An End-to-End Deep Learning Approach to Simultaneous Speech Dereverberation and Acoustic Modeling for Robust Speech Recognition

Bo Wu, Kehuang Li, Fengpei Ge, Zhen Huang, *Student Member, IEEE*, Minglei Yang, Sabato Marco Siniscalchi, *Senior Member, IEEE*, and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*—We propose an integrated end-to-end automatic speech recognition (ASR) paradigm by joint learning of the front-end speech signal processing and back-end acoustic modeling. We believe that "only good signal processing can lead to top ASR performance" in challenging acoustic environments. This notion leads to a unified deep neural network (DNN) framework for distant speech processing that can achieve both high-quality enhanced speech and high-accuracy ASR simultaneously. Our goal is accomplished by two techniques, namely: (i) a reverberation-time-aware DNN based speech dereverberation architecture that can handle a wide range of reverberation times to enhance speech quality of reverberant and noisy speech, followed by (ii) DNN-based multi-condition training that takes both clean-condition and multi-condition speech into consideration, leveraging upon an exploitation of the data acquired and processed with multichannel microphone arrays, to improve ASR performance. The final end-to-end system is established by a joint optimization of the speech enhancement and recognition DNNs. The recent REverberant Voice Enhancement and Recognition Benchmark (REVERB) Challenge task is used as a test bed for evaluating our proposed framework. We first report on superior objective measures in enhanced speech to those listed in the 2014 REVERB Challenge Workshop on the simulated data test set. Moreover, we obtain the best single-system word error rate (WER) of 13.28% on the 1-channel REVERB simulated data with the proposed DNN-based pre-processing algorithm and clean-condition training. Leveraging upon joint training with more discriminative ASR features and improved neural network based language models, a low single-system WER of 4.46% is attained. Next, a new multi-channel-condition joint learning and testing scheme delivers a state-of-the-art WER of 3.76% on the 8-channel simulated data with a single ASR system. Finally, we also report on a preliminary yet promising experimentation with the REVERB real test data.**

***Index Terms*—End-to-end system, dereverberation, robust ASR, deep learning, joint training, microphone array.**

Manuscript received April 1, 2017; revised July 27, 2017, August 25, 2017, and September 6, 2017; accepted September 17, 2017. Date of publication September 26, 2017; date of current version November 16, 2017. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Kate Knill. *(Corresponding author: S. M. Siniscalchi).*

B. Wu and M. Yang were with the Georgia Institute of Technology, Atlanta, GA 30332 USA. They are now with the National Laboratory of Radar Signal Processing, Xidan University, Xi'an 710126, China (e-mail: rambowu11@gmail.com; mlyang@xidian.edu.cn).

F. Ge was with the Georgia Institute of Technology, Atlanta, GA 30332 USA. He is now with the Key Lab of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100864, China (e-mail: gefengpei@hccl.ioa.ac.cn).

S. M. Siniscalchi is with the Faculty of Architecture and Engineering, University of Enna "Kore," Enna 94100, Italy, and also with the Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: marco.siniscalchi@unikore.it).

K. Li, Z. Huang, and C.-H. Lee are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: kehlekernel@gmail.com; huangzhenee@gmail.com; chl@ece.gatech.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2756439

I. Introduction

THE reverberation phenomenon introduces echoes and spectral distortions into the observation signal [2] to a point that speech may become difficult to be understood, especially for hearing-impaired and elderly people. Moreover, the temporal and spectral smearing caused by room reverberation often makes the speech signal useless for post-processing applications, such as automatic speech recognition (ASR), and source localization. This so called *robustness* problem severely limits the widespread deployment of applications and services. In the recent REverberant Voice Enhancement and Recognition Benchmark (REVERB) Challenge [3], [4], for instance, it has been observed that a severe drop in the recognition accuracy can be caused by strong reverberation conditions even in simulated reverberant conditions. To be specific: A WER of 3.50% can be delivered on clean data using a conventional context-dependent-Gaussian mixture model-hidden Markov models (CD-GMM-HMMs) ASR systems trained on clean data, and adopting a 3-gram model. However, a severe drop in the recognition performance is observed when testing this same system in strong reverberation condition, and a final WER of 87.48% was delivered. Interestingly, no performance improvement was gained using context-dependent-Deep neural network-hidden Markov models (CD-DNN-HMMs) trained in clean conditions. It is therefore not surprising that speech dereverberation has become a research topic of growing importance due to applications that encourage hands-free operation [5], where an effective dereverberation would be beneficial.

In principle, the mismatch between training and testing conditions due to reverberation can be *separately* viewed in the signal space, the feature space, and the model space, where $D_1()$, $D_2()$, and $D_3()$ characterize the possible distortions in the three spaces, respectively, as shown in Fig. 1. Feature compensation, and model adaptation are methods that work at the feature and model spaces, respectively. The idea is to reduce the mismatch between the observed utterance and the original

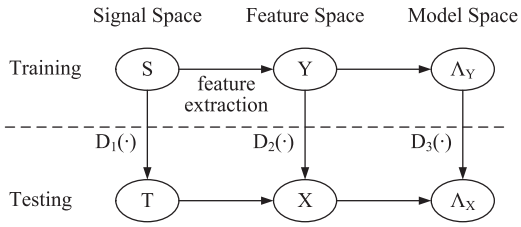


Fig. 1. Mismatches between training and testing conditions (from [1]).

speech models during recognition of the utterance. In feature compensation, we map the distorted features to an estimate of the original features, so that the original acoustic models can be used, e.g., [6]–[8]. In model adaptation, we map the original acoustic models to a transformed model that better match the observed utterance, e.g., [9]–[18]. The interested readers are referred to [19], [20] for more backgrounds. In the signal space, many dereverberation techniques have been proposed over the years, e.g., [21]–[25]. In inverse filtering of the room impulse response (RIR) [26], the dereveberberated signal is estimated convolving the reverberant signal with the inverse filter. However, a minimum phase assumption is often needed, which is almost never satisfied in practice [26]. Some efficient and effective solutions to inverse filtering problems have proposed in [21]–[23], [27], [28], yet the RIR can also be varying in time and hard to estimate [5]. Few studies attempted to separate speech and reverberation via homomorphic transformation [29], [30]. In [31], non-negative matrix factorization (NMF) was used to factorize spectrograms into non-negative speech and noise dictionaries and their non-negative activations; then the clean speech signal can be estimated from the product of speech dictionaries and their activations, e.g., [32], [33]. Recently, deep neural networks (DNNs) have proven to be a reliable vehicle to attain state-of-the-art speech enhancement results due to their strong mapping capabilities, e.g., [34]–[37]. DNNs have also been effective for speech dereverberation [24], [38] in a single-channel scenario. However, DNN speech signal processing area can create a new direction in multi-channel dereverberation. Multi-channel signal processing methods include spatial filtering and channel selection. When the signals from the individual microphone with a known geometry are suitably combined, the array can function as a spatial filter (*a.k.a.* beamforming) for suppressing noise and reverberation [39]. However, if the positions of the microphones are neither known nor fixed, beamforming approaches become less effective. Channel selection is then an alternative approach, which was investigated in [40] for channel selection in multi-microphone environments, and in [41] for distant speech recognition.

