Performance Comparisons of the Integrated Parallel Model Combination Approaches with Front-End Noise Reduction

Guanghu Shen¹, Soo-Young Suk², Hyun-Yeol Chung¹

¹Dept. of ICE, School of EECS, Yeungnam University, Gyeongsan, Gyeongbuk, Republic of Korea
²Speech Processing Group, Information Technology Research Institute, AIST, Tsukuba, Japan

guanghosin@ynu.ac.kr, sy.suk@aist.go.jp, hychung@ynu.ac.kr

Abstract

In this paper, to find the best noise robustness approach, we study on approaches implemented at both-end (i.e. front-end and back-end) of speech recognition system. To reduce the noise with lower speech distortion at front-end, we investigate the Two-stage Mel-warped Wiener Filtering (TMWF) in the integrated Parallel Model Combination (PMC) approach. Furthermore, the first-stage of TMWF (i.e. One-stage Mel-warped Wiener Filgering (OMWF)), as well as the well-known Wiener Filtering (WF), is effective to reduce the noise, so we integrate PMC with those front-end noise reduction approaches. From the recognition performance, TMWF-PMC shows improved performance comparing with the well-known WF-PMC, and OMWF-PMC also shows a comparable performance in all noises.

Index Terms: Parallel model combination, Wiener filtering, Two-stage mel-warped wiener filtering

1. Introduction

The performance of a speech recognition system is often severely degraded in the presence of noise. Noise introduces mismatches between the pre-trained acoustic models and the input features. If we divide a speech recognition system into front-end processing for speech feature extraction and back-end processing for HMM decoding, then methods to compensate for noise can be implemented at front-end or back-end or both. The front-end processing at front-end is to reduce the noise from the noise-corrupted input speech and get more robust speech features, while back-end processing is to compensate for noise and adapt the parameters of acoustic models.

In this paper, we focus on approaches implemented at both-end (i.e. front-end and back-end) of speech recognition system. For back-end processing, Parallel Model Combination (PMC) is well-known as an effective model compensation approach that can improve speech recognition performance in noisy environments. However, recognition performance of PMC is not sufficient if the Signal-to-Noise Ratio (SNR) of the speech is low, because the speech features are buried in the noise. To improve the performance of PMC, Satoshi proposed a Wiener Filtering-Parallel Model Combination (WF-PMC) approach which combines WF at front-end and PMC at back-end [1]. The main idea of WF-PMC is that firstly raises the SNR of the noise-corrupted input speech with WF, then compensates the remaining noise that was not removed completely at front-end, by applying PMC for the acoustic models at back-end. Here, we could conclude that the characteristic of the both-end integrated approach is that the remaining noise after reducing the noise at front-end might be compensated with PMC at back-end, so it is possible to improve the recognition performance in noisy environments. And we also find that in the integrated approach most important thing is noise reduction with low speech distortion. It means the better input speech (i.e. lower speech distortion) might show a higher recognition performance.

To reduce the noise with low speech distortion, in this paper we like to investigate the Two-stage Mel-warped Wiener Filtering (TMWF) which was adopted as a Advanced Front-End (AFE) in ETSI [2], and we also apply it at front-end of our recognition system to raise the SNR of input speech. Since TMWF is an improved Wiener filtering that enhances the noise-corrupted speech in the perceptual domain (i.e. mel-frequency), and speech enhancement in the perceptual domain is advantageous for speech recognition [3][4], we could expect that TMWF-PMC will improve the recognition performance. Especially, TMWF removes the noise in two stages, at first stage coarsely reduce the noise and whiten residual noise, and at second stage wipe out the residual noise by exploiting the correlation characteristics between the speech signal and white noise. According to above characteristic of the integrated approach, we further want to investigate the performance of the first stage of TMWF (i.e. One-stage Mel-warped Wiener Filtering (OMWF)), as well as the integrated PMC with OMWF, referred as OMWF-PMC. Since the residual noise by OMWF is possible handled by PMC at back-end, we could consider that the integrated OMWF-PMC approach might be shown a good recognition performance in noisy environments.

