Isolated Word Recognition Based on Combination of Multiple Noise-Robust Techniques

Noboru Hayasaka[1], Non-member

ABSTRACT

Although many noise-robust techniques have been presented, the improvement under low SNR condition is still insufficient. The purpose of this paper is to achieve the high recognition accuracy under low SNR condition with low calculation costs. Therefore, this paper proposes a novel noise-robust speech recognition system that makes full use of spectral subtraction (SS), mean variance normalization (MVN), temporal filtering (TF), and multi-condition HMMs (MC-HMMs). First, from the results of SS with clean HMMs, we obtained the improvement from 46.61% to 65.71% under 0 dB SNR condition. Then, SS+ MVN+TF with clean HMMs improved the recognition accuracy from 65.71% to 80.97% under the same SNR condition. Finally, we achieved the further improvement from 80.97% to 92.23% by employing SS+MVN+TF with MC-HMMs.

Keywords: Noise-Robust Speech Recognition, Spectral Subtraction, Feature Normalization, Temporal Filtering, Multi-Condition HMM

1. INTRODUCTION

There are high hopes of a voice interface as a method of resolving digital divide problem. However, there are few products with the voice interface due to degrade recognition accuracy in noisy environments. Many systems deal with this by methods such as using several microphones or a headset microphone. However, these methods have disadvantages in costs or usability. For those reasons, improvements in performance when using a single stand-type microphone present a huge challenge.

Noise reduction methods typified by spectral subtraction (SS) [1] or a Weiner filter (WF) [2] are widely used to improve the recognition accuracy. In particular, WF has been adopted into the ETSI ES 202 050 (ETSI-AFE), standardized by the European Telecommunications Standards Institute (ETSI) [3]. WF, however, has high calculation costs in comparison with SS.

As another approach, noise-robust feature extraction methods have been proposed. Especially, Cepstral mean normalization [4], mean variance normalization (MVN) [5] have been widely used for compensating for statistical mismatches during training and testing conditions. These methods are called feature normalization. In addition, MVA processing [6] can also improve the recognition accuracy. MVA emphasizes important changes by applying an autoregressive-moving-average (ARMA) filer to the features followed by MVN. The authors have also already proposed a finite impulse response (FIR) filter applied to the features [7]. These methods are generally called temporal filtering (TF).

Solutions by means of an acoustic model have also been proposed (e.g., hidden Markov model (HMM) com-position [8] and multi-condition training (MC-training) [9]). The HMM composition method creates a new HMM composing both the HMMs of speech and noise in the linear spectrum domain. Then, MC-training is a training method that generate an HMM (MC-HMM) from speech signals under various environments.

Although many approaches have been presented, the improvement under low SNR condition is still insufficient. Moreover, few systems with combination of them have been presented. With these backgrounds, in this paper, we propose a novel noise-robust speech recognition system that makes full use of SS as the noise reduction, MVN and TF as the robust feature extraction, and MC-HMMs as the robust acoustic models. The purpose of this study is to realize the high recognition accuracy under low SNR condition with low calculation costs.

First, we mention SS and its performance in Section 2. Next, we describe the noise-robust feature extraction and its performance in Section 3. Then, we report MC-HMMs and its performance in Section 4. Finally, we summarize and talk of future challenges in Section 5.

2. NOISE REDUCTION

2.1 Spectral Subtraction

To analyze an observed signal, the short-time Fourier transform is applied to it with a frame width and period. When $X_i(t)$, $S_i(t)$, and $N_i(t)$ (where i is the frequency bin) represent the discrete Fourier

Manuscript received on July 31, 2011 ; revised on November 1, 2011.

[1] The author is with the Graduate School of Engineering Science, Osaka University, Japan , E-mail: hayasaka@sys.es.osaka-u.ac.jp

transform on the observed signal, a speech signal, and a noise signal of the t^{th} frame, respectively, the power spectrum on the observed signal can be expressed as

$$\begin{aligned}|X_i(t)|^2 = |S_i(t)|^2 + |N_i(t)|^2 \\ + 2|S_i(t)||N_i(t)|\cos\theta_i(t),\end{aligned} \tag{1}$$

where $\cos\theta_i(t)$ is correlation term between $S_i(t)$, and $N_i(t)$. The correlation term assumes 0 in a lot of systems.

$$|X_i(t)|^2 = |S_i(t)|^2 + |N_i(t)|^2 \ , \tag{2}$$

$$|X_i^{SS}(t)|^2 = \begin{cases} |X_i(t)|^2 - \alpha|\hat{N}_i(t)|^2 \\ if \ \ |X_i(t)|^2 - \alpha|\hat{N}_i(t)|^2 > \beta \\ \beta|X_i(t)|^2 \ , \qquad \text{otherwise} \end{cases} \tag{3}$$

where α and β are an over estimated coefficient and a flooring coefficient, respectively. In this paper, $|\hat{N}_i(t)|^2$ is estimated noise spectra, which calculated from 200 ms immediately before the observed signal is inputted. α copes with variances of noise spectra, β prevents that $|X_i^{SS}(t)|$ become negative value. SS has been widely used, but its recognition performance depends on those parameters. Therefore, they properly need selecting.