All of the above methods rely on complex pipelines composed of multiple algorithms and hand-engineered processing stages that are difficult to reproduce. End users have to follow a very strict set of protocols to effectively utilize spoken language applications. The technology is in practice too fragile that careful designs have to be rigorously practiced to hide technology deficiencies. Moreover, speech quality and intelligibility issues are addressed disjointly and tackled independently in the three different spaces depicted in Fig. 1. In practice, a large gain in speech quality can translate to a negligible improvement in transcription accuracy, and vice-versa. For example, speech enhancement techniques may introduce distortions, such as *musical noise* [42], that can cause additional troubles in post-processing. Moreover, one of the most powerful method for reverberant speech recognition is the use of multi-condition training, as demonstrated, for example by many results from different research groups at the REVERB challenge, which disregard quality aspects.

In this study, we start from the key intuition that robust ASR could not be solved by separate signal pre-processing and model post-processing. Instead, an integrated end-to-end paradigm by jointly modeling the front-end and back-end is needed. Before a perfect front-end could be designed and speech quality and intelligibility could be simultaneously improved, we could gain ASR performance from a joint design. That is especially needed when the front-end, and back-end need different speech features for best performance. Indeed, log-power spectra (LPS) [43] features have been demonstrated to be the best features in tackling speech enhancement [34], [35]. In [44], log Mel-filterbank (LMFB) [45] features have proven to be the best choice for accomplishing speech recognition within the deep learning framework. A unified framework for distant speech processing that can simultaneously tackle speech dereverberation and acoustic modeling is thus proposed. Our goal is accomplished by adopting two techniques, namely DNN-based regression to enhance reverberant and noisy speech, followed by DNN-based multi-condition training. To improve speech enhancement and recognition, we also exploit specific characteristics of the problem at hand and investigate: (i) a reverberation-time-aware (RTA) DNN based speech dereverberation architecture that can handle a wide range of reverberation times without the need of ad-hoc algorithms but simply properly crafting the DNN input layer, and (ii) a simultaneous exploitation of the data acquired with a multi-channel microphone array for improved recognition performance, which outperforms specialized techniques, such as beamforming [46] that has been shown to partially enhance reverberant signals [47]. The proposed new data utilization strategy based on multi-channel data, which could be referred to as multi-channel-condition learning, leverages upon the complementary information captured in microphone array speech to jointly train dereverberation and acoustic deep models. Its key goal is to discover rich and complex interactions in the signals without any ad-hoc pre-processing but with only data. Different from beamforming and channel selection, where expert knowledge has to be involved in order to reach the desired result, the DNN will eventually boost the signal in the direction of the desired source, and/or possibly ignore/deemphasize some of the available channels using information available in the data.

The REVERB Challenge, which includes both speech enhancement and speech recognition tasks with both simulated and real data, is used as a test bed for evaluating our unified framework based on deep learning. The noise mismatch between simulated training and real testing conditions appears to be real problem with real data; therefore, we heavily focus our present research on the simulated data because addressing the dereveberation problem is our main task. However, a

preliminary set of results is also available for real data condition. We first show that improvement at perceptual level can be tightly paired with ASR improvements if a careful design of a DNN dereverberation technique is executed. We then show that in clean-condition training, we obtain the best word error rate (WER) of 13.28% on the 1-channel REVERB simulated evaluation data with the proposed DNN-based pre-processing scheme. We attain a competitive 1-channel WER of 8.75% on simulated evaluation data with the proposed multi-condition training strategy based on clean, enhanced and reverberant data and the same less-discriminative LPS features used in the enhancement stage. A WER of 8.65% is attained on the same evaluation simulated data but with more discriminative LMFB features. A WER of 7.92% is achieved by the proposed joint training technique on LMFB features and using only clean and multi-condition training data provided with the challenge. This result compares favorably with ROVER of five systems reported in [48]. A WER equal to 4.10% [49] is delivered through improved neural network language models [50] and system combination [51], in the 1-channel setup and yet without using any new training data. Exploiting the proposed multi-channel-condition learning scheme, which leverages upon complementary information captured in microphone array speech with eight-channel information to jointly train dereverberation and recognition DNN models, and using a new multi-channel-condition testing scheme, a state-of-the-art WER of 3.76% on the 8-channel simulated data with a single ASR system. Finally, we provide some preliminary results on real evaluation data.

As aforementioned, this work can be considered as a successful step forward toward implementing our long-term vision, namely robust ASR could not be solved by separate signal pre-processing and model post-processing, and an integrated end-to-end paradigm is required. Therefore, some of the ideas discussed in this paper have been previously explored to solve other tasks. Specifically, RTA-DNN dereverberation was introduced in [38] to tackle speech dereverberation, and joint training was evaluated in [52] to tackle the ASR task in noisy conditions. Furthermore, this work also organises our preliminary ideas and findings on end-to-end robust speech recognition [49] in a systematic way and extends them with new analyses and additional experimental evidences that support the feasibility of our proposed framework. In particular, the RTA-DNN dereveberation module was evaluated for speech enhancement only in matched testing conditions in [38], and the authors evaluated it on artificially reverberated TIMIT data. In this paper, RTA-DNN is also evaluated in mismatched testing conditions (referred to as mismatched base-DNN in Section IV-B), and the effectiveness of the RTA-DNN dereveberation pre-processing is assessed against the robust ASR task. In [52], only ASR performance was evaluated on the AURORA 4.0 corpus, base-DNN enhancement was evaluated, and speech enhancement (SE) assessment was not carried out. In contrast, both SE and ASR results are discussed in this paper, and all perceptual and recognition performances are reported on the 2014 REVERB CHALLENGE data. The latter aspect is quite important and is a unique contribution of the present work because readers can now relate with ease and better appreciate SE and ASR improvements: That was not possible in previous works because SE results were reported on simulated, reverberant TIMIT data in [38], but ASR results were reported on AURORA 4.0 data [52]. The enhancement-aware DNN training scheme was not available in [52]. Moreover, the joint multi-channel-condition technique and related experiment evidence were not available in [49], and multi-channel information was not exploited in [52]. We also describe how to use multi-channel information gathered from multiple sensors using a lattice-based post-processing combination scheme, which allow us to assume unknown microphone array configurations. Finally, preliminary results on real evaluation data were not available in [49]. In sum, we critically revise basic ideas explored in previous works, extend them systematically, and properly assess these ideas experimentally.

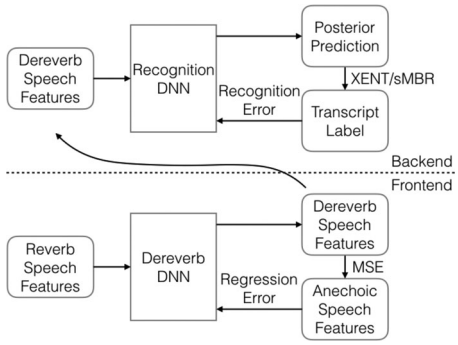


Fig. 2. Disjoint training: deep learning speech dereverberation (*front-end*) followed by robust automatic speech recognition (*back-end*).

II. End-to-End Deep Learning for Speech Dereverberation & Speech Recognition

Figure 2 shows the two key modules in our framework. In the bottom panel, DNN-based dereverberation is illustrated. In the training stage, a regression DNN [35] is trained by a set of reverberant and anechoic speech pairs represented by LPS. In the dereverberation stage, the well-trained DNN is fed with the LPS features of input speech to generate the corresponding enhanced LPS features. The dereverberated waveforms are reconstructed from the estimated spectral magnitude and the reverberant speech phase with an overlap-add method [53]. In the top panel of Fig. 2, the back-end recognition module is shown. Short-time spectral representation, extracted from the dereverberated waveforms at the output of the front-end model, are fed into the DNN-based recognition module using tied, context-dependent phone units, namely senones [54], to compute posterior probabilities, which are in turn used in an hybrid CD-DNN-HMM acoustic model for final decoding.

