IMPROVING ROBUSTNESS AGAINST REVERBERATION FOR AUTOMATIC SPEECH RECOGNITION

Vikramjit Mitra, Julien Van Hout, Wen Wang, Martin Graciarena, Mitchell McLaren, Horacio Franco, Dimitra Vergyri

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA
vikramjit.mitra@sri.com

ABSTRACT

Reverberation is a phenomenon observed in almost all enclosed environments. Human listeners rarely experience problems in comprehending speech in reverberant environments, but automatic speech recognition (ASR) systems often suffer increased error rates under such conditions. In this work, we explore the role of robust acoustic features motivated by human speech perception studies, for building ASR systems robust to reverberation effects. Using the dataset distributed for the "Automatic Speech Recognition In Reverberant Environments" (ASpIRE-2015) challenge organized by IARPA, we explore Gaussian mixture models (GMMs), deep neural nets (DNNs) and convolutional deep neural networks (CDNN) as candidate acoustic models for recognizing continuous speech in reverberant environments. We demonstrate that DNN-based systems trained with robust features offer significant reduction in word error rates (WERs) compared to systems trained with baseline mel-filterbank features. We present a novel time-frequency convolution neural net (TFCNN) framework that performs convolution on the feature space across both the time and frequency scales, which we found to consistently outperform the CDNN systems for all feature sets across all testing conditions. Finally, we show that further WER reduction is achievable through system fusion of n-best lists from multiple systems.

Index Terms— *time-frequency convolution nets, deep convolution networks, robust feature combination, robust speech recognition, reverberation robustness, system fusion.*

1. INTRODUCTION

With the introduction of deep learning techniques [1], ASR systems have seen a phenomenal reduction in error rates [2]. But ASR systems are quite sensitive to speech-signal degradations, such as reverberation, noise, and channel mismatch, which can result in significantly reduced speech recognition accuracy. ASR systems perform exceptionally well under matched conditions, but a subtle difference between the testing and training conditions reveals their vulnerability [3].

Reverberation is a major source of performance degradation for ASR systems [4], where performance degradation (usually represented by WER increases with the increase in reverberation time (usually represented by the RT60 value in seconds). Typically, the environment where the speech sample is collected defines the degree of reverberation and its effect on speech. Reverberation is usually the effect of multiple reflections of the source sound on the ambient enclosure. The time (typically in seconds) required for the reflections of a direct sound to decay to 60 dB is defined as the RT60 value of reverberation. Typically, the higher the RT60 values are, the more distorted the reverberated speech sounds, and vice versa. Such multiple reflections or reverberation seriously degrade speech-signal quality. Approaches to circumvent reverberation effects on speech are now an important research area, with microphone-array processing [5]; echo cancellation [6]; robust signal processing [7]; and speech enhancement [8] being major research thrusts.

Reverberation introduces acoustic mismatch between training and testing conditions, which usually degrades ASR performance. ASR systems trained purely on clean data (i.e., data without any distortion and/or artifacts) typically suffer a substantial increase in error rates when deployed in reverberant conditions. Training the ASR system with reverberated data can mitigate this effect. The results from different research groups at the 2014 REVERB Challenge workshop [4] indicated that using an increased diversity of reverberation conditions during multi-conditioned training usually improves the robustness of acoustic models by reducing acoustic-condition mismatch between the training and testing data.

In ASR systems, robustness against reverberation is usually improved through signal-processing techniques and de-reverberation strategies as explored in [9-14]. These studies demonstrated that using suitable acoustic features improves robustness against reverberation for ASR systems.

Recent advances in deep learning technology have redefined acoustic modeling in ASR systems, with more accurate discriminative learning techniques (such as neural networks) replacing traditional generative learning techniques (such as GMMs). DNNs have simplified many steps for ASR systems (e.g., primitive filterbank energies replace cepstral features [1]). Widely used speaker-normalization techniques, such as vocal tract length normalization (VTLN) [15], no longer offer significant gains in speech recognition accuracy, as DNNs can learn speaker-invariant data representations [16]. Specifically, it was observed [16] that VTLNs make much less impact on ASR accuracy for CDNNs [17] than for traditional DNNs. Recent results [16, 18] also showed that CDNNs are more robust to noise and channel degradations than DNNs. Traditionally, a single layer of convolutional filters are used on the input contextualized feature space to create multiple feature maps that, in turn, are fed to fully connected DNNs. However in [19], it was shown that adding multiple convolutional layers (usually up to two) potentially improves the performance of CDNN systems beyond their single-layer counterparts.

In this work, we explore a set of robust acoustic features in different acoustic modeling setups and analyze their performance in a speech recognition task under reverberant conditions. We compare the performance of traditional GMM-based acoustic

models with more recent DNN and CDNN architectures, and we demonstrate how deep learning can improve ASR performance. We analyze the number of hidden layers and the hidden layer size on a held-out development set to obtain the neural network parameters.

