J Sign Process Syst
https://doi.org/10.1007/s11265-017-1295-x

A Speaker-Dependent Approach to Single-Channel Joint Speech Separation and Acoustic Modeling Based on Deep Neural Networks for Robust Recognition of Multi-Talker Speech

Yan-Hui Tu[1] · Jun Du[1] · Chin-Hui Lee[2]

Received: 5 February 2017 / Revised: 23 August 2017 / Accepted: 21 September 2017

Abstract We propose a novel speaker-dependent (SD) multi-condition (MC) training approach to joint learning of deep neural networks (DNNs) of acoustic models and an explicit speech separation structure for recognition of multi-talker mixed speech in a single-channel setting. First, an MC acoustic modeling framework is established to train a SD-DNN model in multi-talker scenarios. Such a recognizer significantly reduces the decoding complexity and improves the recognition accuracy over those using speaker-independent DNN models with a complicated joint decoding structure assuming the speaker identities in mixed speech are known. 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-DNN based acoustic models. Experimental results on Speech Separation Challenge (SSC) small-vocabulary recognition show that the proposed approach under multi-condition training achieves an average word error rate (WER) of 3.8%, yielding a relative WER reduction of 65.1% from a top performance, DNN-based pre-processing only approach we proposed earlier under clean-condition training (Tu et al. 2016). Furthermore, the proposed joint training DNN framework generates a relative WER reduction of 13.2% from state-of-the-art systems under multi-condition training. Finally, the effectiveness of the proposed approach is also verified on the Wall Street Journal (WSJ0) task with medium-vocabulary continuous speech recognition in a simulated multi-talker setting.

Keywords Multi-talker speech recognition · Speaker-dependent model · Single-channel speech separation · Deep neural networks · Joint training

1 Introduction

With recent wide spreads of smart mobile devices, automatic speech recognition (ASR) technologies are being utilized in more and more speech enable applications. Moreover, ASR is deployed in many new scenarios, such as in a living room environment, in which multiple talkers are speaking at the same time while music or television programs are ongoing in the background. Speech pre-processing techniques to reduce noise, separating competing sources and speech dereverberation are becoming critically important for speech applications to be used in such challenging conditions. In this study we focus our attention on speech recognition of a target speaker in multi-talker scenarios in which target speech is often mixed with competing speech from known and unknown speakers that speech separation [2] usually needs to be performed first before ASR could be effectively used. Even with the availability of dual-microphone settings in today's mobile devices, the speech separation performance is still unsatisfactory.

✉ Jun Du
jundu@ustc.edu.cn

Yan-Hui Tu
tuyanhui@mail.ustc.edu.cn

Chin-Hui Lee
chl@ece.gatech.edu

[1] University of Science and Technology of China, Hefei, China

[2] Georgia Institute of Technology, Atlanta, GA, USA

Published online: 04 October 2017

As early as 2006, the PASCAL Speech Separation Challenge (SSC) [3] was launched that was focused on recognizing speech of a target speaker corrupted by an interfering talker with a prior knowledge about the vocabulary and grammar. Historically, all the approaches to solving this problem could be mainly divided into three categories. First, a factorial hidden Markov model (FHMM) [4–7] for separation was used to model an interaction between the target and competing speech signals with their temporal dynamics, followed by a joint decoding strategy for ASR. Second, nonnegative matrix factorization (NMF) [8, 9] was adopted for single-channel speech separation. Finally, approaches based on computational auditory scene analysis (CASA) [10] to estimate a time-frequency mask of each speaker were proposed in [11–13]. Among all the submissions to SSC, the IBM superhuman system [4], belonging to the first category, performed the best and even exceeded human listeners on this challenge task. It consisted of three main components, namely a speaker recognizer, a supervised speech separator, and a speech recognizer, all based on Gaussian mixture models (GMMs) [14].

Recent studies have shown that deep learning [15, 16] have led to a great success in many speech processing areas. For example, in ASR, a hybrid DNN-HMM structure [17–19] was widely adopted in place of the traditional GMM-HMM for acoustic modeling. In single-channel speech separation, approaches based on the DNNs [20–22] were proposed to separate each target speaker from mixed speech. Furthermore, for single-channel ASR of target speech, one work in [23] utilized novel DNN architectures to jointly model the two mixing speakers with a weighted finite-state transducer (WFST) [24] based decoder. It was shown to outperform the IBM superhuman system.

Nonetheless, both state-of-the-art approaches in [4, 23] utilize a joint decoding framework which requires an additional computational complexity. Meanwhile, those methods cannot be easily extended to scenarios with more than two interfering speakers. To alleviate these difficulties, we devote our attention here to extracting information of the target speaker in a semi-supervised mode [20] which is more relevant in source separation and ASR in multi-talker scenarios. In [20], speaker-dependent (SD) DNN models were designed for speech separation as a pre-processing module for the subsequent ASR task of the target speaker using clean-condition trained GMM-HMMs, yielding a better recognition accuracy than that of the IBM system, and without the need of using IBM's joint decoding scheme which is rather complicated. In this paper, the speaker-dependent concept is extended from speech separation to target speech recognition in multi-talker scenarios.

