The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 226-240

Speech enhancement based on the integration of fully convolutional network, temporal lowpass filtering and spectrogram masking

Kuan-Yi Liu[1], 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, sypdbhee@gmail.com, yu.tsao@citi.sinica.edu.tw, jwhung@ncnu.edu.tw

Abstract

In this study, we focus on the issue of noise distortion in speech signals, and develop two novel unsupervised speech enhancement algorithms including temporal lowpass filtering (TLP) and relative-to-maximum masking (RMM). Both of these two algorithms are conducted on the magnitude spectrogram of speech signals. TLP uses a simple moving-average filter to emphasize the low modulation frequencies of speech signals, which are believed to contain richer linguistic information and exhibit higher signal-to-noise ratios (SNR). Comparatively, in RMM we apply a mask that is directly multiplied with the speech spectrogram in a point-wise manner, and the used masking value is directly proportional to the magnitude of each temporal-frequency (T-F) point in the spectrogram. The preliminary experiments conducted on a subset of TIMIT database show that the two novel methods can promote the quality of noise-corrupted speech signals significantly, and both of them can be integrated with a well-known supervised speech enhancement scenario, namely fully convolutional network, to achieve even better perceptual speech quality values.

Keywords: temporal lowpass filtering, relative-to-maximum masking, moving-average filter, speech enhancement

1. Introduction

Nowadays the technologies of communication have been developed quite quickly and they have changed and influenced our life a lot. In particular, speech communication such as speech signal transmission and reception through a wired or a wireless network, has been a widespread use in our daily life [1-2]. Therefore, high speech quality and intelligibility during communication gradually becomes a prerequisite.

However, in the transmission environment of speech signals, there exist lots of distortions, such as additive noise, channel mismatch and reverberation, which inevitably decrease the speech signal quality/intelligibility seriously. To overcome these distortions in speech communication, a lot of researchers in recent decades have been devoted to developing speech enhancement (SE) techniques. These SE algorithms can be classified based on whether a learning/training process is involved. For example, if the noise statistics in spectral-subtractive

SE algorithms [3-6] and the basis of the clean signal subspace in subspace SE algorithms [7-9] are learned via a training set with explicit labels, then the corresponding SE methods are supervised. Comparatively, in the unsupervised methods, such as spectral subtraction (SS) [10], Wiener filtering [11], short-time spectral amplitude (STSA) estimation [12] and short-time log-spectral amplitude estimation (logSTSA) [13], does not employ prior information about speech and/or noise.

In this study, we develop two novel learning-free SE methods. One is called temporal low-pass filtering (TLF) and the other is relative-to-maximum masking (RMM). Briefly speaking, TLF borrows the idea of Mod-WD [14], a learning-free SE method, while it can be implemented significantly more simply than Mod-WD, and the mask used in RMM is totally data-driven, viz. it is determined by the signal being processed and has nothing to do with a training set. We then examine the SE capability of the presented novel methods, and see if they are additive to an advanced SE framework based on a deep learning-based fully convolutional network (FCN) [15] to provide even better speech quality for noise-distorted speech signals.

The remainder of this paper is organized as follows: Section 2 presents the details of two novel SE methods, TLF and RMM. The experimental setup is given in Section 3, and Section 4 exhibits the experimental results together with their discussions. Finally, a concluding remark is provided in Section 5.

2. The presented novel SE methods

In this section, we present two novel speech enhancement methods, which are named temporal lowpass filtering (TLF) and relative-to-maximum masking (RMM), respectively. Both of these two methods modify the input utterances in the spectro-temporal (spectrogramic) domain, and they do not require a learning (training) procedure.

2.1 Temporal lowpass filtering

It has been revealed that the important information helpful for human intelligibility and automatic recognition is mainly dwelled in the relatively low-varying components of a speech temporal stream [16-18]. 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. The ModWD algorithm discussed in the previous section follows this trend and factorizes the spectrogram of a noisy signal and then decrease the resulting detail (high half modulation-frequency) part.

Experimental results have revealed that ModWD can moderately improve the speech quality and it can be also well additive to some well-known SE method.

Partially inspired by the aforementioned concept, in this study we present using a simple moving-average filter to process the time series of the spectrogram of noise-corrupted utterances. The presented scheme is analogous to ModWD in emphasizing the low-varying component of the acoustic spectra along the temporal axis.

The block diagram of TLF is shown in Figure 3.1, which consists of the following three steps:

Step 1: Create the spectrogram $\{X[m,k], 0 \leq m \leq M-1, 0 \leq k \leq K-1\}$ for a given time-domain signal $x[n]$, where m and k are respectively the indices of frame and acoustic frequency, and M and K are the total numbers of frames and acoustic frequency points, respectively.