We present a modified version of the conventional CDNN that deploys two separate convolution layers, one layer operating across frequency, and the other layer operating across time, and we name this new variant the time-frequency convolution neural network (TFCNN). Finally, we demonstrate that using TFCNNs can further reduce ASR error rates compared to the best-performing CDNN systems.

Recently [20], i-vectors [21] have been used to perform speaker adaptation in DNNs. In this work, we explored both utterance-level and 20 second window-based i-vectors, and used them in addition to robust features for training the DNN systems. We used the data distributed through the ASpIRE 2015 challenge [22] to train and evaluate our systems.

2. TASK AND DATASET

The goal of the ASpIRE challenge was to perform the ASR task on native American English speakers where mismatch between the training and test conditions was quite high [22]. For the ASpIRE evaluation, each participant was expected to build an ASR system for conversational telephone English speech that would be robust to a variety of unknown acoustic environments and recording scenarios. The participants were given English Fisher conversational telephone speech (CTS) to train their models, and ad-hoc corruption of the training data was allowed. Evaluation for ASpIRE was performed with support from the MIT Lincoln Laboratory. During the submission process, the participants were required to provide written documentation and to deliver output from their Speech-to-Text system on the evaluation data to the ASpIRE organizers.

The Fisher-CTS dataset contained single-speaker utterances recorded at an 8 kHz sampling rate. The training corpus was artificially reverberated using 12 different room conditions (split equally among small, medium, and large rooms) with RT60s of approximately 0.5 and room signal-to-noise ratios (SNRs) between 10 to 20 dB using the setup of the REVERB2014 Challenge [4]. The noisy and reverberant training data (NR-train) was obtained by combining 25% of the training dataset after adding reverberation with 12.5% of the clean training data (mutually non-overlapping). The clean training data (CL-train) was comprised of 37.5% of the clean Fisher-CTS data. Unless otherwise mentioned, all acoustic models were trained with the NR-train data. For only two systems, we used the full, reverberated Fisher-CTS training data, where no clean data was used, and we name that training set the NR-train-full set.

The development dataset is partitioned into a development (dev) set and a development-test (dev-test) set containing real recordings distributed through the ASpIRE challenge. Because the dev data came with references, we artificially reverberated that data and used the reverberated data to evaluate the performance of our system. The performance on the dev-test set can only be obtained by uploading the ASR outputs to the ASpIRE challenge website or by emailing the organizers. Because the dev data came with manual segmentation, we used it to produce a *manually segmented* dev set, and created two versions by artificially reverberating one with a fixed RT60 of 0.5 (RT60_0.5) and the other with a fixed RT60 of 0.7 (RT60_0.7). In addition to the manually segmented dev data, we used a speech activity detector (SAD) [23] developed for noisy channel data under the DARPA RATS project (without further optimizing on reverberated data), to automatically segment the dev data (we call this as *SAD segmented* data). We also added reverberation to the dev data at two RT60 values (0.5s and 0.7s) and then performed SAD segmentation. In this paper, we report our results on the different versions of the dev, dev-test, and eval data, with the performances reported in terms of word error rates (WERs).

The evaluation data was collected and transcribed by the organizers specifically for the ASpIRE evaluation and was called the Mixer 8 Pilot corpus [22]. The data was recorded for Intelligence Advanced Projects Activity (IARPA) by the Linguistic Data Consortium (LDC) and transcribed by Appen Butler Hill. The data was collected by using multiple simultaneous microphones placed in a wide range of locations in seven different rooms (some classrooms and some office space) with various different shapes, sizes, surface properties, and noise sources. Speakers were also recorded from several different positions in each room. Two specific evaluation conditions were included: (a) *Single Microphone* condition and; (b) *Multiple Microphone* condition. This work focuses on only on the single microphone evaluation condition. In (a) the participants were allowed to train and test their systems on data from a single microphone only, where the single microphone (selected randomly) test data came from sessions recorded across seven different rooms

3. ACOUSTIC FEATURES

We explored an array of robust features for our experiments, and they are briefly outlined in this section.

3.1 Non-negative Matrix Factorization Enhanced Mel-Filterbank Features (NMF-MFB)

Using the approach outlined in [24], non-negative matrix factorization (NMF) was used to estimate relatively cleaner speech from the reverberated speech. The default NMF parameters as outlined in [24] were used in our experiments. We applied NMF to the training and testing data to produce their possibly enhanced versions and extracted 40-dimensional mel-filterbank (MFB) energies from them. These 40-D features were used as the NMF-MFB features in our experiments reported here.

3.2 Damped Oscillator Coefficients (DOC)

In DOC processing, the hair cells within the human ear are modeled as forced damped oscillators [25]. DOCs try to model the dynamics of the hair-cell oscillations to auditory stimuli within the human ear. The hair cells detect the motion of incoming sound waves and excite the neurons of the auditory nerves, which then transduce the relevant information to the brain. In DOC processing, the incoming speech signal is analyzed by a bank of gammatone filters that split the signal into bandlimited subband signals. We used 40 gammatone filters that were equally spaced on the equivalent rectangular bandwidth (ERB) scale. The bandlimited subband signals from these 40 gammatone filters served as the forcing functions to an array of 40 damped oscillators (more details in [25]) whose response was used as the acoustic feature. We analyzed the damped oscillator response by using a Hamming window of 26 ms with a frame rate of 10 ms. The power signal from the damped oscillator response was computed, then root compressed using the 15^{th} root, and the resulting 40-

dimensional features were used as the DOC feature in our experiments.