Accordingly, a novel SD-DNN for joint modeling of front-end speech separation and back-end acoustic modeling is proposed to simultaneously separate and recognize speech of a target speaker. The main contributions are in three aspects as follows. First, a multi-condition training strategy based on synthesizing a large-scale training set with very limited target speaker speech data is adopted to boost the SD ASR performance in the multi-talker setting, 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 SD regression DNN for mapping the log mel-filterbank (LMFB) [25] features of mixed speech to LMFB features of a target speaker is adopted as the front-end, which is different from pur recently proposed pre-processing DNN using log-power spectra features [20]. Finally, the SD front-end DNN can be seamlessly concatenated and jointly trained with the SD back-end DNN for acoustic modeling as a hybrid DNN architecture, which explicitly normalizes the variability from other interfering speakers and significant boost the ASR performance in multi-speaker speech recognition.

Experimental results on the small vocabulary SSC task show that our proposed SD approach is quite robust to the interference of a competing speaker even in low target-to-masker ratio (TMR) conditions. The best configured SD multi-condition system achieves an average WER of 3.8% across different TMRs, yielding a relative WER reduction of 65.1% from our already top performing system under clean-condition training with only DNN-based speech separation at the front-end [20]. In addition, we also applied the proposed SD approach is also applied to the Wall Street Journal (WSJ0) database [26], a medium-vocabulary continuous speech recognition task, in a new and similar simulated multi-talker setting, and our experimental results also demonstrate the effectiveness of the proposed joint training strategy. The final SD multi-condition system achieves an average WER of 13.2%, or a relative WER reduction of 86.6% from a clean-condition trained system (with an average WER of 98.2%).

This study is an extension of the recently disclosed version [20, 27] with more technical detail and new experimental results on the WSJ0 task with a larger vocabulary. The remainder of the paper is organized as follows. In Section 2, a system overview is given. In Section 3, we detail our proposed SD approach for speech recognition of a target speaker in multi-talker, mixed speech scenarios. In Section 4, we report our experimental results and finally we conclude our findings in Section 5.

2 System Overview

In Fig. 1, we illustrate a work flow of the proposed SD-DNN multi-talker recognition system. In the upper part of Fig. 1, a SD multi-condition (SD-MC) training set is first prepared for a target speaker. Then in the upper portion of the middle