In line with a recent effort in end-to-end modeling [55], we can stack the front-end and back-end modules back-to-back, skipping the waveform reconstruction stage, to let the ASR system adopt state-of-the-art features, namely LMFB [44] to improve the ASR accuracy and robustness. The system in Fig. 3 indeed represents the proposed end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition.

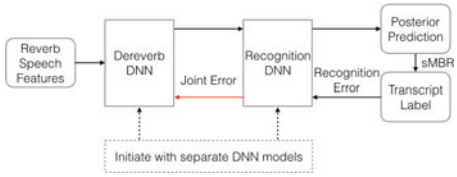


Fig. 3. Joint training: the proposed end-to-end robust ASR system.

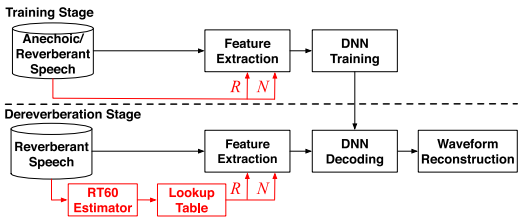


Fig. 4. Block diagram of the proposed RTA-DNN dereverberation module.

A. Reverberant Speech Characteristics

In signal processing of reverberant speech, different room conditions will result in distinct superpositions in the time domain and inter-frame correlations, which is often neglected by most dereverberation algorithms [21]–[24]. In [38], an environment-aware approach, to take into account the superposition characteristics and frame-wise temporal correlations in distinct reverberant situations was successfully deployed. It was demonstrated that a wide context-expansion is not always beneficial, and a high time resolution is needed in weaker reverberation conditions. Reflected sounds travel a less distance to a microphone [56] for weak reverberation, resulting in an intensive superposition in the time domain. A dense sampling, not needed for strong reverberation, would therefore desirable to provide a high temporal resolution, and a varying temporal resolution, dependent of the reverberation condition, should be adopted.

The use of more acoustic context information can improve the continuity of enhanced speech [35]; however, the inter-frame correlation, in reverberant speech, depends upon RT60. At low RT60, the temporal correlation of the consecutive reverberant frames will become weaker. Thus using more context will introduce uncorrelated frames and unnecessary burden in DNN learning [34]. An approach that exploits the continuity of the speech spectra at different RT60s was proposed in [38] and used in this work.

B. Dereverberation Module

The DNN-based dereverberation module can be made reverberation-time-aware by adopting two RT60-dependent parameters, namely frame shift (R) in speech framing and acoustic context size (N) at the DNN input for feature extraction before feeding the log-power spectrum features to the DNNs. We refer to the proposed dereverberation approach as reverberation-time-aware DNN (RTA-DNN). A block diagram of the proposed RTA-DNN system is illustrated in Fig. 4, adopted from [38].

In the training stage, the DNN regression module is trained by a set of multi-condition data, consisting of pairs of reverberant and anechoic speech represented by LPS. A linear activation function at the output layer of the DNN is used [38]. We also globally normalize the target features over all the target utterances into zero mean and unit variance [57]. In training, R and N depend on the utterance-level RT60. The utterances at distinct RT60s are thus enframed by the optimal frame-shift that is extracted according to the results on the training data. The optimal number of frame expansion at each RT60 is also obtained using the training data. Due to the limitation that DNN could only take fixed length vectors as input, the normalized input features of unequal sizes were extended to 11 frames expansion by symmetrically padding zeros to the beginning and the end of the input vectors.

In dereverberation, the unknown RT60s is estimated [58] followed by a lookup table, which was built at the training stage, is used to determine R and N accordingly. Next, the trained DNN is fed with the LPS features of input speech to generate the corresponding enhanced LPSs. The required phase is directly extracted from reverberant speech, because human ears are considered to be not sensitive to such information [59]. Finally the dereverberated waveform is reconstructed from the estimated spectral magnitude and the reverberant phase with an overlap-add method [53]. By skipping the red links and blocks in Fig. 4, R, and N become RT60 independent, and a baseline DNN dereverberation module (base-DNN) is established.

C. Recognition Module

There are two acoustic modeling schemes used in the REVERB Challenge, namely clean- and multi- conditions. Our key goal is to reduce the mismatch between the reverberant test data and the clean training data. We accomplished that through DNN-based speech dereverberation. Multi-condition training is also an effective technique to reduce the mismatch between training and testing data. When combined with carefully designed processing strategies, it facilitates a successful deployment of robust ASR, as discussed in [49].

In Fig. 2, a DNN-based recognition module is shown. However, we have also designed a CD-GMM-HMM speech engine in the clean condition setup. In [36], the authors discussed that the conventional GMM-based acoustic models were good enough to serve as the benchmark for the clean condition training scheme, and sequential (discriminative) training was not needed. In contrast, we want to demonstrate that we can regain the advantages of deep modeling and sequential (discriminative) training due to the strong dereverberation capabilities of our signal-space approach, which reduces the gap between the clean and corrupted data. CD-DNN-HMM acoustic models are always used in multi condition training.

DNNs have been trained by maximizing the log posterior probability over the training frames, which is equivalent to minimizing the cross-entropy (CE) objective function. In [60], [61], it has been shown that sequence training can significantly boost ASR performance. Thus we follow [61], and the minimum Bayes risk objective function at a state-level (sMBR) is used. The input features are LMFB energies [44].

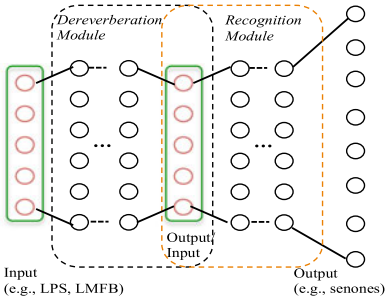


Fig. 5. Proposed DNN architecture for joint training.

D. End-to-end Dereverberation & Robust Speech Recognition

The *end-to-end* design paradigm [62] is of growing interest in several research areas, such as speech recognition, e.g., [63]–[67], object recognition, e.g., [68], [69], machine translation, e.g., [70], keyword spotting, e.g., [71], etc., since systems built under this framework are designed to supersede the above mentioned sophisticated multi-stage processing approaches, which consists of many small sub-components that are tuned separately. An integrated end-to-end paradigm by jointly modeling the front-end and back-end is therefore a desirable solution for improving speech robustness in a reverberant environment, especially because we are far from designing a perfect front-end that can fully recover clean speech from the corrupted signal. Indeed, Du *et al.* [72] have shown that speech robustness can be improved leveraging upon DNN-based speech enhancement, and an error rate reduction by 50% from a baseline system for clean-condition training was demonstrated on the Aurora-4 task. Additional improvement can be observed with joint training of the front-end and back-end models [52]. Moreover, speech robustness can now be boosted through source separation based on signal-space DNN modeling [73], and clean acoustic models were again utilized.

Here, we adopt a hybrid DNN framework to perform joint training of DNNs for both feature mapping and acoustic modeling as shown in Fig. 5. We directly stack the acoustic modeling layers on top of the feature mapping layers. The output layer of feature mapping becomes the input layer for acoustic modeling, which was also a hidden layer of the whole network. The output linear layer, and the hidden non-linear layers of the dereverberation block become hidden layers of the whole neural model. Then the same object functions used to train the recognition module in Section II-C can be adopted to fine-tune all weights.

