LEARNING SPECTRAL MAPPING FOR SPEECH DEREVERBERATION

Kun Han, Yuxuan Wang, DeLiang Wang

Department of Computer Science and Engineering
& Center for Cognitive and Brain Sciences
The Ohio State University
Columbus, OH 43210-1277, USA
{hank,wangyuxu,dwang}@cse.ohio-state.edu

ABSTRACT

Reverberation distorts human speech and usually has negative effects on speech intelligibility, especially for hearing-impaired listeners. It also causes performance degradation in automatic speech recognition and speaker identification systems. Therefore, the dereverberation problem must be dealt with in daily listening environments. We propose to use deep neural networks (DNNs) to learn a spectral mapping from the reverberant speech to the anechoic speech. The trained DNN produces the estimated spectral representation of the corresponding anechoic speech. We demonstrate that distortion caused by reverberation is substantially attenuated by the DNN whose outputs can be resynthesized to the dereverebrated speech signal. The proposed approach is simple, and our systematic evaluation shows promising dereverberation results, which are significantly better than those of related systems.

Index Terms— Speech Dereverberation, Deep Neural Networks, Spectral Mapping

1. INTRODUCTION

In real-world environments, the sound reaching the ears comprises the original source (direct sound) and its reflections from various surfaces. These attenuated, time-delayed reflections of the original sound combine to form a reverberant signal. In reverberant environments, speech intelligibility is degraded substantially for hearing impaired listeners [6], and normal hearing listeners when reverberation is severe [13]. Reverberation also causes significant performance degradation in automatic speech recognition (ASR) [5] and speaker identification (SID) systems [15]. Given the prevalence of reverberation, a solution to the dereverberation problem will benefit many speech technology applications.

Reverberation corresponds to a convolution of the direct sound and the room impulse response (RIR), which distorts the spectrum of speech in both time and frequency domains. Thus, dereverberation may be treated as inverse filtering. The magnitude relationship between an anechoic signal and its reverberant version is relatively consistent in different reverberant conditions, especially within the same room. This relationship inspires us to utilize supervised learning to model the characteristics of the reverberation process. In other words, we propose to "learn" an inverse filter, or learn the mapping from the reverberant speech to its anechoic version. The mapper can be trained where the input is the spectral representation of the reverberant speech and the desired ouput is that of the anechoic speech.

Deep neural networks (DNNs) have shown strong learning capacity [4]. Stacked denoising autoencoder (SDA) [16] is a deep learning approach, and it is trained to reconstruct the raw clean data using the noisy data, where hidden layer activations are used as learned features. Although the primary goal of SDAs is to improve generalization, the idea motivated us to utilize DNNs to learn the mapping from the corrupted data to clean data. A recent study [18] used DNNs to denoise acoustic features in each time-frequency unit for speech separation. Our approach, on the other hand, deals with reverberant speech and the mapping is directly based on frame-level spectral features.

In the next section, we discussed related speech dereverberation studies. We then describe our approach in detail in Section 3. The experimental results are shown in Section 4. We conclude our study in the last section.

2. RELATION TO PRIOR WORK

Many previous approaches have been proposed to deal with speech dereverberation [10]. Inverse filtering is one of the commonly used techniques [9]. Since the reverberation effect can be described as a convolution of clean speech with room impulse response, the inverse filtering based approach first needs an inverse filter that reverses the effects of the room response, and then estimates the anechoic signal by convolving the reverberant signal with the inverse filter. However,

This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NIDCD grant (R01 DC012048), and the Ohio Supercomputer Center.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

in many situations, the inverse filter cannot be determined directly and must be estimated, which is not a trivial problem. Further, this approach assumes that the RIR function is a minimum-phase function that is often not satisfied in practice [11]. Wu and Wang [20] utilized a two-stage approach including inverse filtering and spectral subtraction to deal with early reverberation and late reverberation separately, which relies on accurate estimate of the inverse filter in one microphone scenario. Some other studies dealt with dereverbration by exploiting the properties of speech such as modulation spectrum [1], harmonic structure [19], etc.

Recent studies show that the ideal binary mask (IBM) can be extended to suppress reverberation and have been proven to improve speech intelligibility [6, 12, 13]. The IBM based approaches treat the direct sound or direct sound plus the early reflections as the target and the rest as the masker, and the dreverberated signals are resynthesized from the binary mask. Therefore, the IBM can be considered as a computational goal for dereverberation. Hazrati *et al*. [3] proposed to estimate a binary mask based on a single variance-based feature against an adaptive threshold and yielded intelligibility improvements for hearing impaired listeners.