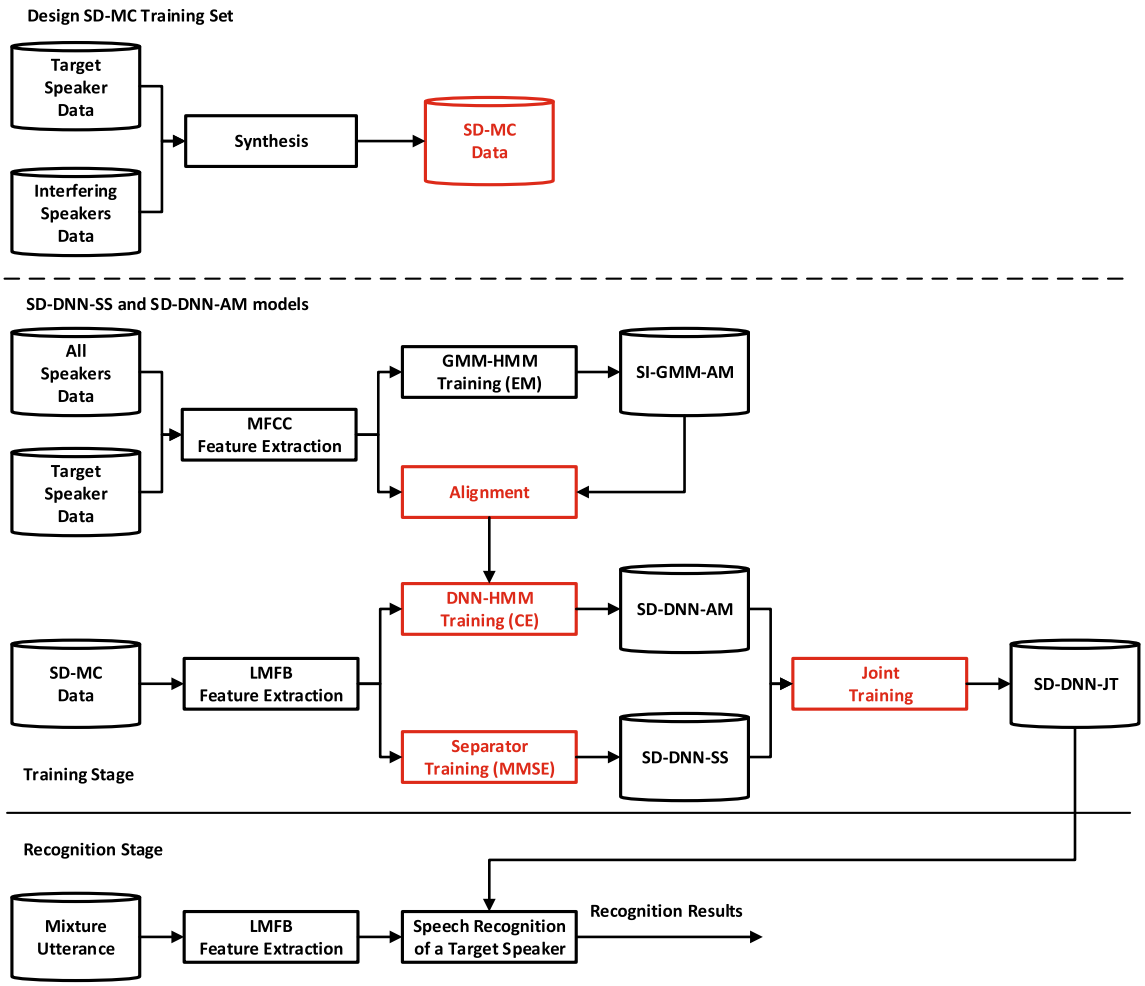


Figure 1 SD recognition system in multi-talker scenarios.

part of Fig. 1, the MFCC features are extracted from speech data of all training speakers which are then utilized to train a speaker-independent (SI) clean-condition acoustic model, denotes as SI-GMM-AM. This trained model is then used to align the MFCC features extracted from the target speaker data to generate the phone state labels of the target speaker. Next in the lower portion of the middle part of Fig. 1, the LMFB feature pairs of the mixed speakers and the target speaker are extracted from the training data samples and 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. And the training stage can be described as follows:

Step 1: Design SD-MC Training Set

Step 2: Train SI-GMM-AM Model and Obtain the Labels

1. The MFCC features are extracted from speech data of all training speakers.

2. Train a speaker-independent (SI) clean-condition acoustic model, denotes as SI-GMM-AM, with the MFCC feature.

3. Align the MFCC features extracted from the target speaker data to generate the phone state labels of the target speaker.

Step 3: Joint Train SD-DNN-SS and SD-DNN-AM Models

1. The LMFB feature pairs of the mixed speakers and the target speaker are extracted from the training data samples.

2. Train speaker-dependent DNNs for speech separation and acoustic modeling, denoted as SD-DNN-SS and SD-DNN-AM, respectively.

3. A hybrid DNN (SD-DNN-JT) is generated.

In the recognition stage, in the lower part of Fig. 1 as in most conventional ASR procedures, the LMFB features of the mixture utterance are directly fed into to the hybrid SD-DNN-JT to generate the recognized sentence accordingly. In

the next section, we will elaborate the highlighted modules in red in Fig. 1, namely the design of the SD-MC training set and the proposed joint training procedure.

3 Training of Speaker-Dependent DNNs

3.1 Design of a Large-scale SD-MC Training Set

In most conventional SI ASR systems, the multi-condition training strategy (e.g., [28]) is widely adopted to improve the ASR robustness in noisy environments. However in the multi-talker scenarios, this concept may not be directly applicable because it is difficult to differentiate the interfering speakers from the target talker. So in the IBM superhuman system, the clean-condition trained GMM-HMMs are adopted with two streams of each speaker from the separation module for subsequent joint decoding. Only in a recent work [23], a DNN architecture to simultaneously model two speakers at the output layer of a DNN could accommodate the multi-condition training strategy for SI recognition systems. 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, the ambiguity between speakers has been reduced by focusing on the target speaker. The required 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 available, e.g., about 15 minutes of SD training speech for each target speaker in the SSC task.

3.2 Labeling of the Mixture Utterances

For training of SD-DNN-AM with the synthesized SD-MC training set in multi-talker scenarios, the SI GMM-HMMs can not correctly classify the silence segments which are dominated by the interfering speakers in mixed speech. So the labels of the mixture utterances used in acoustic modeling should correspond to those of the underlying target speaker utterances. In this way, the HMMs of the speech units are guided to learn the phonetic information of the target speaker while the speech segments dominated by other interfering speakers are forced to be aligned to the "non-speech" units, like the filler segments in keyword spotting [29, 30]. With this SD recognizer, it can perform a "selective" recognition of speech segments corresponding to the target speaker and ignore the segments dominated by other interfering speakers.