2.2 Experimental conditions

We performed noisy speech recognition experiments in order to evaluate SS. We used the 100 place names database supplied from Japan Electronics and Information Technology industries Association (JEITA), and each data sample was a noise-free speech signal at a sampling frequency of 16 kHz with 16-bit quantization. We used the 17 kinds of noise listed in Table 1 from JEITA database, each type of noise was added to the noise-free speech signals at 20,

Table 1: *Noisy environments*

Running car (2000 cc)	Crowd
Running car (1500 cc)	Train (bullet train)
Exhibition hall (booth)	Train (conv. line)
Exhibition hall (in aisle)	Computer room (mid-size)
Station	Computer room (w. s.)
Telephone box	Air-conditioner (large)
Factory	Fan-coil
Sorting site	Elevator hall
Highway	

Table 2: *Analysis conditions*

Pre-emphasis	1-0.97z^{-1}
Frame length / Frame period	25ms / 10ms
Window function	Hamming

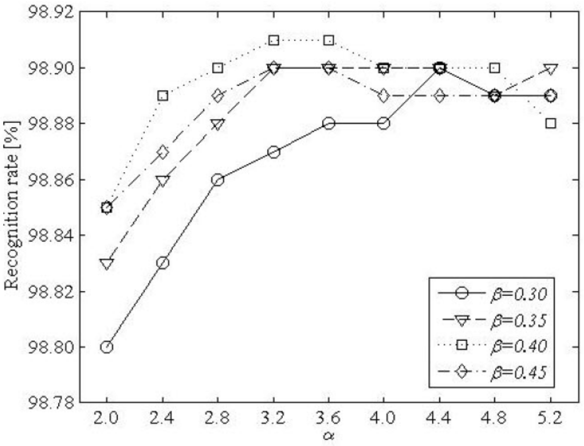

Fig.1: *Results of SS under 20 dB SNR condition*

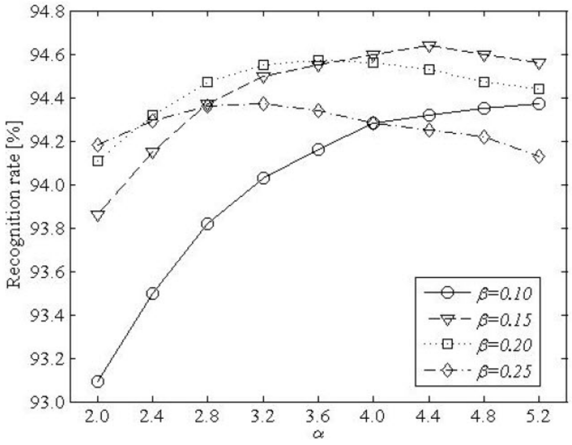

Fig.2: *Results of SS under 10 dB SNR condition*

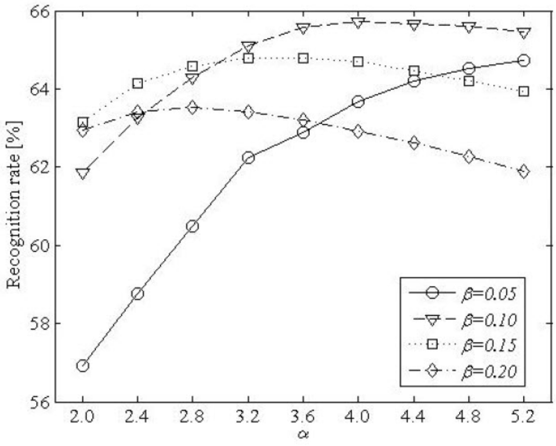

Fig.3: *Results of SS under 0 dB SNR condition*

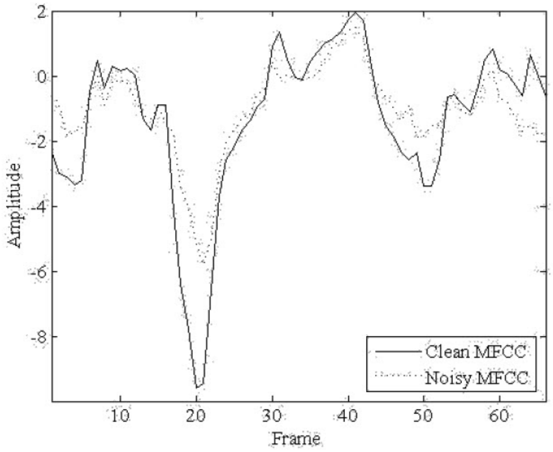

Fig.4: *Time-series of 1st-MFCCs*

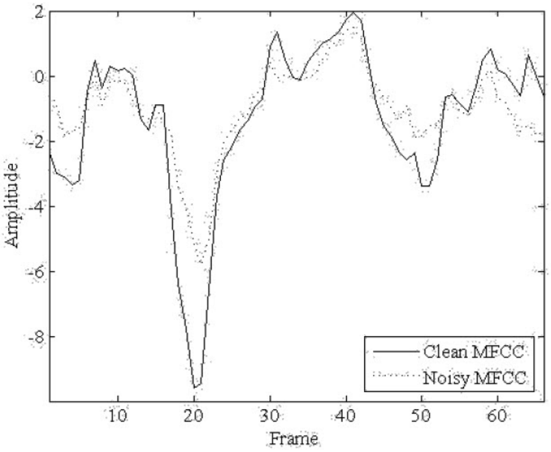

Fig.5: *Time-series of 1st-MFCCs followed by MVN*

10, and 0 dB SNR levels. Features consisted of 13-dimensional MFCCs (including 0^{th}-MFCC), and their first and second order time-derivatives. The analysis conditions are shown in Table 2. The HMMs were modelled in word units. The number of states was determined based on the word length (22 to 70 states), and the number of mixtures for each state was set to 2.

2.3 Experimental results of SS

Figures 1-3 show the results. These figures are plotted with α as the horizontal axis and the recognition rate as the vertical axis. We notice that the best result under each SNR condition was obtained from different α and β. In other words, it was difficult to obtain the best results under all SNR conditions with the same parameters. Then,