3. ALGORITHM DESCRIPTION

3.1. Spectral features

An input signal $s(t)$ is first passed through a 64-channel gammatone filterbank spanning from 80 Hz to 5000 Hz. The response of each filter channel is then divided into 20-ms time frames with 10-ms frame shift, forming a cochleagram [17]. We use $X(m, c)$ to denote the energy in the time-frequency (T-F) unit for frequency channel c and time frame m. Therefore, in the cochleagram domain, each frame can be represented as a vector $\mathbf{x}(m)$:

$$\mathbf{x}(m) = [X(m, 1), X(m, 2), \ldots, X(m, 64)]^T \quad (1)$$

Spectral feature mapping is based on the cochleagram in each frame $\mathbf{x}(m)$. Since room reverberation leads to characteristic temporal variations of the direct signal, we will include the spectral features of neighboring frames into a feature vector for training. Therefore, the input feature vector for the DNN feature mapping is:

$$\tilde{\mathbf{x}}(m) = [\mathbf{x}(m - d), \ldots, \mathbf{x}(m), \ldots, \mathbf{x}(m + d)]^T \quad (2)$$

where d denotes the number of neighboring frames in each side and is set to 5 in this study. So the dimensionality of the input is $64 \times 11 = 704$. Note that, although the reverberant distortion is primarily caused by previous signal in the time domain, the signal after the current frame provides useful information for the spectral mapping. We have conducted preliminary experiments using only the previous neighboring frames, but the performance is not as accurate as that of using the frames in both sides.

The desired output of the neural network is the cochleagram of clean speech in the current frame m, denoted by a 64-dimensional vector $\mathbf{y}(m)$, whose elements correspond to the energies in T-F units.

3.2. DNN based spectral mapping

The DNN in this study includes three hidden layers, as shown in Fig. 1. The input for each training sample is the cochleagram in a window of 11 frames, corresponding to 704 input units, and the output is the cochleagram in the current frame, corresponding to 64 output units. Each hidden layer includes 1024 hidden units. The parameters of the number of hidden layers and hidden units are chosen from a development set.

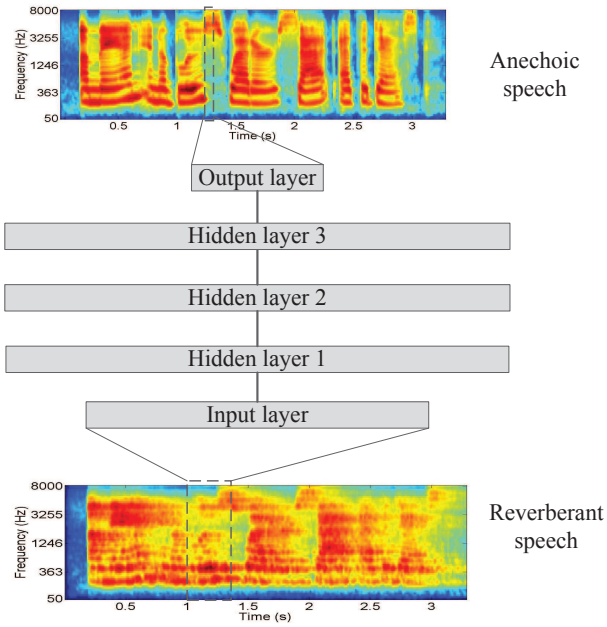


Fig. 1: Structure of the DNN for dereverberation.

We first pre-train a DNN using a stack of restricted Boltzmann machines (RBM) in an unsupervised fashion [4]. The RBM is a generative model and is trained to maximize the likelihood of the input data distribution. The resulting weights are then finetuned by the backpropagation algorithm. We also tried to turn off the RBM pre-training and the results are slightly worse than with pre-training.

The objective function for the backpropagation is based on the mean square error. In addition, we found that we achieve the best performance when using two regularization terms into the objective function. The first regularization is the sparsity of the units in the last hidden layer [7]. We regularize the mean hidden activations to be a small positive num-

ber using Kullback-Leibler (KL) divergence:

$$\mathbf{KL}(\hat{\rho}||\rho) = \sum_k \rho \log \frac{\rho}{\hat{\rho}} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}} \quad (3)$$

where k indexes hidden units, ρ and $\hat{\rho}$ are the expected and the actual mean hidden activation levels, respectively. ρ is set to 0.1 in this study.