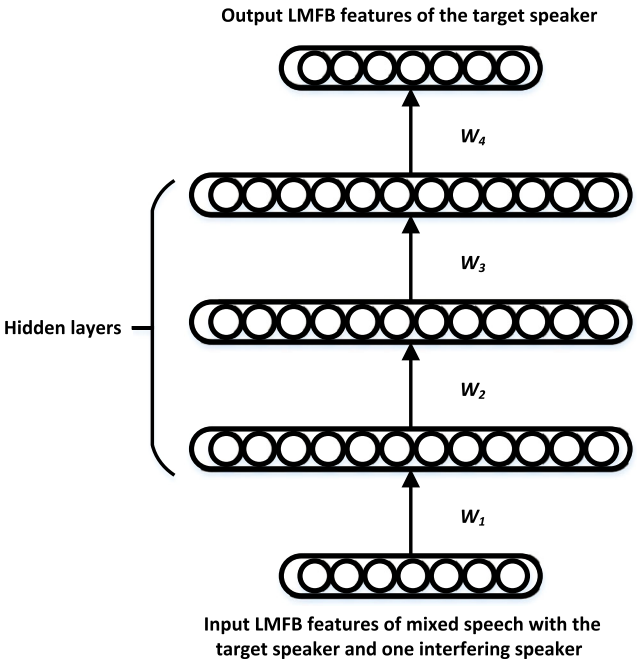


Figure 2 An illustration of SD-DNN-SS.

3.3 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 [27, 31], here we adopt the DNN as a regression model to directly map the ASR features of the mixed speakers to those of the target. It can be considered as an irrelevant variability normalization step [32, 33] for the SD recognizer. As shown in Fig. 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 [31]. An unsupervised pre-training step via the restricted Boltzmann machine (RBM) [34] 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 [35] between the DNN output and the reference LMFB features of the target speaker:

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

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 neighboring 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.

3.4 Acoustic Modeling and Joint Training

To build the acoustic model SD-DNN-AM, we follow the recipe in [17, 18, 36]. First, GMM-HMMs trained on clean utterances of the target speakers are used to generate the frame-level senone [17] (shared HMM state) labels for the SD-MC data set. Then layer-by-layer generative pre-training [18] followed by discriminative pre-training [36] is conducted. Finally, the cross-entropy (CE) criterion [37] is adopted to update all the parameters.

So far, SD-DNN-SS and SD-DNN-AM are separately learned using different objective functions. Since the output layer of SD-DNN-SS can be directly fed into the input layer of SD-DNN-AM, it is straightforward to concatenate the two DNNs to form a hybrid DNN (SD-DNN-JT) with the learning objective to maximize the recognition performance as in conventional DNN-based ASR systems. Therefore the proposed joint training procedure as illustrated in Fig. 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.

To illustrate the differences of recognizing mixed speech under clean-condition and multi-condition training, Fig. 4 gives spectrograms together with the recognition results of a test utterance using the two recognizers. Fig. 4a and b show the spectrograms of target and interfering speech with the correct texts, respectively. For a comparison, we insert the segment of silence into the end of interfering speech to make the two utterances with an equal length. Fig. 4c is the spectrogram of mixture utterance with the recognition results using clean-condition and multi-condition models. The speech segments of the mixture utterance marked by the yellow rectangle boxes, from either the target or the interfering speaker, are both recognized by the clean-condition model, while the proposed SD multi-condition model only recognizes the segments belonging to the target speaker. As for the speech segments marked by the green rectangle boxes, where the speech segments of the target and interfering speakers

