Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

A Unified Approach of Compensation and Soft Masking Incorporating a Statistical Model into the Wiener Filter

Byung-Ok Kang and Ho-Young Jung

Speech/Language Information Reasearch Center,
Electronics and Telecommunications Research Institute, Daejeon, Korea

PACS: 43.72.DV, 43.72.NE

ABSTRACT

In this paper, we present a new single-channel noise reduction method that integrates compensation and soft masking into the same statistical model assumptions for noise-robust speech recognition. By utilizing a Gaussian mixture model(GMM) as a pre-knowledge of speech and added noise signals, the proposed method can effectively restore clean speech spectra and separate out ambient noises from a target speech in the Wiener filter framework. The soft mask methods originally attempted to separate out the speech signal of the speaker of interest from a mixture of speech signals. In the proposed method, by using pre-trained speech and noise models, the soft mask techniques can be applied to separate out added noises from the target speech. Combined with the model-based Wiener filter performing compensation on the power spectrum, the technique can efficiently reduce distortions caused by non-stationary noises and finally reconstruct clean speech spectra from noise-corrupted observation. By applying the result in order to infer the a priori SNR of the Wiener filter, we can estimate the clean speech signal with greater accuracy. While the conventional Wiener filter causes inevitable distortions owing to noise reduction and does not solve non-stationary noises overlapped with speech presence periods, the proposed method can considerably solve these problems through compensation and softmasking based on speech and noise GMMs. The results evaluated in a practical speech recognition system for car environments show that the proposed method outperforms conventional methods.

INTRODUCTION

One of the largest obstacles to the commercialization of speech recognition systems is performance degradation due to background noise. It is commonly known that a speech recognition system trained in a clean environment cannot achieve good performance when working in real environments.

Many approaches have been proposed to address this problem. Most of them can be classified into signal-processing-based spectral enhancement techniques and statistical-model-based model adaptation methods. Among speech enhancement techniques, spectral subtraction, Wiener filter, and minimum mean-square error short-time spectral amplitude (MMSE STSA) estimator [1] are the most widely used approaches. Statistical-model-based model adaptation methods, instead of enhancing the input signal, transforms acoustic models to represent the noisy environment. For this method, proposed techniques include the Maximum a Posteriori (MAP) and Parallel Model Combination (PMC) [2] methods. MAP estimation is a model compensation method based on the amount of adaptation data and priori density.

The Wiener filter is known to be very effective in reducing stationary noise from input speech signals and a Wiener filter in the form of 2 stages is adopted as the standard of ETSI [3]. However, the performance of the Wiener filter often degrades because a Wiener gain function is given by the expectation value of the speech estimated directly from the distorted observation signal.

In this paper, we introduce a modified Wiener filter in which speech and noise GMMs are used to achieve noise reduction based on compensation and soft masking in the same model assumptions.

The soft mask methods [4] originally attempted to separate out the speech signal of the speaker of interest from a mixture of speech signals. In the proposed method, by using pre-trained speech and noise models, the soft mask techniques can be applied to separate out added noises from the target speech. Combined with the model-based Wiener filter performing compensation on the power spectrum, the technique can efficiently reduce distortions caused by non-stationary noises and finally reconstruct clean speech spectra from noise-corrupted observation. By applying the result in order to infer the *a priori* SNR of the Wiener filter, we can estimate the clean speech signal with greater accuracy. While the conventional Wiener filter causes inevitable distortions owing to noise reduction and does not solve non-stationary noises overlapped with speech presence periods, the proposed method can considerably solve these problems through compensation and soft-masking based on speech and noise GMMs.

The remainder of this paper is organized as follows: Firstly, works related to the proposed method are introduced. Namely, the proposed model-based Wiener filter method and Soft mask method for source separation is briefly described. And then, we explain the unified approach of compensation and soft masking. Next section describes performance evaluation. The final section concludes this paper.

