MODEL-BASED WIENER FILTER FOR NOISE ROBUST SPEECH RECOGNITION

Takayuki Arakawa, Masanori Tsujikawa and Ryosuke Isotani

Media and Information Research Laboratories, NEC Corporation, Japan
t-arakawa@cp.jp.nec.com, tujikawa@cb.jp.nec.com, r-isotani@bp.jp.nec.com

ABSTRACT

In this paper, we propose a new approach for noise robust speech recognition, which integrates *signal-processing-based spectral enhancement* and *statistical-model-based compensation*. The proposed method, *Model-Based Wiener filter (MBW)*, takes three steps to estimate clean speech signals from noisy speech signals, which are corrupted by various kinds of additive background noise. The first step is the well-known spectral subtraction (SS). Since the SS averagely subtracts noise components, the estimated speech signals often include distortion. In the second step, the distortion caused by SS is reduced using the minimum mean square error estimation for a Gaussian mixture model representing pre-trained knowledge of speech. In the final step, the Wiener filtering is performed with the decision-directed method. Experiments are conducted using the Aurora2-J (Japanese digit string) database. The results show that the proposed method performs as well as the ETSI advanced front-end in average and the variation range of the recognition accuracy according to the kind of noise is about one third, which demonstrates the robustness of the proposed method.

1. INTRODUCTION

Noise robust speech recognition is desired as an input method for mobile equipment and car equipment. Various kinds of background noise exist in the real world. Therefore robustness against various kinds of noise is quite important. Several approaches have been proposed to deal with this issue [1]-[6].

One approach is *signal-processing-based spectral enhancement*. Examples are the spectrum subtraction (SS) method [1] and the Wiener filter with the decision-directed (DD) method [2]. The ETSI advanced front-end (AFE) [3] using the 2-step Wiener filter shows high noise reduction performance for strongly stationary noise through the use of strong smoothing. This approach needs less computational costs, but needs many tuning costs depending on the kind of noise and signal-to-noise ratio (SNR).

Another approach is *statistical-model-based noise adaptation*. In this approach, instead of removing the noise component, an acoustic model i.e., a hidden Markov model (HMM), is adapted to the noisy environment. The parallel model combination (PMC) method [4] is well known as an example of this kind of approach. This approach is robust against various kinds of noise, because the effect of noise on each Gaussian distribution composing the HMM is considered. However it needs huge computational costs to adapt these distributions to a noisy environment.

Another approach is *statistical-model-based compensation* [5][6]. A mismatch function is estimated in this approach. The function is the difference between clean speech and noisy speech in a logarithmic spectrum domain. The function is derived as a weighted average of moving vectors caused by adapting Gaussian distributions in a Gaussian mixture model (GMM), which is trained by clean speech in advance, to the noisy environment. This approach is robust against various kinds of noise in the same way as the former approach and its computational cost is smaller because the number of the Gaussian distributions to be adapted to the noise can be limited by using the GMM instead of the HMM. However, the computational cost is still much more than that of the *signal-processing-based spectral enhancement*.

In this paper, we propose an approach which is more robust against various kinds of noise than the *signal-processing-based spectral enhancement* and whose computational cost is much smaller than the *statistical-model-based compensation* through the integration of both approaches. The paper is organized as follows. In section 2, Model-Based Wiener filter (MBW) is proposed. The experiments and results on Aurora2-J database are presented in section 3, and our work is summarized in section 4.

2. PROPOSED METHOD

In the following, we show the procedure of the MBW algorithm (see Fig.1). The noisy speech signal $\boldsymbol{X}(t)$ is modeled as

$$\boldsymbol{X}(t) = \boldsymbol{S}(t) + \boldsymbol{N}(t), \tag{1}$$

where $\boldsymbol{X}(t), \boldsymbol{S}(t)$, and $\boldsymbol{N}(t)$ denote the vectors of spectrum at t-th short time frame of noisy input speech, clean component and noise component respectively.

1. A noise component $\overline{\boldsymbol{N}(t)}$ is estimated from a noisy input speech $\boldsymbol{X}(t)$. We use a weighted noise estimation method [7]. In this method, the estimated noise com-

ponent is updated with a weighted value depending on SNR.

2. The SS method is performed, and a temporary clean speech $\widehat{\boldsymbol{S}_0}(t)$ is estimated,

$$\widehat{\boldsymbol{S}_0}(t) = \max(\boldsymbol{X}(t) - \overline{\boldsymbol{N}(t)}, \alpha \boldsymbol{X}(t)), \quad (2)$$

where α indicates a *flooring parameter*.

3. The temporary clean speech is transformed into cepstrum,

$$\widehat{\boldsymbol{C}_0}(t) = \mathrm{DCT}\left[\log(\widehat{\boldsymbol{S}_0}(t))\right], \quad (3)$$

where DCT $[\cdot]$ means the discrete cosine transform.