Another regularization term is the L_2 norm of the weights $||\mathbf{w}||_2$. Therefore, the final objective function is:

$$\mathcal{L}(\mathbf{y},\mathbf{x};\mathbf{w}) = \frac{1}{2}||\mathbf{y}-f(\mathbf{x};\mathbf{w})||_2^2 + \alpha||\mathbf{w}||_2^2 + \beta\mathbf{KL}(\hat{\rho}||\rho) \quad (4)$$

where $\mathbf{y}$ and $\mathbf{x}$ are the desired output and input of the DNN, respectively. $f(\cdot)$ denotes the nonlinear mapping defined by the weights $\mathbf{w}$ of the neural network. α and β are the regularization parameters chosen from a development set.

The output of the DNN is the estimated cochleagram of the corresponding anechoic speech. With the capacity of learning internal representations, DNN promises to be able to encode the spectral transformation and help to restore the cochleagram of anechoic speech.

Fig. 2 shows an example of the cochleagram mapping for a female sentence "A man in a blue sweater sat at the desk". Figs. 2(a) and (b) show the cochleagram of the reverberant speech with reberberation time $T_{60} = 0.9$ s and the corresponding DNN outputs. As shown in Fig. 2 (b), the smearing energy caused by reverberation is largely removed or attenuated, and the boundaries between voiced and unvoiced frames are restored. The DNN output is a very good estimate of the cochleagram of the anechoic speech, which is shown in Fig. 2(c).

With the estimated cochleagram, it is straightforward to compute a ratio mask for each T-F unit:

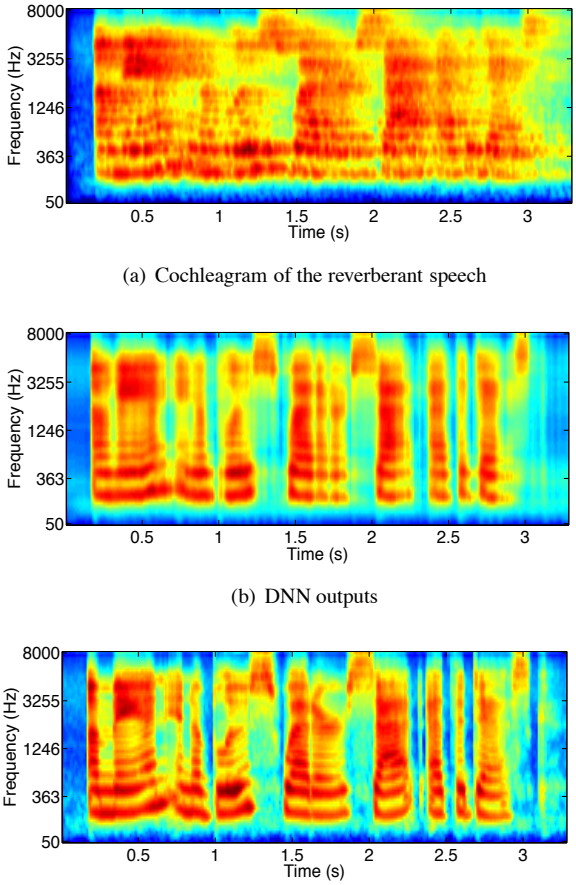
$$M(m,c) = \frac{E_{out}(m,c)}{E_{rev}(m,c)} \quad (5)$$

where, E_{out} and E_{rev} are the DNN outputs and the energy of T-F unit in the reverberant speech, respectively. The dereverberated time-domain signal is then resynthesized using the mask in Eq. (5) and the original reverberant signal [17].

It is worth mentioning that we have attempted to train a DNN on the spectrogram based on short-time Fourier transformation. The DNN also outputs accurate spectrogram estimates. However, the speech quality of the resynthesized time-domain signal is not as good as using the cochleagram mapping, partly due to the low resolution in the low frequency range in the spectrogram. We will compare the evaluation results for the spectrogram and the cochleagram mappings in the next section.