Figs.1-3 gave us the fact that α should be selected in proportion to β. In Fig.2, for example, $\alpha = 5.2$ for $\beta = 0.10$, $\alpha = 4.4$ for $\beta = 0.15$, $\alpha = 3.2$ for $\beta = 0.20$, and $\alpha = 2.8$ for $\beta = 0.25$ were proper values. In addition, the smaller β tended to be preferred under lower SNR condition. We presume that this tendency was caused by the over-subtraction under high SNR condition.

3. NOISE-ROBUST FEATURE EXTRACTION

3.1 Feature normalization

MVN compensates for channel distortion and additive noise by normalizing a mean and a variance of each feature. Now, we rewrite Eq. (2) in order to consider the channel distortion. The equation is given by

$$|X_i(t)|^2 = |H_i(t)|^2(|S_i(t)|^2 + |N_i(t)|^2), \qquad (4)$$

where $H_i(t)$ is the channel distortion. To extract mel-frequency cepstral coefficients (MFCCs), which are useful features for speech recognition, we in turn perform mel-frequency transformation, logarithmic transformation, and discrete cosine transformation (DCT). The representation of Eq. (4) in feature space is given as

$$C_n^X(t) = C_n^H(t) + C_n^{S,N}(t), \qquad (5)$$

where $C_n^X(t)$, $C_n^H(t)$, and $C_n^S, N(t)$ are the n^{th}-MFCC of the observed signal, the channel distortion, and mixed signal (i.e., $C_n^S, N(t)$ corresponds to $|S_i(t)|^2 + |N_i(t)|^2$ in Eq.(4)), re-spectively. Since the channel distortion slowly fluctuate in frame-time domain, we can assume that a derivative of each $C_n^H(t)$ is nearly 0. Therefore, the effect of the channel distortion can be reduced by subtracting the frame-time average of each $C_n^X(t)$. Then, it is known that the additive noise compresses a dynamic range of each $C_n^X(t)$ [10]. Fig. 4 shows 1st-MFCCs for an utterance. In this figure, the “Clean MFCC” denotes 1st-MFCC on a noise-free speech signal, and the “Noisy MFCC”

