode an utterance with T frames, the acoustic scores information can be encoded into a $(T+1)$-state WFSA U where both input and output labels of each arc between states are HMM transition identifiers and the cost is the corresponding acoustic score, *i.e.*, $p(x_t|s_t)$, under Log or Tropical semiring. Decoding the utterance is essentially finding the best path in the $U \circ HCLG$ graph. But the graph $U \circ HCLG$ is never explicitly constructed as it is extremely large without pruning. Instead, finding the best path is done via token passing on the HCLG graph: At frame t, each active token is associated with one state in the HCLG graph; Then we consume one more speech frame by passing all the active tokens through the arcs and accumulate the corresponding acoustic scores; At frame $(t{+}1)$, all the tokens that fall outside the beam-width are cut off. After the token passing is done, we can then perform a trace back to find best sequence from the information stored in the dynamic programming table.

A. *Joint Token Passing on the* HCLG *Graphs*

The task now is to find best two-state sequence in the 2-D joint state space whose size will be the square of the size for each individual speaker's state space. The key part of our proposed decoding algorithm is joint token passing on the two HCLG decoding graphs in conjunction with the acoustic scores accumulations using instantaneous high and low energy DNN models. The main difference in token passing between our joint decoding and conventional decoding is that now each token is associated with two states rather than one in the decoding graph. Denote by θ^{H} and θ^{L} instantaneous high and low energy signal DNN models trained as described in Section III-D. The joint decoder is to find best two state sequence such that the sum of each joint state-sequence log-likelihood is maximized,

$$(\mathbf{s}^{1*}, \mathbf{s}^{2*}) = \underset{(\mathbf{s}^1,\mathbf{s}^2)\in\{\mathbf{s}^1\times\mathbf{s}^2\}}{\operatorname{argmax}} \Big\{p(x_{1:T}|\mathbf{s}^1;\theta^{\mathrm{H}},\theta^{\mathrm{L}})p(\mathbf{s}^1) \cdot p(x_{1:T}|\mathbf{s}^2;\theta^{\mathrm{H}},\theta^{\mathrm{L}})p(\mathbf{s}^2)\Big\}. \qquad (8)$$

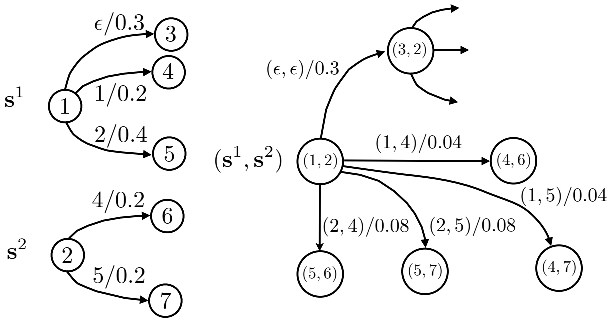


Fig. 1: A toy example illustrating the joint token passing on the two WFST graph: $\mathbf{s}^1$, $\mathbf{s}^2$ denote state space corresponds to one of two speakers; $(\mathbf{s}^1, \mathbf{s}^2)$ represent the joint state space.

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 we will pass the tokens through are 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 trained 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,

$$\begin{aligned}\mathcal{AC} = \max\{&p(x_t|s_t^1;H_t=1)\cdot p(x_t|s_t^2;H_t=1),\\ &p(x_t|s_t^1;H_t=2)\cdot p(x_t|s_t^2;H_t=2)\}\\ = \max\{&p(x_t|s_t^1;\theta^{\mathrm{H}})\cdot p(x_t|s_t^2;\theta^{\mathrm{L}}),\\ &p(x_t|s_t^1;\theta^{\mathrm{L}})\cdot p(x_t|s_t^2;\theta^{\mathrm{H}})\},\end{aligned} \qquad (9)$$

where H_t is the index of speaker who has higher energy. Note that after we compare the joint likelihood in two possible cases, i.e. $H_t = 1$ and $H_t = 2$ using the equation above, we can tell which speaker has higher energy in the corresponding signal at certain frame t along this search path. For the arcs with ϵ input labels, the expansion process is a 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). And for each newly created joint states, we repeat the same joint token passing process until all the tokens are processed. Note that although the nominal size of the joint WFST search space will be $O(|\mathbf{s}^1||\mathbf{s}^2|)$ and the complexity of evaluating the state likelihoods is $O(k|\mathbf{s}^1||\mathbf{s}^2|)$, the actual decoding cost is much lower with beam-pruning during the search.