The rest of this paper is organized as follows. We describe several noise reduction approaches and PMC in section 2 and section 3, respectively. In section 4, the integrated approaches are described in detail. In section 5, analysis conditions and experimental results are presented. Finally, conclusion is presented in section 6.

2. Noise reduction approaches

2.1. Wiener Filtering

Assuming an additive noise model, the noisy signal \( y(t) \) is given by \( x(t) + d(t) \), where \( x(t) \) is the clean speech signal which is assumed to be independent of the noise \( d(t) \) and \( t \) is time index. A short-time Fourier analysis is applied to the input signal by computing the Discrete Fourier Transform (DFT) of overlapping windowed frames. In the frequency domain we have \( Y_k(t) = X_k(t) + D_k(t) \), where \( k \) denoting the frequency bin index. A Wiener Filtering (WF) is the optimal Bayesian linear filter that minimizes the expected mean-squared error \( E[(X_k(t) - X_k(t))^2] \) for the noise corruption model [5][6]. In the frequency domain, the Wiener filter is constructed as in Eq.1.
$$H_k(t) = \frac{|X(t)|^2}{|X(t)|^2 + |D(t)|^2}$$ (1)

The output spectrum \(\hat{X}_k(t)\) is computed by \(H_k(t)Y_k(t)\).

2.2. Two-stage Mel-warped Wiener Filtering

Noise reduction is based on Wiener filtering theory and it is performed in two stages. TMWF has been approved as the main part of the ETSI AFE for Distributed Speech Recognition (DSR). The block diagram of TMWF is shown in Fig.1, the details of each block can be found in [2][7].

2.2.1. The first stage

The first stage coarsely reduces the noise and whitens residual noise. Noise reduction is performed on a frame-by-frame basis. After framing the input signal, the linear spectrum of each frame is estimated in the Spectrum Estimation block. In Power Spectral Density (PSD) Mean block, the signal spectrum is smoothed along the time index. Then, in the WF Design block, Wiener filter coefficients in frequency domain are calculated using both the current frame spectrum estimation and the noise spectrum estimation. The noise spectrum is estimated from noise frames, which are detected by a Voice Activity Detector module (VADNest). Linear Wiener filter coefficients are further smoothed along the frequency axis by using a Mel Filter Bank, resulting in a mel-warped frequency domain Wiener filter. The impulse response of this mel-warped Wiener filter is obtained by applying a Mel-warped Inverse Discrete Cosine Transform (Mel-IDCT). Finally, the input signal of each stage is filtered in the Apply Filter block.

2.2.2. The second stage

The second stage wipes out the residual noise by exploiting the correlation characteristics between the speech signal and white noise. In the framework, the output of the first stage is entered as the input signal in the second stage, then blocks of the bottom part of Fig.1 are conducted. Finally, at the end of noise reduction, the DC offset of the noise-reduced signal is removed in the off block. Additionally, in the second stage, the aggression of noise reduction is controlled by Gain Factorization block.

![Block diagram of two-stage mel-warped wiener filtering](image)

Figure 1: Block diagram of two-stage mel-warped wiener filtering.

3. Parallel Model Combination

PMC was first developed by Varga and Moore [8] and was later refined by Gales and Young [9]. PMC generates noisy-speech-like models by combining clean acoustic models (i.e. clean speech HMM) and a noise model (i.e. noise HMM). The HMM’s cepstral parameters must be mapped to the linear spectral domain for combination. The detailed procedures can be summarized as the following steps, where \(\mu^c\) and \(\Sigma^c\) are the clean speech HMM’s mean and variance respectively in the cepstral domain, \(\mu^s\) and \(\Sigma^s\) are in the log spectral domain, \(\mu\) and \(\Sigma\) are in the linear spectral domain.