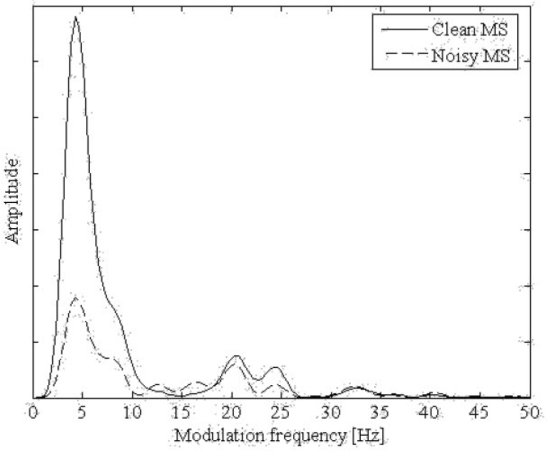

Fig.6: *Modulation spectra of 1st-MFCCs followed by MVN*

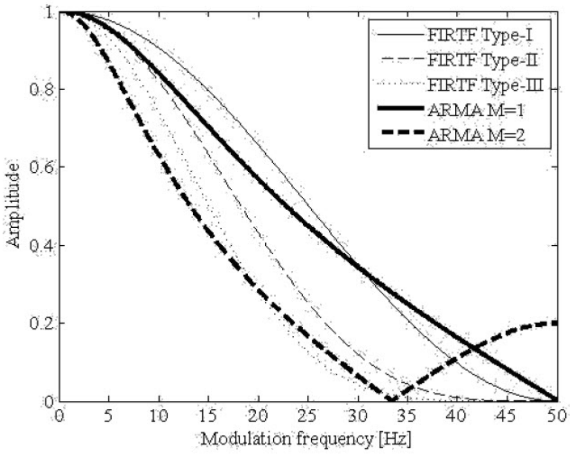

Fig.7: *Amplitude responses of temporal filters*

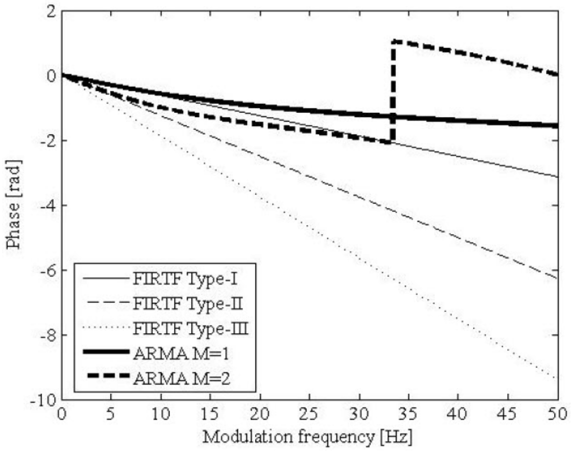

Fig.8: *Unwrapped phase responses of temporal filters*

denotes that on a noisy speech signal. From Fig.4, we can verify that the dynamic range of "Noisy MFCC" is compressed by the additive noise. The effect of the additive noise can be reduced by normalizing the variance of each $C_n^X(t)$. The features followed by MVN are calculated by

$$C_n^{MVN}(t) = \frac{C_n^X(t) - \mu_n}{\sigma_n}, \tag{6}$$

where $C_n^{MVN}(t)$ is an n^{th}-MFCC followed by MVN, μ_n and σ_n are given by

$$\mu_n = \frac{1}{T}\sum_{t=1}^{T} C_n^X(t), \tag{7}$$

$$\sigma_n = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\{C_n^X(t) - \mu_n\}^2}, \tag{8}$$

where T is the frame number of the utterance. Here, Fig. 5 shows 1^{st}-MFCCs followed by MVN. The difference of the range is compensated for, and we can confirm that these MFCCs have high similarity. Moreover, it can be expected that MVN reduces the effect of the over-subtraction in SS because we can assume that the over-subtraction is the mismatch during training and testing.

3.2 Temporal filtering

It is important that we effectively deal with frequency-domain signals of time-series of features or spectra. The frequency-domain signals are called modulation spectra, and its frequencies are called modulation frequencies. It has been reported in [10] that a part of modulation spectra have been important for speech recognition. Especially, modulation frequencies of 1 to 16 Hz are an essential band corresponding to change of syllable. Figure 6 shows clean and noisy modulation spectra of 1^{st}-MFCCs followed by MVN. This figure is plotted with the modulation frequencies as the horizontal axis and its amplitude as the vertical axis, and "Clean MS" and "Noisy MS" indicate clean and noisy modulation spectrum, respectively. We can find out a modulation frequency band with large (1 to 10 Hz) and small (30 to 50 Hz) amplitude in this figure. The band with large amplitude is very important for recognition. In addition, it can be found out that the important band in "Noisy MS" was attenuated by noise. Therefore, a significant improvement can be expected by emphasizing the important band using TF.