Step 2: Pass the magnitude spectral sequence $\{|X[m,k]|, 0 \leq m \leq M-1\}$ for each acoustic frequency (with index k) through a length-L moving-average filter. The resulting new magnitude sequence is:

$$\left|\hat{X}[m,k]\right| = \frac{1}{L}\sum_{\ell=0}^{L-1}|X[m-\ell,k]|, \qquad (3.1),$$

where $\left|\hat{X}[m,k]\right|$ is the updated magnitude spectral sequence.

Step 3: Construct the new time-domain signal $\hat{x}[n]$ by applying the inverse STFT to the updated spectrogram, which consists of the new magnitude spectrogram $\left\{\left|\hat{X}[m,k]\right|\right\}$ and the original phase spectrogram $\{\angle X[m,k]\}$.

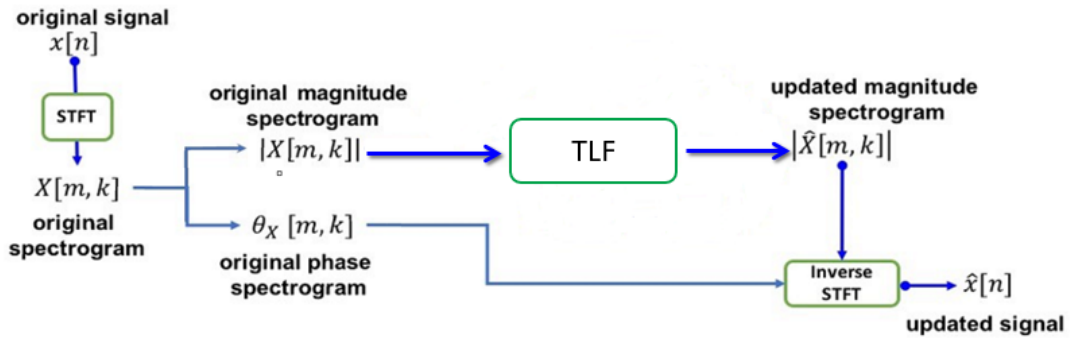

Figure 3.1: The block diagram of TLF.

Since this new method applies a simple lowpass filter (i.e., the moving-average filter) along

the temporal domain of the spectrogram, it is named as temporal lowpass filtering with a short-hand notation "TLF". Some major underlying characteristics of this new method TLF are stated as follows:

1. The used moving-average filter is to emphasize the relatively low modulation frequency portion of the acoustic (magnitude) spectral time series, which is believed to contain helpful linguistic information and be more energy concentrated with a higher signal-to-noise ratio (SNR) compared to the high modulation portion.
2. The greater the length of the employed moving-average filter, the smoother the resulting magnitude spectral curve. However, the filter length needs to be carefully determined in order to diminish the possibly harmful high-frequency part while avoid over-smoothness that ruins the low-varying part.
3. In comparison with the method ModWD that uses DWT and inverse DWT (which consists of at least four filtering processes together with down-sampling and up-sampling), this new approach just applies a filter and is thus simpler in implementation.

2.2 Relative-to-maximum masking

The speech enhancement methods based on time-frequency (T-F) masking have received much attention in the recent decade partially due to its simplicity in computation as well as high capability in segregation speech signals from noise. Among these mask-wise SE methods, a general ideal binary mask (IBM) [19,20] method uses a zero-one masking matrix performing on the spectrogram such that the instantaneous T-F unit for the spectrogram is kept unchanged if it is greater than a threshold that depends on a local SNR criterion (LC), and is set to zero otherwise. By contrast, the method of ideal ratio mask (IRM) [21] applies a soft mask for each instantaneous T-F unit, with the RMM value within the range of zero and one which somewhat reflects the probability of the T-F unit to be speech-wise.

In both methods of IBM and IRM, a key procedure is to estimate the instantaneous signal-to-noise ratio (SNR) of the processed signal. Furthermore, in the recent studies a deep neural work (DNN)-based scenario is used to learn the mask coefficients in IBM and IRM, which inevitably requires a training data set, which contains a great number of noisy-clean signal pairs.

Partially motivated by the ideas of IBM and IRM, in this study we propose a novel RMM scheme which aims to enhance the spectrogram of noise-corrupted utterances. This novel RMM scheme requires no SNR estimation, nor a training stage. The used mask coefficients are totally determined by the utterance being processed. The block diagram of RMM is shown in Figure 3.2, which consists of the following three steps:

Step 1: Create the spectrogram $\{X[m,k], 0 \le m \le M-1, 0 \le k \le K-1\}$ for a given time-domain signal $x[n]$, where m and k are respectively the indices of frame and acoustic

frequency, and M and K are the total numbers of frames and acoustic frequency points, respectively.