Those parameters for noise-corrupted speech models and indicated by the notation “’” and those for noise models by the notation “”. Thus, for instance, \(\mu^s\) and \(\Sigma^s\) are the mean vectors for the noise-corrupted speech models and noise models, respectively, in the cepstral domain.

1. Inverse DCT transformation

\[
\mu^l = C^{-1}\mu^c, \quad \Sigma^l = C^{-1}\Sigma^c(C^{-1})^T
\] (2)

where \(C\) is the matrix for DCT.

2. Exponential transformation

\[
\mu_i = \exp\left(\mu^l_i + \frac{\Sigma^l_{ii}}{2}\right), \quad \Sigma_{ij} = \mu_i\mu_j[\exp(\Sigma^l_{ij}) - 1]
\] (3)

3. Composition

\[
\hat{\mu} = \mu + \hat{\mu}, \quad \hat{\Sigma} = \Sigma + \hat{\Sigma}
\] (4)

4. Logarithm transformation

\[
\hat{\mu}^l_i = \log(\mu_i) - \frac{1}{2}\log\left(\frac{\hat{\Sigma}^l_{ii}}{\hat{\mu}^l_i} + 1\right), \quad \hat{\Sigma}_{ij}^l = \log\left(\frac{\hat{\Sigma}^l_{ij}}{\hat{\mu}^l_i\hat{\mu}^l_j} + 1\right)
\] (5)

5. DCT transformation

\[
\hat{\mu}^c = C\hat{\mu}^l, \quad \hat{\Sigma}^c = C\hat{\Sigma}^lC^T
\] (6)

4. The Integrated Approaches

Usually, noise reduction approach is not able to remove noise completely at low SNR and causes new problems in that insufficient or excessive reduction processing leads to remaining noise or speech distortion, respectively. On the contrary, model compensation approach is also not sufficient when SNR is low, because the feature of speech is buried in the noise. Therefore, an approach that integration of noise reduction with model compensation becomes a possible solution to improve recognition performance in noisy environments, especially at low SNR. Based on the analysis of the integrated PMC approach, we could find its recognition performance most depends on the quality of the noise-reduced input speech at front-end. It also means that the enhanced speech with lower distortion should be shown a higher recognition performance.

In this paper, to improve the performance of the integrated PMC approach, we focus on finding a suitable front-end processing to reduce the noise from the noise-corrupted input speech. Firstly, we could consider TMWF which is an improved version of Wiener filtering and also has been approved as the main part of the AFE in ETSI for DSR. Especially, TMWF is based in perceptual domain (i.e. mel-frequency), so TMWF is possible to raise the SNR of input noise with low speech distortion and then to show an improved recognition performance comparing with the existing well-known WF-PMC, especially at low SNR.
5. Experiments

5.1. Experimental Conditions

To evaluate the performance of the integrated PMC approaches, the speech data used are KLE (center for Korean Language Engineering) 452 database, which is a phonetically balanced isolated word database. The corpus includes 38 male speakers and each speaker pronounces 452 words for one time. Those are divided into two subsets. One has 15,280 words of 35 speakers’ utterance, which used for training clean HMMs. Another has 1,356 words of 3 speakers’ utterance, which used for testing. To set up the noise-corrupted speech data for testing, we add 3 kinds of noises (i.e. subway, car and exhibition noise) to above utterance regions of the noise-reduced speech. The noise model used in PMC framework, a model of one state with one Gaussian distribution, is trained on the noise from the non-utterance region of all available testing data. Here, the number of noise frames extracted from the non-utterance region is 20.

In addition, we also integrate PMC with the first stage of TMWF (i.e. OMWF-PMC) in this paper. Even OMWF coarsely reduces the noise and remains the white residual noise, but if it happens a small speech distortion, the integrated OMWF-PMC is also possible to perform a good recognition performance, because PMC can handle the residual noise at back-end. To investigate its performance, we conduct recognition experiments and compare it with WF-PMC, TMWF-PMC.

