IMPROVING SPEECH RECOGNITION IN REVERBERATION USING A ROOM-AWARE DEEP NEURAL NETWORK AND MULTI-TASK LEARNING

*Ritwik Giri**

University of California, San Diego
rgiri@ucsd.edu

Michael L. Seltzer, Jasha Droppo, Dong Yu

Microsoft Research, Redmond
{mseltzer,jdroppo,dongyu}@microsoft.com

ABSTRACT

In this paper, we propose two approaches to improve deep neural network (DNN) acoustic models for speech recognition in reverberant environments. Both methods utilize auxiliary information in training the DNN but differ in the type of information and the manner in which it is used. The first method uses parallel training data for multi-task learning, in which the network is trained to perform both a primary senone classification task and a secondary feature enhancement task using a shared representation. The second method uses a parameterization of the reverberant environment extracted from the observed signal to train a room-aware DNN. Experiments were performed on the single microphone task of the REVERB Challenge corpus. The proposed approach obtained a word error rate of 7.8% on the SimData test set, which is lower than all reported systems using the same training data and evaluation conditions, and 27.5% on the mismatched RealData test set, which is lower than all but two systems.

Index Terms— Multi-task learning, deep neural network, reverberation, room impulse response.

1. INTRODUCTION

Substantial improvements in speech recognition accuracy have been made in recent years using acoustic models based on deep neural networks (DNN). Despite this progress, speech recognition in distant-talking scenarios remains a significant challenge [1]. In such scenarios, the speech signal must be captured by one or more microphones located at a distance from the user, which makes it susceptible to distortion from additive noise and reverberation.

Many methods of dereverberation have been proposed that operate on the signal or the recognition features, e.g. [2, 3, 4]. Recently, the REVERB Challenge was held which provided a common corpus and evaluation framework to benchmark various approaches to improving speech recognition in far-talking scenarios [5]. Among the top performing systems, a wide variety of approaches were used, including a linear prediction-based dereverberation algorithm [6], a Long Short term Memory (LSTM)-based acoustic model, [7], learning-based speech enhancement [8], and the use of multiple front-ends and i-vector speaker adaptation [9].

In this paper we are interested in making the DNN inherently more robust to reverberation and propose two methods to do so. Both methods are similar in that they exploit auxiliary information in training the DNN but differ in the type of information and the way it is used by the network. In the first method, the DNN is trained with multi-task learning, in which the network uses a shared representation to learn to both classify the observations into senones and perform feature enhancement. The second method is inspired by recent work in noise-aware training [10] and i-vector features for speaker adaptation [11]. In this work, a feature vector that characterizes the reverberation is extracted from the signal and input to the network, to create a "room-aware" DNN. These features are extracted from an estimate of the room impulse response obtained via non-negative matrix factorization (NMF) [4]. Both methods were evaluated using the REVERB Challenge corpus, and significant improvements over a conventional DNN were observed. Furthermore, the proposed approach outperformed all reported systems that used the same training data and evaluation conditions on the SimData test set and all but one system on the RealData test set.

The rest of the paper is organized as follows. In Section 2, multi-task learning is introduced and we show how it can be applied to DNN-based acoustic models to improve robustness to reverberation and noise. In Section 3, we show how features that characterize the reverberation in the observed signal can be used to create a room-aware DNN. The efficacy of these two approaches in shown through a series of experiments in Section 4, and finally, we present our conclusions in Section 5.

2. MULTI-TASK LEARNING

Multi-task learning is a technique in which a model is trained to solve multiple tasks simultaneously, using a shared set of parameters. It is an approach to inductive transfer, or transfer learning, in which knowledge gained from one task can be applied to other tasks [12, 13]. The idea of multi-task learning is that internal representations learned for one task can be helpful for the other tasks, and vice versa. By learning multiple tasks in parallel, the model can learn additional information about the domain using the training signals of the related tasks. MTL allows the internal features discovered by one task to be used by the other tasks and also enables the learning of representations that would not have been discovered by training the network on any of the tasks in isolation.