MODEL-BASED WIENER FILTER METHOD

In this section, we describe the model-based Wiener filter method (MBW) [5]. The MBW combines a signal processing-based spectral enhancement technique and a spectrum compensation method based on the manner of MMSE estimation using statistical models. By performing compensation on the power spectrum with a clean speech GMM, MBW provides a sensible improvement in the estimation of the_*a priori* SNR in the Wiener filter framework.

The MBW performance procedure is as follows:

1. At the current input frame, the noise component $\overline{N(t)}$ is estimated from an input noisy speech *Y(t)*. $\overline{N(t)}$ is updated for speech absence based on a voice activity detector (VAD). To improve the VAD performance, we adopt a soft decision method instead of an energy-based approach [6].

2. The initial estimate for a clean speech spectrum, $S_0(t)$, is found through a pre-Wiener filter. Because it may include an inevitable distortion due to reducing the noise component not updated in speech presence frames, $S_0(t)$ is temporary clean speech roughly estimated using pre-filtering.

3. Using the pre-trained speech GMM and $S_0(t)$, an expected value of clean speech is obtained by MMSE estimation as

$$p(k \mid \left|\hat{S}_0(t)\right|^2) = \frac{P(k)P(\left|\hat{S}_0(t)\right|^2 \mid k)}{\sum_{k^{'}=1}^{K} P(k^{'})P(\left|\hat{S}_0(t)\right|^2 \mid k^{'})}$$

$$\overline{\left\langle \left|S(t)\right|^2 \right\rangle} = \sum_{k}^{K} p(k \mid \left|\hat{S}_0(t)\right|^2)\,\mu_k \tag{1}$$

where K denotes the number of frequency bins, and μ_k is the mean value of the speech GMM in the power spectrum.

4. Next, the *a priori* SNR with the decision-directed method [1] is calculated by

$$\eta(t) = \beta\frac{\left|\hat{S}(t-1)\right|^2}{\left|N(t-1)\right|^2} + (1-\beta)\frac{\overline{\left\langle \left|S(t)\right|^2 \right\rangle}}{\left|N(t)\right|^2} \tag{2}$$

where β is the smoothing parameter between 0 and 1.

5. Finally, the amplitude of the clean speech is estimated as

$$\left|\hat{S}(t)\right| = \frac{\eta(t)}{1+\eta(t)}\left|Y(t)\right| \tag{3}$$

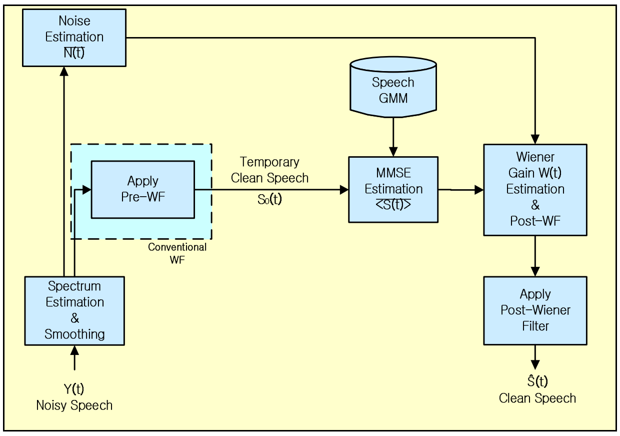


Figure 1. Block diagram of model-based Wiener filter

Figure 1 shows a block diagram of the model-based Wiener filter method. The amplitude of the clean speech signal $\hat{s}(t)$, which is estimated by the proposed MBW, is transferred to the speech recognition system.

SOFT MASK METHOD FOR SOURCE SEPARATION

In the problem of single-channel speaker separation, the speech signal of the speaker of interest is separated out from monaural recordings of multiple concurrent speakers. Most current techniques are based on the principle of masking. The soft mask method reported by Reddy *et al.* computes the probability in which any time-frequency component of the input signal is dominated by the target speaker using speaker models [4].

When the speech and noise GMMs are prepared, the concept of a soft mask can therefore be applied to the problem of noise reduction from the input signal. Namely, if a correct model for the distribution of the noise is provided, the soft mask method can estimate the clean speech spectra by separating a desired speech signal from noisy input signals.