III. Experimental Setup

We evaluate our proposed methods on the official datasets of the REVERB Challenge [3], which covers single-speaker utterances acquired with one-channel, two-channel, or eight-channel circular microphone arrays. Training, development, and testing sets are available. The training dataset consists of (i) a clean data extracted from the WSJCAM0 [74] training set, consisting of approximately 17.5 hours (7861 training utterances) and (ii) a multi-condition (MC) training set, which was generated from the clean WSJCAM0 training data by convolving the clean utterances with 24 measured room impulse responses (with reverberation times from 0.1 sec to 0.8 sec) and then adding recorded background noise at an SNR of 20 dB. The development and testing material provided by the REVERB Challenge organizers has been deployed to allow the assessment of de-reverberation algorithms (i) in practicality realistic conditions and (ii) against a wide range of reverberant conditions, i.e., robustness. Therefore, both development and testing data are designed to consist of simulated data (**SimData**) and real recordings (**RealData**).

SimData utterances are extracted from the WSJCAM0 corpus, and reverberation effects were artificially introduced. Specifically, six different reverberation conditions have been simulated: Three rooms with different volumes (small, medium, and large) and two distances between a speaker and a microphone array (near = 50 cm, and far = 200 cm). Hereafter, the rooms are referred to as Room 1, Room 2, and Room 3. The simulated development set has 742 utterances in each of far and near microphone conditions, almost equally spread in three room types (1, 2, and 3). SimData testing set contains 1088 utterances in each of the far and near microphone conditions, each of which were split into three room conditions (1, 2 and 3). The real data is extracted from the MC-WSJ-AV corpus [75], which consists of utterances recorded in a noisy and reverberant room. RealData development and testing sets contain 172, and 372 utterances, respectively, split equally between near and far microphone conditions, respectively. The room reverberation time is about 0.7 sec. The interested reader is referred to [3] for more information.

In this study, we target reverberant speech recognition. We have noticed that the noisy condition between training and simulated evaluation data are similar. In contrast, noise discrepancies between simulated training and real testing recordings are quite severe, and noise reduction would no longer be a secondary problem with respect to dereverberation. Our intuition is confirmed by the experimental investigation carried out in [28], where an additional extended multi condition training set was carefully crafted in order to cover additional signal-to-noise (SNR) conditions not seen in the original simulated multi-condition training data. The authors in [28] observed a 1.10% absolute reduction of the average WER on SimData by replacing the original multi condition training set with their own extended multi-condition training set. A 5.00% absolute improvement was instead observed on the RealData task using the same strategy (see Table 3 in [28]). Since the extended and original training data cover the same reverberation times, we can conclude that noise reduction accounts for the big boost performance observed in RealData. In [35], it was already shown that data augmentation is a viable path for improving noise robustness in DNN-based speech enhancement. For our current work, we use only the SimData evaluation set as the chosen vehicle to demonstrate the viability and effectiveness of our proposed approach.

A. Dereverberation Module Configuration

To emphasize upon the importance of proper signal space dereverberation, we deploy the dereverberation module in both *mismatched* and *matched* training/testing conditions. Specifically, we started our analyses using base-DNN and RTA-DNN dereverberation models trained on the TIMIT dataset without

adding ambient noise. This setup is already available to us [38], and it helps simulating mismatched conditions between training and testing environment. In [38], the simulated room dimension was 6 by 4 by 3 meters (length by width by height), and the positions of the loudspeaker and the microphone were at (2, 3, 1.5) and (4, 1, 2) meters, respectively. Then RIRs were simulated using an improved image-source method (ISM) [76] with RT60 ranging from 0.1 to 1.0 s. The reverberant environments of REVERB Challenge are however quite different [3]. For example, reverberant speech is simulated by convolving clean speech with RIRs and subsequently adding measured noise signals with SNR of 20 dB. Room sizes, RT60, and distances a speaker and a microphone are also different. We refer to these models as mismatched base-DNN and mismatched RTA-DNN. The matched base-DNN and RTA-DNN are instead built training with REVERB Challenge training data. For signal analysis, speech was sampled at 16 kHz. The frame length was 32 ms. A 512-point DFT of each overlapping windowed frame is computed. Then 257-dimension log-power spectra feature vectors [53] were used to train DNNs having 3 hidden layers, and 2048 nodes per layer using KALDI [77]. The number of pre-training epochs for each RBM layer was 1, and the learning rate was 0.4. As for fine-tuning, the learning rate and the maximum number of epochs were 0.00008 and 30, respectively. The mini-batch size was set to 128. The configuration parameters were found in [35]. Input and target features of DNN were globally normalized to zero mean and unit variance. Moreover, for the base-DNN, the frame shift, R, and the acoustic context size, N, were fixed to 16 ms and 7 frames, respectively. For RTA-DNN, these two parameters were determined by the estimated utterance-level RT60s [38].

B. Recognition Module Configuration

KALDI was also used to build all ASR systems. In clean condition training, 40-dimensional "linear discriminant analysis (LDA) + maximum likelihood linear transform (MLLT)" features extracted from MFCCs were used to train both the CD-GMM-HMM and CD-DNN-HMM systems. A 9-frame context window was considered during DNN training. In multi-condition training, CD-DNN-HMMs were trained with 72-dimensional LMFB features with a 9-frame context window. All DNNs have 6 hidden layers, and each layer has 2048 sigmoid units if not conversely stated. The output softmax layer has 2085 units. The DNNs were initialized with the stacked RBM-based pre-training, and the cross-entropy loss function was used for frame-level fine-tuning. Sequential training (sMBR) was performed to further enhance the ASR accuracy, e.g., [60]. The standard 3-gram language model (LM) provided in the REVERB Challenge was used in decoding if not conversely stated. Maximum likelihood estimation was used to train the CD-GMM-HMM systems.

IV. Experimental Results

A. Speech Dereverberation Results

Following the REVERB Challenge evaluation instructions, we had our system fully tested on both the SE and ASR tasks.

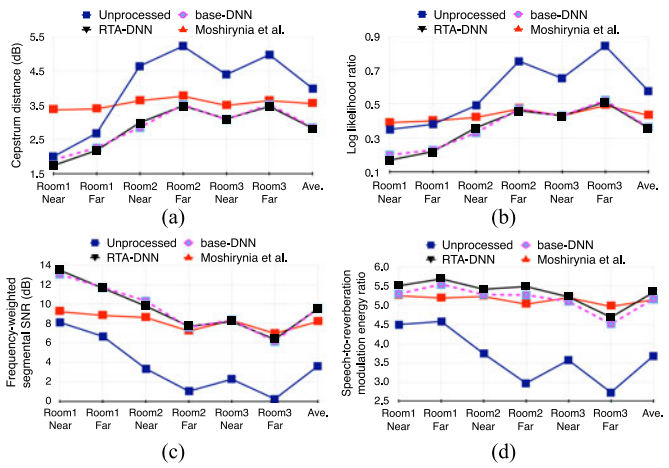


Fig. 6. Objective measures on different test situations. For a fair comparison, we report on and compare against top results obtained with the evaluation tools for SE provided by the 2014 REVERB Challenge organizers [4]. (a) CD, (b) LLR, (c) FWSegSNR, (d) SRMR.