In this work, the primary task is senone classification, and we use feature enhancement as the additional task. Specifically, the second task is to minimize the squared error between the observed features that are corrupted by reverberation and noise and the underlying clean speech features. If $\mathbf{y}$ represents the observed reverberant and noisy speech feature vectors and $\mathbf{x}$ represents the underlying clean speech feature vectors, the MTL objective function used to train this model is

$$\mathcal{J} = \alpha \sum_t p(s|\mathbf{y}_t) + (1-\alpha) \sum_t (\hat{\mathbf{x}}_t - \mathbf{x}_t)^2 \qquad (1)$$

*This work was performed while the author was an intern at Microsoft Research.

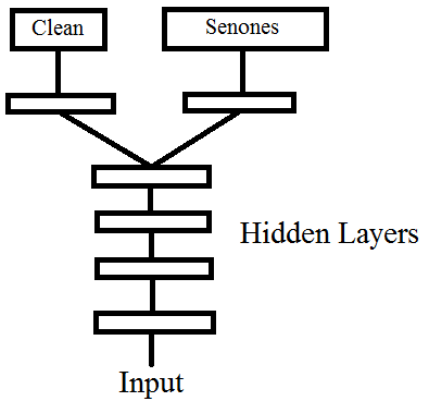

Fig. 1. Network architecture used for multi-task learning

where the first term is the primary cross entropy task and the second term is the secondary feature enhancement task, and α is the weight parameter which determines how much importance the secondary task should get.

In original multi-task learning algorithm, the entire network was shared among the tasks and only the output weights between the final hidden layer and the network outputs were different. In this work we chose to share the lower hidden layers of the network and have a single separate higher hidden layer for each task. This is shown in Figure 1.

Multi-task learning can be viewed as a form of regularization. Consider that there are several ways the network can learn to properly classify the observed reverberant training data. For example, with enough capacity, the network can simply memorize the training data. By using MTL with an enhancement task, the network must denoise and dereverberate the features while trying to classify them. That is, the network is learning that representations good for producing clean speech should be easier to classify. One significant advantage of multi-task learning is that it only impacts training complexity. At runtime, only the primary task is performed and the parameters associated with the secondary tasks can be discarded.

In Table 1, the results of a preliminary experiment on the development set of the REVERB Challenge corpus to demonstrate the performance of multi-task learning. The table compares the performance a DNN trained with conventional single task learning using a cross-entropy objective function to a network trained with the proposed multi-task learning strategy. The results for different values of α in (1) are shown. We can see that when the test data is similar to the training data (SimData), MTL is highly beneficial, while negligible improvement is observed when there is significant mismatch between the test and training data (RealData). The value of $\alpha = 0.91$ that gave the best results will be used in all further experiments in Section 4.

3. ROOM-AWARE DNN

Recent work has shown that the performance of DNN-based acoustic models can be improved by augmenting the standard spectral features typically input to the network with auxiliary information extracted from the signal or environment. For example, in noise-aware training (NAT), improvements were obtained by appending an estimate of the additive noise corrupting the utterance to the log mel filterbank features [10]. Similarly, improvements have been obtained by augmenting the input features with an i-vector [11] or learned code [14] that characterizes the speaker. In this work, our goal is to provide the network with a characterization of reverberation in which the speech was captured in order to build room-awareness into the model.

Table 1. WER on the REVERB Challenge development set

Model	SimData WER (%)	RealData WER (%)
Baseline	13.63	32.12
MTL ($\alpha = 0.85$)	12.72	32.52
MTL ($\alpha = 0.9$)	12.58	32.48
MTL ($\alpha = 0.91$)	**12.41**	**31.97**
MTL ($\alpha = 0.95$)	12.70	32.53

3.1. Characterizing the room

Unlike noise estimates for noise robustness or i-vectors for speaker adaptation, there is no standard method of encoding the important features of a room. However, there are a number of metrics that describe various aspects of reverberation [2]. In this work, we focus on two measures in particular. The first is the reverberation time, which reflects the time it takes the energy of an impulse to decay 60 dB. Reverberation time, denoted as T_{60}, can be easily measured from a measured room impulse response by plotting its energy decay curve (EDC).