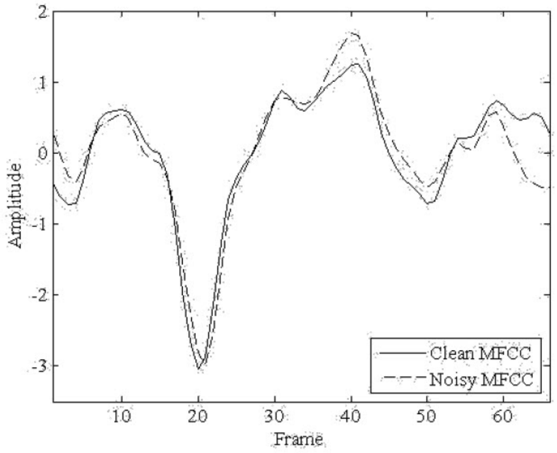

Fig.9: *Time-series of 1^{st}-MFCCs followed by MVN+TF*

For example, Chen, et al. in [6] have proposed the method that applies an ARMA filter to the features followed by MVN. The ARMA filter is defined by

$$H^{ARMA}(z) = z^M \frac{\sum_{m=0}^{M} z^{-m}}{2M + 1 - \sum_{m=1}^{M} z^{-m}}, \tag{9}$$

where M is the order of the ARMA. It has been reported in [6] that the good results have been obtained with low calculation costs ($M = 2$). However, the ARMA filter can cause phase-distortion due to nonlinear-phase. As a result, the potential performance cannot be reduced.

Therefore, we employ an FIR filter as TF for overcoming the disadvantage. An FIR filtering can exclude the effect of the phase-distortion, and an FIR filter can flexibly be designed. In this paper, we prepare three different the FIR filters with the similar amplitude response to that of the ARMA filter. Moreover, the FIR filters are also designed to have as few filter-taps as possible. Figures 7 and 8 show the amplitude responses and the phase responses, respectively. "FIRTF" indicates the characteristics of the FIR filters. First, all amplitude responses in Fig. 7 are gentle. The filters have few filter-taps, and consequently we can realize the low calculation costs. Then, in Fig. 8, we can find out that the nonlinear-phase of "ARMA" and the linear-phase of "FIRTF".

Now, Fig. 9 shows 1st-MFCCs followed by MVN and TF. TF smoothes these MFCCs, which present the higher similarity than those in Fig. 5.

3.3 Experimental results with respect to noise-robust features

In this section, we report the results of experiments with respect to the noise-robust features. The experimental conditions were the same as those in Section 2.2.

First, we performed the experiments to confirm the effectiveness of TF. The results are shown in Table 3. Under lower SNR conditions, "ARMA" led to improvements in comparison with "no-filter", and "FIRTF" presented further improvements. Additionally, the filters with high attenuation (i.e.,"FIRTF Type-III" and "ARMA $M = 2$") tended to be preferred in "Average".

Then, we carried out the experiments combined MVN+FIRTF (Type-III) with SS. The results are shown in Figs. 10-12. The improvements were obtained under all SNR conditions. Especially, under 0 dB SNR condition, we obtained the improvements from 77.38% to 80.97% in the best case (i.e., parameter set $\{\alpha,\beta\}$ is $\{2.4, 0.05\}$). Furthermore, we can notice that the best performance under all SNR conditions can be obtained from the similar param-

Table 3: *Recognition results [%] of noise-robust features with clean HMMs*

SNR	MVN+ no-filter	MVN+ ARMA M=1	MVN+ ARMA M=2	MVN+ FIRTF Type-I	MVN+ FIRTF Type-II	MVN+ FIRTF Type-III
20 [dB]	**99.18**	**99.18**	99.14	99.17	99.16	99.14
10 [dB]	95.92	96.50	96.27	96.42	**96.54**	96.47
0 [dB]	70.17	76.58	76.92	75.76	77.04	**77.38**
Average	88.42	90.75	90.77	90.45	90.91	**91.00**

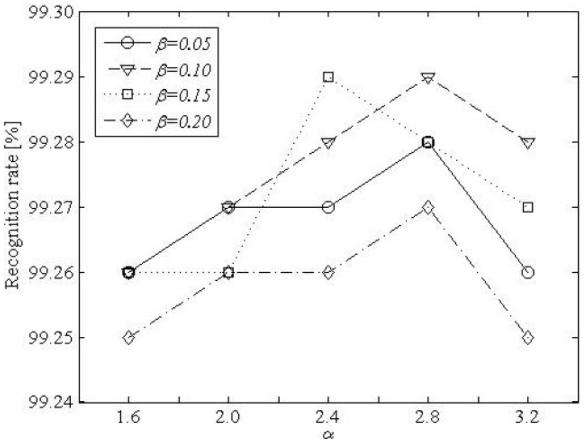

Fig.10: *Results of SS+MVN+FIRTF (Type-III) with clean HMMs under 20 dB SNR condition*