Step 2: Compute the RMM coefficients by

$$S_{mk} = \frac{|X[m,k]|}{\max_{m,k}\{|X[m,k]|\}}, \quad 0 \le m \le M-1, 0 \le k \le K-1, \tag{3.2}$$

where S_{mk} is the mask value that will apply to the magnitude spectrogram $|X[m,k]|$ at the m^{th} time frame and k^{th} acoustic frequency bins. From Eq. (3.2), we see that the mask value is simply the ratio of the instantaneous T-F magnitude to the maximum T-F magnitude over the whole spectrogram of the utterance. Thereafter, the new magnitude spectrogram is determined by

$$\left|\hat{X}[m,k]\right| = S_{mk}|X[m,k]|, \qquad 0 \le m \le M-1, 0 \le k \le K-1. \tag{3.3},$$

Step 3: Construct the new time-domain signal $\hat{x}[n]$ by applying the inverse STFT to the updated spectrogram, which consists of the new magnitude spectrogram $\{|\hat{X}[m,k]|\}$ and the original phase spectrogram $\{\angle X[m,k]\}$.

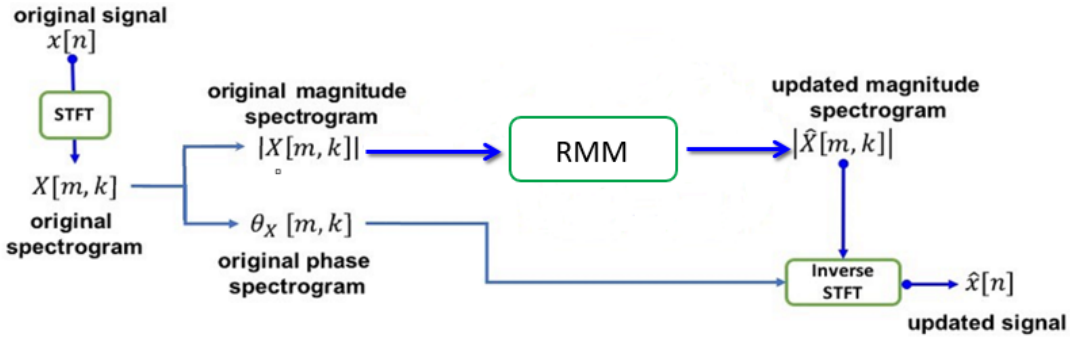

Figure 3.2: The block diagram of RMM.

The new RMM algorithm stated above is termed relative-to-maximum masking, abbreviated by RMM, since the RMM value for each T-F entity is determined by Eq. (3.2). The origin of RMM is a quite simple and naïve idea: a larger magnitude entity in the spectrogram often comes with a high signal-to-noise (SNR) ratio, and it deserves a higher confidence score which is reflected by a larger RMM value. Compared with the other two types of RMM methods, IBM and IRM, RMM does not require noise estimation, nor a supervised/unsupervised learning stage to determine the applied mask. The used mask in RMM is completely data-driven, viz. it totally depends on the test utterance being processed at the present.

3. Experimental setup

As for our evaluation experiments, we use a subset of the TIMIT database [22] to prepare the test set for all SE methods and the training set for the method FCN. TIMIT contains American English utterances produced by 630 speakers. From TIMIT we select 700 utterances pronounced by male speakers and recorded at a sampling rate of 16 Hz. Among these selected 700 utterances, 600 utterances are used to be the training set, while the remaining 100 utterances serve as the test set.
Next, all of the utterances in the training and test sets are manually corrupted by additive noise at various signal-to-noise ratios (SNRs). The numbers of noise types that corrupt the training set and test set are 5 and 3, respectively.

Four speech enhancement (SE) methods, including fully convolutional network (FCN), modulation-domain wavelet denoising (ModWD), temporal lowpass filtering (TLP) and relative-to-maximum masking (RMM), will be evaluated here. As for these four SE methods, FCN is the only one supervised learning method, which requires the training set to learn the associated network parameters. Some important setup factors about the used FCN here are as follows:

1. The FCN model consists of eight convolutional layers with padding, each layer containing filters each with a filter size of 11.
2. The activation function of for each layer output is parametric rectified linear units (PReLUs).
3. The FCN model parameters are trained using Adam optimization algorithm, in which batch normalization is applied so as to minimize the mean square error between the output of the final layer and desired clean time-domain utterance.