If this is not possible, it must be estimated blindly from the observed signal itself. This remains a difficult research challenge, particularly in the presence of noise [15]. Furthermore, because the acoustic properties of a given room are typically frequency dependent, it is often beneficial to evaluate the reverberation time in several subbands. Knowledge of the reverberation time is crucial for several dereverberation signal processing approaches, and the difficulty of blindly estimating it limits the widespread use of these algorithms.

While T_{60} is an important feature to describe reverberation, it is independent of the position of the source and the microphone. Therefore, it does not provide any indication of the amount reverberant energy that is in the captured signal compared to the desired direct path signal. This can be obtained from the direct-to-reverberant ratio (DRR), which is the ratio of the energy in the direct path to the energy from all reflected paths that cause the reverberation. If n_d denotes the sample that divides the direct path of the room impulse response from the reverberant part, the DRR can be computed as

$$DRR = \log_{10} \frac{\sum_0^{n_d} h^2(n)}{\sum_{n_d+1}^{\infty} h^2(n)} \quad (2)$$

Typically, for a given room, the DRR is lower as user moves farther from the microphone. Note that DRR is closely related to measures such as $C30$ and $C50$, where n_d is fixed at 30 ms or 50 ms respectively.

3.2. Estimating the room parameters

Blind extraction of these room parameters from an utterance is a challenging problem. State-of-the-art signal processing methods of extracting T_{60} from an utterance do not perform accurately in the presence of noise. In this work we have used a method based on non-negative matrix factorization, inspired by [4]. This approach estimates a non-negative representation of the clean speech signal

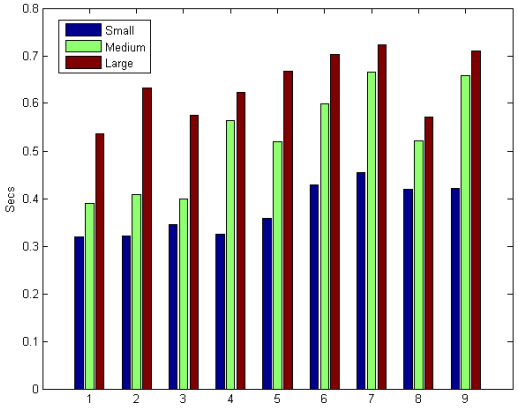

Fig. 2. T_{60} estimates

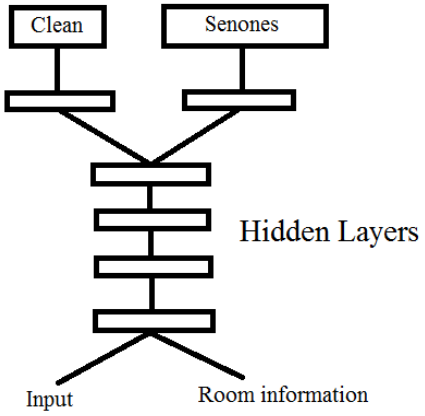

Fig. 3. Room-aware DNN with multi-task learning

and the room impulse response directly from the reverberant speech. If we define the magnitude of the reverberant observation as $Y_k[n]$, where k represents the frequency index and n represents the frame index, we assume that

$$Y_k[n] = X_k[n] * H_k[n] \tag{3}$$

where $H_k[n]$ is the magnitude representation of the room impulse response, $X_k[n]$ is the magnitude of the underlying clean speech signal, and $*$ represents convolution. Note that this relationship is approximate because of the error in representing reverberation as a magnitude-domain convolution and the possible presence of noise in the observation. To obtain an estimate of H_k, the following objective function is minimized with suitable non-negativity and sparsity constraints, which leads to NMF-style multiplicative updates [4],