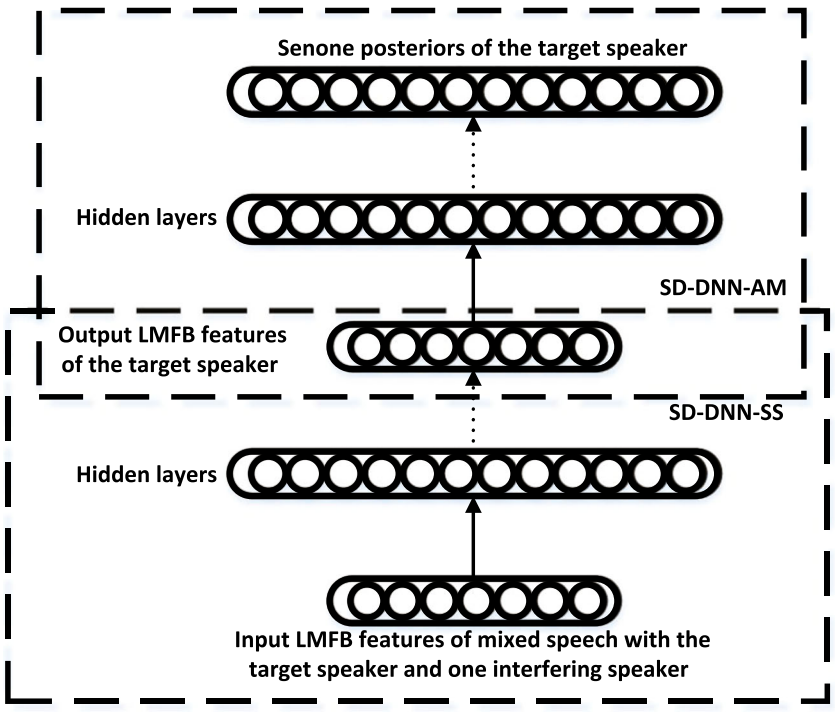
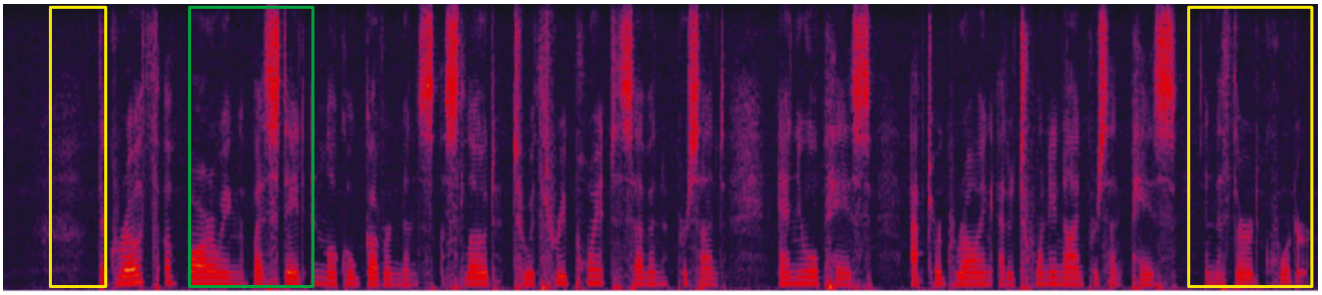
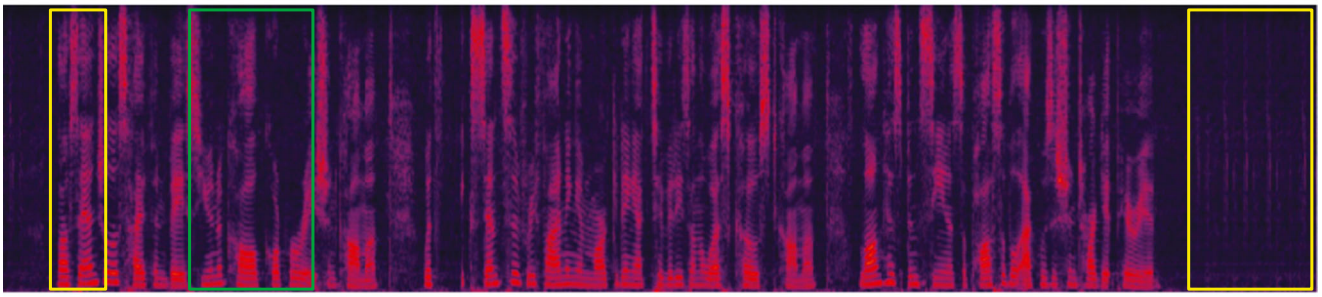


Figure 3 An illustration of the joint training procedure.



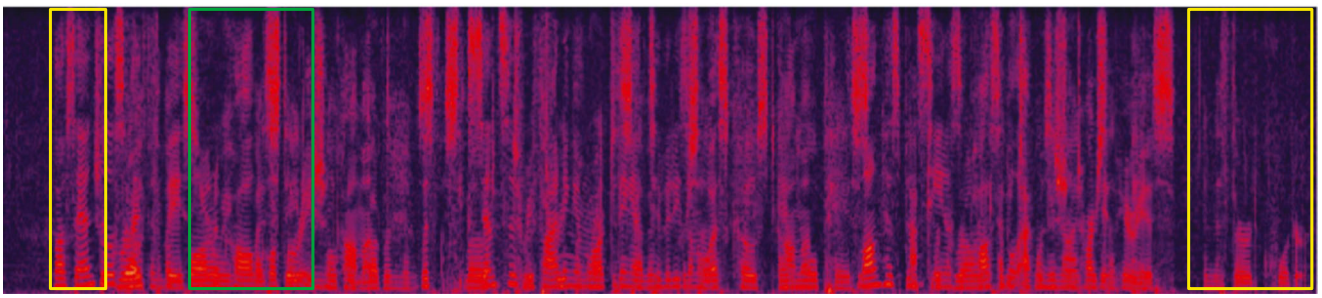
The correct text of target speech:
THE GROWTH OF BORROWING BY STATE AND LOCAL GOVERNMENTS SLOWED TO A THREE POINT SEVEN PERCENT ANNUAL PACE AFTER GROWING AT A TWENTY NINE PERCENT PACE IN THE THIRD QUARTER

(a) The spectrogram of target speech



The correct text of interfering speech:
LOS ANGELES BASED UNOCAL CORPORATION COMMA WITH EXTENSIVE REFINING AND MARKETING ACTIVITIES ON THE WEST COAST COMMA LONG HAS BEEN SEEN AS A POSSIBLE ACQUISITION TARGET PERIOD

(b) The spectrogram of interfering speech



Results of recognizer under clean-condition:
LOS ANGELES <UNK> STORAGE AND CALM OF A <UNK> SEASON SAYS REFINING AND MARKETING ACTIVITIES WEST COAST COMMON PAYS LONG HAS BEEN POSSIBLE ACQUISITION TARGET AREAS IN THE THIRD QUARTER

Results of recognizer under multi-condition:
THE GROWTH OF BORROWING BY STATE AND LOCAL GOVERNMENTS SLOWED TO A THREE POINT SEVEN PERCENT ANNUAL PACE AFTER GROWING AT A TIME WHEN THE NINE PERCENT PACE IN THE THIRD QUARTER

(c) The spectrogram of mixture speech

Figure 4 Spectrograms with the recognition results of an utterance from the test set using clean-condition and multi-condition trained recognizers.

are overlapped in the mixture utterance, the recognition results of the clean-condition trained model are wrong, while the proposed SD multi-condition model can correctly recognize mixed speech. Clearly the proposed SD multi-condition training approach is quite robust to the interfering speaker for multi-talker recognition.