TABLE I
Clean Condition Training: WERs (in %) on SIMDATA Eval

Clean Condition Training		Mismatched DNN DeReverberation						
		Room 1		Room 2		Room 3		Ave
Acoustic Model	Test Data	Near	Far	Near	Far	Near	Far	
CD-GMM-HMM	Rev.	10.22	17.89	27.29	74.39	39.63	87.48	42.81
	Base DeRev.	10.04	18.19	27.36	35.19	29.00	41.61	27.56
	RTA DeRev.	11.86	15.06	25.70	30.81	29.09	41.68	25.70
		Matched DNN DeReverberation						
CD-GMM-HMM	Base DeRev.	9.61	11.69	12.54	22.99	14.38	25.78	16.17
	RTA DeRev.	9.44	12.13	13.97	22.91	15.92	26.27	16.77
CE CD-**DNN**-HMM	Base DeRev.	7.56	9.28	10.87	19.84	12.21	23.93	13.95
	RTA DeRev.	6.57	8.62	12.12	20.13	14.18	24.07	14.28
sMBR CD-**DNN**-HMM	RTA DeRev.	5.90	7.37	11.54	19.85	12.86	22.14	13.28

Results in the upper part are attained with a DNN dereverberation module trained on *mismatched* data. Results gathered in *matched* conditions are given in the lower part.

The tools given by the organizer was used so as to have a fair comparison with submitted results on all objective measures, including cepstrum distance (CD), log-likelihood ratio (LLR), frequency-weighted segmental signal-to-noise ratio (FWSegSNR), and speech-to-reverberation modulation energy ratio (SRMR). The objective measures for SE are given in Fig. 6. It is shown that for CD and LLR, the proposed DNN systems outperform all other methods in all situations - the best performing method listed in the REVERB Challenge is shown for ease of comparison. As for FWSegSNR, the proposed DNN systems have significant improvements in the first three situations when compared with the other methods, and have the best performance in the last three situations. Furthermore, the proposed system has also attained top SRMR scores.

B. ASR Results with Clean-condition Training

In Table I, we list all experimental results in clean condition training of the acoustic models, using information coming from

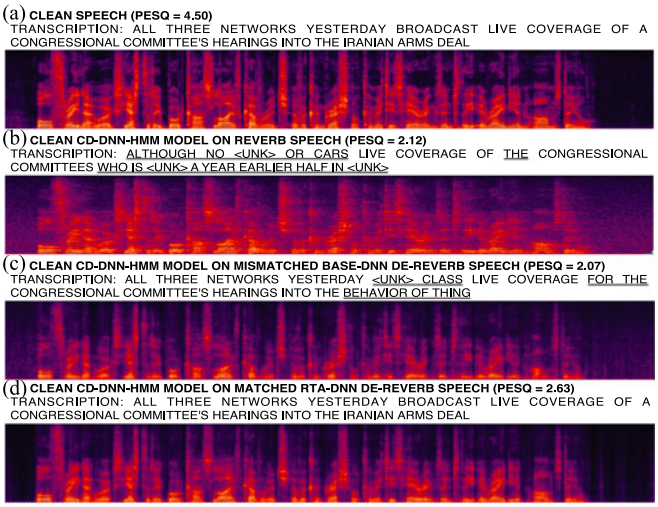


Fig. 7. Spectrograms of a test utterance in simulated *room2 (far-condition)*: (a) clean speech; (b) unprocessed reverberant; (c) dereverberated with DNN trained on the mismatched TIMIT data with no noise; and (d) dereverberated with RTA-DNN trained on the matched REVERB data.

the reference channel (1-channel) only and testing on SimData. A swift visual inspection of the table immediately reveals that speech robustness can be improved by a proper reduction of the reverberation noise in the signal space. Indeed, the WER is drastically reduced from the initial 42.81% on average attained with a CD-GMM-HMM ASR system down to 27.56% and 25.70% using our base-DNN and RTA-DNN dereverberation modules, respectively, which were built in mismatched conditions (see Section III-A). Although the performance boost is already meaningful and competitive–WER of 28.30% is the best reported in clean condition and same task [36], signal space robustness can be boosted by a dereverberation module trained in matched condition. WERs equal to 16.17% and 16.77% are attained using matched base-DNN, and RTA-DNN, respectively, corresponding to a relative improvement of about 60%. Finally, better recognition results can be attained using a CD-DNN-HMM acoustic model, and a WER equal to 13.28% is delivered employing clean CD-DNN-HMMs along with sequential discriminative training. The latter results contrast the idea that GMM-based acoustic models are good enough to serve as the benchmark for the clean condition [36].

Fig. 7 displays spectrograms of a test utterance in simulated *room2-far* condition with transcriptions shown on the top of the corresponding spectrograms and with recognition errors underlined. The clean signal is displayed in Fig. 7(a), and it has a very high PESQ of 4.50. The reverberant signal (Fig. 7(b)) was severely corrupted by reverberation (RT60 = 0.5 s) and noise (SNR = 20 dB), resulting in a low PESQ of 2.12. The base-DNN (Fig. 7(c)), trained on mismatched data set, did not improve PESQ due to the presence of noise which was not considered in the base-DNN. It motivated us to train a matched RTA-DNN resulting in the spectrogram of dereverberated speech (Fig. 7(d)) which was noted to be a very close match to the spectrogram of clean speech in Fig. 7(a), achieving a considerable PESQ increase from 2.12 to 2.63. Furthermore, we got a perfect recognition.

C. ASR Results with Multi-Condition Training

In Table II, we list all experimental results on SimData test data when multi condition training data acquired with the reference channel only is exploited during training. A WER equal to 10.76% can be achieved by training a CD-DNN-HMM backend with a multi condition training scheme and LMFB features, which accounts already for a 74.86% WER reduction (WERR) with respect to the clean CD-GMM-HMM system. In extending the training data by systematically adding clean, base-DNN and RTA-DNN dereverberated data, we can build two different systems, namely S2 and S3, with WERs equal to 9.43%, 9.25%, respectively. By fine-tuning through discriminative sMBR training, we have S4 and S5, attaining a competitive single-system WER of 8.75% in S5 with the same LPS features used in front-end enhancement. A close look at the bottom two rows confirms that a slight yet consistent improvement can be observed by replacing LSP with LMFB features even with less training data. If boosting ASR accuracy is the final goal, the LMFB features appear to be more suitable for the final task. We therefore incorporated LMFB in joint training to be discussed next.

In Table III, we demonstrate that our joint-training solution is not only viable but also outperforms state-of-the-art systems on the same task. The training material is limited to clean and multi condition data provided by the REVERB Challenge organizers in order to clearly demonstrate that improved robustness is conferred to the ASR engine by the beneficial effect on dereverberation and acoustic modeling of the proposed end-to-end training scheme. Indeed, the system, S6, is build through joint training on LMFB features and attains a WER down to 7.92%, which represents the best 1-channel ASR performance with no extra speech training data. If we rescored the S6 lattices with RNN-LM [78], S10, a WER equal to 5.88% was achieved. LSTM-LM [50] based rescoring allowed us to deliver a WER of **4.46**%, which sets a new state-of-the-art. Next, we rescored the lattices generated with disjoint modules based on LPS features, namely S4 and S5, using either RNN- or LSTM-based LMs, with corresponding systems referred to as S8, S9, S11, and S12. As expected, a WER drop, on average, was observed in every case. We then combined RNN-LM systems, S7, S8, and S9 as indicated in [51], and a WER equal to 5.03% was observed. By combining S10, S11, and S12, a WER of **4.10**% was attained, representing another new state-of-the-art. This result was attained using only 1-ch data, and it is remarkable considering that top 1-, 2-, and 8-ch WERs were 5.20%, 4.40%, and 4.20% [28] with lots of extra data.