B. *Penalties on Energy Switching*

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. Fig.2 shows an

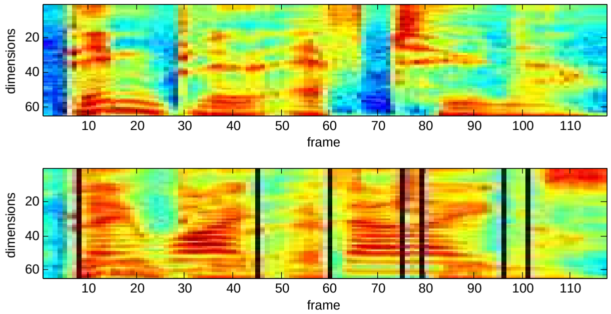


Fig. 2: An example illustrating the energy switching frames in a co-channel mixed speech utterance. Upper: the mean and variance normalized mel-scale filter-bank features of the sample utterance under the clean condition; Bottom: the mean and variance normalized mel-scale filter-bank features under the 0 dB condition, the vertical lines in the fig show the locations where the energy switching occurs.

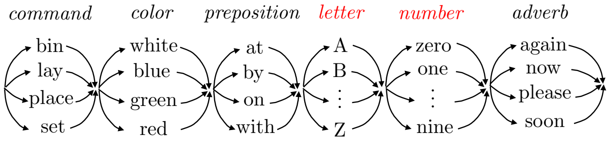


Fig. 3: The grammar of 2006 monaural speech separation and recognition challenge contains six parts: command, color, preposition, letter (with W excluded), number, and adverb; The evaluation metric is the WER on letters and numbers spoken by the target speaker.

System/Method	IBM [4]	Human	Next best [5]
WER	21.6%	22.3%	34.2%

TABLE I: Overall keywords WERs of three systems/methods on the 2006 challenge task. This is a rather difficult task as indicated by the poor human recognition accuracy and the fact that the second best system performs significantly worse than the human performance.

example where the energy switching points occur only 7 times in a 117-frame co-channel mixed speech utterance under the 0 dB condition. This issue can be overcome by introducing a constant penalty in each searching path when the speaker-DNN association has changed in the decoder from the previous frame so that the state sequences with both relatively high likelihood and low speaker switch frequency will survive in the end. Recall that when we compute the acoustic cost during joint decoding we know which speaker is considered by the decoder as the louder speaker at each frame along the search path since we can keep this information when doing token passing and beam searching. If the louder speaker has changed from last frame, we add a constant cost to this search path.

Alternatively, we can estimate the probability that a certain frame is the energy switching point and let the value of the penalty adaptively change 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 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,

$$\mathcal{P} = -\alpha \cdot \log p(y_t | x_t; \theta^S), \tag{10}$$

where y_t is a binary variable which indicates whether energy switching occurs at frame t. With the penalty value evaluated in the above equation, the modifications to regular joint-token passing process on HCLG graph is minor, we just need to add this value to the accumulative cost of each active joint-token at each frame.

V. Experiments

In this section, we report our experimental results with all systems we have discussed in previous sections on the 2006 monaural speech separation and recognition challenge data [2].

A. The Challenge Task and Scoring Procedure

The main task of The 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 the GRID corpus [34]. As shown in Fig 3, the fixed grammar contains six parts: command, color, preposition, letter (with W excluded), number, and adverb, *e.g.* ”place white at L 3 now”. 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). 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.

The 2006 monaural speech separation and recognition challenge is a rather difficult task as shown in Table I. Even human performs poorly on the task. Nevertheless, the best system submitted by IBM [4] can beat the human performance with an overall keyword WER of 21.6% which is very impressive.

B. 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 a 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 generative pre-training [35] followed by discriminative pre-training [21]. 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 frame accuracy improvement is less than 0.5%, we shrink the learning rate

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 II: WERs (%) of baseline GMM-HMM and DNN-HMM systems: both systems are trained on the clean training data.

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 II. As can be seen, the DNN-HMM system trained only on clean data performs poorly in all SNR conditions except the clean condition, motivating the use of multi-style training.