3.3 Normalized Modulation Coefficients (NMC)

Studies [26, 27] have shown that amplitude modulation (AM) of subband speech signals plays an important role in human speech perception and recognition. The NMC feature tries to capture and use the AM information from bandlimited speech signals. NMCs were obtained by using the approach outlined in [28], where the features are generated from tracking the AM trajectories of subband speech signals in a time domain by using a Hamming window of 26 ms with a frame rate of 10 ms. The speech signal was analyzed by using a time-domain gammatone filterbank with 40 channels equally spaced on the ERB scale. The subband signals were then processed by using a modified version of the discrete energy separation algorithm (DESA) (outlined in [28]), which produced instantaneous estimates of AM signals. The powers of the AM signals were then root compressed using the 15^{th} root. The resulting 40-dimensional feature vector was used as the NMC feature in our experiments.

3.4 Modulation of Medium Duration Speech Amplitudes (MMeDuSA)

MMeDuSA [29] is similar in essence to the NMC features, where it tracks the subband AM signals of speech by using a medium duration analysis window. On top of tracking the subband AM signals, MMeDuSA also tracks the overall summary modulation information. The summary modulation plays an important role in both tracking voiced speech and locating events such as vowel prominence/stress, etc. Unlike NMCs, MMeDUSA does not use the DESA algorithm to track the AM signals, but instead directly uses the nonlinear Teager energy operator [30] to crudely estimate the AM signal from the bandlimited subband signals. The MMeDuSA-generation pipeline used a time-domain gammatone filterbank with 40 channels equally spaced on the ERB scale. The MMeDuSA pipeline used a Hamming analysis window of ~51 ms with a 10 ms frame rate. The powers were root compressed, and the resultant information was used as the acoustic feature in our experiments. More details regarding MMeDuSA feature extraction can be obtained in [29].

3.5 Gammatone Filterbank (GFBs) Energies

Gammatone filters are a linear approximation of the auditory filterbank of the human ear. We used a time-domain implementation of the gammatone filters, where a bank of 40 gammatone filters equally spaced on the equivalent rectangular bandwidth (ERB) scale was used to analyze speech signal. The power of the bandlimited time signals from the gammatone filters were analyzed by using a Hamming window of ~26 ms with a 10 ms frame rate. The subband powers were then root compressed using the 15^{th} root, and the resulting 40-dimensional feature vector was used as the GFBs.

3.6 I-Vectors

I-vectors of 200 dimensions were extracted from a subspace trained on all the training data. The subspace was based on a universal background model (UBM) with 512 components trained with mel-frequency cepstral coefficients (MFCCs) (20 cepstra including C_0 with both Δ and $\Delta\Delta$ appended to produce 60D MFCCs). Speech activity detection (SAD) for extracting the i-vectors was based on a GMM SAD (using MFCCs with 13 cepstra including C_0 with both Δ and $\Delta\Delta$ appended) that was robust to microphone and telephone noise. The i-vector was extracted for a single whole conversation side of the data and appended to all stacked DOC features for the same conversation side, resulting in the DOC-IV features.

In addition to the utterance-level i-vectors, we also explored short-term dynamic i-vectors (we name these as sIV), which were based on 20 seconds of speech (as detected by the SAD). The 20-second i-vectors should better capture the dynamic nature of conversations (i.e., excitement, anger, disinterest, etc.) over time, compared to those extracted from the whole conversation side. The short-term i-vectors were appended with the NMC features, resulting in NMC-sIV features. Note that the i-vector extractor was trained using the CL-train data.

3.7 Max Kurtosis-Based Enhancement

Reverberation is typically caused by delayed reflections of sound from ambient surfaces. Reverberation usually affects the excitation signal, or in a linear prediction (LP) model, the LP residual. As reverberation effect increases, the kurtosis of the LP residual decreases [31]. The goal of the enhancement used in this processing was to maximize the kurtosis of the LP residual of reverberated speech, assuming that the LP residuals were unaffected by background contamination. An LMS-like adaptive filtering was used to maximize the kurtosis of the LP residual, and the resultant was used to generate the enhanced speech. The enhanced speech was used to produce DOC features, and we name the resulting features as DOC-Kurtosis.

Note that all filterbank features used in the experiments reported in this paper were mean and variance normalized before being fed to the acoustic models for training and testing. The i-vectors were length normalized and then appended with the other features.

