Learning Spectral Mapping for Speech Dereverberation and Denoising

K. Gayathri
PG Scholar, Dept of DECS, Sri Padmavathi Mahila University, AP, India, E-mail: kgayatri096@gmail.com.

Abstract: In real-world environments, human speech is usually distorted by both reverberation and background noise, which have negative effects on speech intelligibility and speech quality. They also cause performance degradation in many speech technology applications, such as automatic speech recognition. Therefore, the dereverberation and denoising problems must be dealt with in daily listening environments. In this paper, we propose to perform speech dereverberation using supervised learning, and the supervised approach is then extended to address both dereverberation and denoising. Deep neural networks are trained to directly learn a spectral mapping from the magnitude spectrogram of corrupted speech to that of clean speech. The proposed approach substantially attenuates the distortion caused by reverberation, as well as background noise, and is conceptually simple. Systematic experiments show that the proposed approach leads to significant improvements of predicted speech intelligibility and quality, as well as automatic speech recognition in reverberant noisy conditions. Comparisons show that our approach substantially outperforms related methods.

Keywords: Deep Neural Networks (DNNs), Denoising, Dereverberation, Spectral Mapping, Supervised Learning.

I. INTRODUCTION

In real-world environments, the sound reaching the ears comprises the original source (direct sound) and its reflection 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[9], and normal hearing listeners when reverberation is severe [2]. In addition, room reverberation when combined with background noise is particularly disruptive for speech perception. Reverberation and noise also cause significant performance degradation in automatic speech recognition (ASR) [17] and speaker identification systems. Given the prevalence of reverberation and noise, a solution to the dereverberation and denoising problems 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. Even when reverberant speech is mixed with background noise, it is still possible to restore speech to some degree from the mixture, because speech is highly structured. These properties motivate us to utilize supervised learning to model the reverberation and mixing process. In this paper, we propose to learn the spectral mapping from reverberant speech to its anechoic version.

The mapper is trained where the input is the spectral representation of reverberant speech and the desired output is that of anechoic speech. We then extend the spectral mapping approach to perform both dereverberation and denoising. Deep neural networks (DNNs) have shown strong learning capacity [8]. A stacked denoising auto encoder (SDA) is a deep learning method, and it can be trained to reconstruct the raw clean data from the noisy data, where hidden layer activations are used as learned features. Although SDAs were proposed to improve generalization, the main idea behind SDAs motivated us to utilize DNNs to learn the mapping from the corrupted data to clean data. A recent study used DNNs to denoise acoustic features in each time-frequency unit for speech separation. In addition, Xu et al. proposed a regression based DNN method for speech enhancement. Unlike these studies, our approach deals with reverberant and noisy speech, which is a substantially more challenging task. We note that an earlier version of our study dealing with just reverberation is published in [6] (more on this in Section V). The paper is organized as follows. In the next section, we discuss related speech dereverberation and denoising studies. We then describe our approach in detail in Section III. The experimental results are shown in Section IV. We discuss related issues and conclude the paper in the last section.

II. RELATION TO PRIOR WORK

Many previous approaches have been proposed to deal with speech dereverberation[4], [6]. Inverse filtering is one of the commonly used techniques[3]. Since the reverberation effect can be described as a convolution of clean speech with the room impulse response, the inverse filtering based approach first determines inverse filter that can reverse the effects of the room response, and then estimates the anechoic signal by convolving the reverberant signal with the inverse filter. However, in many situations, the inverse filter cannot be determined directly and must be estimated, which is a hard problem. Further, this approach assumes that the RIR function is minimum-phase that is often not satisfied in practice [7]. Wu and Wang utilized a two-stage approach including inverse
filtering and spectral subtraction to deal with early reverberation and late reverberation separately, which relies on an accurate estimate of the inverse filter in one microphone scenarios. Other studies dealt with dereverberation by exploiting the properties of speech such as modulation spectrum [2], power spectrum, and harmonic structure. Recent studies show that the ideal binary mask (IBM) can be extended to suppress reverberation and improve speech intelligibility. The IBM based approaches treat the direct sound or direct sound plus the early reflections.

III. ALGORITHM DESCRIPTION

We describe the algorithm in this section, including three subsections: feature extraction, model training, and post-processing.

A. Spectral Features

We first extract features for spectral mapping. Given a time domain input signal \( s(t) \), we use the short time Fourier transform (STFT) to extract features. We first divide the input signal into 20-ms time frames with 10-ms frame shift, and then apply fast Fourier transform (FFT) to compute log spectral magnitudes in each time frame. For a 16 kHz signal, we use 320-point FFT and therefore the number of frequency bins is 161. We denote \( x(m,k) \) the log magnitude in the \( k \)th frequency and the \( m \)th frame as. Therefore, in the spectrogram domain, each frame can be represented as a vector \( x(m) \).

\[
x(m) = [X(m,1), X(m,2), \ldots, X(m,16)]^T
\]

(1)

Fig. 1. (Color online) Structure of the DNN based spectral mapping. The inputs are the log spectra of the current frame and its neighboring frames, and the outputs are the log spectra of the current frame.