$$E_k = \sum_n (Y_k[n] - \sum_m X_k[m]H_k[n-m])^2 + \lambda \sum_n (X_k[n])^p$$

where,

$$X_k[n] \geq 0, H_k[n] \geq 0, \sum_n H_k[n] = 1$$

After minimizing E_k, we obtain $\hat{H}_k$, an estimate of the magnitude representation of the room impulse response for frequency bin k. Because we are using mel-scale features in our DNN acoustic model, we perform this NMF estimation on the mel-scale representation of the magnitude spectrum. This produces a magnitude-domain estimate of the room impulse response in each of the mel subbands. The estimate of T_{60} in each mel subband was obtained by estimating by constructing the EDC of the corresponding impulse response and using Schroeder's backward integration. The result was one T_{60} estimate for each of the mel subbands, 24 in our case. From these 24 estimates, 9 were selected as most discriminative using the training data. To see if these estimates effectively convey information about the room, we clustered the T_{60} estimates from a set of training data from 3 different room sizes. Figure 2 shows the cluster centroids of the 9 selected mel subbands for the three room sizes: large, medium, and small. As the figure shows, the NMF-based estimates of the room impulse response are providing T_{60} estimates that can differentiate between different sized rooms.

To estimate the DRR, we averaged the estimated room impulse responses obtained from the 24 mel subbands and then compared the energy in the first frame to the energy in the remainder. Thus, the final feature vector used to characterize the room had 10 dimensions, consisting of 9 subband T_{60} estimates and an estimate of the DRR.

This 10-dimensional room characterization vector was then used as a feature in addition to the usual context window of log mel filterbank features. Note that this room vector was estimated for each utterance individually. This room-aware DNN is combined with the multi-task learning strategy described in Section 2 to create the network architecture shown in Figure 3.

4. EXPERIMENTAL EVALUATION

To evaluate the efficacy of the proposed approaches, we performed a series of experiments on the REVERB Challenge corpus [5]. This corpus is a farfield version of the WSJCAM0 corpus. The multi-condition training data was created by convolving the utterances from the WSJCAM0 training set with a series of impulse responses recorded from different several rooms. This reverberant speech was then corrupted by additive noise captured in these rooms at an SNR of 20 dB. There are two test sets for evaluation. The first set, called SimData, is created in the same manner as the training set, although the rooms (and hence the impulse responses and noise samples) are different. The SimData test set includes three different rooms, with each room having a near and far condition, which reflects the distance from the user to the microphone. The second set, called RealData, is a subset of the MC-WSJ-AV corpus, and consists of utterances from WSJ captured from real users in a meeting room. This set is recorded in a single room but also has a near and far condition. A development set of both SimData and RealData is also provided. In the original REVERB Challenge, there were separate evaluation tracks for one microphone, two microphones, and eight microphones. All the experiments in this work use the data from one microphone only.

In all experiments, the primary features input to the neural network were 24-dimensional log mel filterbank coefficients, along with their delta and delta-delta coefficients. Utterance-level mean normalization and corpus-based variance normalization were applied. A symmetric context window of 11 frames was input to the network, resulting in an input vector with a dimensionality of 792. There were 3168 senones in the system and the labels for the training data were derived from forced alignment of the clean training data using

Table 2. WER over Evaluation data

Model	Sim Room 1 Near	Sim Room 1 Far	Sim Room 2 Near	Sim Room 2 Far	Sim Room 3 Near	Sim Room 3 Far	Sim Avg	Real Room 1 Near	Real Room 1 Far	Real Avg
Baseline	5.78	6.73	7.03	12.26	8.52	13.38	8.95	27.56	29.00	28.28
MTL ($\alpha = 0.91$)	5.91	6.35	7.09	11.52	7.76	12.80	8.57	26.92	28.49	27.70
MTL+Room information	5.61	6.37	7.12	11.71	7.78	12.18	8.45	27.79	28.33	28.06
MTL+Room information + PT	5.03	6.29	6.40	10.66	6.91	11.36	**7.77**	27.63	27.38	**27.50**

a GMM-HMM system trained with maximum likelihood. A single decoding pass was performed using the standard WSJ trigram language model. All neural networks were trained using the Computational Network Toolkit (CNTK) [16].