4. An expected value of the clean speech is derived from the below equations.

In the proposed method, a GMM with K Gaussian distributions is used as the knowledge of clean speech in the cepstrum domain

$$\mathrm{P}(\boldsymbol{C}) = \sum_{k=1}^{K} \mathrm{P}(k)\mathrm{P}(\boldsymbol{C}|k), \quad (4)$$

$$\boldsymbol{C} = \mathrm{DCT}\left[\log \boldsymbol{S}\right], \quad (5)$$

where $\mathrm{P}(k)$ is the mixture weight of the k-th Gaussian distribution (*a priori* probability), $\mathrm{P}(\boldsymbol{C}|k)$ is a Gaussian distribution $\mathrm{P}(\boldsymbol{C}|k) = \mathcal{N}(\boldsymbol{C}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. All of these parameters are trained with a sufficient amount of clean speech in advance.

A posteriori probability of the k-th Gaussian distribution for the temporary estimated speech $\widehat{\boldsymbol{C}_0}(t)$ is derived,

$$\mathrm{P}(k|\widehat{\boldsymbol{C}_0}(t)) = \frac{\mathrm{P}(k)\mathrm{P}(\widehat{\boldsymbol{C}_0}(t)|k)}{\sum_{k=1}^{K} \mathrm{P}(k)\mathrm{P}(\widehat{\boldsymbol{C}_0}(t)|k)}. \quad (6)$$

An expected value of the clean speech is calculated as the manner of the minimum mean square error (MMSE) estimation [5].

$$\overline{\langle \boldsymbol{S}(t) \rangle} = \exp\left[\sum_{k=1}^{K} \mathrm{P}(k|\widehat{\boldsymbol{C}_0}(t))\boldsymbol{\mu}_k^{\log}\right], \quad (7)$$

where $\boldsymbol{\mu}_k^{\log} = \mathrm{IDCT}[\boldsymbol{\mu}_k]$, IDCT$[\cdot]$ is the inverse discrete cosine transform.

The concept of Eq.(7) is described in Fig.2. In this figure, $\boldsymbol{\mu}_k$ indicates a mean value of a Gaussian distribution of clean speech GMM. The temporary value, which is roughly estimated by the SS, is pulled back into a region which is considered to be clean speech.

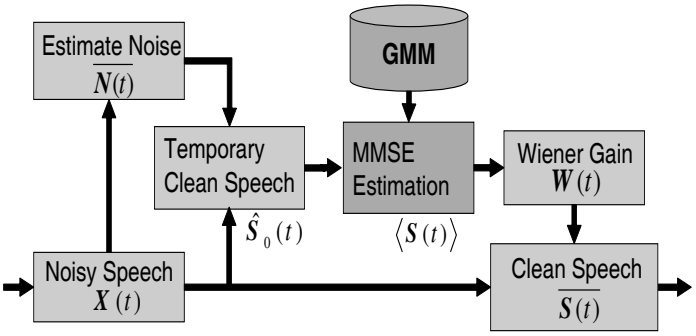

Fig. 1. MBW algorithm

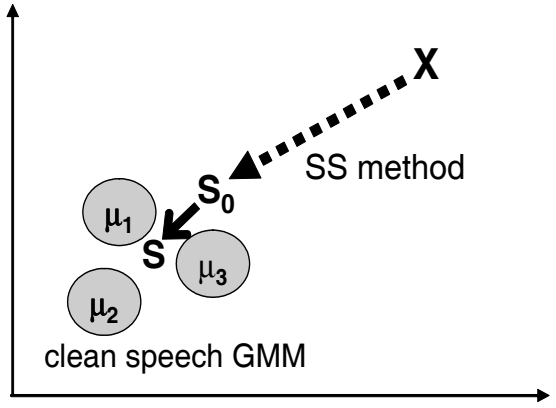

Fig. 2. MMSE estimation

5. Wiener gain with the DD method is solved in a spectrum domain.

$$\boldsymbol{W}(t) = \frac{\boldsymbol{\eta}(t)}{\boldsymbol{\eta}(t)+1}, \quad (8)$$

$$\boldsymbol{\eta}(t) = \beta\boldsymbol{\eta}(t-1) + (1-\beta)\frac{\overline{\langle \boldsymbol{S}(t) \rangle}}{\overline{\boldsymbol{N}(t)}}, \quad (9)$$

where β is a smoothing parameter.

6. We get the final estimated clean speech by multiplying Eq.(8) by the noisy input speech

$$\overline{\boldsymbol{S}(t)} = \boldsymbol{W}(t)\boldsymbol{X}(t). \quad (10)$$

If Eq.(7) is set to be a final result, particularly when the number of Gaussian distributions is small, an excessive amount of noisy input speech information is lost. For this reason, the Wiener gain is calculated in Eq.(10). Moreover, we can perform more correct estimation by iterating procedures 3 to 6.