B. DNN Based Spectral Mapping

We train a deep neural network to learn the spectral mapping from reverberant, or reverberant and noisy, signals to cleansignals. The DNN in this study includes three hidden layers, as shown in Fig. 1. The input for each training sample is the log magnitude spectrogram in a window of frames, and the number of input units is the same as the dimensionality of the feature vector. The output is the log magnitude spectrogram in the current frame, corresponding to 161 output units. Each hidden layer includes 1600 hidden units. We use cross validation on a development set to train neural networks to choose the number of hidden layers and hidden units. The objective function for optimization is based on mean square error

\[
\mathcal{L}(y; x; \theta) = \sum (y_c - f_c(x))^2
\]

(2)

where \( C=161 \) corresponds to the index of the highest frequency bin, is the desired output vector, and is the actual output of the \( c \)th neuron in the output layer, denotes the parameters we need to learn. To train the neural network, the input is normalized to zero mean and unity variance over all feature vectors in the training set, and the output is normalized into the range of \([0,1]\). The activation function in the hidden layers is the rectified linear function and the output layer uses the sigmoid function, shown in and respectively.

\[
f(x) = \max(0, x)
\]

(3)

The output of DNN is the estimated log magnitude spectrogram of clean speech. With the capacity of learning internal representations, DNN promises to be able to encode the spectral transformation from corrupted speech to clean speech and help to restore the magnitude spectrogram of clean speech.

B. Post-Processing

After the DNN generates magnitude spectrogram estimates, we need to resynthesize time-domain signals using the inverse FFT process. A straightforward method to reconstruct time-domain signals is to directly apply inverse the short-time Fourier transform (iSTFT) using the DNN-generated magnitude and the phase from unprocessed time-domain signals as shown in Fig. 2. However, the original phase of noise-free speech is corrupted, and the corruption usually introduces perceptual disturbances and leads to negative effects on sound quality. In addition, the STFT is computed by concatenating Fourier transforms of overlapping frames of a signal, and thus is a redundant representation of the time-domain signal. For a spectrogram-like matrix in the time-frequency domain, it is not guaranteed there exists a time-domain signal whose STFT is equal to that matrix [4]. In other words, the magnitude spectrograms of the resynthesized time-domain signal could be different from the one we intended to resynthesize a signal from. This inconsistency should be taken into account for synthetic or modified spectrograms, like our DNN generated magnitudes.

IV. EXPERIMENTS

A. Metrics and Parameters

We quantitatively evaluate our approach by two objective measurements of speech intelligibility: frequency-weighted segmental speech-to-noise ratio (SNRfw) [2] and short-time objective intelligibility measure (STOI) [4]. Specifically, is a speech intelligibility indicator, computing a signal-to-noise estimate for each critical band:
Learning Spectral Mapping for Speech Dereverberation and Denoising

We compare the proposed approach with two dereverberation algorithms. Hazrati et al. [7] recently proposed a dereverberation approach, utilizing a 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 [4] used estimated inverse filters and spectral subtraction to attenuate early reverberation and late reverberation, respectively. We perform dereverberation using the IBM with the relative criterion dB suggested in [3]. Since the IBM is generated from the anechoic speech against the reverberant speech, the results can be considered as a ceiling performance of binary masking systems. In Fig. 3, we show the evaluation results in terms of frequency-weighted SNR, STOI, and PESQ, as well as those of the comparison systems. For results shown in Fig. 3(a), without iterative reconstruction post-processing, the DNN approach using reverberant phase without post-processing. “DNN-post” denotes the proposed spectral mapping approach with iterative signal reconstruction as shown in Fig. 4. To further evaluate generalization of our approach, we conduct a cross-corpora experiment, i.e., we directly evaluate the system on the TIMIT corpus [4] without retraining. We randomly choose ten utterances from ten different female speakers (one utterance from one speaker) from the TIMIT corpus and generate reverberant signals using the same three RIRs used as in the above experiments. Fig. 5 shows the comparison results. As shown in the figure, although the DNN model is trained on the IEEE corpus with only one speaker, it generalizes well to another corpus with multiple speakers. The relative improvements are smaller, but our supervised learning based approach with no retraining outperforms other approaches in terms of PESQ.

Fig. 2. (Color online) DNN dereverberation results. (a) Log magnitude spectrogram of clean speech. (b) Log magnitude spectrogram of reverberant speech with T60=0.6 (c) DNN outputs. (d) Log magnitude spectrogram of resynthesized signal. (e) Log magnitude spectrogram of resynthesized signal with post-processing.

Fig. 3. DNN based dereverberation results: (a), (b) STOI, (c) PESQ. “Unproc” denotes the results for unprocessed reverberant speech. “Hazrati et al.” and “Wu-Wang” denote two baselines as described. “IBM” denotes the dereverberation results using the IBM. “DNN” denotes the proposed spectral mapping.