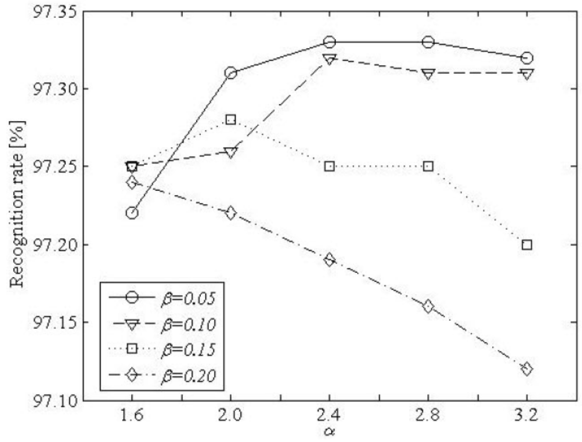

Fig.11: *Results of SS+MVN+FIRTF (Type-III) with clean HMMs under 10 dB SNR condition*

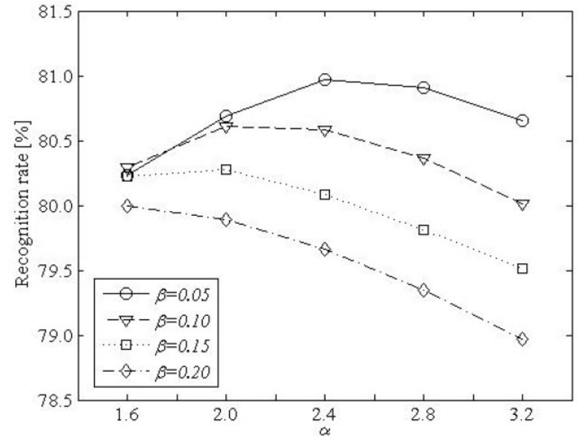

Fig.12: *Results of SS+MVN+FIRTF (Type-III) with clean HMMs under 0 dB SNR condition*

eter set (i.e., $\{\alpha, \beta\}$ is $\{2.4,\ 0.05\}$ or $\{2.4,\ 0.15\}$). In other words, the effects of the over-subtraction were reduced by combining the noise-robust features with SS. However, since the recognition accuracy is still insufficient, we introduce multi-condition training (MC-Training).

4. NOISE-ROBUST ACOUSTIC MODEL

4.1 Multi-condition HMM

MC-training has been proposed as a training method to improve the recognition performance in noisy environments [9]. The method generates an acoustic model from noise-free and noisy speech signals. The acoustic models are called MC-HMMs. The schematic of MC-HMMs is shown in Fig. 13. If the speech signals for training are recorded from single environment (i.e., Noise-free, Noise-A, and so on), the HMMs specialized in the environment are generated. To train MC-HMMs, therefore, we should use the speech signals recorded under various environments. Consequently, MC-HMMs can acquire the ability to cope with speech signals under various environments. Here, Table 4 shows the detailed results in the previous experiments. In this paper, a training set for MC-HMMs consists of both the noise-free speech signals and noisy speech signals. To generate the noisy speech signals, “Station”, “Factory”’, and “Sorting site” were artificially added to the noise-free speech signals at 20 and 10 dB SNR levels. These types of noise are the worst 3 in Table 4.

4.2 Experimental results with respect to MC-HMMs

We inform the results of experiments with respect to MC-HMMs in this section. In the experimental conditions, the number of mixtures for each state was set to 3, and all the others were the same as those in Section 2.2.

Table 4: Detailed results [%] of SS+MVN+FIRTF (Type-III) with clean HMMs

Noise \ SNR	0[dB]	10[dB]	20[dB]
Running car (2000 cc)	98.90	99.47	99.60
Running car (1500cc)	98.27	99.40	99.50
Exhibition hall (booth)	75.23	97.07	99.23
Exhibition hall (in aisle)	83.07	97.83	99.10
Station	**58.17**	**94.43**	**98.83**
Telephone box	98.40	99.40	99.57
Factory	**60.93**	**95.23**	**99.10**
Sorting site	**62.73**	**94.87**	**99.13**
Highway	78.17	97.20	99.30
Crowd	88.03	98.20	99.23
Train (bullet train)	97.27	99.27	99.50
Train (conv. line)	73.83	95.07	99.27
Computer room (mid-size)	80.60	98.10	99.47
Computer room (w. s.)	83.97	97.63	99.17
Air-conditioner (large)	84.43	98.13	99.40
Fan-coil	89.83	98.60	99.23
Elevator hall	64.63	94.70	98.97
Average	**80.97**	**97.33**	**99.27**

Table 5: Recognition results [%] of noise-robust features with MC-HMMs

SNR	MVN+ no-filter	MVN+ ARMA M=1	MVN+ ARMA M=2	MVN+ FIRTF Type-I	MVN+ FIRTF Type-II	MVN+ FIRTF Type-III
20 [dB]	**99.58**	99.52	99.38	99.55	99.50	99.48
10 [dB]	**98.91**	98.86	98.51	**98.91**	98.83	98.74
0 [dB]	88.91	90.53	89.43	90.51	**90.61**	90.40
Average	95.80	96.30	95.77	**96.32**	96.31	96.21