4. ACOUSTIC MODEL

For acoustic modeling, we used traditional GMMs, DNNs, and CDNNs in our experiments. The GMM-HMM acoustic model training was performed by using SRI's DECIPHER® LVCSR system, which used 13 NMC Cepstral features (NMCCs) (these are the cepstral features obtained from discrete cosine transform (DCT) of the NMC features) and their Δs, Δ^2s, and Δ^3s. Global mean and variance normalization was performed on the acoustic features prior to acoustic model training, and heteroscedastic linear discriminant analysis (HLDA) transform was performed similar to [13]. The acoustic models were trained as cross-word triphone HMMs with decision-tree-based state clustering using a recipe similar to [13].

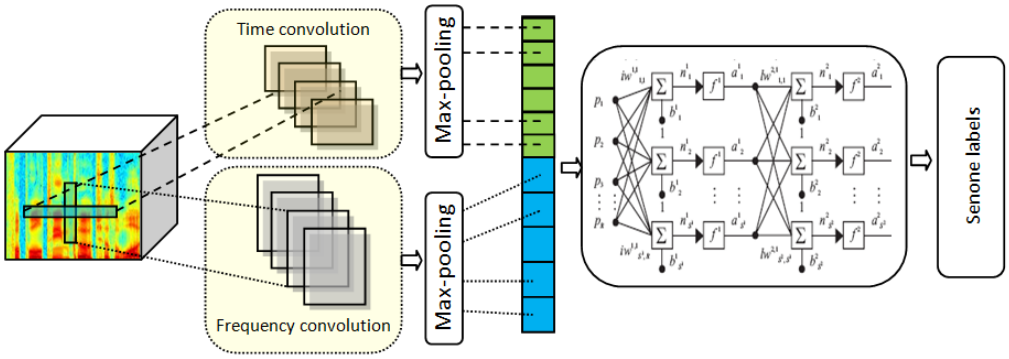

Figure 1. Block diagram showing time-frequency convolution neural nets (TFCNN). The top dotted block shows convolution filters working across time, and the bottom dotted block shows convolution filters across frequency. The max-pooled outputs of these convolution filters are fed to a fully connected 4-layered deep neural net.

To generate alignments to train the DNN and CDNN systems, a Kaldi Speech Recognition toolkit [40] based GMM-HMM model using MFCCs (standard 39 dimenisons) was trained using the CL-train data to produce the senone labels. The reverberated and noisy data of NR-train were time-aligned with their clean counterparts, such that the alignments from the clean data could be used as alignments for the noisy and reverberated training data. Altogether, the GMM-HMM system produced 7827 senones in our experiment. The DNN/CNN systems were trained using Theano.

The input layer of the DNN/CDNN systems was formed by using a context window of 15 frames (7 frames on either side of the current frame). The DNN/CDNN acoustic model was trained by using cross-entropy on the alignments from the GMM-HMM system, where NR-train was used. The input features were filterbank energy coefficients with a context of seven frames from each side of the center frame for which predictions were made. We used 200 convolution filters of size eight in the convolution layer and set the pooling size to three without overlap. Note that only one convolution layer was used in the CDNN. The resulting CDNN included four hidden layers with 1024 nodes each and an output layer with 7827 nodes representing the senones. The networks were trained by using an initial few iterations with a constant learning rate of 0.008, followed by learning-rate halving based on cross-validation error decrease. Training stopped either when no further significant reduction in the cross-validation error was noted or when the cross-validation error started to increase. Backpropagation was performed using stochastic gradient descent with a mini-batch of 256 training examples. The HMM decoding of the DNN/CNN lattices was performed using Kaldi [40].

We also explored the time-frequency convolution neural nets (TFCNN) [32], shown in Figure 1. In TFCNNs, two levels of convolution are performed on the input contextualized feature space, where a context of 17 frames was used. We used 75 filters to perform time convolution and 200 filters to perform frequency convolution. The filter band sizes were eight in both cases. A max-pooling over three samples was used for frequency convolution, while max-pooling over five samples was used for time convolution. The feature maps after both the convolution operations were concatenated and then fed to the fully connected neural net, which had 1024 nodes and four hidden layers. The delayed acoustic reflections in reverberation usually introduce distortion or artifacts along the time axis, and hence the motivation behind time-convolution and max-pooling across the time axis is to mitigate that distortion. Figure 1 briefly outlines the TFCNN architecture.

All training data was used for training the language model (LM). SRILM [33] was used to train the 4-gram LM by using modified Kneser-Ney smoothing, which produced about 650K 4-grams, 1.39M 3-grams, 2.46M 2-grams, and 38K 1-grams. We also used an approach that explores n-gram statistics to extract multiwords from the language model training data. The details of the approach can be found in [34]. In addition, a recurrent neural net (RNN)-based LM was also used. An RNN-LM [35] has a recursive structure that predicts a current word w_j given the previous word w_{j-1} and previous hidden state vector h_{j-1}. An RNN-LM can be learned by using backpropagation through time to maximize the log-likelihood of the training sentences. To take advantage of full sentence context, we employed a backward RNN-LM $p(w_j|w_{j+1},h_{j+1})$ trained with sentences in reverse word order. We used the same language model training text for the baseline word n-gram language models to train forward and backward RNN-LMs with 500 hidden nodes, and applied the forward and backward RNN-LMs to rescore n-best lists extracted from Kaldi lattices. RNN-LM scores were used for n-best ROVER.