Fig. 4. Generalization results in different T=60s. “DNN” denotes the DNN based spectral mapping approach without post-processing, and “Unprocessed” the results for original reverberant speech.

Fig. 5. Cross-corpora dereverberation results. The DNN model is trained on the IEEE corpus, but tested on the TIMIT corpus: (a), (b) STOI, (c) PESQ. See Fig. 3 caption for notations.

Although mild to moderate reverberation does not significantly impact speech perception for normal hearing listeners, an adverse effect occurs when reverberation is severe [3]. We have also conducted dereverberation experiments for strong reverberation conditions, when is greater than 1.0s. Similar to the above experiment, we use the same utterances to generate reverberant sentences with set to 1.2 s, 1.5 s, and...
1.8 s. The training and test sets use different utterances and different RIRs. Experimental results are shown in Fig. 6. Comparing with unprocessed sentences, the DNN based methods improve and STOI scores. Note that, unlike moderate reverberation conditions as shown in Fig3, PESQ scores are boosted in each reverberation time as shown in Fig. 6(c). In these conditions, the post-processing achieves consistently better performance.

B. Dereverberation and Denoising

Our approach can deal with not only reverberation but also background noise. We can use the same supervised approach to perform dereverberation and denoising simultaneously. In this situation, the input to the neural network is the log magnitude spectrogram of reverberant and noisy speech, and the output is the log magnitude spectrogram of anechoic clean speech. We conduct experiments for dereverberation and denoising. We generate a simulated room corresponding to a specific and randomly create a set, representing the locations of the target, the interference and the microphone inside the room, respectively [14]. From these locations, a reverberant mixture is constructed, where, and are the RIR of the target and the interference at the microphone.

We simulate three acoustic rooms and their s are 0.3, 0.6, and 0.9 s, respectively. The training set contains reverberant mixtures including 200 utterances convolved with 3 RIRs and mixed with 3 noise types: speech-shaped noise, factory noise and babble noise [16] at 0 dB SNR. Here, the SNR is computed as the ratio of the energy of the reverberant noise-free signal to that of the reverberant noise-only signal. To test the system, T60 new reverberant utterances are mixed with the three training noises and three new noises, white noise, cocktail party noise, and crowd noise in playground [10], under each T60 but using different RIRs.

C. Robust Speech Recognition

The above evaluations show that our DNN based spectral mapping significantly attenuates reverberation and noise and produces good estimates of magnitude spectrogram of clean speech. One would expect our approach to improve ASR performance in reverberant and noisy conditions. In this evaluation, we use the second CHiME challenge corpus (track 2) to evaluate ASR performance [36]. In the CHiME-2 corpus, the utterances are taken from the speaker-independent 5k vocabulary subset of the Wall Street Journal (WSJ0) corpus. Each utterance is convolved with one recorded binaural room impulse response corresponding to a front position at a distance of 2 m, and then mixed with binaural recordings of real room noise over a period of days in the same family living room at 6 SNRs of -6, -3, 0, 3, 6, 9 dB. Since To perform ASR, the proposed approach is treated as a front-end to enhance all sentences in both training and test datasets. We first randomly choose 3000 out of 7138 sentences from the CHiME-2 training set to train our DNN based dereverberation and denoising model. With this trained DNN model.

PESQ scores for new noises:: (a) white noise, (b) cocktail-party noise, (c) crowd noise in playground. We use the Kaldi toolkit [8] to train two ASR systems, using original sentences and processed sentences, respectively. The first ASR system is a standard GMM-HMM based system using MFCC features with triphone three-state models. Speaker adaptive training [1] is performed during the training stage. The second is a hybrid ASR system, which uses alignments achieved from the GMM-HMM system and then trains DNNs with
Learning Spectral Mapping for Speech Dereverberation and Denoising

VI. REFERENCES


An RNN aims to capture long-term temporal dynamics time-delayed self-connections and is trained sequentially. We have trained RNN models for spectral mapping, and yielded around 0.2 dB improvement in terms of . Although this improvement is not significant, it is worth exploring RNNs in future work, for example, long short-term memory (LSTM) [9]. In our experiments, we train the DNN model using the IEEE corpus, which includes only one female speaker. In order to test generalizability, we have also conducted a cross-corpus experiment using the TIMIT corpus, where multiple speakers are contained in the test dataset. In summary, we have proposed to use DNNs to learn a spectral mapping from corrupted speech to clean speech for dereverberation, and dereverberation plus denoising. To our knowledge, this is the first study employing supervised learning to address the problem of speech dereverberation. Conceptually simple, our supervised learning approach significantly improves dereverberation, as well as denoising, performance in terms of predicted speech intelligibility and quality scores, and boosts ASR results in a range of reverberant and noisy conditions.
Author Profile: 

Ms. K. Gayathri received B.Tech degree from KMMITS, A.P, India and M.Tech degree in Digital Electronics and communication systems from Sri Padmavathi Mahila University, in AP India. Her interested areas are wireless Communications and DSP.