C. Speech Separation Based on Average Energy and Pitch

To investigate the use of multi-style training for the high and low energy signal models, we generated two mixed-speech training datasets:

I. 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, 6 dB, 3 dB, 0 dB. (17,000 × 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 0 dB, -3 dB, -6 dB, -9 dB. (17,000 × 15).

Then we use these two training sets to train two DNN models, DNN-HI and DNN-LO, for recognizing higher and lower energy signals in the mixed-speech, respectively. We used the DNN-HI to recognize mixed-speech for the 0-6 dB cases and DNN-LO to recognize mixed-speech for the -9-0 dB cases and list the results in Table III. From the table, we can see that the results are surprisingly good, especially in the cases where two mixing signals have large energy level difference, *i.e.* 6 dB, -6 dB, -9 dB. In the DNN HI+LO setup, two DNNs recognize the mixed-speech independently and the one whose result contains the color white is treated as the target speaker and its result is used as the final result. We can observe that the combined DNN HI+LO system achieves 25.4% WER compared to 67.4% obtained with the DNN trained only on clean data. However, the DNN HI+LO system still underperforms the state-of-the-art IBM superhuman system. The main cause 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 we adopted to train the high and low energy speech has the inherent label ambiguity problem as we discussed in Section III-C.

For the high and low pitch signal models, we first estimate the pitch for each speaker from the clean training set using a robust pitch tracking algorithm [36] in Voicebox. Then we combine the Train Set I and Train Set II to form Train Set III (17,000 × 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 same rule that the target speaker always says the color 'white'. We list the WERs in Table IV. 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 0 dB case, but worse in the other cases and on average.

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-HI	**4.5**	**16.8**	56.8	-	-	-	-
DNN-LO	-	-	52.6	33.6	**18.4**	**17.4**	-
IBM [4]	15.4	17.8	**22.7**	**20.8**	22.1	30.9	**21.6**
DNN HI+LO	**4.5**	16.9	49.8	39.8	21.7	19.6	25.4

TABLE III: WERs (%) of the DNN systems trained to recognize the higher and lower energy signals in the mixed-speech; DNN HI: multi-style trained DNN for the high energy signals; DNN LO: multi-style trained DNN for the low energy signals; DNN HI+LO: the combined system using the rule that the target speaker is the one who speaks color 'white'.

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

TABLE IV: WERs (%) of the DNN systems for high and low pitch signals; DNN PITCH: multi-style trained DNN for high and low pitch signals.

D. Speech Separation with Energy-Dependent Denoiser

In this section we show results on the multi-style trained deep denoiser. With the same training Set I, we trained a DNN as a front-end denoiser as described in Section III-B. Note that there is no softmax layer in the denoiser and the training is carried out to minimize the mean square error between the clean speech and the cleaned speech estimated by the DNN. To reduce variability the features are mean and variance normalized before fed into the denoiser DNNs. With the deep denoisers, we tried two different setups. In the first one, the denoised features are fed into the DNN trained on the clean data. In the second one, we retrained another DNN on the denoised multi-condition training data. We list the results of both setups in Table V. The results indicate that a DNN trained to predict senone labels directly (i.e., the DNN-HI setup) is slightly better on average than the one that performs denoising prior to classification (i.e., Denoiser HI+DNN setup). This implies that DNN is capable of learning robust representations automatically and there may be no need to extract enhanced features in the front-end. Also note that the energy-dependent denoiser has the same label ambiguity problem (except here the label is the clean speech used as the target of the denoising operation) we have observed in the high/low energy DNN case when the two speech signals have similar energy levels.

To take an insightful look at the denoised features, we select features of one test sample speech utterance under different conditions and feed them into the deep denoiser trained to

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

TABLE V: WERs (%) of deep denoisers for high energy signals; Denoiser HI + DNN: the denoised features are fed into the DNN train on clean data. Denoiser HI + DNN (retrained): the classification DNN was retrained on the denoised feature.