Regarding the other three SE methods, ModWD, TLF and RMM, which mainly process the magnitude spectrogram of each test utterance, the general arrangements are listed below:

1. Each test utterance is split into overlapped frames. The frame duration and frame shift are set to be 64 ms and 10 ms, respectively, and thus the frame rate is 100 Hz, which covers the modulation frequency range [0, 50 Hz] for the analyzed speech feature streams.
2. A Hamming window is then applied to each frame signal.
3. The size of the discrete Fourier transform applied to each frame signal is 512, and thus the first 257 frequency bins of the resulting spectrum are used.
4. The biorthogonal 3.7 wavelet basis is used for the DWT and inverse DWT of ModWD.

5. The length of the moving-average filter in TLF is set to 2, with the purpose to cover the modulation frequencies 0-25 Hz approximately, which is highly correlated with linguistic information.
6. Unless otherwise specified, in the RMM method the mask derived with the original or the enhanced test utterance is always applied to the spectrogram of the original (unprocessed) version of the test utterance.

Finally, to evaluate the denoising capability of the aforementioned four SE methods, we employ the well-known objective measure metric, perceptual estimation of speech quality (PESQ) [23], which ranks the 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 tested utterance is closer to its clean counterpart.

4. Experimental results and discussions

4.1 Each single SE method

At the outset, we would like to investigate the SE behavior for any individual of the SE methods, which include fully convolutional network (FCN), temporal lowpass filtering (TLP), modulation-domain wavelet denoising (Mod-WD) and relative-to-maximum masking (RMM). Tables 4.1 list the PESQ scores obtained from the baseline and these SE methods with averaging three noisy cases "Engine", "White" and "Crowd". From this tables, we have the following observations:

1. The PESQ score degrades as the signal-to-noise ratio (SNR) of the environment becomes worse, and thus it is believed to be a good metric to reflect the quality of speech utterances.
2. About the cases of the SNR greater than -6 dB for the three noise types, FCN performs the best, closely followed by RMM, and then TLF and ModWD. Notably, as for the two low-pass filtering methods, TLF behaves moderately better than ModWD while it can be implemented in a simpler manner. As mentioned earlier, TLF just uses a moving-average filter, while ModWD requires a DWT-IDWT procedure, which involves both filtering, down-sampling and up-sampling.

To briefly conclude, FCN gives better PESQ scores than the other three methods at moderate noise levels, RMM works quite well for almost all SNR cases, while ModWD and TLF give rise to relatively slight improvement. Since these SE methods are developed along different

directions, it is natural to assume that the combination of two or three of them might cause further improvement relative to each component method. This part will be examined in the subsequent two sub-sections.

Finally, we evaluate different SE methods in the domain of magnitude spectrogram. A speech signal corrupted by engine noise at 0 dB SNR is individually processed by any of these SE methods, and the corresponding magnitude spectrograms are shown in Figure 4.1. From this figure, it is clear that FCN brings an optimal denoising performance compared with the other methods. The novel presented RMM behaves also quite well, but it seems to over-eliminate the portion of high acoustic frequencies partially because these frequencies possess significantly low energy and cause low masking values.

Table 4.1: The PESQ scores obtained from the baseline, and any of four SE methods as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	Baseline	FCN	ModWD	TLF	RMM
-15dB	1.017	0.989*	1.035	1.035	**1.122**
-12dB	1.078	1.076*	1.087	1.087	**1.228**
-6dB	1.283	1.504	1.302	1.310	**1.600**
0dB	1.592	**2.024**	1.611	1.623	1.976
6dB	1.973	**2.494**	1.992	2.005	2.388
12dB	2.391	**2.855**	2.407	2.423	2.737
18dB	2.811	**3.145**	2.819	2.838	2.943