D. ASR Results with Multi-channel-condition (MCC) Training

Multi-condition training has been demonstrated to be a powerful method for reverberant speech recognition in REVERB Challenge [4]. However, the multi-condition scenario has been limited to collecting data at different SNRs, and/or RT60. We here extend the concept of "multi-condition training" to include multi-channel data, which together with their pre-processed signals are treated as new training materials to be used for acoustic modeling. A large-scale diversified training scenario can thus be used, and shown to be more effective than the conventional one

TABLE II
MULTI CONDITION BACK-END TRAINING: WERS (IN %) ON SIMDATA EVAL

Training Scheme	Sys.	Training Data				Eval. Data	Room 1		Room 2		Room 3		Ave.
		Cln.	Rev.	Base	RTA		**Near**	**Far**	**Near**	**Far**	**Near**	**Far**	
CE	S1		X			Rev.	6.95	8.11	8.68	14.61	9.96	16.29	10.76
	S2	X	X	X		Base	6.15	6.98	7.83	12.91	9.11	13.60	9.43
	S3	X	X	X	X	RTA	6.25	6.45	7.62	12.46	8.85	13.87	9.25
sMBR	S4	X	X	X		Base	6.10	6.56	7.46	11.36	9.09	12.88	8.91
	S5	X	X	X	X	RTA	5.70	6.00	7.10	11.50	8.80	13.40	**8.75**
	S5-LMFB	X	X			RTA	5.18	5.95	6.54	11.78	8.67	13.75	8.65

Dereverberation module trained on *matched* data. In the table: ASR system, test, clean, reverberant, base-DNN dereverberated, and RTA-DNN dereverberated data are abbreviated as Sys., Eval. Data, Cln., Rev., Base. and RTA, respectively. All results are generated using CD-DNN-HMMs.

TABLE III
WER (IN %) WITH JOINT TRAINING W/O ADVANCED ASR SCHEMES

System Configurations			Room 1		Room 2		Room 3		Ave.
Acoustic Model	System	Training + LM	Nearw	Far	Near	Far	Near	Far	
sMBR CD-DNN-HMM	S6	LMFB + Joint Training + 3-gram	5.05	5.86	6.00	10.63	7.73	12.27	7.92
	S7	LMFB + Joint Training + RNN-LM	3.11	4.01	4.42	7.75	5.82	9.74	5.88
	S10	LMFB + Joint Training + LSTM-LM	2.49	2.80	3.18	6.43	4.32	7.52	**4.46**
	S8	S4 + RNN-LM	4.30	4.71	5.45	8.42	6.43	9.66	6.50
	S11	S4 + LSTM-LM	2.90	3.39	4.29	6.51	4.98	8.14	**5.04**
	S9	S5 + RNN-LM	3.74	4.40	5.48	9.11	6.19	10.10	6.50
	S12	S5 + LSTM-LM	2.71	3.15	3.95	6.99	4.71	8.38	**4.98**
System Combination [51]		S7 + S8 + S9	3.29	3.71	4.18	6.43	4.88	7.71	5.03
		S10 + S11 + S12	2.57	2.71	3.46	5.50	3.87	6.48	**4.10**

TABLE IV
WER RESULTS (IN %) WITH MULTICHANNEL DATA AND LSTM-LM

System Configurations		Room 1		Room 2		Room 3		Ave.
Acoustic Model	System	Near	Farw	Near	Far	Near	Far	
2-ch	S13	2.27	2.91	3.30	6.14	4.28	6.72	**4.27**
BF	S14	2.93	3.18	4.89	7.14	5.29	7.11	5.09
8-ch	S15	2.35	2.57	2.97	5.82	4.37	6.23	**4.05**

Results obtained through beamforming are indicated with BF.

based only on beamforming. We call this new strategy multi-channel-condition (MCC) training and we used only multi-condition training data provided with REVERB Challenge in our proposed joint-training solution. To make the experimental results comparable to those in the previous sections, we first tested all systems on speech captured by the reference microphone (ch-1). We will show how to leverage upon all microphones in the next section. In the following experimental investigation, we report only LSTM-LM based results, since it has already been clearly established that the LSTM-LM gives the best recognition accuracy.

First, we pooled the 2-channel speech data to fine-tune the dereverberation and recognition modules separately. Then, we jointly trained the two modules on the same 2-channel training material. Results in Table IV confirm that better accuracies can be attained with information from more channels and joint training. This can be better understood by comparing S13 in Table IV against S9 in Table III. It could be argued that lower WERs could be obtained through a beamforming pre-processing step. To this end, we first beamformed the signal coming from the two channels and generated a new speech stream named "12beamformed" for both training and testing. Next, we trained the dereverberation and recognition modules on this new stream. Finally, we moved to joint training, and we refer to the corresponding ASR system as S14 in Table IV. The results reported in the upper part of Table IV demonstrate that joint training can better exploit the information in distinct channels. In the bottom part of Table IV, the multi-channel-condition scenario is extended to all available eight channels in REVERB Challenge and system S18 demonstrates that speech robustness can be further improved. Together with our proposed joint training strategy, this single system delivered a new state-of-the-art WER of 4.05%.

E. ASR Results with MCC Training and MCC Testing

Next, we test our multi-channel-condition joint training technique on a new MCC testing scenario, leveraging upon speech from multiple sensors, assuming unknown microphone array configurations. The recognition results for each of the eight channels are reported in Table V. We can see that WERs among different channels are comparable. Next, we can combine the word lattices generated for each copy of the input speech signal using the previously adopted system combination scheme [51], assuming each word lattice has been generated by a differ-

TABLE V
WER RESULTS (IN %) WITH MULTI-CHANNEL COMBINATION AND LSTM-LM ON THE 8 DIFFERENT AVAILABLE CHANNELS

System Configurations		Room 1		Room 2		Room 3		Ave.
Acoustic Model	System	Near	Far	Near	Far	Near	Far	
multi-channel-condition	ch-1	2.35	2.57	2.97	5.82	4.37	6.23	**4.05**
	ch-2	2.27	2.68	2.96	5.74	4.44	6.26	**4.06**
	ch-3	2.39	2.78	3.17	6.17	4.16	6.47	**4.19**
	ch-4	2.35	2.64	3.25	6.27	4.11	6.48	**4.18**
	ch-5	2.41	2.54	2.93	5.82	4.40	6.12	**4.04**
	ch-6	2.32	2.66	2.78	6.12	4.42	6.70	**4.17**
	ch-7	2.42	2.57	3.07	6.01	4.32	6.86	**4.21**
	ch-8	2.42	2.52	3.05	6.00	4.09	6.55	**4.11**
Channel Combination		2.20	2.52	3.09	5.08	4.18	5.46	**3.76**

In the last row, we combine word lattices generated by decoding each copy of the input spoken utterance acquired with one of the eight microphone.

TABLE VI
WERS (IN %) ON THE REALDATA EVALUATION SET

System Configurations			Room 1		Ave.
#channels	System	LM	Near	Far	
1-ch	NTT (WPE) [28]	3-gram	27.90	27.70	27.80
	S6	3-gram	25.52	28.60	27.06
	S10 + Unsupervised Adaptation	LSTM-LM	15.23	17.56	**16.40**
8-ch	NTT (WPE) [28]	3-gram	25.20	24.20	24.70
	S19	3-gram	24.05	25.57	24.81
	S21 + Unsupervised Adaptation	LSTM-LM	20.70	24.60	22.65

NTT results are extracted from [28] and refers to ASR systems trained using original multi condition training data (referred to as AM 1 in [28]). WPE stands for weighted prediction error algorithm.

ent ASR system. In doing so, we obtain another state-of-the-art WER of **3.76**% shown in the bottom row with only a single system. This is a very interesting outcome, since it allows us to avoid performing beamforming which is not easy especially in situations with unknown or inexact array configurations, and postpone information fusion to a post-processing stage to combine recognition results.