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 HI+LO	**4.5**	**16.9**	49.8	39.8	**21.7**	**19.6**	25.4
IBM [4]	15.4	17.8	22.7	20.8	22.1	30.9	21.6
Joint Decoder	18.3	19.8	**19.3**	21.3	23.2	27.4	21.5
Joint Decoder + SP	16.1	18.7	20.5	19.6	23.6	26.8	20.9
Joint Decoder + ASP	16.5	17.1	19.9	**18.8**	22.5	25.3	**20.0**

TABLE VI: WERs (%) of the DNN systems based on the instantaneous energy and joint decoders; Joint Decoder: the joint decoder system without the energy switching penalties; Joint Decoder + SP: the joint decoder system with the constant energy switching penalties inserted; Joint Decoder + ASP: the joint decoder system with the adaptive switching penalties.

clean the high energy signals to compare the original input features and the denoised output features. In Fig 4a-4d we show pairs of input and output filter-bank features under clean, 6 dB, 3 dB, 0 dB conditions respectively (note that the input features are mean and variance normalized). The squared errors averaged over all time-frequency bins under four conditions are 0.11, 0.19, 0.28 and 0.52, respectively. From the figures, we can also tell that the denoiser works fairly well in 6 dB and 3 dB conditions where the energy in some time-frequency bins belonging to the interference speaker can be removed. However, the residuals under 0 dB become very severe because when two mixing signals have similar energy levels the DNN has conflict requirements of generating an output that is close to both signals.

E. Speech Separation with Instantaneous Energy and Joint Decoder

Finally, we use training Set III to train two DNN models for instantaneous high and low energy signals as described in Section III-D. With these two trained models, we perform a joint decoding as described in Section IV. The results of this Joint Decoder approach are shown in Table VI. The last two systems correspond to the cases where we introduce the energy switching penalties. The Joint Decoder + SP is the system with the constant energy switching penalty and Joint Decoder + ASP is the system with adaptive switching penalty. To get the value of the energy switching penalties as defined in (10), we trained a DNN to estimate an energy switching probability for each frame.

From Table VI we can observe that all approaches using the joint decoder outperform the IBM's superhuman system. With the adaptive switching penalty the joint decoding system cuts error by 1.6 absolute over the IBM's superhuman system. From Table VI, we can also see that the DNN HI+LO 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 + ASP 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 level differences between the two signals. Note that if the SNRs of the input speech signals are available, system combinations can be directly done by selecting either the multi-trained DNN HI+LO or the joint decoder system. We list the oracle WERs with this assumptions in Table VII and show that the optimal average we can achieve is 16.9%. If the SNR levels are not available, we need to introduce the mechanisms to determine which system to use. In Section V-F and V-G, we will present two ways to do system combinations based on the front-end denoisers and lattice confidence scores.

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 HI+LO	**4.5**	16.9	49.8	39.8	21.7	**19.6**	25.4
IBM [4]	15.4	17.8	22.7	20.8	22.1	30.9	21.6
Joint Decoder + ASP	16.5	17.1	19.9	18.8	22.5	25.3	20.0
Combined I	16.0	16.6	**19.7**	18.8	23.0	24.1	19.7
Combined II	11.1	**15.9**	22.5	**21.3**	**20.7**	21.3	**18.8**
Combined (oracle)	4.5	16.9	19.9	18.8	21.7	19.6	16.9

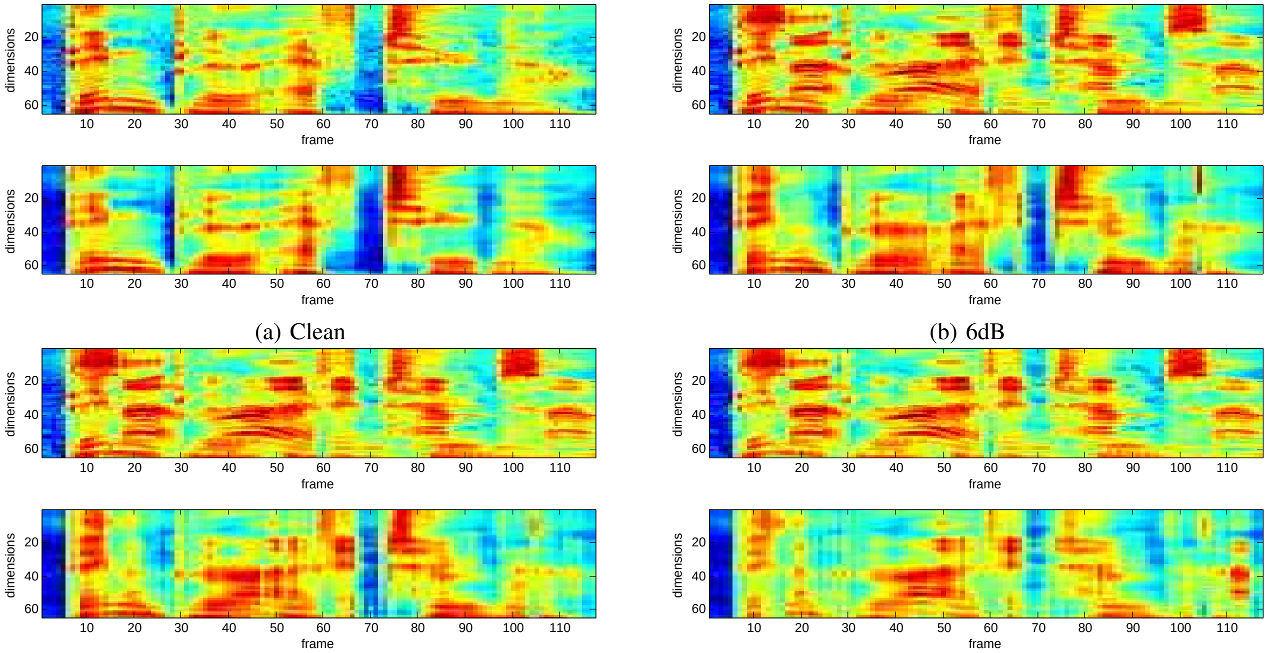
TABLE VII: WERs (%) of the combined systems using DNN HI+LO and Joint Decoder + ASP; Combined (oracle): the oracle combined system under the assumption that the SNR information is available; Combined I: the combined system based on the energy level estimation using the deep denoisers; Combine II: the combined system based on the confidence level estimation.