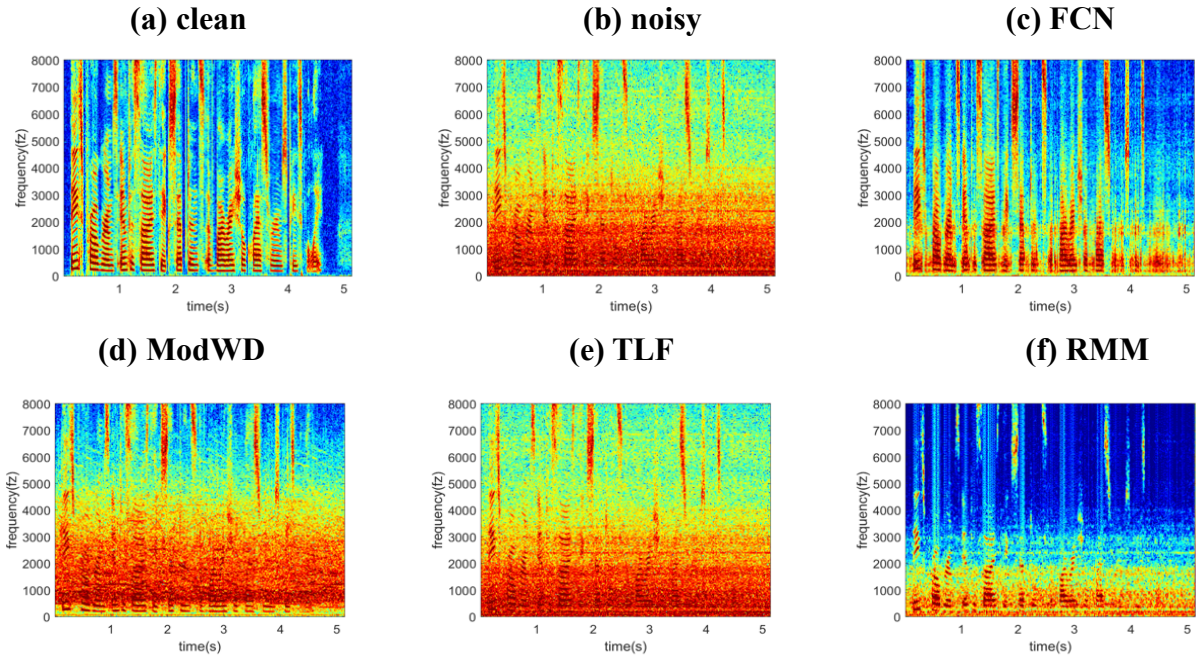

Figure 4.1: The magnitude spectrograms of (a) a clean-noise free signal $\boldsymbol{x}$ (b) the noisy counterpart, $\tilde{\boldsymbol{x}}$, of $\boldsymbol{x}$, which contains 0-dB engine noise, (c) the FCN-enhanced version of $\tilde{\boldsymbol{x}}$, (d) the ModWD-enhanced version of $\tilde{\boldsymbol{x}}$, (e) the TLF-enhanced version of $\tilde{\boldsymbol{x}}$, (f) the RMM-enhanced version of $\tilde{\boldsymbol{x}}$

4.2 The pairing of two SE methods

As mentioned before, almost any of the used four SE methods (FCN, ModWD, TLF and RMM) can enhance the distorted utterances, while the SNR cases for each method to work best are different. In addition, these SE methods might process different parts of the distorted utterances with different goals in speech enhancement. For example, FCN directly minimizes the discrepancy of the noisy speech and its clean counterpart in the training set, ModWD and TLF alleviates the high modulation frequency portions in noisy speech, and TLF emphasizes the high-energy temporal-spectral bins. With this in mind, we would like to investigate whether the cascade of two or three of these SE methods can behave better than each constituent method.

First of all, the cascade of FCN and either of ModWD and TLF is evaluated, in which the test utterances at the three noise conditions are first processed by FCN, and the resulting spectrogram is lowpass filtered by ModWD or TLF. The corresponding PESQ scores are listed in Tables 4.2 and 4.3. From these two tables, we find that both combinations, "FCN plus ModWD" and "FCN plus TLF", give rise to even better results than each component method at almost all SNR cases (except 18 dB for FCN plus ModWD). The amount of PESQ improvement is more significant at lower SNRs. In addition, "FCN plus TLF" outperforms "FCN plus ModWD", which further reveals the advantage of TLF over ModWD, since TLF behaves better with a lower computational cost.

Table 4.2: The PESQ scores obtained from the baseline, FCN, ModWD and the pairing of FCN and ModWD as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	Baseline	FCN	ModWD	FCN+ModWD
-15dB	1.017	0.989*	1.035	**1.126**
-12dB	1.078	1.076*	1.087	**1.216**
-6dB	1.283	1.504	1.302	**1.629**
0dB	1.592	2.024	1.611	**2.116**
6dB	1.973	2.494	1.992	**2.539**
12dB	2.391	2.855	2.407	**2.856**
18dB	2.811	3.145	2.819	3.112

Table 4.3: The PESQ scores obtained from the baseline, FCN, TLF and the pairing of FCN and TLF as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	Baseline	FCN	TLF	FCN+TLF
-15dB	1.017	0.989*	1.035	**1.159**
-12dB	1.078	1.076*	1.087	**1.263**
-6dB	1.283	1.504	1.310	**1.672**
0dB	1.592	2.024	1.623	**2.144**
6dB	1.973	2.494	2.005	**2.559**
12dB	2.391	2.855	2.423	**2.888**
18dB	2.811	3.145	2.838	**3.149**