The baseline system was a 5-layer DNN with 2048 sigmoidal hidden units per layer. The network was trained using a cross entropy objective function with model parameters randomly initialized. Training was performed using stochastic gradient descent in minibatches of 1024 with a learning rate per sample of 0.002. We then evaluated the performance of the multi-task learning approach described in Section 2. The MTL system consisted of 4 hidden layers shared by both tasks and 1 additional task-specific hidden layer for each of the two tasks. We used the same α value of 0.91, which had the best performance on the development set, as shown in Table 1. Comparing the results of the baseline system and the MTL system shown in Table 2, MTL reduces the average WER of the SimData test set from 8.95% to 8.57%, a relative WER reduction of 4.2%. A smaller 2% WER reduction was obtained on the RealData test set, reflecting the mismatched nature of this set.

We then evaluated the effectiveness of adding room-awareness to the network by introducing the 10-dimensional feature vector comprised of T_{60} and DRR estimates to the network. These features were estimated for each utterance individually using the NMF-based approach described in Section 3. These features were used along with the 792-dimensional input vector of log mel filterbank features to create a room-aware DNN. The room-aware network was trained as follows. First the MTL network described previously was trained for 25 epochs. Then the additional room-related features were attached to the first hidden layer with randomly initialized weights, creating the network shown in Figure 3. This network was then trained for an additional 10 epochs. The benefit of adding room awareness to the DNN network can be seen from the third line of Table 2. Lastly, we retrained the final system using generative layer-by-layer pre-training to initialize the weights, rather than random initialization. As the results indicate, this brings a significant additional improvement for all test conditions. We also attempted to learn data-driven codes to describe the rooms, similar to the approach taken in [14]. However, this did not perform as well as the proposed reverberation features.

Finally we compared our best results with the top systems from the 2014 REVERB Challenge. Because we are interested in evaluating the effectiveness of the proposed acoustic model, we limited our comparison the results that used the same training data, trigram language model, and evaluation conditions. We would like to point out that other systems that used additional training data, RNN language models, and large batch adaptation observed additional improvements. Table 3 compares the average WER of the SimData and RealData test sets for the top five performing systems from the challenge and the proposed approach. These systems represent a wide variety of approaches. The proposed approach outperforms all other systems on the SimData test set and all but one of the system on the RealData system.

It is interesting to note that several of these systems used some form of signal or feature dereverberation algorithm prior to acoustic modeling while our proposed approach was able to exceed the performance of almost all of these systems without any explicit dereverberation. In addition, we note that the multi-task learning system alone, without the room-aware features, is capable of running in real time with no added latency.

Table 3. Comparison of top performing single microphone systems from the REVERB Challenge

Systems	SimData WER (%)	RealData WER (%)
Linear Prediction Dereverb + DNN [6]	8.00	27.8
Spectral Subtraction Dereverb + GMM-SGMM-DNN ROVER [17]	8.51	**23.70**
Multi-FE + i-vector + DNN ROVER [9]	10.0	27.1
DNN Dereverb + DNN [8]	11.21	32.45
Multistream GMM + BLSTM [7]	13.75	36.78
Proposed	**7.77**	27.50

5. CONCLUSION

In this paper we proposed a novel framework to augment auxiliary information to a deep neural network acoustic model to improve performance in reverberant environments. First, a multi-task learning approach was introduced in order to help the network learn a shared representation that was beneficial for both classification and feature enhancement. Then we added a vector of room-related features to the network that characterized the reverberant environment the utterance. These features were derived from T_{60} and DRR estimates extracted blindly from the observed signal. Experimental results showed that a room-aware DNN trained with multi-task learning significantly improves the speech recognition performance over a conventional DNN, and resulted in state of the art performance on the REVERB Challenge corpus. Future directions of this work include improving the way in which room information is extracted and incorporated into the network, incorporated adaptation in order to improve the performance on the RealData test set, and extending this approach to multiple microphones.