4. EXPERIMENTS

We evaluate our approach in this section. To simulate room acoustics, we generate a simulated room corresponding to



(a) Cochleagram of the reverberant speech

(b) DNN outputs

(c) Cochleagram of the anechoic speech

Fig. 2: DNN dereverberation results. (a) Cochleagram of a reverberant speech. (b) DNN outputs. (c) Cochleagram of the corresponding anechoic speech.

a specific T_{60} [2] and randomly create a set of room impulse responses (RIRs) under this T_{60} condition. To train the system, we generate three reverberation times ranging from 0.3, 0.6, and 0.9 s, and for each T_{60} we generate 2 different RIRs. We use 200 anechoic utterances from the IEEE corpus [14] to generate the training set. Therefore, there are $200 \times 3 \times 2 = 1200$ reverberant sentences in the training set. The test set includes 60 reverberant sentences, corresponding to 20 speech utterances, three T_{60}s, and one RIR. Neither of the utterances nor RIRs are used in the training set.

We quantitatively evaluate the dereverberation results by an objective measurements: frequency-weighted segmental speech-to-reverberation ratio (SRR_{fw}), which is a speech intelligibility indicator [8], computing signal-to-reverberation estimations for each critical band:

$$\text{SRR}_{fw} = \frac{10}{M}\sum_{m=1}^{M}\frac{\sum_{k=1}^{K} W(k)\log_{10}\frac{|S(m,k)|^2}{|S(m,k)-\hat{S}(m,k)|}}{\sum_{k=1}^{K} W(k)} \quad (6)$$

where $W(k)$ is the weight placed on the kth frequency band, K is the number of bands, M is the total number of frames in the signal, $S(m,k)$ is the critical-band magnitude of the clean signal in the kth frequency band at the mth frame, and $\hat{S}(m,k)$ is the corresponding spectral magnitude of the enhanced signal in the same band.

We compare the proposed approach with three baselines. Hazrati *et al.*[3] proposed a state-of-the-art dereverberation approach, utilizing a single variance-based feature from the reverberant signal and comparing its value against an adaptive threshold to compute a binary mask for dereverberation. Wu and Wang[20] used estimated inverse filters and speech subtraction to attenuate early reverberation and late reverberation separately. We also perform dereverberation using the IBM with the relative criterion -5 dB suggested by [13]. Since the IBM is generated from the anechoic speech against the reverberant speech, the results can be considered as a ceiling performance of the binary mask based systems. In Fig. 3, we show the results for both the spectrogram mapping and the cochleagram mapping. Both methods improve SRR_{fw} relative to the unprocessed reverberant speech by more than 2 dB. The cochleagram based mapping achieves higher SRR_{fw} than the spectrogram mapping under all T_{60} conditions. Compared with other approaches, our spectral mapping based approaches achieve the best performances in all reverberant conditions.

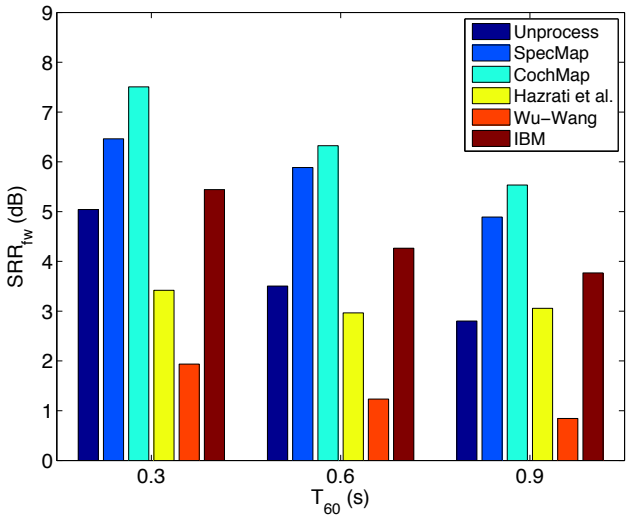


Fig. 3: SRR_{fw} comparisons. "Unprocess" denotes the SRR_{fw} results for unprocessed reverberant speech. "SpecMap" and "CochMap" denote the proposed spectral mapping approach using spectrogram and cochleagram, repectively. "Hazrati et al.", "Wu-Wang", and "IBM" denote three baselines as described.

Since our approach is a supervised learning method, it is important to explore its generalizability. We generate another set of RIRs with T_{60} from 0.2 to 1.0 s, with a step of 0.1 s. Note that, none of RIRs in this experiment are seen in the training set because they are created from different rooms. Fig. 4 shows the generalization results for different T_{60}s. Compared with the unprocessed reverberant speech, the proposed approach substantially boosts SRR_{fw} in each T_{60}s and the advantage is increasingly larger as the T_{60} increases, demonstrating that our approach generalizes well to new reverberant environments in a certain range. We also show the DNN processed results for the anechoic speech, corresponding to $T_{60} = 0$ s in the figure.