Next, we examine the integration of RMM with the other three methods. The mask used in RMM is created by any of FCN-, ModWD- and TLF-preprocessed test utterances, which is then applied to the "original" (unprocessed) test utterance counterpart. The respective PESQ results are shown in Tables 4.4, 4.5 and 4.6. From the three tables and figure we have several findings listed below:

1. As for the method "FCN plus RMM", it performs better than FCN and RMM at low SNRs, -15 dB, -6 dB and 0 dB, while it gets worse with the increase of the SNR. One possible reason for the performance degradation at higher SNRs is the phase mismatch in the complex-valued spectrograms of the original and FCN-processed utterances. As we know, FCN updates a test utterance in the time domain, and thus changes both the magnitude and phase parts of the respective spectrogram. However, only the FCN-processed magnitude part is used to create the mask in RMM, which is applied to the original magnitude part. Accordingly, the FCN-processed phase part is discarded in the whole process.

2. Regarding the two combinative methods "ModWD plus RMM" and "TLF plus RMM", the associated PESQ scores are always much higher than the single ModWD and TLF, indicating that for the original noisy spectrogram, the masking operation (with spectrogram masks created by ModWD- and TLF-processed signals) are more effective than the operations of ModWD and TLF. In addition, "ModWD plus RMM" and "TLF plus RMM" outperforms RMM for the SNRs less than 18 dB. At a high SNR as 18 dB, the Mod-WD/TLF-wise masks might over-smooth the spectrogram, and thus are less helpful than the mask created by the nearly clean signal.

Table 4.4: The PESQ scores obtained from the baseline, FCN, RMM and the pairing of FCN and RMM as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	Baseline	FCN	RMM	FCN+RMM
-15dB	1.017	0.989*	1.122	**1.039**
-12dB	1.078	1.076*	**1.228**	1.190
-6dB	1.283	1.504	1.600	**1.718**
0dB	1.592	2.024	1.976	**2.118**
6dB	1.973	**2.494**	2.388	2.390
12dB	2.391	**2.855**	2.737	2.566
18dB	2.811	**3.145**	2.943	2.719*

Table 4.5: The PESQ scores obtained from the baseline, ModWD, RMM and the pairing of ModWD and RMM as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	Baseline	ModWD	RMM	ModWD+RMM
-15dB	1.017	1.035	1.122	**1.140**
-12dB	1.078	1.087	1.228	**1.252**
-6dB	1.283	1.302	1.600	**1.615**
0dB	1.592	1.611	1.976	**1.998**
6dB	1.973	1.992	2.388	**2.403**
12dB	2.391	2.407	2.737	**2.741**
18dB	2.811	2.819	**2.943**	2.926

Table 4.6: The PESQ scores obtained from the baseline, TLF, RMM and the pairing of TLF and RMM as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	Baseline	TLF	RMM	TLF+RMM
-15dB	1.017	1.035	1.122	**1.132**
-12dB	1.078	1.087	1.228	**1.257**
-6dB	1.283	1.302	1.600	**1.628**
0dB	1.592	1.611	1.976	**2.012**
6dB	1.973	1.992	2.388	**2.416**
12dB	2.391	2.407	2.737	**2.755**
18dB	2.811	2.819	**2.943**	2.939

Finally, the results for all combinative methods are summarized in Table 4.7 and Figure 4.2. From these results, we find that the method "FCN plus TLF" behaves the best, except for the case of -6 dB-SNR, showing that a simple lowpass filtering is quite additive to FCN to alleviate the noise effect. Comparatively, the two well-behaved methods, FCN and RMM, do not necessarily exhibit the most complementary effect.

Table 4.7: The PESQ scores obtained from several combinative methods as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	FCN+ModWD	FCN+TLF	FCN+RMM	ModWD+RMM	TLF+RMM
-15dB	1.126	**1.159**	1.039	1.140	1.132
-12dB	1.216	**1.263**	1.190	1.252	1.257
-6dB	1.629	1.672	**1.718**	1.615	1.628
0dB	2.116	**2.144**	2.118	1.998	2.012
6dB	2.539	**2.559**	2.390	2.403	2.416
12dB	2.856	**2.888**	2.566	2.741	2.755
18dB	3.112	**3.149**	2.719	2.926	2.939