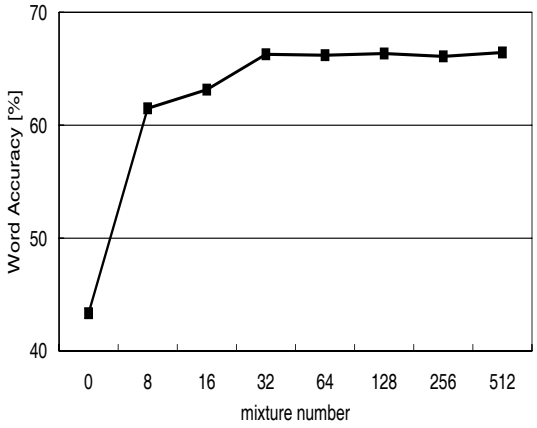

Fig. 3. The Word Accuracy as a function of the mixture number of the GMM for 5dB restaurant noise

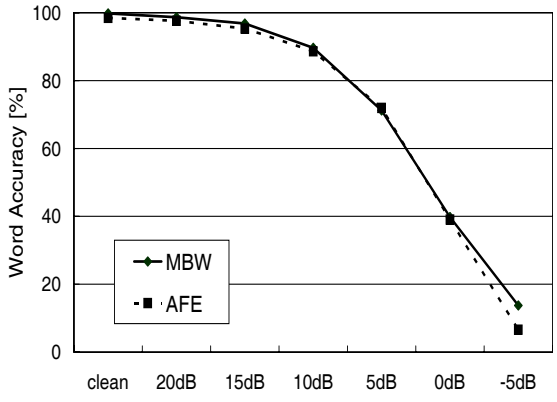

Fig. 4. Word Accuracy averaged for various kinds of noise for each SNR

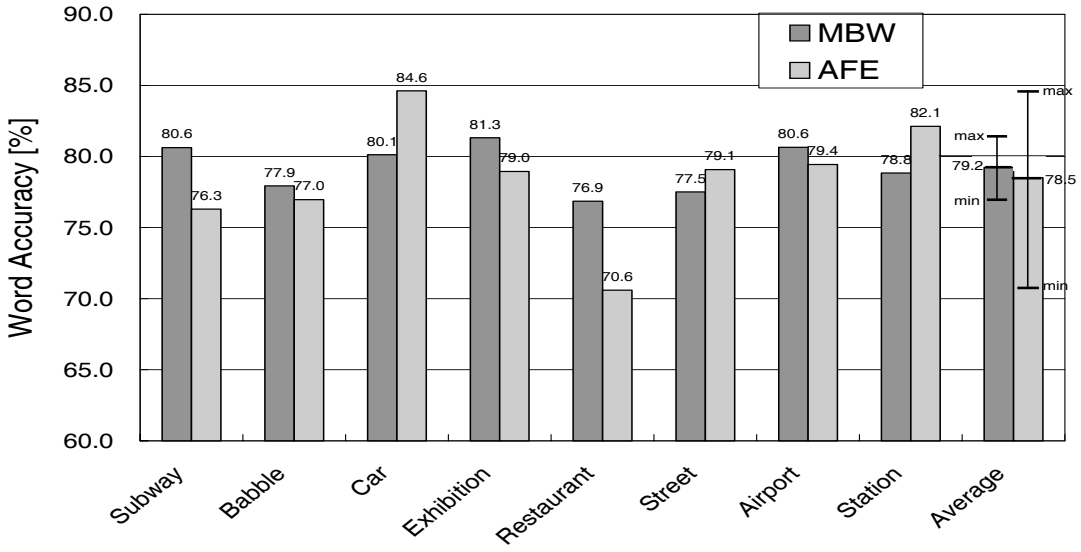

Fig. 5. Word Accuracy averaged over the SNR for each kind of noise

3. EXPERIMENTS AND RESULTS

3.1. EXPERIMENTAL CONDITIONS

We performed experiments with the MBW and the AFE under the same conditions. The Mel-frequency cepstral coefficients (MFCC) and their 1st and 2nd derivatives are used as feature value of speech. A cepstrum dimension is set at 13 including $zero$-th MFCC, so we use a 39-dimension feature value. This feature value is used for training the HMM and recognizing test data. Aside from this, we prepared a GMM for noise reduction. The feature value for this GMM is composed of a 13-dimensional MFCC only. This GMM is pre-trained with the same data as that for the HMM. The flooring parameter in Eq.(2) is set at 0.1, and the smoothing parameter in Eq.(9) is set at 0.98. In these experiments, the procedures 3 to 6 in the former section are iterated twice.