5. RESULTS

The acoustic models were trained by using batch processing, for which no prior information about the speakers, room conditions, or background noise was used. Table 1 presents the results for the six systems broken down by the four feature types: (1) the MFCC-GMM system trained with the CL-train data (note that this is the only model that was trained only with CL-data in our experiments); (2) the NMCC-GMM system trained with the NR-train data; (3) the DNN/CDNN systems trained using mel-filterbank (MFB); and (4) the DNN/CDNN systems trained using the NMC features extracted from the NR-train data.

Table 1 presents the WERs from these six different systems and shows that the CDNN-based system performed much better than either of the GMM and DNN systems, which we have also observed in our earlier ASR experiments [13, 16, 36] using noise- and reverberation-corrupted speech data.

Table 1 show that the WERs for the CDNN/DNN systems were significantly lower than the GMM systems. It also shows that using robust features (NMCs for DNN/CDNN, and NMCC for GMM) reduced the WER significantly compared to those of the systems using mel-filterbank features (MFCC for GMM, and MFB

for DNN/CDNN). A relative 9% and 11% reduction in WER was noted for the DNN and CDNN systems, respectively, when NMCs replaced MFBs.

Table 1. WERs from different baseline systems using reverberated dev (*manually segmented* using transcription) data for decoding.

System	Feature	WER (%)		
		RT60-0.5	RT60-0.7	Avg.
GMM	MFCC	85.5	91.2	88.4
GMM	NMCC	72.3	79.9	76.1
DNN	MFB	66.1	70.6	68.4
CDNN	MFB	63.8	68.1	65.9
DNN	NMC	59.5	64.5	62.0
CDNN	NMC	**56.1**	**61.0**	**58.6**

In Table 2, we present the detailed results from all the features discussed in Section 3 for the DNN systems trained with the NR-train data. Table 2 shows that all but the NMF-MFB- and DOC-Kurtosis-based systems gave lower WERs for all the conditions compared to the MFB features. DOC performed the best for all the conditions, showing 10% or more absolute reduction in WER compared to the MFBs. Note that the forced damped oscillators [25] used in the DOC feature-generation pipeline have a long-term memory, whereas the other features treat speech as a piece-wise, independent signal. The DOCs' long-term memory might help them to efficiently cope with the temporal artifacts introduced by the background reverberations. Interestingly, the IV-based fused feature systems did not show any improvement beyond their individual counterparts, and hence, we did not use the IV-based fused features for training the CDNN systems. Table 3 shows the performance of the different features in the four-layered CDNN systems, trained with the NR-train data. Tables 2 and 3 indicate that the robust features help in reduction of WERs compared to baseline mel-filterbank features in noisy and reverberated conditions. An independent study [39] using multi-channel noisy data processed through beamforming and using MMeDuSA and DOC robust features, demonstrated similar reduction in WERs compared to beamformed MFB features.

Table 2. Dev (*Manually* segmented and *SAD* segmented) WERs from the DNN systems trained with different features.

Feature	WER (%) for dev				
	Manually segmented		*SAD segmented*		
	RT60-0.5	RT60-0.7	RT60-0.5	RT60-0.7	no_rev
MFB	66.1	70.6	69.2	73.2	67.5
NMC	59.5	64.5	61.3	65.5	47.9
DOC	**56.2**	**60.7**	**58.4**	**62.1**	**46.2**
GFB	58.6	63.7	60.7	64.9	47.3
MMeDuSA	57.2	61.5	59.3	62.6	48.4
NMF-MFB	69.6	73.8	70.6	73.8	58.4
DOC-Kurtosis	71.6	75.8	73.6	76.9	59.2
DOC-IV	58.0	61.0	62.9	65.2	48.4
NMC-sIV	62.9	65.2	68.8	68.0	49.1

From Table 3, we observe that the CDNN systems always produced lower WERs compared to their DNN counterparts for all the features. Note that, although the CDNN systems had only four hidden layers, the DNN systems had five layers, except the DOC-IV system, which had six layers. Also note that the CDNN systems had only one convolution layer applied to the first layer of the network. Tables 2 and 3 confirm our prior observations from ASR experiments on noise- and channel-degraded speech [13, 16, 36], where we observed that (1) the CDNN systems always perform better than the DNN systems and (2) the robust features always gave a sizeable performance gain compared to the MFB features.

Table 3. Dev (*Manually* segmented and *SAD* segmented) WERs from the CDNN systems trained with different features.

Feature	WER (%) for dev				
	Manually segmented		*SAD segmented*		
	RT60-0.5	RT60-0.7	RT60-0.5	RT60-0.7	no_rev
MFB	63.8	68.1	67.7	71.8	66.6
NMC	56.1	61.0	60.9	66.7	45.7
DOC	**54.4**	**58.8**	**56.8**	**60.4**	**44.9**
GFB	56.5	61.0	58.2	61.6	45.6
MMeDuSA	54.8	58.7	57.3	60.8	47.1
NMF-MFB	67.8	72.1	69.3	72.0	56.5
DOC-Kurtosis	70.3	74.7	72.2	75.3	57.6

We evaluated the top-performing features from our CDNN experiments on TFCNNs. The TFCNNs used four hidden layers with 1024 neurons in each layer. Table 4 shows the WERs obtained from the TFCNN systems.