F. A Preliminary Investigation with Real Recordings

For results in Table VI, either 3-gram or LSTM-LM was used in decoding. On this real task, unsupervised model adaptation of all connectionist parameters is performed, since mismatch between SimData training material and RealData testing set is very severe and originates not only from noise and reverberation but also from other factors related to the spoken utterances such as speaking style. Unsupervised adaptation accounts for using a seed ASR engine to decode un-transcribed data for a specific testing condition, e.g., new testing data. New acoustic models, which should better resemble the testing condition, are then built using these automatic transcripts as the label during acoustic model adaptation. A key advantage of the DNN framework is that model adaptation of the recognition module can be deployed by fine-tuning of the connectionist parameters with available testing data and corresponding target labels obtained through speech recognition with the unadapted ASR system. It may interesting to remark, that adaptation is extended to the dereverberation module in our system.

For comparison, we also report WERs shown in [28] using only the original multi condition training data along with the weighted prediction error (WPE) algorithm. The WPE approach with a 3-gram LM delivered an average WER of 27.80%, and 24.70% on ReadData in 1-channel and 8-channel scenarios, respectively, as shown in the first and fourth rows in Table VI. Average WERs of 27.06%, and 24.81%, respectively, were attained with our proposed multi-channel-condition joint training approaches with a 3-gram LM shown in the second and fifth rows. Rescoring with an advanced LSTM-LM allow us to attain a new state-of-the-art average WER of 16.40% (see third row) on RealData in the 1-channel scenario in contrast to the 17.40% WER reported in [28] in the same scenario but using the extended training data set. In the 8-channel scenario, LSTM-LM rescoring also boosts the recognition performance slightly, and an average WER of 22.65% was attained, as shown in the bottom row in Table VI. This was not as expected as reported in the 1-channel test case discussed earlier. Nonetheless, noise reduction seems to play an even more important role in multi-channel speech recognition than LSTM-LM. We believe a better ASR performance can be delivered by expanding the multi-channel training set to include better simulation data that are closer to real-world reverberant conditions than the training set we were currently using as done in some early studies [28] for SimData. This will also rely on extending the current room simulation formulation to new theories to be explored in the future.

On one hand, we see that we can obtain competitive results against the standard WPE technology with the same simulated training data and without using advanced LMs. On the other hand, we know from [28] that expanding the training set to cover more SNR conditions, the ASR performance on RealData can be significantly boosted. Nonetheless, top performance are again obtained by combining serval techniques in a pipeline and leveraging upon advanced LMs. We believe that a good theory is needed to generate better simulation training data closer resembling real-world conditions to extend our success in SimData to RealData evaluations.

V. CONCLUSION

Recent advances in deep learning with DNNs had spurred on new research efforts in DNN-based classification, such as automatic speech recognition and image object recognition. In this paper, we have instead focused on developing high-dimensional nonlinear regression approaches to a classical speech dereverberation problem. In this work, we have demonstrated through a comprehensive set of experiments that a better ASR performance is attained through the proposed joint training approach. In particular, joint training with more discriminative ASR features and improved DNN based language models allowed us to obtain a WER as low as of 4.46% with a single system. By adopting a new multi-channel-condition (MCC) joint learning scheme, coupled with MCC testing of all the 8-channel SimData, a further error reduction was observed with a new state-of-the-art WER of 3.76% with a single ASR system. Finally

we have also shown a preliminary yet promising study with the REVERB RealData, which leads us to believe that a good theory to generate better simulation training data for real-world conditions is needed.

REFERENCES

[1] A. Sankar and C.-H. Lee, "A maximum likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 3, pp. 190–202, May 1996.

[2] J. Benesty, S. Makino, and J. D. Chen, Eds, *Speech Enhancement*. Berlin, Germany: Springer, 2005.

[3] K. Kinoshita *et al.*, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2013, pp. 1–4.

[4] K. Kinoshita *et al.*, "A summary of the reverb challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 7, no. 1, pp. 1–19, 2016.

[5] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. London, U.K.: Springer, 2010.

[6] V. V. Digalakis, D. Rtischev, and L. G. Neumeye, "Speaker adaptation using constrained estimation of gaussian mixtures," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 357–366, Sep. 1995.

[7] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *Comput., Speech, Lang.*, vol. 12, pp. 75–98, 1998.

[8] X. Lei, J. Hamaker, and X. He, "Robust feature space adaptation for telephony speech recognition," in *Proc. INTERSPEECH*, 2006, pp. 773–776.

[9] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, 1995.

[10] J. Neto *et al.*, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," in *Proc. Eurospeech*, 1995, pp. 2171–2174.

[11] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–989, 1998.

[12] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[13] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 276–287, Mar. 2001.

[14] S. M. Siniscalchi, J. Li, and C.-H. Lee, "Hermitian polynomial for speaker adaptation of connectionist speech recognition systems," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 21, no. 10, pp. 2152–2161, Oct. 2013.

[15] P. Karanasou, Y. Wang, M. Gales, and P. Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors," in *Proc. Interspeech*, 2014, pp. 2180–2184.

[16] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 11, pp. 1938–1949, Nov. 2015.

[17] L. Samarakoon and K. C. Sim, "Factorized hidden layer adaptation for deep neural network based acoustic modeling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 12, pp. 2241–2250, Dec. 2016.

[18] Z. Huang, S. M. Siniscalchi, and C.-H. Lee, "Bayesian unsupervised batch and online speaker adaptation of activation function parameters in deep models for automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 64–75, Jan. 2017.

[19] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," in *Proc. IEEE*, vol. 88, no. 8, pp. 1241–1269, Aug. 2000.

[20] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[21] M. Wu and D. L. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 774–784, May 2006.

[22] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 534–545, May 2009.

[23] S. Mosayyebpour, M. Esmaeili, and T. A. Gulliver, "Single-microphone early and late reverberation suppression in noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 322–335, Feb. 2013.

[24] K. Han *et al.*, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015.

[25] O. Schwartz, S. Gannot, and E. Habets, "An expectation-maximization algorithm for multi-microphone speech dereverberation and noise reduction with coherence matrix estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1495–1510, Sep. 2016.

[26] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, vol. 66, no. 1, pp. 165–169 1979.

[27] W. C. Sabine, *Collected Papers on Acoustics*. London, U.K.: Harvard Univ. Press, 1922.

[28] M. Delcroix, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge," in *Proc. REVERB Challenge Workshop*, 2014.

[29] Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[30] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[31] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Inf. Process. Syst.*, 2001, pp. 556–562.

[32] T. V. J. F. Gemmeke and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2067–2080, Sep. 2011.

[33] H. Kallasjoki, J. Gemmeke, K. J. Palomäki, A. Beeston, and G. Brown, "Recognition of reverberant speech by missing data imputation and NMF feature enhancement," in *Proc. REVERB Challenge Workshop*, 2014.

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

[35] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[36] X. Xiao *et al.*, "The ntu-adsc systems for reverberation challenge 2014," in *Proc. REVERB Challenge Workshop*, 2014.

[37] M. Mimura, S. Sakai, and T. Kawahara, "Reverberant speech recognition combining deep neural networks and deep autoencoders augmented with a phone-class feature," *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 1, 2015, Art. no. 62.

[38] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 102–111, Jan. 2017.

[39] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer, 2001.

[40] M. Wolf and C. Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Commun.*, vol. 57, pp. 170–180, 2014.