F. System Combination Using Deep Denoisers

One way to combine the two systems is based on the energy level difference of two mixing speech signals. To get energy level difference between two mixing signals, we can use the front-end 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. Note that since both the input and denoised features of the front-end deep denoisers are mean and variance normalized, we need to transform the denoised features back to the unnormalized ones when calculating the energy level difference. We first tune and obtain an 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 HI+LO to decode the test utterance, otherwise the system Joint Decoder + ASP will be used. The results are listed in Table VII as Combined I with an average WER of 19.7%.

G. System Combination Using Confidence Scores

As shown in Section V-D, the denoised features have severe residuals when the two mixed speech signals have similar energy levels, which will leads to very inaccurate estimations of the energy ratio. Therefore, we propose another approach to do the system combination based on the confidence score, i.e.,



(a) Clean

(b) 6dB

(c) 3dB

(d) 0dB

Fig. 4: Upper: mean and variance normalized mel-scale filter-bank features of the sample utterance under the clean, 6dB, 3dB and 0dB condition; Bottom: the reconstructed filter-bank features from the high energy denoisers.

state sequence posteriors derived from the decoding lattice. The main idea is that if one system generates the results with the high state sequence posteriors $p(W|x_{1:T})$, the system should work better with relatively high confidence. So we can use it to determine whether we use DNN HI+LO or Joint Decoder system. The sequence posteriors are derived from the decoding lattices using the equation

$$p(W|x_{1:T}) = \frac{p(x_{1:T}|W)p(W)}{\sum_{W'} p(x_{1:T}|W')p(W')}. \tag{11}$$

We first use the posteriors generated by either high or low energy signal models alone. The assumption is that if one of two speakers in the speech signal dominates, either high or low energy signal model would generate the fairly high $p(W|x_{1:T})$. But we find the system combination based on this strategy does not perform well. Alternatively, we try to derive the sequence posteriors from the lattices generated by the joint decoder working with the instantaneous high and low energy models. Note that in this case each path in the lattice will contain the state sequences for both speakers. For the efficient evaluation, we use the following equation as the sequence posteriors for two speakers together,

$$\mathcal{C} = \frac{p(x_{1:T}|W^1)p(W^1)p(x_{1:T}|W^2)p(W^2)}{\sum_{W'^1,W'^2} p(x_{1:T}|W'^1)p(W'^1)p(x_{1:T}|W'^2)p(W'^2)}. \tag{12}$$

With this confidence score, we first tune and obtain an optimal threshold on the development set, and use it for the system combination. If the confidence exceeds the threshold the joint decoder approach is used. Otherwise the DNN HI + LO is used. The results are listed in Table VII as Combined II. With this new system combination approach, we obtain the lowest 18.8% average WER.