Table 4. Dev (*Manually* segmented and *SAD* segmented) WERs from the TFCNN systems trained with different features.

Feature	WER (%) for dev				
	Manually segmented		*SAD segmented*		
	RT60-0.5	RT60-0.7	RT60-0.5	RT60-0.7	no_rev
NMC	55.8	60.1	57.7	61.6	45.3
DOC	**53.7**	**58.0**	**56.1**	**59.8**	**44.1**
GFB	55.6	59.7	57.6	60.9	45.4

Table 4 shows that the TFCNN systems always gave lower WERs compared to their CDNN counterparts for all features, indicating that time-convolution helps increase the robustness of the acoustic models.

Next, we performed n-way ROVER [37] combination of all the GMM, DNN, CDNN and TFCNN systems trained in this work. We observed consistent improvement in performance from system combination, and the results are shown in Table 5. The rationale behind system combination is that different portions of the n-best lists from different sub-systems may be correctly recognized, and these portions can be combined to produce a better hypothesis by using the system-combination technique called ROVER [37]. Stolcke et al. [38] extended ROVER to n-best lists from multiple systems. We applied the n-best ROVER implemented in the SRILM toolkit [33], to n-best lists generated for each utterance from the multiple subsystems. Note that in Table 5, the results from the DOC DNN, CDNN, and TFCNN systems have improved from those reported in Tables 2, 3, and 4. This is because, to have a fair comparison with the ROVER results, we ran forced alignment with the DOC DNN and CDNN 1-best hypotheses and the ROVER output using a PLP-GMM model trained on English broadcast news audio data, and then scored the CTM files against the STM references. The improvement happened because the PLP-GMM model could discard some hypothesized words in the hypothesis that could not align well with the audio, which resulted in reduced insertions and, hence, better WERs. In table 5 we present two system fusion results from 3-way ROVER combination, where ROVER 1 consists of NMC-CDNN, MMeDuSA-CDNN and NMC-sIV-DNN systems and ROVER 2 consists of DOC-TFCNN,

MFB-CDNN and DOC-CDNN systems. Note that we have exhaustively performed up to 3-way ROVER combinations and are presenting the top two systems from that experiment in Table 5.

Table 5. SAD segmented dev WERs from the best DNN, CDNN, and TFCNN systems, and the best n-way ROVER combinations.

System	WER (%) for SAD segmented dev		
	RT60-0.5	RT60-0.7	no_rev
DOC-DNN	56.6	61.1	42.6
DOC-CDNN	55.3	59.5	41.4
DOC-TFCNN	54.4	58.9	40.7
3-way ROVER 1	53.4	**57.0**	**40.5**
3-way ROVER 2	**53.3**	57.5	40.6

As evident from the Tables 2, 3, and 4, the DOC-TFCNN system was the top-performing system and, hence, was selected as one of the candidate systems during ROVER combination. Note that although the DOC-IV system did not perform as accurately as some systems using other features, it was surprisingly found to help during ROVER combination. We also explored using RNN-LM during system rescoring and observed that for the DOC-CDNN system, it reduced the WER by 0.2% to 0.5% absolute compared to the standard LM.

Finally in table 6 we show the recognition performance on the dev-test and eval datasets. Our best system on the eval data came from a 4-way ROVER combination (shown in table 6), where the subsystems were DOC-CDNN, NMC-CDNN, NMF-MFB-CDNN and DOC-IV-DNN systems. Interestingly, we observed (see table 6) that the performances in dev-test and eval datasets did not correlate well, for example the 3-way ROVER-1 system gave lowest WER amongst the 3 ROVER combined systems shown in table 6, but it produced the highest WER on eval data. On the other hand the best performing system, which is the 4-way ROVER combined system, gave lower WER on eval data but highest WER on dev-test. This may indicate that the eval data is much different than either of the dev and dev-test datasets and hence performing system fusion or system selection based on dev-test or dev datasets may not be optimal for the ASpIRE task.

Table 6. dev-test and eval WERs from the top performing systems.

System	WER (%)	
	dev-test	eval
DOC-TFCNN	39.5	51.9
3-way ROVER 1	38.7	51.6
3-way ROVER 2	38.8	51.3
4-way ROVER	39.4	50.7

6. CONCLUSION

In this paper, we presented our results for IARPA's ASpIRE Challenge evaluation. Using artificially reverberated development data we demonstrated that CDNNs can significantly reduce WERs compared to other standard acoustic modeling techniques such as GMMs or DNNs. We presented a time-frequency CNN (TFCNN) and demonstrated that under reverberant conditions TFCNNs gave lower WERs than traditional CNNs. We investigated several robust acoustic features and found that they help in reducing the WERs compared to the baseline MFB energies. We obtained further reduction in WER by combining multiple systems. Interestingly, we found that even if the NMC-IV-DNN system produced quite high WERs compared to other systems, it was selected in ROVER combination, indicating that suboptimal systems are relevant for system combination, as they provide sufficient complementary information with respect to the top-performing systems. With ROVER, we observed 2%, 3%, and 0.5% relative reductions in WERs compared to the best-performing DOC-TFCDNN system for the SAD segmented reverberated dev data with RT60s 0.5 and 0.7, and the SAD segmented non-reverberated dev data, respectively.