4 Experiments and Result Analysis

First, we demonstrate the effectiveness of our proposed approach using the SSC database [38], where the task aims at the small vocabulary with a command-sentence grammar. Second, we apply our approach to more challenging

recognition task using WSJ0 database [26], which is a medium-vocabulary continuous speech recognition task.

4.1 Experiments on the SSC Task

4.1.1 Setup

Our experiments were first conducted on the SSC corpus. The challenge task was to recognize the keywords from simple *target* sentences when presented with a simultaneous *masker* sentence with a very similar structure [3]. All the training and test materials were drawn from the GRID corpus [39]. There were 34 speakers for both 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 "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 system configurations in [23]. First, 64-dim 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, denoting that the size was 576 $(64 \times 9, \tau = 4)$ at the input layer, with 2048 units 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 system parameter settings can be found in [31, 40, 41].

4.1.2 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, implying a need of for multi-condition training.

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

4.1.3 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 2 lists a WER comparison of the SD DNN-HMM systems under clean-condition and multi-condition training on the test set of 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 6 TMR levels in the test set, ranging from -9 dB to 6 dB with an increment of 3 dB, 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 indicated 6 randomly selected speakers had a good representation of all 34 speakers and would be used for the subsequent experiments. 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 6 selected target speakers under multi-condition training with different

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

Table 3 WER comparison 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

amounts of training data. 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 (500×34×6) training utterances. 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 (500×34×21) 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 a WER of 28.1%.

4.1.4 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 [27], 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 5.6%) was much better than the best performance of the pre-processing DNN approach at 6 dB (a WER of 7%).

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
SND-DNN[27]	7	8.5	9.2	11.3	12.7	16.9	10.9

4.2 Experiments on the WSJ0 Task

4.2.1 Setup

The medium-vocabulary continuous speech recognition task based on WSJ0 database [26] is adopted for the following experiments. There are 83 speakers providing short-term data for SI training in the database. We selected 2 speakers (1 female and 1 male) as the target speakers and other 10 speakers (5 females and 5 males) as the interfering speakers to construct the experimental data set. There were about 90 utterances for each speaker, with an average duration of about 7 s. For each speaker, we selected 10 utterances as the test data and the remaining utterances as the training data. For clean-condition training, all clean utterances of the 83 speakers excluding 20 test utterances of two target speakers were used. For multi-condition training, each training utterance of each target speaker was corrupted with speech segments randomly selected from utterances of other 10 interfering speakers normalized by 5 different TMR levels, ranging from -6 dB to 6 dB with an increment of 3 dB, generating a set of 4000 (80×10×5) training utterances approximately equal to about 50 hours of speech data. The test utterances of each target speaker were mixed with 4 selected interfering speakers (2 females and 2 males) at five different TMR levels, ranging from -6 dB to 6 dB with an increment of 3 dB. Therefore, at each TMR level, there are $2 \times 4 \times 10 = 80$ test utterances.

As for front-end and back-end processing, the configurations were similar to Section 4.1.1. First, 40-dimensional LMFB features (appended with first-order and second-order derivatives) with a context window of 11 frames were adopted to train both the SD-DNN-SS and SD-DNN-AM components. The architecture of SD-DNN-SS was 1320-2048-2048-2048-1320, which denoted that the size was 1320 (40 ×3×11, τ=5) at the input layer, 2048 for the 3 sigmoidal hidden layers, and 1320 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 1985 units corresponding to the tied states of HMMs. The mini-batch size was set to 256. For the decoding, 3-gram language model with 5k-word vocabulary was adopted.

Table 5 WER comparison of SI DNN-HMM system under clean-condition training on the test set of one male and female target speakers with different TMRs.

System	6dB	3dB	0dB	−3dB	−6dB	Avg.
Male	70.3	83.5	94.5	115.4	121.4	97.0
Female	78.4	92.2	98.8	112.6	115.0	99.4
Avg.	74.4	87.9	96.7	114.0	118.2	98.2

4.2.2 Experiments Under Clean-condition Training

Table 5 shows a WER comparison of SI DNN-HMM systems on the test set of one male and female target speakers under clean-condition training. Obviously, it was still a mismatch testing scenario under clean-condition training. And for this medium-vocabulary continuous speech recognition task, the WERs at low TMRs were over 100% due to many insertion errors, where speech segments of interfering speakers were also recognized as the target speech, just like the example in Fig. 4.