UNIFIED APPROACH OF COMPENSATION AND SOFT MASKING FOR NOISE ROBUST ASR

The MBW performs compensation on the power spectrum and thus estimates more accurately the clean speech spectrum using the GMM of the clean speech signal. If a statistical model of the noise distribution is known and the noise characteristics of the training and test environment are similar, the statistical model of noise can also provide good prior knowledge.

Under this background, we integrate the soft mask method in order to separate out noise components based on pre-noise models. Morris *et al.* have shown that the speech recognition performance obtained with soft masks is significantly better than the performance obtained with binary masks [7]. Because we can assume that speech and noise are uncorrelated and the assumption of log-max approximation is applied, the soft mask can also be applied to separate out speech from ambient noise if a proper noise model is provided for corresponding noisy environments.

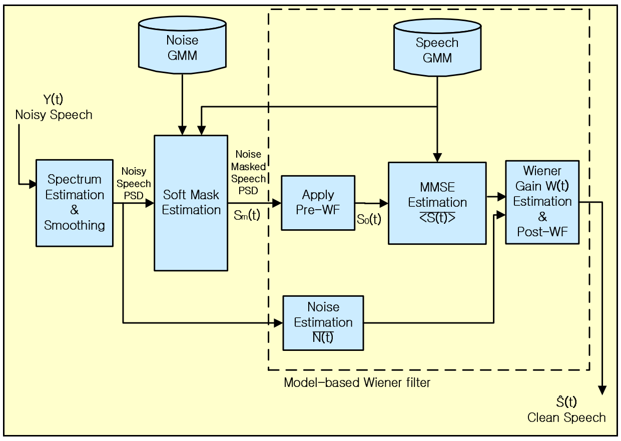


Figure 2. Block diagram of proposed unified approach.

Figure 2 shows a block diagram of the proposed strategy. For the noise-corrupted observation signal, power spectrum estimation and smoothing over time and frequency is carried out in the Wiener filter framework. Then, for log-max approximation of the soft mask method, the logarithm of the power spectrum is calculated. Using noise and speech GMMs, the soft mask for enhancing the input noisy speech is estimated. A noise masked speech power spectrum calculated by soft mask estimation is integrated with the MBW in order to improve the noise reduction performance. The noise masked speech poser spectrum is represented by

$$S_m(t) = P(s > n \mid y) Y(t) \tag{4}$$

$$P(s > n \mid y) = \sum_{k_s, k_n} P(k_s, k_n \mid y) \frac{P_s(y \mid k_s) C_n(y \mid k_n)}{P(y \mid k_s, k_n)}$$

where $P(s > n \mid y)$denotes the soft mask on the log spectrum domain that identifies the contribution of speech in an observation, and k_s and k_n indicate the indexes of speech and noise GMMs.

The effect of eliminating the distortions cased by non-stationary noises and restoring the clean speech spectra flows to the succeeding operation of the MBW. Under the same speech GMM, MBW performs the compensation of musical noise artifacts and restores the final clean spectrum from the noise-masked speech spectrum.

EXPERIMENTAL RESULTS

The unified approach of compensation and soft masking was evaluated via the task of recognizing 46,000 point-of-interests (POIs) for a car navigation system. We used a triphone-based HMM that is a tied-state model with 1150 states, where each state is a mixture of sixteen Gaussian components. The feature vectors consist of 39-dimensional vectors with 13 mel-frequency cepstral coefficients (MFCCs), including C0 and their first and second derivatives.

The training DB for the acoustic model consists of three speech corpuses. The first corpus comprises 130k POI utterances recorded from 623 speakers using a microphone attached to a car in various driving environments. The second consists of 50k utterances recorded from 1800 speakers in a silent office environment and includes a phonetically optimized word set. The final corpus comprises 10k POI utterances recorded from 100 speakers using a navigational system in a car.