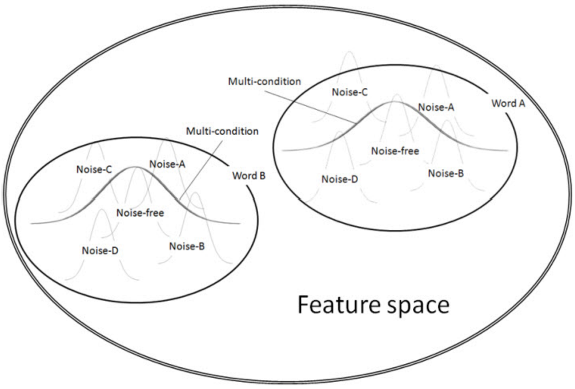

Fig.13: Schematic of MC-HMMs

First, we evaluated the performance of MC-HMMs with the noise-robust features. The results are shown in Table 5. TF did not effectively work under 20 and 10 dB SNR conditions, although TF was able to improve the accuracy under 0 dB SNR. TF can emphasize the important modulation spectra attenuated by noise but damage modulation spectrum of speech in high modulation frequency band. Under the SNR conditions considered in MC-training, TF normally should be skipped. However, it is difficult to know SNR for an input signal. Among these filters, therefore, “FIRTF Type-I” is prospective TF be-cause the filter maintained the performance under 20 and 10 dB SNR conditions.

Then, we ran the experiments introduced SS to MVN +FIRTF (Type-I) with MC-HMMs. The results are shown in Figs. 14-16. Although we achieved further improve-ments from 90.51 to 92.23 under 0 dB SNR condition by introducing SS, the performance did not advance under the higher SNR. This tendency was also caused by the above-mentioned reason. It should be noted that introduction of SS improved the performance under the unconsidered SNR condition without degrading the performance under the considered SNR conditions.

Finally, Table 6 shows the detailed results of SS+

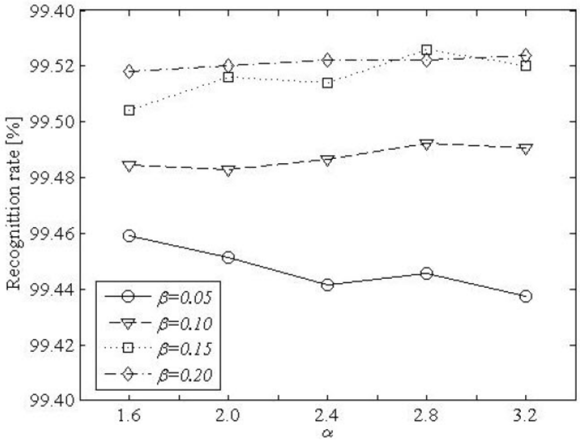

Fig.14: *Results of SS+MVN+FIRTF (Type-I) with MC-HMMs under 20 dB SNR condition*

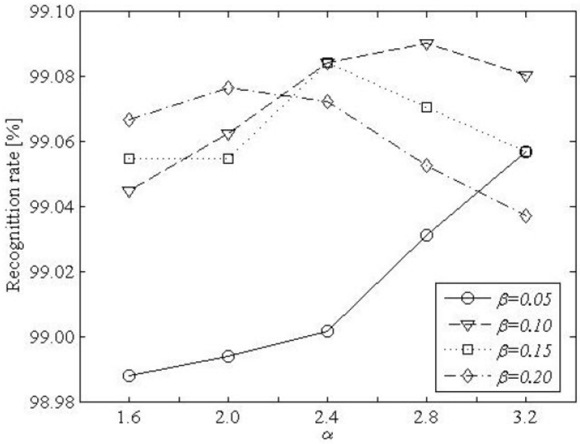

Fig.15: *Results of SS+MVN+FIRTF (Type-I) with MC-HMMs under 10 dB SNR condition*