From our experiments, we observed convincingly that the CDNN systems always performed better than the DNN. Adding an additional convolution layer for performing time-convolution was found to be useful. Studies [19] have shown that using multiple convolution layers typically improves ASR performance compared to using only one layer and in a separate study [32] we have validated this fact on a different reverberated speech dataset. In the future, we wish to explore using multiple convolution layers and explore fusing that with the time convolution presented in this work. The NMF-based speech enhancement used in our experiments was built specifically to combat reverberation effects; however, we did not observe that this enhancement offered sufficient performance improvement compared to other features. One possible reason for this performance gap might be because NMF distorts the speech spectra, which, in turn, results in less-sharp models.

Finally, we observed that the WERs from the dev-test and eval datasets for different ROVER combined system did not correlate well. This indicates that system fusion and system selection based on dev and dev-test may have been a poor choice. This may also indicate that the use of suitable adaptation techniques may help in improving WERs on eval data. In future we intend to explore adaptation techniques and report results.

7. ACKNOWLEDGMENT

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0037. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

8. REFERENCES

[1] A. Mohamed, G.E. Dahl and G. Hinton, "Acoustic Modeling Using Deep Belief Networks," *IEEE Trans. on ASLP*, vol. 20, no. 1, pp. 14 –22, 2012.

[2] F. Seide, G. Li and D. Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks," *Proc. of Interspeech*, 2011.

[3] M. Seltzer, D. Yu, and Y. Wang, "An Investigation of Deep Neural Networks for Noise Robust Speech Recognition," *Proc of ICASSP*, 2013.

[4] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot and B. Raj, "The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[5] M. S. Brandstein and D. B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. New York: Springer Verlag, 2001.

[6] R. Martin and P. Vary, "Combined Acoustic Echo Cancellation, Dereverberation and Noise Reduction: A Two Microphone Approach," *Journal of Annales des Télécommunications*, vol. 49, iss. 7–8, pp. 429–438, 1994.

[7] K. Ohta and M. Yanagida, "Single Channel Blind Dereverberation Based on Auto-Correlation Functions of Frame-Wise Time Sequences of Frequency Components," *Proc. of IWAENC*, pp. 1–4, 2006.

[8] M. Wu and D.L. Wang, "A Two-Stage Algorithm for One-Microphone Reverberant Speech Enhancement," *IEEE Trans. Aud. Speech & Lang. Process.,* vol. 14, no. 3, pp. 774–784, 2006.

[9] A. Sehr and W. Kellermann, "A New Concept for Feature-Domain Dereverberation for Robust Distant-Talking ASR," *Proc. of ICASSP*, pp. 369–372, 2007.

[10] M. Delcroix and S. Watanabe, "Static and Dynamic Variance Compensation for Recognition of Reverberant Speech with Dereverberation Preprocessing," *IEEE Trans. on Aud. Speech & Lang. Process.,* vol. 17, no. 2, pp. 324–334, 2009.

[11] Md. J. Alam, V. Gupta, P. Kenny, P. Dumouchel, "Use Of Multiple Front-Ends and I-Vector-Based Speaker Adaptation for Robust Speech Recognition," in *Proc. of REVERB Challenge*, 2014.

[12] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, I. Nobutaka, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, "Linear Prediction-Based Dereverberation with Advanced Speech Enhancement and Recognition Technologies for the REVERB Challenge," in *Proc. of REVERB Challenge*, 2014.

[13] V. Mitra, W. Wang, Y. Lei, A. Kathol, G. Sivaraman, C. Espy-Wilson, "Robust Features and System Fusion for Reverberation-Robust Speech Recognition," in *Proc. of REVERB Challenge*, 2014.

[14] V. Mitra, J. Van Hout, M. McLaren, W. Wang, M. Graciarena, D. Vergyri and H. Franco, "Combating Reverberation in Large Vocabulary Continuous Speech Recognition," *Proc. of Interspeech 2015*.

[15] P. Zhan and A Waibel, "Vocal Tract Length Normalization for LVCSR," in *Tech. Rep. CMU-LTI-97-150*. Carnegie Mellon University, 1997

[16] V. Mitra, W. Wang, H. Franco, Y. Lei, C. Bartels, M. Graciarena, "Evaluating Robust Features on Deep Neural Networks for Speech Recognition in Noisy and Channel Mismatched Conditions," in *Proc. of Interspeech*, 2014.