The training DB for the noise GMM consists of 40 minutes of car noise DB. It was recorded in a Rezzo, a mini MPV manufactured by Daewoo Motors, and an Accent, a compact car made by Hyundai Motors, using a navigation system equipped in each vehicle. In the experiment, we used a noise GMM that has a mixture of eight Gaussian components. The training DB for the speech GMM consists of 4k utterances recorded from 100 speakers inside idling cars. We used a speech GMM of eight Gaussian components for the experiment.

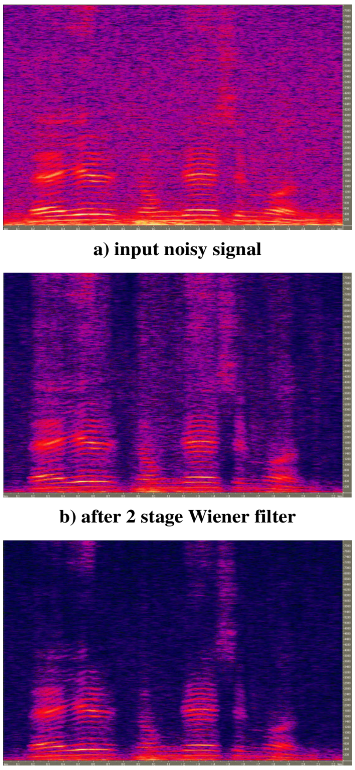
The test DB comprises 1,252 POI utterances recorded from 30 speakers in the high-speed driving environment. It is characterized by added heavy noises such as car audio signal and various car noises which aggravate the performance of the speech recognition system. It was recorded using a navigational system equipped in C and D segment sedans and sports utility vehicles.

The proposed unified method and MBW are evaluated and compared with the conventional 2 stage Wiener filter method. The testing environment is so noisy and the word correction rate performance of the baseline speech recognition system with no enhancement of input signals is 7.5%.

Table 1. Word correction rate(%) after each enhancement methods

2-stage Wiener Filter	Model-Based Wiener Filter	Unified approach
76.42	84.58	86.26

Table 1 shows the speech recognition results of the target task for each of the methods. It can be observed that the proposed unified method shows an error rate reduction (ERR) performance of 42% compared to the conventional 2 stage Wiener filter.



a) input noisy signal

b) after 2 stage Wiener filter

c) after the proposed unified approach

Figure 3. Comparison of signal enhancement results.

Figure 3 shows that the proposed method more effectively removes non-stationary noise overlapped with the desired speech signal compared with the conventional 2 stage Wiener filter.

CONCLUSION

In this paper, we have proposed a novel method for noise reduction that provides compensation and soft masking, incorporating a statistical model into the Wiener filter.

Provided that the pre-statistical model for background noise is reliable, to add to the success of the MBW, the proposed method can reduce inevitable distortions owing to noise reduction and address non-stationary noises overlapped with periods of speech presence. For car noise environments that maintain well the relative correlations between training and test environments, the proposed approach has outperformed conventional methods.

REFERENCES

1 Y. Ephraim and D. Malah, “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, 1984, pp. 1109-1121

2 M.J.F Gales and S.J. Young, “Robust Continuous Speech Recognition Using Parallel Model Combination,” *IEEE Trans. Speech and Audio Processing*, vol. 4, 1999, pp. 352-359

3 ETSI Std. ES 202 050 v1.1.1, “Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms,” 2002

4 A. M. Reddy and B. Raj, “Soft Mask Methods for Single-Channel Speaker Separation,” *IEEE Trans. Audio, Speech, Language Processing*, vol. 25, no. 6, 2007, pp. 1766-1776

5 B. O. Kang, H. Y. Jung, and Y. Lee, “Model Based Wiener Filter for Processing Dynamic Noise,” *Proc. KSPS*, Nov 2007, pp. 104-107

6 J. Sohn, N. S. Kim, and W. Sung, “A Statistical Model-Based Voice Activity Detection,” IEEE Signal Processing Letters, vol. 6, no. 1, 1999, pp. 1-3

7 A.C. Morris, J. Barker, and H. Bourlard, “From missing data to maybe useful data: Soft data modeling for noise robust ASR,” *Proc .Workshop UI SP,* Apr 2001, pp. 153-164