5.2. Experimental Results

Figure 3 shows the recognition performance in terms of different SNRs in subway, car and exhibition noise, respectively. In all figures about experimental results, the bar chart represents the performance of noise reduction approaches, while the line graph represents those of the integrated PMC approaches.

First, for bar chart in Fig. 3, comparing the performance of TMWF, OMWF with the well-known WF, we could find that the performance is very similar at high SNR (i.e. 20dB, 15dB), while at low SNR (i.e. 10dB, 5dB) they are big, even OMWF also shows a better performance. It means that both TMWF and OMWF are good noise reduction approaches as well as at low SNR, since those are based on mel-frequency domain that might advantageous for robustness speech recognition.

Second, for line graph in Fig. 3, the most low line graph (i.e. black line) is about the recognition results of PMC. From the results, PMC seems to have benefits only at low SNR (10dB, 5dB) comparing with noise reduction approaches at front-end. In general, PMC shows a similar performance with WF, except in subway noise.

Third, comparing the performance of the integrated PMC approaches (i.e. line graph), it is easy to know that the integrated PMC approaches show a big improved performance comparing with PMC, especially at low SNR (10dB, 5dB). Among the integrated PMC approaches, TMWF-PMC shows the best performance, while OMWF-PMC and WF-PMC are shown a very closed performance.

Above the analysis of Fig. 3 are focus on comparing the visual performance differences of each approach, thus most of them are about low SNR (10dB, 5dB) conditions. To show the completed performance comparisons, we extract the performance results at high SNR (20dB, 15dB) from Fig. 3, and show the enlarged view of those in Fig. 4. In Fig. 4, we could find that the integrated PMC approaches degrade the performance at most conditions, especially at 20dB. It means that the accuracy of noisy HMM generated by PMC is not good enough, further also means the noise used in PMC framework is different with the remaining noise in noise-superposed utterance regions after noise reduction processing. Therefore, to satisfy PMC works well even at high SNR, we consider that the remaining noise in non-utterance regions and the remaining noise in noise-superposed utterance regions should be identical as much as possible.

In addition, we also show the averaged performance over four SNRs (20dB-5dB) in different noise in Fig. 5. From Fig. 5, we could find that the integrated PMC approaches show the significant improvements comparing with other approaches in most noise conditions. For example, comparing with the well-known WF-PMC, TMWF-PMC gives 2.6%, 1.6%, 5.5% averaged improvements, and OMWF-PMC gives -0.3%, 0.5%, 2.2% averaged improvements in subway noise, car noise and exhibition noise, respectively. Therefore, we could confirm that the integrated PMC approach is effective for speech recognition in noisy environments, especially at low SNR.

Table 1: Analysis conditions for speech data.

<table>
<thead>
<tr>
<th>Analysis Type</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-processing</td>
<td>8 KHz Sampling rate, 16 bits, 25 ms Hamming Window, 10 ms frame shift</td>
</tr>
<tr>
<td>Speech Features</td>
<td>12 Delta MFCCs + Delta LogE (total 26 dim.)</td>
</tr>
<tr>
<td>Pre-emphasize</td>
<td>1 − 0.97\text{z}^{-1}</td>
</tr>
</tbody>
</table>
6. Conclusions

In this paper, to find the best noise robustness approach, we study on approaches implemented at both-end (i.e. front-end and back-end) of speech recognition system. To find a suitable front-end noise reduction approach, we investigate TMWF in the integrated PMC approach, referred as TMWF-PMC. As a result, TMWF-PMC shows improved performance comparing with the well-known WF-PMC. Furthermore, we investigate the first-stage of TMWF (i.e. OMWF) for the integrated PMC (i.e. OMWF-PMC) approach, its performance is also very good and comparable to WF-PMC. We consider that the recognition improvements of the integrated PMC approach come from two factors, one is the enhancement of SNRs of input speech by noise reduction approaches at front-end and this cause a good accuracy of noise compensation for acoustic models by PMC at back-end.

7. References