3.2. AURORA2-J TASK

The MBW method was tested on the Aurora2-J task [8]. This task contains utterances (in Japanese) of consecutive digit string recorded in clean environments. There are two different experiments that can be conducted for this task: a clean-condition training scenario and a multi-condition training scenario. In our work, we performed the clean-condition training scenario. A total of 8,440 clean digit string utterances spoken by 110 speakers were used for training, and 1,001 digit string utterances spoken by 104 speakers were used for testing. These utterances are recorded at 16kHz sampling rate and down-sampled to 8kHz. Several kinds of noise (subway, babble, car, etc.) were added to these utterances in several SNRs from

-5dB to 20dB. After that, these utterances were filtered with G.712 to consider the realistic frequency characteristics of terminals and equipment in the telecommunication area [9]. The recognition system is provided and based on the HMM, which is trained with the HMM Tool Kit (HTK) software [10].

3.3. PRELIMINARY EXPERIMENTS

We performed preliminary experiments to determine the mixture number of the GMM. Figure 3 shows the Word Accuracy (W.A.) as a function of the mixture number of the GMM. This experiment was performed on 5dB restaurant noise condition. The restaurant noise includes mixed murmuring of voices and clatter of dishes, and it is not stationary. When the mixture number is 0, the MBW method is equivalent to the usual Wiener filter with the DD method. As the mixture number increases, the performances of the MBW method improves. At the point of 128 or 256, it becomes saturated. Therefore, the mixture number is set at 256 in the following experiments.

3.4. RESULTS

The results are summarized in Figs.4-5. Figure 4 shows the Word Accuracy for each SNR, which is averaged over the kind of noise. The average performance of the proposed method is almost equivalent to that of the AFE. Figure 5 shows the Word Accuracy averaged over the SNR for each kind of noise. The scores of the AFE varies widely with the kind of noise. The AFE performs well for highly stationary noise like car noise, but not so well for a non-stationary noise like restaurant noise. On the other hand, the scores of the MBW method is almost constant. The average and the range of the scores are shown in the right side of this figure. Though the average value is almost the same in both methods, the range of the MBW method is almost one-third that of the AFE. These results show that the proposed method is much more robust than the AFE against various kinds of noise.

4. SUMMARY

In this paper, we have proposed the *Model-Based Wiener filter (MBW)* method, a new approach for noise robust speech recognition which integrates *signal-processing-based spectral enhancement* and *model-based compensation*. The MBW method first roughly estimates clean speech signals using SS and compensates them using a GMM trained on clean speech to improve robustness against non-stationary noise. The compensated speech signal is used to calculate the Wiener gain with DD method to obtain the clean speech for use in speech recognition. Wiener filtering strongly suppresses stationary noise. The MBW method was compared with the AFE on Aurora2-J database. The results show that the proposed method performs as well as the ETSI AFE in average and the variation range of the recognition accuracy according to the kind of noise is about one third that of the AFE. These results demonstrate that the proposed method is robust against various kinds of noise. In future work, we are planning to evaluate the proposed method on English and other tasks.

5. REFERENCES

[1] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. ASSP 27, pp.113-120, 1979.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," IEEE Trans. ASSP 32, pp.1109-1121, 1984.

[3] ETSI ES 202 050 v1.1.1, "Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms," 2002.

[4] M. J. F. Gales and S. J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination", IEEE Trans. Speech and Audio Processing, Vol.4, pp.352-359, 1996.

[5] J. C. Segura, A. de la Torre, M. C. Benitez and A.M. Peinado, "Model-Based Compensation of the Additive Noise for Continuous Speech Recognition. Experiments Using AURORA II Database and Tasks," EuroSpeech'01, Vol.I, pp.221-224, 2001.

[6] M. Fujimoto and Y. Ariki, "Robust Speech Recognition in Additive and Channel Noise Environments using GMM and EM Algorithm," ICASSP'04, Vol.I, pp.941-944, 2004.

[7] M. Kato, A. Sugiyama and M. Serizawa, "A Family of 3GPP-standard Noise Suppressors for the AMR Codec and the Evaluation Results," ICASSP'03, Vol.I, pp.868-871, 2003.

[8] S. Nakamura, K. Yamamoto, K. Takeda, S. Kuroiwa, N. Kitaoka, T. Yamada, M. Mizumachi, T. Nishiura, M. Fujimoto, A. Sasou and T. Endo, "Data Collection and Evaluation of Aurora2-J Japanese Corpus," ASRU'03, pp.619-623, 2003.

[9] ITU recommendation G.712, "Transmission Performance Characteristics of Pulse Code Modulation Channels," 1996.

[10] HTK Web site:
http://htk.eng.cam.ac.jpk/