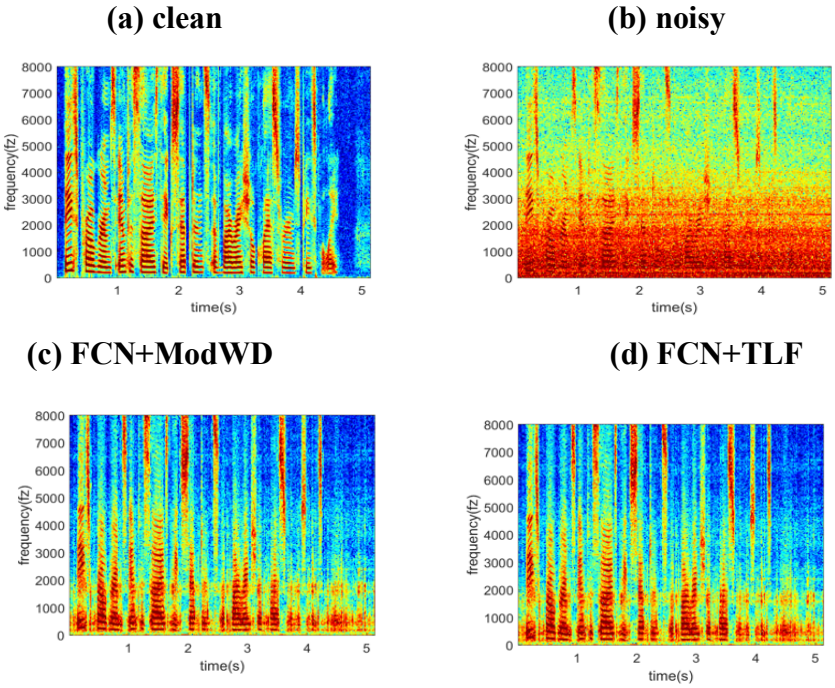

Figure 4.2: The magnitude spectrograms of (a) a clean-noise free signal $\boldsymbol{x}$ (b) the noisy counterpart, $\tilde{\boldsymbol{x}}$, of $\boldsymbol{x}$, which contains 0-dB engine noise, (c) the FCN-plus-ModWD-enhanced version of $\tilde{\boldsymbol{x}}$, (d) the FCN-plus-TLF-enhanced version of $\tilde{\boldsymbol{x}}$

Figure 4.2 shows the magnitude spectrograms for the clean and noisy signals, plus the signals

processed by two combinative methods, "FCN+ModWD" and "FCN+TLF", which have been revealed to promote the PESQ scores apparently. From this figure, we reconfirm that these two combinative methods can reduce the noise effect a lot in the distorted signal and thus bring the recovery of the embedded clean noise-free part.

4.3 The integration of three SE methods

Following the trend in the previous two sub-sections, here we would like to investigate what happens if we use the concatenation of three SE methods to process the test utterances. For simplicity, we use two forms of concatenation: one is FCN followed by ModWD and RMM in turn, denoted by "FCN+ModWD+RMM", and the other is FCN followed by TLF and RMM successively, denoted by "FCN+TLF+RMM". Therefore, these two forms differ in the used lowpass processing method at the median stage. The corresponding PESQ scores are listed in Tables 4.11and 4.12. For the ease of comparison, the results of FCN, FCN plus ModWD/TLF, and ModWD/TLF plus RMM are also listed in these tables. According to the results, we have two findings:

1. The two forms of three-method concatenation outperform the single FCN and the other two-method concatenations at the SNRs of -12 dB, -6 dB and 0 dB. When the SNR becomes higher, adding RMM at the final stage fails to increase the PESQ scores, which is probably due to an effect of over-adjustment.
2. When used in the intermediate or final stage, TLF always behaves superior to ModWD. This again confirms the advantage of TLF over ModWD.

Table 4.11: The PESQ scores obtained from FCN, FCN plus ModWD, ModWD plus RMM, and FCN plus ModWD and RMM, as for the utterances in the three noise environments **"Engine", "White" and "Crowd"**.

SNR	FCN	FCN+ModWD	ModWD+RMM	**FCN+ModWD+RMM**
-15dB	0.989	1.126	**1.140**	1.103
-12dB	1.076	1.216	1.252	**1.262**
-6dB	1.504	1.629	1.615	**1.780**
0dB	2.024	2.116	1.998	**2.157**
6dB	2.494	**2.539**	2.403	2.410
12dB	2.855	**2.856**	2.741	2.567
18dB	**3.145**	3.112	2.926	2.707

Table 4.12: The PESQ scores obtained from FCN, FCN plus TLF, TLF plus RMM, and FCN plus TLF and RMM, as for the utterances in the three noise environments **"Engine",**

"White" and "Crowd".