VI. Summary and Discussion

In this work, we investigated DNN-based systems for single-channel multi-talker speech recognition with a multi-style training strategy. 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 a competing speaker. The best setup of our proposed systems achieves 18.8% overall WER which improves upon the results obtained by the IBM superhuman system by 2.8% absolute, with fewer assumptions and lower computational complexity. Five techniques contributed to this result: a multi-style training strategy on artificially mixed speech data to enable the DNN to generalize to similar patterns in the test data, a separate DNN to estimate senone posterior probabilities of the louder and softer speakers, a WFST-based two-talker decoder to jointly estimate and correlate the speaker and speech, a speaker switching penalty estimated from the energy pattern change in the mixed-speech, and a confidence based system combination strategy.

Although our system outperformed IBM's superhuman system and human performance, there are still a lot of efforts needed to solve the multi-talker speech recognition problem. In fact, the 2006 speech separation and recognition challenge is a synthesized challenge set. It is different from the real data in which the SNR level is often over 6 dB. But more importantly, the challenge is a small vocabulary task which favors automatic speech recognition systems due to tight search space and less confusion. At the same time, the grammar used in the challenge is very different from actual sentences people would normally speak and thus artificially degraded people's performance on the dataset due to mismatched language model. However, by using this challenge dataset, we are able to compare our proposed approach to many other different approaches on the same dataset and investigate the pros and cons of the proposed techniques.

The technique proposed in this paper can be further improved in many aspects. For example, the searching criterion used in the joint decoder is based on the joint sequence likelihood alone. A better searching criterion would combine the state sequence likelihood with other criterion such as the speech separation score, or a completely different searching criterion may be proposed. In the current implementation we used simple switching penalty together with the joint decoder to trace speakers. This can be improved by exploiting other information sources such as pitch. Our future work will focus on these areas and on real data.

VII. Acknowledgements

We would like to thank Geoffrey Zweig, Frank Seide at Microsoft Research for their valuable suggestions and Kun Han from The Ohio State University for the valuable discussions.

References

[1] C. Weng, D. Yu, M. Seltzer, and J. Droppo, "Single-channel mixed speech recognition using deep neural networks," in *ICASSP*. IEEE SPS, May 2014.

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

[4] 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*. ISCA, 2006.

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

[6] 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.

[7] J. Barker, N. Ma, A. Coy, and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 94–111, Jan. 2010.

[8] 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*. ISCA, 2006.

[9] Y. Shao, S. Srinivasan, Z. Jin, and D. 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.

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

[11] 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.

[12] S. Rennie, J. Hershey, and P. Olsen, "Single-channel multitalker speech recognition," *Signal Processing Magazine, IEEE*, vol. 27, no. 6, pp. 66–80, Nov 2010.

[13] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. 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.

[14] G. 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, jan. 2012.

[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] J. Li, D. Yu, J.-T. Huang, and Y. Gong, "Improving wideband speech recognition using mixed-bandwidth training data in cd-dnn-hmm." in *SLT*. IEEE, 2012, pp. 131–136.

[17] 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," *CoRR*, vol. abs/1301.3605, 2013.

[18] A. Maas, Q. Le, T. O'Neil, O. Vinyals, P. Nguyen, and A. Ng, "Recurrent neural networks for noise reduction in robust asr," in *Proceedings of INTERSPEECH*, 2012.

[19] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The munich 2011 chime challenge contribution: Nmf-blstm speech enhancement and recognition for reverberated multisource environments," in *Proc. Machine Listening in Multisource Environments (CHiME 2011), satellite workshop of Interspeech 2011, ISCA, Florence, Italy*, 2011.

[20] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist speaker normalization and adaptation," in *in Eurospeech*, 1995, pp. 2183–2186.

[21] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *ASRU*, 2011, pp. 24–29.

[22] M. J. Hunt and C. Lefebvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *Proc. ICASSP1989*, 1989.

[23] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[24] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, vol. 3, 2000, pp. 1635–1638 vol.3.

[25] O. Vinyals, S. Ravuri, and D. Povey, "Revisiting Recurrent Neural Networks for Robust ASR," in *ICASSP*, 2012.

[26] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *INTERSPEECH*, 2011, pp. 237–240.

[27] T. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 4153–4156.

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

[29] C. Weng, D. Yu, S. Watanabe, and B.-H. Juang, "Recurrent deep neural networks for robust speech recognition," in *ICASSP*, Florence, Italy, 2014.