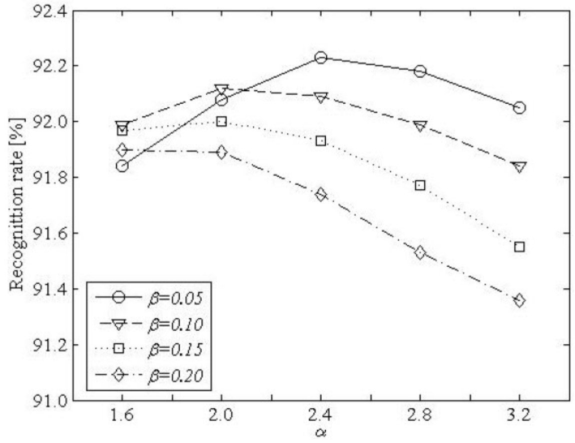

Fig.16: *Results of SS+MVN+FIRTF (Type-I) with MC-HMMs under 0 dB SNR condition*

Table 6: *Detailed results [%] of SS+MVN+FIRTF (Type-I) with MC-HMMs*

Noise \ SNR	0[dB]	10[dB]	20[dB]
Running car (2000 cc)	99.10	99.30	99.37
Running car (1500cc)	98.87	99.33	99.50
Exhibition hall (booth)	90.77	99.03	99.43
Exhibition hall (in aisle)	94.43	99.07	99.40
Station	**82.47**	**98.77**	**99.30**
Telephone box	98.93	99.33	99.47
Factory	**85.03**	**98.50**	**99.43**
Sorting site	**85.83**	**98.70**	**99.53**
Highway	91.30	99.03	99.43
Crowd	95.90	99.03	99.43
Train (bullet train)	98.50	99.27	99.37
Train (conv. line)	84.27	98.43	99.27
Computer room (mid-size)	93.37	99.27	99.43
Computer room (w. s.)	94.83	99.13	99.53
Air-conditioner (large)	93.87	99.27	99.53
Fan-coil	96.57	99.30	99.47
Elevator hall	83.80	98.27	99.27
Average	**92.23**	**99.00**	**99.42**

MVN+FIRTF (Type-I) with MC-HMMs. Compared with the results in Table 3, MC-HMMs brought about the remarkable improvement in not only the considered types of noise (**bold**) but also the others. Particularly, under 20 and 10 dB SNR conditions, we were able to accomplish the accuracy over 98% in all types of noise. On the other hand, under 0 dB SNR, the accuracy was below 90% in some types of noise. Therefore, we should introduce the processing specialized under observed environment.

5. CONCLUSIONS

To achieve the high recognition accuracy under low SNR condition with low calculation costs, we have proposed a novel speech recognition system based on combination of multiple noise-robust techniques. Concretely, the system comprises SS, MVN, FIRTF, and MC-HMMs. Consequently, without degrading the accuracy under the higher SNR conditions, the combined method allowed the accuracy to be improved from 46.61% to 92.23% under 0 dB SNR.

As challenges for the future, we would like to further improve the performance under noisy environment by introducing the processing specialized under observed environment. Moreover, we would like to develop a total speech recognition system that includes both voice activity detection and misrecognition rejection with the aim of making speech recognition systems more popular.

6. ACKNOWLEDGEMENT

I would like to express my deepest gratitude to Prof. Miyanaga who provided helpful supports and comments as guest editor.

References

[1] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech Signal Process.*, vol.ASSP-27, no.2, pp.113–120, 1979.

[2] N. Wiener, *Extraction, Interpolation, and Smoothing of Stationary Time Series*, Wiley, New York, 1949.

[3] ETSI ES 202 050 v.1.1.5, “Speech processing, trans-mission and quality aspects (STQ), advanced distrib-uted speech recognition; front-end feature extraction algorithm; compression algorithms,” Jan. 2007.

[4] B. S. Atal, “Effectiveness of linear prediction charac-teristics of the speech wave for automatic speaker identification and verification,” *J. Acoust. Soc. Am.*, vol.55, no.6, pp.1304–1312, June 1974.

[5] O. Viikki and K. Laurila, “Cepstral domain segmental feature vector normalization for noise robust speech recognition,” *Speech Commun.*, vol.25, pp.133–147, 1998.

[6] C. P. Chen and J. A. Bilmes, “MVA processing of speech features,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 257–270, 2007.

[7] S. Yoshizawa, N. Hayasaka, N. Wada, and Y. Miyanaga, “Cepstral amplitude range normalization for noise robust speech recognition,” *IEICE Trans. Inf. Syst.*, vol.E87-D, no.8, pp.2130–2137, 2004.

[8] M. J. F. Gales and S. J. Young, “Robust continuous speech recognition using parallel model combination,” *IEEE Trans. Speech Audio Process.*, vol.4, no.5, pp.352–359, 1996.

[9] H.G. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions,” *ISCA ITRW ASR2000*, pp.181–188, Sept. 2000.

[10] H. Hermansky and N. Morgan, “RASTA processing of speech,” *IEEE Trans. Speech Audio Process.*, vol.2, no.4, pp.578–589, 1994.

Noboru Hayasaka received the B.S., M.S., and Ph.D. degrees from Hokkaido University, Japan in 2002, 2004, and 2007, respectively. He is currently an Assistant Professor at the Graduate School of Engineering Science, Osaka University. His current research interests are speech processing and speech recognition.