SNR	FCN	FCN+TLF	TLF+RMM	**FCN+TLF+RMM**
-15dB	0.989	**1.159**	1.140	1.141
-12dB	1.076	1.263	1.252	**1.285**
-6dB	1.504	1.672	1.615	**1.797**
0dB	2.024	2.144	1.998	**2.165**
6dB	2.494	**2.559**	2.403	2.411
12dB	2.855	**2.888**	2.741	2.572
18dB	3.145	**3.149**	2.926	2.714

5 Conclusion

To our knowledge, a fully convolutional network (FCN) applied in an SE framework outperforms conventional neural networks like densely connected network and convolutional neural network (convnet) in promoting the quality of distorted speech signals. Compared with an FCN-based SE framework, the two novel learning-free SE algorithms, temporal lowpass filtering (TLF) and relative-to-maximum masking (RMM) presented in this paper are shown to provide even better denoising performance at some particular signal-to-noise ratio (SNR) cases, despite their simplicity in implementation and their irrelevance with pre-training. Furthermore, our experimental results show that TLF is quite complementary to FCN since the paring of FCN and TLF behaves significantly better than FCN alone. We also show that the two novel methods, TLF and RMM, are quite additive to each other.

In the future avenue, we plan to evaluate FCN, TLF and RMM and the respective possible integrations on the other speech databases, which are recorded in environments that contain various distortions such as additive noise, channel mismatch, and reverberation. In addition, we would like to investigate the theoretical reason why RMM can bring about significant speech quality improvement, and further enhance it by tuning the used mask with a learning scenario.

References

[1] D. O' Shaughnessy, "Speech communications: human and machine," *2nd ed., Hyderabad, India: University Press (India) Pvt. Ltd.*, 2007.

[2] Y. Ephraim, H. L. Ari and W. Roberts, "A brief survey of speech enhancement," *Electrical Engineering Handbook, 3rd ed. Boca Raton, FL: CRC*, 2006.

[3] P. C. Loizou, "Speech enhancement: theory and practice," T*aylor and F. Group, Eds. Boca Raton, FL, USA: CRC Press*, 2013.

[4] R. Martin, "Spectral subtraction based on minimum statistics," *In Proc. European Conference Signal Processing, pp. 1182–1185*, 1994.

[5] P. Krishnamurthy, and S. R. M. Prasanna, "Modified spectral subtraction method for enhancement of noisy speech," *in Proc. International Conference Signal, Image Processing, pp.146-150*, Dec. 2005.
[6] S. Ogata and T. Shimamura, "Reinforced spectral subtraction method to enhance speech signal, " *in Proc. International Conference on Electrical and Electronic Technology, vol. 1, pp. 242-245*, 2001.
[7] A. H. Abolhassani, S.-A. Selouani and D. O'Shaughnessy, "Speech enhancement using PCA and variance of the reconstruction error in distributed speech recognition," *in Proc.IEEE ASRU'07,* 2007
[8] B. Nazari, M. Sarkrni and P. Karimi, "A method for noise reduction in speech signal based on singular value decomposition and genetic algorithm," *In Proc. of IEEE EUROCON'09*, 2009
[9] S. J. Rennie, J. R. Hershey and P. A. Olsen, "Efficient model-based speech separation and denoising using non-negative subspace analysis," *In Proc. of ICASSP'08*, 2008.
[10] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *in Proc. ICASSP*, 2002.
[11] P. Scalart, J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *in Proc. ICASSP*, 1996.
[12] 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.
[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", *IEEE Trans. Acoustics., Speech and Signal Process.*, 1985
[14] S.-k. Lee, S.-S. Wang, Y. Tsao and J.-w. Hung, "Speech enhancement based on reducing the detail portion of speech spectrograms in modulation domain via discrete wavelet transform," in ISCSLP, 2018
[15] S. -W. Fu, Y. Tsao, X. Lu and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," *in Proc. APSIPA-ASC*, 2017
[16] N. Kanedera, T. Arai, H. Hermansky and M. Pavel, "On the importance of various modulation frequencies for speech recognition," *in Proc. Eurospeech*, 1997.
[17] C. Chen and J. Bilmes, "MVA processing of speech features," *IEEE Trans. on Audio, Speech, and Language Processing*, 2006.
[18] X. Xiao, E. S. Chng and H. Li, "Normalization of the speech modulation spectra for robust speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, 2008
[19] G. Kim and P. Loizou, "Improving speech intelligibility in noise using a binary mask that is based on magnitude spectrum constraints," *IEEE Signal Processing Letters, vol. 17 no. 12 pp. 1010-1013*, 2010.
[20] R. Koning, N. Madhu and J. Wouters, "Ideal time frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners," *IEEE Trans. Biomed. Eng. vol. 62 no. 1 pp. 331-341*, 2015.
[21] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," *in Proc. ICASSP*, 2013
[22] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database" *NIST Tech Report*, 1988
[23] A. Rix, J. Beerends, M. Hollier and A. Hekstra, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *Int. Telecommun. Union, T Recommendation, Art. no. 862,* 2001