[17] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, "Applying Convolutional Neural Networks Concepts to Hybrid NN-HMM Model for Speech Recognition," *Proc. of ICASSP*, pp. 4277 –4280, 2012.

[18] J. van-Hout, V. Mitra, Y. Lei, D. Vergyri, M. Graciarena, A. Mandal and H. Franco, "Recent Improvements in SRI's Keyword Detection System for Noisy Audio," in *Proc. of Interspeech*, pp. 1727-1731, Singapore, 2014.

[19] T. Sainath, A. Mohamed, B. Kingsbury and B. Ramabhadran, "Deep Convolutional Neural Network for LVCSR," *Proc. of ICASSP*, 2013.

[20] G. Saon, H. Soltau, D. Nahamoo and M. Picheny, "Speaker Adaptation of Neural Network Acoustic Models Using I-vectors," in *Proc. ASRU*, 2013.

[21] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. on Speech and Audio Processing*, 2011, 19, 788–798.

[22] M. Harper, "The Automatic Speech Recognition in Reverberant Environments (ASpIRE) Challenge," *Proc. of ASRU*, 2015.

[23] M. Graciarena, A. Alwan D. Ellis, H. Franco, L. Ferrer, J.H.L. Hansen, A. Janin, B-S. Lee, Y. Lei, V. Mitra, N. Morgan, S.O. Sadjadi, T.J. Tsai, N. Scheffer, L.N. Tan and B. Williams, "All for One: Feature Combination for Highly Channel-Degraded Speech Activity Detection," *Proc. of Interspeech*, pp. 709–713, Lyon, 2013.

[24] K. Kumar, R. Singh, B. Raj, R. Stern, R., "Gammatone Sub-Band Magnitude-Domain Dereverberation for ASR," *Proc. of ICASSP*, pp. 4604–4607, 2011.

[25] V. Mitra, H. Franco and M. Graciarena, "Damped Oscillator Cepstral Coefficients for Robust Speech Recognition," *Proc. of Interspeech*, pp. 886–890, 2013.

[26] R. Drullman, J. M. Festen, and R. Plomp, "Effect of Reducing Slow Temporal Modulations on Speech Reception," *J. Acoust. Soc. of Am*., vol. 95, no. 5, pp. 2670–2680, 1994.

[27] O. Ghitza, "On the Upper Cutoff Frequency of Auditory Critical-Band Envelope Detectors in the Context of Speech Perception," *J. Acoust. Soc. of America*, vol. 110, no. 3, pp. 1628–1640, 2001.

[28] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized Amplitude Modulation Features for Large Vocabulary Noise-Robust Speech Recognition," *Proc. of ICASSP*, pp. 4117–4120, 2012.

[29] V. Mitra, H. Franco, M. Graciarena, D. Vergyri, "Medium Duration Modulation Cepstral Feature For Robust Speech Recognition," *Proc. of ICASSP*, Florence, 2014.

[30] H. Teager, "Some Observations on Oral Air Flow During Phonation," in *IEEE Trans. ASSP*, pp. 599–601, 1980.

[31] B. Yegnanarayana and P.S. Murthy, "Enhancement of Reverberant Speech Using LP Residual Signal," *IEEE Trans. Speech and Aud. Processing*, 8(3), pp. 267–281, 2000.

[32] V. Mitra, H. Franco, "Time Frequency Convolution Nets for Robust Speech Recognition," submitted to ASRU 2015.

[33] A. Stolcke, "SRILM—An Extensible Language Modeling Toolkit," *Proc. of ICSLP 2002*, pp. 901–904, 2002.

[34] X. Lei and W. Wang and A. Stolcke, "Data-Driven Lexicon Expansion for Mandarin Broadcast News and Conversation Speech Recognition," *Proc. of ICASSP*, 2009.

[35] T. Mikolov, S. Kombrink, D. Anoop, L. Burget, and J. Cernocky, "RNNLM—Recurrent Neural Network Language Modeling Toolkit," *Proc. of ASRU*, 2011.

[36] V. Mitra, W. Wang and H. Franco, "Deep Convolutional Nets and Robust features for Reverberation-Robust Speech Recognition," in *Proc. of SLT*, pp. 548–553, 2014.

[37] J. G. Fiscus, "A Post-Processing System to Yield Reduced

Word Error Rates: Recognizer Output Voting Error Reduction. (ROVER)," *Proc. of ASRU*, pp. 347–354, 1997.

[38] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V.R. Rao Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, W. Weng, J. Zheng, "The SRI March 2000 Hub-5 Conversational Speech Transcription System," *Proc. of NIST Speech Transcription Workshop*, 2000.

[39] T. Hori, Z. Chen, H. Erdogan, J.R. Hershey, J. Le Roux, V. Mitra, S. Watanabe, "The MERL/SRI system for the 3rd Chime Challenge using Beamforming, Robust feature extraction, and Advanced Speech Recognition," submitted to ASRU 2015.

[40] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlıcek, Y. Qian, P. Schwarz et al., "The kaldi speech recognition toolkit," in Proc. ASRU, 2011.