[41] I. Himawan, P. Motlicek, S. Sridharan, D. Dean, and D. Tjondronegoro, "Channel selection in the short-time modulation domain for distant speech recognition," in *Proc. Interspeech*, 2015, pp. 741–745.

[42] P. Scalart, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf.*, 1996, pp. 629–632.

[43] Quatieri and T. F, *Discrete-Time Speech Signal Processing: Principles and Practice*. Delhi, India: Pearson Edu. India, 2002.

[44] A. Mohamed, G. E. Dahl, and G. E. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 14–22, Nov. 2012.

[45] V. Tyagi and C. Wellekens, "On desensitizing the mel-cepstrum to spurious spectral components for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. 529–532.

[46] B. D. van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

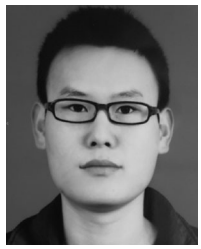
[47] K. Eneman and M. Moonen, "Multimicrophone speech dereverberation: Experimental validation," *EURASIP J. Audio Speech, Music Process.*, 2007, Art. no. 136.

[48] Y. Tachioka, T. Narita, F. J. Weninger, and S. Watanabe, "Dual system combination approach for various reverberant environments with dereverberation techniques," in *Proc. REVERB Challenge Workshop*, 2014, pp. 1–8.

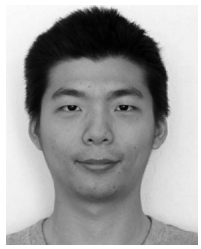
[49] B. Wu, K. Li, Z. Huang, S. M. Siniscalchi, M. Yang, and C.-H. Lee, "A unified deep modeling approach to simultaneous speech dereverberation and recognition for the reverb challenge," in *Proc. Hands-free Speech Commun. Microphone Arrays*, 2017, pp. 36–40.

[50] M. Sundermeyer, R. Schluter, and H. Ney, "LSTM neural networks for language modeling," in *Proc. Interspeech*, 2012, pp. 194–197.

[51] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Comput. Speech Lang.*, vol. 25, pp. 802–828, 2011.
[52] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4375–4379.
[53] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *Proc. Interspeech*, 2008, pp. 569–572.
[54] M.-Y. Hwang and X. Huang, "Shared-distribution hidden markov models for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 4, pp. 414–420, Oct. 1993.
[55] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 4960–4964.
[56] H. Kuttruff, *Room Acoustics*. London, U.K.: Spon Press, 2009.
[57] B. Wu, K. Li, M. L. Yang, and C.-H. Lee, "A study on target feature activation and normalization and their impacts on the performance of DNN based speech dereverberation systems," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016.
[58] A. Keshavarz, S. Mosayyebpour, M. Biguesh, T. A. Gulliver, and M. Esmaeili, "Speech-model based accurate blind reverberation time estimation using an LPC filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 6, pp. 1884–1893, Aug. 2012.
[59] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, no. 4, pp. 679–681, Aug. 1982.
[60] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 3761–3764.
[61] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech*, 2013, pp. 2345–2349.
[62] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 369–376.
[63] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.,* 2014, pp. II-1764–II-1772.
[64] A. Maas, Z. Xue, D. Jurafsky, and A. Ng, "Lexicon-free conversational speech recognition with neural networks," in *Proc. NAACL*, 2015, pp. 345–354.
[65] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, 2015, pp. 167–174.
[66] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. ICML'16*, New York, NY, USA, 2015, pp. 173–182.
[67] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech, recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 4945–4949.
[68] D. C. CIresan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," in *Proc. Int. Joint Conf. Neural Netw.*, 2011, pp. 1918–1921.
[69] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[70] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Representations*, 2015.
[71] K. Audhkhasi, A. Rosenberg, A. Seth, B. Ramabhadran, and B. Kingsbury, "End-to-end ASR-free keyword search from speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 4840–4844.
[72] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Proc. Interspeech*, 2014, pp. 616–620.
[73] J. Du, Y. Tu, L.-R. Dai, and C.-H. Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1424–1437, Aug. 2016.
[74] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAMO: A british english speech corpus for large vocabulary continuous speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, 1995, pp. 81–84.
[75] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments," in *Proc. IEEE Autom. Speech Recognit. Understanding*, 2005, pp. 357–362.
[76] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 269–277, 2008.
[77] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. Autom. Speech Recognit. Understanding*, 2011.
[78] M. Thomáš, "Statistical language models based on neural networks," Ph.D. dissertation, Brno Univ Technol, Brno, Czech Republic, 2012.



Bo Wu received the B.Eng. degree in electronic information engineering from Southwest University, Chongqing, China, in 2012. He is currently working toward the Ph.D. degree in National Laboratory of Radar Signal Processing, Xidian University, Xi'an, China. From September 2014 to October 2016, he was a visiting student in Center for Signal and Information Processing, Georgia Institute of Technology, Atlanta, GA, USA. His current research interests include signal processing, machine learning, and speech dereverberation.



Kehuang Li received the B.S. degree in information engineering, and the M.S. degree in communication and information system from Shanghai Jiao Tong University, Shanghai, China. He received the M.S. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA. He is currently working toward the Ph.D. degree at the Georgia Institute of Technology. His research interests include signal processing, machine learning, and speech recognition.

Fengpei Ge received the B.Eng. degree from Tianjin University, Tianjin, China, in 2005, and the Ph.D. degree in signal and information processing, and speech recognition, from the Institute of Acoustics, Chinese Academy of Sciences, Beijing, China, in 2010. She is an Associate Professor at the Institute of Acoustics, Chinese Academy of Sciences. From 2010 to 2013, she was an Assistant Professor at the Institute of Acoustics, Chinese Academy of Sciences. She received a scholarship under the State Scholarship Fund to pursue study as a visiting scholar at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

Zhen Huang (S'16) received the B.S. degree in electrical and computer engineering from Southeast University, Nanjing, China, in 2009. He received the dual M.S. degree in electrical and computer engineering from Shanghai JiaoTong University, Shanghai, China, and Georgia Institute of Technology, Atlanta, GA, USA, in 2012. His research interests include areas of speech recognition, deep learning, general machine learning, multimedia information retrieval, and image processing, currently more focus on deep learning based speech recognition and adaptation.

Minglei Yang received the B.Eng. degree in electronic engineering and the Ph.D. degree in signal and information processing both from Xidian University, Xi'an, China, in 2004 and 2009, respectively. Since June 2009, he has been working as an Associate Professor in the National Laboratory of Radar Signal Processing, Xidian University. His current research interests include but not limited to signal processing, parameter estimation, and polarization information processing.

Sabato Marco Siniscalchi (SM'16) received the Laurea and Doctorate degrees in computer engineering from the University of Palermo, Palermo, Italy, in 2001 and 2006, respectively. He is an Associate Professor at the University of Enna "Kore," Enna, Italy and affiliated with the Georgia Institute of Technology. He is currently an associate editor of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. His research interests include speech processing, in particular automatic speech and speaker recognition, and language identification.

Chin-Hui Lee (F'97) is currently a Professor in the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Before joining academia in 2001, he had 20 years of industrial experience ending in Bell Laboratories, Murray Hill, NJ, as a Distinguished Member of Technical Staff, and the Director of the Dialogue Systems Research Department. He has published more than 450 papers, and 30 patents, and was highly cited close to 30 000 times for his original contributions with an h-index of 65 on Google Scholar. He received numerous awards, including the Bell Labs President's Gold Award in 1998. He received the IEEE Signal Processing Society'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 received the International Speech Communication Association Medal in scientific achievement for "Pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition." He is a Fellow of ISCA.