4.2.3 Experiments Under Multi-condition Training

Table 6 gives a WER comparison of SD DNN-HMM systems on the test set of one male and female target speakers under multi-condition training. By comparing the results of Tables 5 and 6, the SD multi-condition training not only reduced the mismatch between training and testing, but also made the speech recognizer distinguish the target speaker and the interfering speaker. The SD DNN-HMM system achieved an average WER of 15.4%, representing a relative WER reduction of 84.3% from SI DNN-HMM system under clean-condition training with a WER of 98.2%.

4.2.4 Experiments with Jointly Trained DNN Models

Finally, we also examine the effectiveness of our proposed jointly trained SD-DNN-JT system on WSJ0 database as shown in Table 7. Similar performance gains could be observed from the SD-DNN-JT system with an average WER of 13.2%, or a relative WER reduction of 14.3% from the multi-condition trained SD-DNN-AM system (an average WER of 15.4%), which demonstrated that our proposed approach could be still quite effective for the medium-vocabulary continuous speech recognition task.

Table 6 WER comparison of SD DNN-HMM system under multi-condition training on the test set of one male and female target speakers with different TMRs.

System	6dB	3dB	0dB	−3dB	−6dB	Avg.
Male	11.6	12.6	13.6	15.8	20.0	14.7
Female	11.3	12.4	15.5	18.4	22.9	16.1
Avg.	11.4	12.5	14.5	17.1	21.4	15.4

Table 7 WER comparison of SD-DNN-JT system under multi-condition training on the test set of one male and female target speakers with different TMRs.

System	6dB	3dB	0dB	−3dB	−6dB	Avg.
Male	10.6	11.0	12.3	14.4	16.3	12.9
Female	9.2	10.3	12.6	15.7	19.4	13.4
Avg.	9.9	10.7	12.4	15.0	17.9	13.2

5 Conclusion and Future Work

In this paper, we have proposed a novel speaker-dependent approach to simultaneous speech separation and acoustic modeling in one hybrid DNN architecture for single-channel automatic speech recognition of mixture speech in a multi-talker setting. Coupling with a new multi-condition training strategy, we have obtained very promising speech recognition results on the SSC task. In the meantime we have also verified the effectiveness of the proposed framework on medium-vocabulary continuous speech recognition using the WSJ0 database in a simulated multi-speaker situation. We have demonstrated that the conventional multi-condition training framework can be extended to multi-condition training in a speaker-dependent setting with interfering speakers serve as the varying conditions we are exploiting, an ideal scenario in which the information of a particular known speaker is fully utilized for both source separation and speech recognition. As for future work, first we will try to optimize the training criterion, such as use multi-task training where the SS objective as part of the training criterion. Second, we will continue to explore more target speakers and larger database. Finally, we will continue to explore adverse conditions, including background and convolutional noises together with multiple-speaker interferences.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grants No. 61671422.

References

1. Tu, Y., Du, J., Dai, L., & Lee, C. (2016). In *Proc. ISCSLP*.
2. Radfar, M.H., & Dansereau, R.M. (2007). *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(8), 2299.
3. Cooke, M., Hershey, J.R., & Rennie, S.J. (2010). *Computer Speech and Language*, *24*(1), 1.
4. Kristjansson, T.T., Hershey, J.R., Olsen, P.A., Rennie, S.J., & Gopinath, R.A. (2006). In *Proc. annual conference of international speech communication association. (INTERSPEECH)*.