In addition, we point out that we have trained speaker-independent models use the TIMIT database [21] and tested on new speakers. Although limited space does not permit detail results, our approach achieves similar performances to the IEEE database, suggesting that the spectral mapping approach is robust to new speakers.

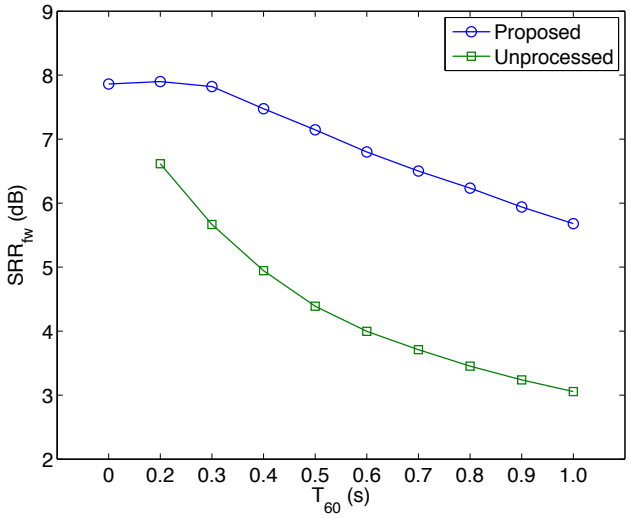


Fig. 4: Generalization results in different T_{60} . "Proposed" denotes the cochleagram mapping approach, and "Unprocessed" the results for original reverberant speech.

5. CONCLUSION

In this paper, we have proposed to use DNNs to learn the spectral mapping from the reverberant speech to the anechoic speech for dereverberation. This novel approach is simple and yet effective. Our supervised learning approach significantly boosts dereverberation performance in a range of reverberant conditions.

Although this paper focuses on the dereverberation problem, the proposed spectral mapping approach can be extended to handle combined dereverberation and denoising. This will be discussed elsewhere in the future.

6. REFERENCES

[1] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in *Proc. of ICSLP 1996*, vol. 2. IEEE, 1996, pp. 889–892.

[2] E. Habets, *Room Impulse Response Generator*, 2010, http://home.tiscali.nl/ehabets/rir_generator.html.

[3] O. Hazrati, J. Lee, and P. C. Loizou, "Blind binary masking for reverberation suppression in cochlear implants," *J. Acoust. Soc. Am.*, vol. 133, pp. 1607–1614, 2013.

[4] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[5] B. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *Proc. of IEEE ICASSP 1997*, vol. 2. IEEE, 1997, pp. 1259–1262.

[6] K. Kokkinakis, O. Hazrati, and P. C. Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants," *J. Acoust. Soc. Am.*, vol. 129, pp. 3221–3232, 2011.

[7] H. Lee, C. Ekanadham, and A. Ng, "Sparse deep belief net model for visual area v2," in *Proc. of NIPS*, 2007, pp. 873–880.

[8] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3387–3405, 2009.

[9] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 36, no. 2, pp. 145–152, 1988.

[10] P. Naylor and N. Gaubitch, Eds., *Speech dereverberation*. Springer, 2010.

[11] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, p. 165, 1979.

[12] N. Roman and J. Woodruff, "Intelligibility of reverberant noisy speech with ideal binary masking," *J. Acoust. Soc. Am.*, vol. 130, p. 2153, 2011.

[13] ——, "Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold," *J. Acoust. Soc. Am.*, vol. 133, no. 3, pp. 1707–1717, 2013.

[14] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoustics*, vol. 17, pp. 227–246, 1969.

[15] S. O. Sadjadi and J. H. L. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proc. of ICASSP 2011*. IEEE, 2011, pp. 5448–5451.

[16] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Machine Learning Res.*, vol. 11, pp. 3371–3408, 2010.

[17] D. L. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, algorithms and applications*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006.

[18] Y. Wang and D. L. Wang, "Feature denoising for speech separation in unknown noisy environments," in *Proc. of IEEE ICASSP 2013*. IEEE, 2013, pp. 7472–7476.

[19] M. Wu and D. L. Wang, "A one-microphone algorithm for reverberant speech enhancement," in *Proc. of IEEE ICASSP 2003*. IEEE, 2003, pp. 844–847.

[20] ——, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 3, pp. 774–784, 2006.

[21] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.