[30] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' Speech Separation and Recognition Challenge: Datasets, tasks and baselines," in *ICASSP*, Vancouver, Canada, 2013.

[31] J. Geiger, F. Weninger, J. Gemmeke, M. Wollmer, B. Schuller, and G. Rigoll, "Memory-enhanced neural networks and nmf for robust asr," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 6, pp. 1037–1046, June 2014.

[32] A. Graves, N. Jaitly, and A. rahman Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *ASRU*, 2013, pp. 273–278.

[33] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, vol. 20, no. 1, pp. 69–88, 2002.

[34] M. Cooke, J. Barker, S. Cunningham, and X. 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.

[35] A. Mohamed, G. 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.

[36] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Klein and K. K. Palival, Eds. Elsevier, 1995.

Chao Weng is currently a Speech Research & Dev. Engineer in Siri team at Apple Inc. Prior to joining Apple, he has been with AT&T Lab Research and Microsoft Research as a Research Intern, working on recurrent neural network language modeling, Mandarin and Japanese speech recognition, deep neural networks for multi-talker speech recognition. Chao contributes to the Kaldi project, a popular open-source speech recognition toolkit that has been widely adopted by academia and industry. His backgrounds lie generally in the areas of speech recognition and natural language processing with special focus on discriminative training and recurrent neural networks for robust speech recognition, weighted finite-state transducers (WFSTs) for speech and language processing. Chao holds a Ph.D. in Electrical and Computer Engineering from Georgia Institute of Technology.

Dong Yu (M'97, SM'06) is a principal researcher at Microsoft Research. He holds a Ph.D. degree in computer science from University of Idaho, an MS degree in computer science from Indiana University at Bloomington, an MS degree in electrical engineering from Chinese Academy of Sciences, and a BS degree (with honor) in electrical engineering from Zhejiang University. His research interests include speech processing, robust speech recognition, and machine learning. He has published two monographs and over 140 papers in these areas and is the inventor/coinventor of near 60 granted/pending patents. His work on deep learning and its application in large vocabulary speech recognition was recognized by the IEEE SPS 2013 best paper award.

Dr. Dong Yu is currently serving as a member of the IEEE Speech and Language Processing Technical Committee (2013-) and an associate editor of IEEE transactions on audio, speech, and language processing (2011-). He has served as an associate editor of IEEE signal processing magazine (2008-2011) and the lead guest editor of IEEE transactions on audio, speech, and language processing - special issue on deep learning for speech and language processing (2010-2011).

Michael L. Seltzer received the Sc.B. degree with honors from Brown University in 1996, and M.S. and Ph.D. degrees from Carnegie Mellon University in 2000 and 2003, respectively, all in electrical engineering. From 1998 to 2003, he was a member of the Robust Speech Recognition Group at Carnegie Mellon University. Since 2003, Dr. Seltzer has been a member of the Speech & Dialog Research Group at Microsoft Research, where he is currently a Senior Researcher. In 2006, he was awarded the IEEE SPS Best Young Author paper award for his work on microphone array processing for speech recognition. While at Microsoft, Dr. Seltzer has made scientific contributions in noise robustness and speech enhancement, and his algorithms are used in several Microsoft products including Bing Voice Search, Windows Phone, Windows Automotive, and Windows Live Messenger. He is currently a member of the IEEE Speech and Language Technical Committee (SLTC) and from 2006-2008, he was Editor-in-Chief of the SLTC e-Newsletter. From 2009-2011, he was an Associate Editor of the IEEE Transactions on Audio, Speech, and Language Processing. His current interests include speech recognition in adverse environments, acoustic modeling and adaptation, neural networks, microphone arrays, and machine learning for speech and audio applications.

Jasha Droppo (M03-SM'07) received the B.S. degree in electrical engineering (with honors) from Gonzaga University, Spokane, WA, in 1994, and the M.S. and Ph.D. degrees in electrical engineering from the University of Washington, Seattle, in 1996 and 2000, respectively. At the University of Washington, he helped to develop and promote a discrete theory for time-frequency representations of audio signals, with a focus on speech recognition. He is best known for his research in robust speech recognition, including algorithms for speech signal enhancement, model-based speech feature enhancement, robust speech features, model-based adaptation, and noise tracking. His current interests include the use of neural networks in acoustic modeling and the application of large data and general machine learning algorithms to previously hand-authored speech recognition components.