5. Virtanen, T. (2006). In *Proc. annual conference of international speech communication association. (INTERSPEECH)*.
6. Weiss, R.J., & Ellis, D.P.W. (2007). In *Proc. IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)* (pp. 114–117).
7. Ghahramani, Z., & Jordan, M.I. (1997). *Machine Learning*, *29*, 245.
8. Schmidt, M.N., & Olsson, R. (2006). In *Proc. annual conference of international speech communication association. (INTERSPEECH)*.
9. Jackson, E.P. (2006). In *Proc. annual conference of international speech communication association. (INTERSPEECH)*.
10. Wang, D., & Brown, G.J. (2006). *Journal of the Acoustical Society of America*, *124*(1), 13.
11. Barker, J., Ma, N., Coy, A., & Cooke, M. (2010). *Computer Speech and Language*, *24*(1), 94.
12. Ming, J., Hazen, T.J., & Glass, J.R. (2010). *Computer Speech and Language*, *24*(1), 67.
13. Shao, Y., Srinivasan, S., Jin, Z., & Wang, D. (2010). *Computer Speech and Language*, *24*(1), 77.
14. Reynolds, D.A., & Rose, R.C. (1995). *IEEE Transactions on Speech and Audio Processing*, *3*(1), 72.
15. Hinton, G.E., & Salakhutdinov, R. (2006). *Science*, *313*(5786), 504.
16. Hinton, G.E., Osindero, S., & Teh, Y. (2006). *Neural Computation*, *18*(7), 1527.
17. Dahl, G.E., Yu, D., Deng, L., & Acero, A. (2012). In *IEEE Transactions on audio, speech, and language processing*.
18. Mohamed, A., Dahl, G.E., & Hinton, G.E. (2012). *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 14.
19. Hinton, G.E., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A.W., Vanhoucke, V., Nguyen, P., Sainath, T.N., & et al. (2012). *IEEE Signal Processing Magazine*, *29*(6), 82.
20. Du, J., Tu, Y., Dai, L., & Lee, C. (2016). *IEEE Transactions on Audio, Speech, and Language Processing*, *24*(8), 1424.
21. Huang, P., Kim, M., Hasegawajohnson, M., & Smaragdis, P. (2015). *IEEE Transactions on Audio, Speech, and Language Processing*, *23*(12), 2136.
22. Zohrer, M., Peharz, R., & Pernkopf, F. (2015). *IEEE Transactions on Audio, Speech, and Language Processing*, *23*(12), 2398.
23. Weng, C., Yu, D., Seltzer, M.L., & Droppo, J. (2015). *IEEE Transactions on Audio, Speech, and Language Processing*, *23*(10), 1670.
24. Mohri, M., Pereira, F., & Riley, M.P. (2002). *Computer Speech and Language*, *16*(1), 69.
25. Nadeu, C., Macho, D., & Hernando, J. (2000). *Speech Communication*, *34*(1), 93.
26. Paul, D.B., & Baker, J.M. (1992). In *Proc. 5th DARPA speech and natural lang. workshop* (pp. 357–362).
27. Tu, Y., Du, J., Dai, L., & Lee, C. (2015). In *Proc. ICASSP* (pp. 61–65).
28. Yu, D., Seltzer, M.L., Li, J., & Seide, F. (2013). In *Proc. CoRR*, Vol. 1301.
29. Zhang, Y., & Glass, J.R. (2009). In *Proc. IEEE automat. Speech recognition and understanding workshop.(ASRU)*.
30. Wilpon, J.G., Lee, C.H., & Rabiner, L.R. (1989). In *Proc. ICASSP* (pp. 254–257).
31. Tu, Y., Du, J., Dai, L., & Lee, C. (2015). In *Proc. ICSP* (pp. 532–536).
32. Gales, M. (1998). *Computer Speech and Language*, *12*(2), 75.
33. Hu, Y., & Huo, Q. (2007). In *Proc. annual conference of international speech communication association. (INTERSPEECH)*.
34. Bengio, Y. (2009). *Foundat. and Trends Mach Learn*, *2*(1), 1.
35. Ephraim, Y., & Malah, D. (1984). *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(6), 1109.
36. Seide, F., Li, G., Chen, X., & Yu, D. (2011). In *Proc. IEEE automat. speech recognition and understanding workshop. (ASRU)*.
37. De Boer, P., Kroese, D.P., Mannor, S., & Rubinstein, R.y. (2005). *Annals of Operations Research*, *134*(1), 19.
38. Cooke, M., & Lee, T.W. (2016). http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm.
39. Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). *Journal of the Acoustical Society of America*, *120*(5), 2421.
40. Xu, Y., Du, J., Dai, L., & Lee, C. (2014). *IEEE Signal Processing Letters*, *21*(1), 65.
41. Hinton, G.E. A practical guide to training restricted Boltzmann machines (University of Toronto, 2010).

Yan-Hui Tu received the B.S. degree from the Department of Electronic Information Engineering, Yunnan University in 2013. He is currently a Ph.D. candidate at University of Science and Technology of China (USTC). His current research interests include speech separation, microphone arrays and robust speech recognition.

Jun Du received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC) in 2004 and 2009, respectively. From 2004 to 2009, he was with iFlytek Speech Lab of USTC. During the above period, he worked as an Intern twice for 9 months at Microsoft Research Asia (MSRA), Beijing. In 2007, he also worked as a Research Assistant for 6 months in the Department of Computer Science, The University of Hong Kong. From July 2009 to June 2010, he worked at iFlytek Research on speech recognition. From July 2010 to January 2013, he joined MSRA as an Associate Researcher, working on handwriting recognition, OCR, and speech recognition. Since February 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing (NEL-SLIP) of USTC.

Chin-Hui Lee is a Professor in the School of Electrical, and Computer Engineering, Georgia Institute of Technology. Before joining academia in 2001 he had 20 years of industrial experience ending in Bell Laboratories, Murray Hill, New Jersey, as a Distinguished Member of Technical Staff, and Director of the Dialog Systems Research Department. Dr. Lee is a Fellow of the IEEE, and a Fellow of ISCA. He has published over 450 papers, and 30 patents, and was highly cited over 30,000 times for his original contributions with an h-index of 66 on Google Scholar. He received numerous awards, including the Bell Labs President's Gold Award in 1998. He also won the SPS's 2006 Technical Achievement Award for "Exceptional Contributions to the Field of Automatic Speech Recognition." In 2012, he was invited by ICASSP to give a plenary talk on the future of speech recognition. In the same year he was awarded the ISCA Medal in scientific achievement for pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition.