Reducing noise and reverberation in speech signals via the integration of denoising autoencoder and temporal lowpass filtering

Kuan-Yi Liu[1], *Shih-kuang Lee*[1,2], *Syu-Siang Wang*[2], *Yu Tsao*[2], *Jeih-weih Hung*[1]

[1]National Chi Nan University, Taiwan
[2]Academia Sinica, Taiwan
s106323508@mail1.ncnu.edu.tw, s105323501@mail1.ncnu.edu.tw, sypdbhee@gmail.com, yu.tsao@citi.sinica.edu.tw, jwhung@ncnu.edu.tw

Abstract

The applications and respective devices regarding speech signals continue increasing in number and expanding in scope. However, various sources of environmental distortions deteriorate speech signals, thus somewhat limiting the capability of the corresponding applications. In this study, we focus on developing a novel architecture that improves speech quality by reducing the detrimental effect caused by additive noise and reverberation. The presented architecture mainly consists of two algorithms, denoising auto-encoder (DAE) and temporal lowpass filtering (TLF); the former is a deep learning-based method based on a supervised learning scenario, while the latter is based on the characteristics of speech signals. One particularity of this study is to investigate the effect of the DAE in reducing reverberant distortion, and this has seldom been evaluated in past research to our knowledge. The evaluation experiments are conducted on a subset of the database Mandarin Hearing In Noise Test (MHINT), and the results reveal that the DAE network trained with additive noise-corrupted utterances can improve the quality of speech signals distorted by both additive noise and reverberations. In addition, the TLF method is able to reduce the harmful effect of the aforementioned two types of distortion. Furthermore, the pairing of DAE and TLF behaves better than each component method and provides more significant speech quality improvement, which shows it is quite a promising scenario for speech enhancement.

Index Terms: additive noise, reverberation, speech enhancement, moving-average filter, spectrogram, noise reduction, temporal processing

1. Introduction

Mobile devices including smartphones and tablets are becoming far more widespread among people nowadays, and a lot of applications have been developed and integrated in these mobile devices. In particular, voice/speech-based functions and applications, such as sound recording, voice communication, and speech recognition, are probably the most essential in mobile devices, and the respective quality is highly demanded. However, there exist various sources of interference that deteriorate speech signals during transmission, and thus they undermine the capability of the aforementioned functions and applications. These interference sources include additive noise, channel distortion and reverberation, among others. A variety of techniques has been developed in recent decades in order to compensate for the interference effect, and they can be mainly split into two schools: unsupervised and supervised. Unsupervised methods, such as spectral subtraction (SS) [1], Wiener filtering [2], short-time spectral amplitude (STSA) estimation [3], and short-time log-spectral amplitude estimation (logSTSA) [4], do not employ prior information about speech and/or noise. By contrast, supervised speech enhancement methods use a training set to learn distinct models for clean speech and noise signals; notable examples include codebook-based approaches [5] and hidden Markov model (HMM) based methods [6]. In particular, some of the supervised SE methods follow the prevalent trend of deep learning. For example, the algorithm of the deep denoising autoencoder (DAE) [7, 8] models the relationship between the distorted speech signal and the embedded clean counterpart with a deep neural network (DNN) architecture, which has been shown to be quite effective in dealing with additive noise.

This study has two main directions. The first is to investigate using the DAE algorithm to enhance speech utterances contaminated with additive noise plus reverberation, and the second is to present a simple while effective SE method, termed temporal low-pass filtering (with a brief notation "TLF") that employs a moving-average filter to smooth the spectral time series of distorted utterances. As mentioned earlier, DAE is a supervised SE method, while it will be shown that TLF has little learning to do and thus does not depend on any training data.

A series of experimental evaluations are conducted and the preliminary results indicate that both the DAE and TLF can effectively promote the quality of the distorted utterances in the test set in terms of the perceptual evaluation of speech quality (PESQ) index, and the cascade of the DAE and TLF can bring even better PESQ scores as compared with either of the DAE and TLF.

The remainder of this paper is organized as follows: Section 2 states the presented SE architecture, which contains the algorithms of DAE and TLF, as well as the respective combinations. Experimental setup is given in Section 3, and Section 4 provides the experimental results and the corresponding discussions. Finally, a brief concluding remark is given in Section 5.

2. Proposed Method

In this study, we propose a novel scenario that integrates the DAE and TLF with the aim of enhancing speech signals contaminated by reverberation and additive noise. In the following, the two algorithms, DAE and TLF, are first briefly described, and then the procedure of the presented scenario is stated, together with the respective discussions.

2.1. Denoising auto-encoder (DAE)

DAE is originated from the auto-encoder (AE) algorithm, which trains a neural network that sets the desired output to be the

same as the respective input, namely an identity mapping system. However, the size (the number of neurons) of the hidden layer(s) for the AE network is smaller than that of the input/output layers, and thus the hidden-layer output reveals an encoded or compressed representation of the input data. In the meanwhile, the flow from the hidden layer to the output layer serves as a decoding/decompression stage so the output data approximates the input data well.

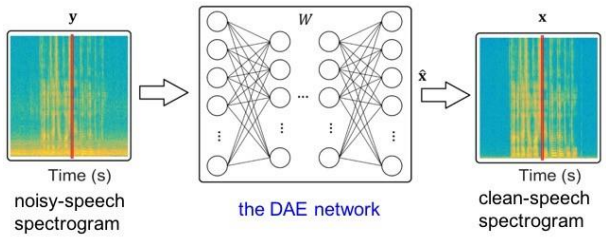

Figure 1: *The flowchart of DAE that processes the spectrogram of a signal with the data unit being the frame-wise magnitude spectrum.*

In comparison with AE, DAE does not require the input and the desired output of the network to be identical. Instead, the input of a DAE network is noise-corrupted data, and the desired output is set to the clean noise-free counterpart of the input data. As a result, a well-trained DAE can be viewed as a denoising neural network that intentionally sets the hidden layer size to be smaller than the input/output layer size.

When DAE is used to enhance a speech signal, the frame-wise magnitude spectrum is often used as the data unit. Stating more clearly, the training set consists of a bunch of distorted utterances and the corresponding clean distortion-free counterparts, and each utterance in the training set is converted to a complex-valued spectrogram (the time series of the frame-wise spectra). Then each of the frame-wise magnitude spectra with respect to distorted utterances is treated as the DAE input, with the clean counterpart as its desired output (target), in order to learn the tunable parameters of the DAE network. Meanwhile, the distorted utterance in the test set is also transformed to its spectrogram, and the magnitude spectrum of each frame is passed through the learned DAE to obtain its enhanced version. A simplified flowchart of the DAE is depicted in Fig. 1.

2.2. Temporal lowpass filtering (TLF)

According to the various research results in the recent decades [9-11], the important information helpful for human intelligibility and automatic recognition is mainly dwelled in the relatively low-varying components of a speech temporal stream. Thus some well-known speech enhancement and noise-robust feature extraction algorithms are developed via emphasizing/diminishing the low/high modulation frequency components of frame-wise speech feature time series. Here, we propose employing a lowpass filter to shape the spectrogram of speech utterance, and expect that the resulting new spectrogram can possess lower deterioration caused by additive noise and/or reverberation.

Let $X[m,k]$ be the complex-valued spectrogram of an arbitrary utterance $x[n]$, in which m and k are the indices of the frames and acoustic frequencies, respectively. The presented TLF applies a moving-average filter $h[m]$ with length L to the magnitude part of $X[m,k]$ along the frame (m) axis with respect to each acoustic frequency as follows:

$$\left|\hat{X}[m,k]\right| = \frac{1}{L}\sum_{\ell=0}^{L-1}|X[m-\ell,k]|, \quad (1),$$

where $\left|\hat{X}[m,k]\right|$ is the updated magnitude spectrogram. We combine $\left|\hat{X}[m,k]\right|$ with the original phase spectrogram to form the new complex-valued spectrogram, which is then converted to the time domain via the inverse STFT to obtain $\hat{x}[n]$, the updated version of $x[n]$.

2.3. The integration of DAE and TLF

A block diagram of the newly presented speech enhancement architecture is depicted in Figure 2. In this figure, $x[n]$ denotes a distorted speech signal, which is a clean speech signal $s[n]$ contaminated by additive noise $d[n]$ plus reverberation that corresponds to an air impulse response $h[n]$. Therefore, it can be written by

$$x[n] = (s[n] + d[n]) * h[n], \quad (2)$$

where "$*$" denotes the convolution operation. The complex-valued spectrogram of $x[n]$ is first created through the STFT, and then the magnitude spectrogram is updated by the cascade of the DAE and TLF, or either of DAE and TLF. The resulting new magnitude spectrogram together with the original phase spectrogram are used to build the enhanced time-domain signal $\hat{x}[n]$.

3. Experimental Setup

The database "Mandarin Hearing In Noise Test (MHINT)" [12] was used for evaluation, among which 400 utterances were used for training (in particular with the DAE) and the other 80 utterances were used for testing. All utterances in this database were pronounced by a male native speaker and recorded at a sampling rate of 16 kHz. The training set for DAE corresponds to seven different noise types and 8 signal-to-noise-ratio (SNR) levels ranging from -16 dB to 16 dB at a 4 dB interval except for 0 dB. As for the test data, different datasets corresponding to either of the 2 noise types and any of the 7 SNR levels were prepared. As for the test set, the two additive noise types were "Car noise idle noise 60mph" and "DowntownStreet", and the 5 SNR levels included -5 dB, 0 dB, 5 dB, 10 dB, 15 dB and 20 dB. In particular, the test set was completely unseen for train set in SNRs and noises.

In addition, to generate the room impulse response $h[n]$ in Eq. (2), we used the image method [13, 14] to simulate the reflection in a room with dimensions 5m×4m×6m (length×width×height). The source position in the room was (2m, 3.5m, 2m) (length, width, height), the receiver position in the room is (2m, 1.5m, 2m) (length, width, height), and the reverberation time T_{60} was set to 0.4 sec.

Furthermore, we have the following arrangements for preparing the spectrogram of the data, the DAE network and the TLF filter:

- Each utterance was split into overlapped frames. The frame duration and frame shift were set to be 64 ms and 10 ms, respectively, and thus the frame rate was 100 Hz, which

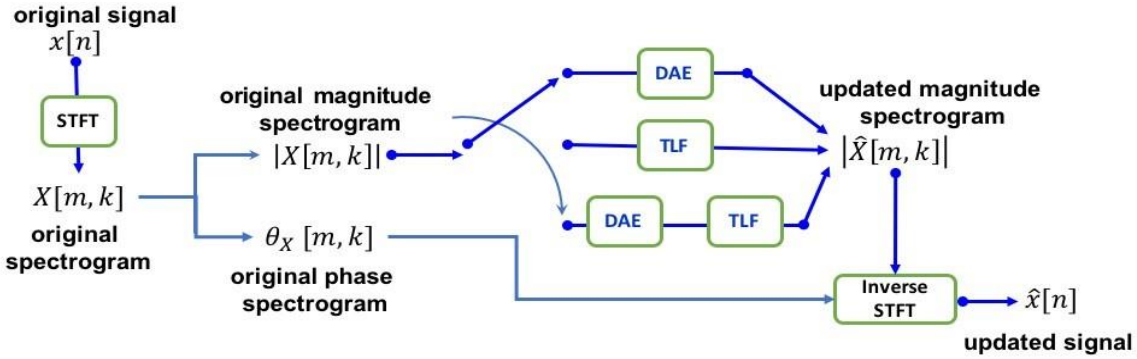

Figure 2: *The block diagram of the presented speech enhancement architecture.*

covers the modulation frequency range [0, 50 Hz] for the analyzed speech feature streams.

- The size of the discrete Fourier transform applied to each frame signal was 512, and thus the first 257 frequency bins of the resulting spectrum were used.
- The used DAE network was set to have 3 hidden layers and 500 neurons in each hidden layer.
- The length of the moving-average filter in TLF was set to any of 3, 5 and 7.

To evaluate the denoising capability of the DAE and the respective combinations with TLF, we used the well-known metric, PESQ [15], which measures the quality level of enhancement for the processed utterances relative to the original noise-free ones. PESQ indicates the quality difference between the enhanced and clean speech signals, and it ranges from -0.5 to 4.5. A higher PESQ score implies that the enhanced utterance is closer to its clean counterpart.

4. Experimental Results and Discussions

First, we evaluate the two enhancement methods, DAE and TLF, in their capability of improving the quality of distorted speech signals. Here, the length of the filter used in TLF was fixed to be 3 here. The upper part of Tables 1 and 2 lists the PESQ values reached by the baseline, DAE and TLF, respectively. From these results, we find that:

1. The PESQ score degrades more obviously as the SNR decreases, and it is further reduced by the introduction of reverberation to noisy signals. These results show that PESQ is a good indicator to reflect the speech quality as well as its distortion level.
2. The DAE trained with noisy signals can effectively enhance both noisy utterances and noisy plus reverberant utterances. It is interesting to note that, although the information about reverberation is not included in the mentioned DAE, it can still improve the signals distorted by reverberation, even though the improvement is just moderate.
3. The TLF method, which simply emphasizes the low modulation frequency part of the spectrogram of speech signals, behaves well for both kinds of distorted utterances. Thus it seems that a simple lowpass operation in the temporal spectral stream of speech signals can reduce the harmful effect of additive noise and reverberation.

Next, we would like to inveterate if the fusion of DAE and TLF can achieve even better results relative to each individual component method. The distorted utterances were first passed through the DAE before being processed by TLF. The corresponding PESQ scores are shown in the lower part of Tables 1 and 2, from which we have the following observations:

1. The DAE and TLF are significantly beneficial to each other since the pairing of them behaves better than either of the DAE and TLF alone in most cases. The possible underlying explanation is that DAE mainly deals with the effect of additive noise, and the subsequent TLF further reduces reverberant distortion.
2. When the SNR of the utterances to be processed increases, the PESQ improvement achieved by the DAE plus TLF becomes less significant. It is possibly caused by the lowpass filtering of TLF, from which the high modulation spectral component that has useful speech information is diminished.

Finally, we varied the length of the moving-average filter used in TLF, as it is integrated with DAE; the results of which are depicted in Figures 3 and 4. From these two figures, we find that increasing the filter length in TLF from 3 to 5 can achieve better PESQ scores in most SNR cases, implying that a length-5 moving average filter is a better choice to alleviate the reverberation effect in speech utterances. However, further increasing the filter length to 7 results in poorer speech quality, which is probably caused by the respective over-smoothing effect for the spectral time series.

Table 1: PESQ results for either of the DAE, TLF, and DAE plus TLF, with respect to the utterances distorted with (1) additive noise and (2) additive noise and reverberation, at various SNRs. The type of additive noise is **"Car noise idle noise 60mph."**.

additive noise						
SNR	**–5**	**0**	**5**	**10**	**15**	**Avg.**
baseline	1.524	1.842	2.194	2.541	2.889	2.198
DAE	1.582	1.988	2.391	**2.725**	**3.009**	2.339
TLF	1.551	1.864	2.210	2.538*	2.856	2.204
DAE+TLF	**1.619**	**2.021**	**2.408**	2.708	2.950	**2.341**
additive noise plus reverberation						
SNR	**–5**	**0**	**5**	**10**	**15**	**Avg.**
baseline	1.167	1.371	1.534	1.644	1.714	1.486
DAE	1.140*	1.401	1.579	1.655	1.679*	1.491
TLF	**1.211**	1.397	1.563	1.666	**1.725**	1.512
DAE+TLF	1.206	**1.450**	**1.617**	**1.690**	1.718	**1.536**

Table 2: PESQ results for either of the DAE, TLF, and DAE plus TLF, with respect to the utterances distorted with (1) additive noise and (2) additive noise and reverberation, at various SNRs. The type of additive noise is **"DowntownStreet."**.

additive noise						
SNR	–5	0	5	10	15	Avg
baseline	1.125	1.356	1.537	1.796	2.120	1.587
DAE	1.330	1.538	1.852	2.223	**2.567**	1.902
TLF	1.054*	1.370	1.526*	1.781*	2.107*	1.568
DAE+TLF	**1.381**	**1.584**	**1.870**	**2.227**	2.558	**1.924**
additive noise plus reverberation						
SNR	–5	0	5	10	15	Avg
baseline	1.121	1.204	1.285	1.425	1.569	1.321
DAE	1.299	1.233	1.351	1.510	1.633	1.405
TLF	1.043*	1.214	1.316	1.448	1.580	1.320
DAE+TLF	**1.306**	**1.284**	**1.367**	**1.551**	**1.659**	**1.479**

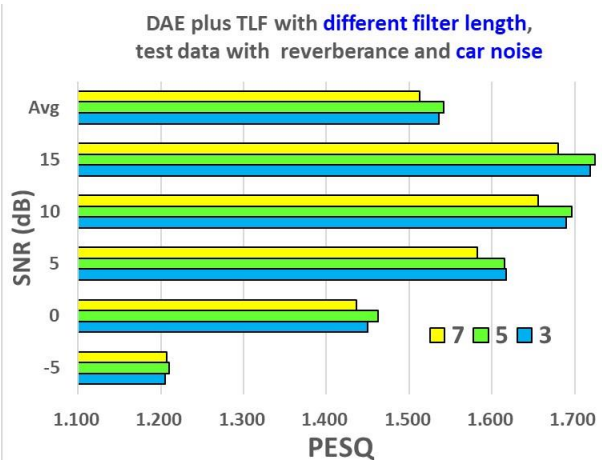

Figure 3: The PESQ scores achieved by the cascade of the DAE and TLF with different filter length for the utterances corrupted by reverberation and additive noise **"Car noise idle noise 60mph."**

5. Conclusions and Future Works

This study proposes a novel speech enhancement scheme that integrates the denoising autoencoder (DAE) and temporal lowpass filtering (TLF) operated on the spectrogram of utterances corrupted by additive noise and reverberation. The preliminary evaluation results indicate that the DAE network trained with the utterances containing additive noise only can improve the quality of the utterances distorted by both additive noise and reverberation. In addition, the newly presented TLF behaves well in enhancing the distorted speech in almost all cases, and it can be well additive to DAE to achieve even higher PESQ scores. However, it is also found that reverberation causes a quite significant quality degradation to speech utterances even if the level of additive noise is slight, and this degradation is just moderately diminished by any of the DAE, TLF, and DAE plus TLF. In the future avenue, we plan to improve the performance of TLF by learning its filter parameters from the training data, and incorporate the temporal information for training the DAE network in order to combat the reverberation better.

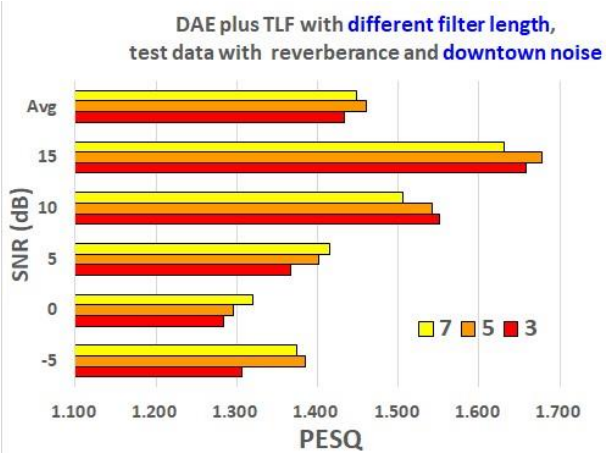

Figure 4: The PESQ scores achieved by the cascade of the DAE and TLF with different filter lengths for the utterances corrupted by reverberation and additive noise **"DowntownStreet."**

References

[1] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *in Proc. ICASSP*, 2002.

[2] P. Scalart, J. V. Filho, "Speech enhancement based on a pri-ori signal to noise estimation," *in Proc. ICASSP*, 1996.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Process.*, 1984.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", *IEEE Trans. Acoust., Speech and Signal Process.*, 1985

[5] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Acoust., Speech and Signal Process.*, 2006.

[6] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Trans. Acoust., Speech, Signal and Lang. Process.*, 2007.

[7] I. Goodfellow, Y. Bengio and A. Courville, "Deep learning," *MIT Press*, 2016

[8] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," *in Proc. Interspeech*, 2013.

[9] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition," *in Proc. Eurospeech*, 1997.

[10] S.-k. Lee, S.-S. Wang, Y. Tsao, J.-w. Hung, "Speech Enhancement based on Reducing the Detail Portion of Speech Spectrograms in Modulation Domain via Discrete Wavelet Transform," *in Proc. ISCSLP 2018*,

[11] C. Chen and J. Bilmes, "MVA processing of speech features," *IEEE Trans. on Audio, Speech, and Language Processing*, 2006.

[12] L. L. Wong, S. D. Soli, S. Liu, N. Han, and M-W. Huang, "Development of the Mandarin hearing in noise test (MHINT)," *Ear and Hearing*, 2007

[13] E. Habets, "Room impulse response generator," https://github.com/ehabets/RIR-Generator.

[14] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal Acoustic Society of America*, 1979.

[15] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," *in Proc. ICASSP*, 2001.