6. REFERENCES

[1] S. Renals and P. Swietojanski, "Neural networks for distant speech recognition," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*, May 2014, pp. 172–176.

[2] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.

[3] V. Leutnant, A Krueger, and R. Haeb-Umbach, "Bayesian feature enhancement for reverberation and noise robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 8, pp. 1640–1652, Aug 2013.

[4] Kshitiz Kumar, Rita Singh, Bhiksha Raj, and Richard Stern, "Gammatone sub-band magnitude-domain dereverberation for asr," in *Proc. IEEE ICASSP*, 2010.

[5] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, A. Sehr, W. Kellermann, S. Gannot, R. Maas, R. Haeb-Umbach, V. Leutnant, and B. Raj, "The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. of WASPAA*, 2013.

[6] Marc Delcroix, Takuya Yoshioka, Atsunori Ogawa, Yotaro Kubo, Masakiyo Fujimoto, Nobutaka Ito, Keisuke Kinoshita, Miquel Esp, Takaaki Hori, Tomohiro Nakatani, and Atsushi Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge," in *Proc. REVERB challenge workshop*, 2014.

[7] Jrgen T. Geiger, Erik Marchi, Bjrn W Schuller, and Gerhard Rigoll, "The tum system for the reverb challenge: Recognition of reverberated speech using multi-channel correlation shaping dereverberation and blstm recurrent neural networks," in *Proc. REVERB challenge workshop*, 2014.

[8] Xiong Xiao, Zhao Shengkui, Duc Hoang Ha Nguyen, Zhong Xionghu, Douglas Jones, Eng-Siong Chng, and Haizhou Li, "The NTU-ADSC systems for reverberation challenge 2014," in *Proc. REVERB challenge workshop*, 2014.

[9] Md. Jahangir Alam, Vishwa Gupta, Patrick Kenny, and Pierre Dumouchel, "Use of multiple front-ends and i-vector-based speaker adaptation for robust speech recognition," in *Proc. REVERB challenge workshop*, 2014.

[10] M.L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE ICASSP*, 2013, p. 73987402.

[11] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 55–59.

[12] R Caruna, "Multitask learning: A knowledge-based source of inductive bias," in *Machine Learning: Proceedings of the Tenth International Conference*, 1993, pp. 41–48.

[13] M.L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proc. IEEE ICASSP*, 2013.

[14] Shaofei Xue, O. Abdel-Hamid, Hui Jiang, and Lirong Dai, "Direct adaptation of hybrid dnn/hmm model for fast speaker adaptation in lvcsr based on speaker code," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 6339–6343.

[15] N. D. Gaubitch, H. W. Lollman, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes, "Performance comparison of algorithms for blind reverberation time estimation from speech," in *International Workshop on Acoustic Signal Enhancement*, Aachen, Germany, sep 2012.

[16] Dong Yu, Adam Eversole, Mike Seltzer, Kaisheng Yao, Zhiheng Huang, Brian Guenter, Oleksii Kuchaiev, Yu Zhang, Frank Seide, Huaming Wang, et al., "An introduction to computational networks and the computational network toolkit," Tech. Rep., Tech. Rep. MSR, Microsoft Research, 2014, http://cntk.codeplex.com, 2014.

[17] Yuuki Tachioka, Tomohiro Narita, Felix J Weninger, and Shinji Watanabe, "Dual system combination approach for various reverberant environments with dereverberation techniques," in *Proc. REVERB challenge workshop*, 2014.