6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ROBUST HMM TO VARIATION OF NOISY ENVIRONMENTS BASED ON VARIANCE EXTENSION OF NOISE MODELS

Hiroshi Matsumoto and Hiroaki Ubukata

Faculty of Engineering, Shinshu University
500 Wakasato, Nagano-shi, Nagano 380, Japan
matsu@sp.shinshu-u.ac.jp

ABSTRACT

This paper addresses the problem of making the PMC-based HMM robust to variation of SNR. The method proposed consists of simply expanding the variace of cepstral coefficients for the noise model in the parallel model combination (PMC). The effect of this technique is examined through speaker independent isolated digit recognition tests using NOISEX-92 noise data. The results show that the variance expansion of several lower order cepstra extremely improves robustness to a wide range of SNR mismatch over the standard PMC. The appropriate expansion factor is such that the expanded variance of the zeroth cepstrum is around 20dB with respect to its geometric mean level.

10.21437/Eurospeech.1999-606

1. INTRODUCTION

In noisy speech recognition, the parallel model combination (PMC) [1] or HMM composition (NOVO) [2] have been proved to be effective for stationary noise environments. In case of non-stationary noise condition, however, these methods are not practical due to computational cost needed to update HMMs for change of noise conditions. In use of hands-free microphone, signal-to-noise ratio ISNR) and speech spectra varies temporally caused by variation of relative position between mouth and microphone even in stationary noise environment. While these variations reflect on variance of noise model through long-term training, the resultant HMMs can not cope with these infrequent environmental changes. Furthermore, in real applications, it is difficult to estimate exact signal-to-noise ratio of input speech, and thus, the resultant HMMs may not match with the actual noise condition. Although PMC itself can deal with non-stationary noise conditions by using multi-state noise model, it requires much computational cost due to increased number of states.

In order to make the composite HMMs robust to mismatch noise conditions, this paper proposes a simple and effective method which expands the variace of cepstral coefficients for the noise model in PMC. This method intends to cope with changes in SNR by simulating a model of nonstationary noise. The effect of the variance expansion through speaker independent isolated digit regcognition tests using several noises taken from NOISEX-92 data [4].

2. VARIANCE EXPANSION OF NOISE MODEL

2.1. Expansion of Noise Variance

In order to cope with mismatch conditions of noise, we assume that noise level and/or spectrum fluctuate around their means much greater than actually observed ones. Since we use Mel-Frequency Cepstrum Coefficients (MFCC) [3] as feature parameters of HMM, this assumed noise is modeled by expanding the variances with the mean vector unchanged.

Thus, for a given actual noise model of a single Gaussian with a mean vector $\boldsymbol{\mu}^c$ and diagonal covariances $D_{iag}(\sigma_i^c)$, the mean vector $\widehat{\boldsymbol{\mu}}^c$ and the i-*th* variance $\widehat{\sigma}_i^c$ of the simulated noise model are given by

$$\widehat{\boldsymbol{\mu}}^c = \boldsymbol{\mu}^c, \quad (1)$$

$$\widehat{\sigma_i}^c = a_i \sigma_i^c \quad (i = 0, 1, \cdots, p), \quad (2)$$

where the superscript "c" refers to parameters on cepstral domain, and a_i is the expansion factor of the i-*th* variance. Figure 1 illustrates the change in level variation by expanding only the zeroth variance. Since these variances are com-

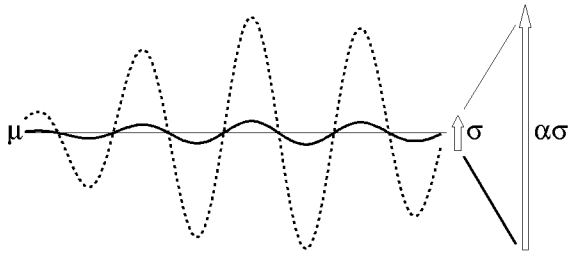

Figure 1. Extension of noise level variation

puted from short-term spectral analysis in speech recognition, the variation of the zeroth cepstrum is caused by not only fluctuation of noise level but also spectral variation for each frame. In this study the appropriate expansion factor $\{a_i\}$ will be experimentally determined.

2.2. Effects of variance expansion

PMC assumes that the spectral components on linear spectral domain is log-normally distributed [1]. Thus, the i-th element of the mean vector $\boldsymbol{\mu}^l$ and the ij-th element of the covariance σ^l_{ij} in log-spectral domain are mapped onto linear spectral domain for both speech and noise by

$$\mu_i = \exp\left(\mu^l_i + \frac{\sigma^l_{ii}}{2}\right), \quad (3)$$

$$\sigma_{ij} = \mu_i \mu_j \left(\exp(\sigma^l_{ij}) - 1\right), \quad (0 \le i, j \le p), \quad (4)$$

where the symbols without superscript represent the variables in linear spectral domain. Since the variance expansion of the zeroth cepstrum of noise makes larger its covariance σ^l_{ij} in log-spectral domain, it increases the power of noise spectrum μ_i and the covariance σ_{ij} in linear spectral domain. After speech and noise addition, the inverse transformation from linear to log-spectral domain is also carried out using the same relationship in equations (4) and (5). Therefore, its effect on the mean vector μ^c_i and σ^c_{ij} is complicated. However, since noisy speech spectrum is affected by additive noise in low SNR frequency bands, it is anticipated that, in low SNR states of HMM, a_0 will change the lower order elments of $\boldsymbol{\mu}^c$ and a_i will make variances larger.

3. EVALUATION

3.1. Data and Analysis

The proposed method is evaluated through gender independent isolated digit recognition tests. Speech data consists of two utterances of isolated 10 digits from each of 40 male and 40 female speakers, which are taken from the JEIDA data [5]. The gender-independent HMM of each digit is trained using 120 tokens from 30 male and 30 female speakers. The test set consists of the remaining 10 male and 10 female speakers for a total of 40 tokens per word. Noise data used are car, speech bubble and factory noises taken from NOISEX-92 data [4] as well as computer generated white noise. The noise model with a single emitting state was trained on all the available noise data, giving about fifty second length. The analysis condition of speech signal is shown in Table 1.

Table 1. Analysis condition and HMM configuration

Sampling Frequency	16kHz
Window	20ms Hamming
Frame Period	5ms
Preemphasis	None
Mel-Filter Bank	17 channels
Spectral Parameter	0 - 16 MFCC
Speech HMM	7 loops with 4 mixtures Diagonal Covariance
Noise HMM	1 loop with 1 mixture Diagonal Covariance

3.2. Effect of the Zeorth Variance

First, the effect of the expansion factor a_0 is examined. In this experiment, the robustness to SNR mismatch is evaluated using noisy speech data of different SNRs from the matched SNR. This condition assumes that the level of nonstationary noise is mostly steady during a word period. Figures 2 (A) to (D) show the recognition accuracies at several test speech SNRs as a function of expansion factor a_0 for white, car, factory, and speech noises, respectively. In these figures, SNRs of car, white, speech bubble and factory noises are set to 0dB, 6dB, 0dB, and 6dB, respectively. As shown in these figures, while the HMMs derived from the standard PMC ($a_0 = 1$), extremly degrade the recognition accuracies at mismatched SNR, the HMMs obtained by PMC with expanded variance of c_0 dramatically increase the accuracies for mismatched SNR conditions. In the case of white noise of 18 dB SNR, the recognition accuracy increases from about 45% for $a_0 = 1$ to 90% for $a_0 = 8$.

The optimal expansion factor a_0 depends mainly on the kind of noise and also on the degree of SNR mismatch between HMM and input speech. The larger SNR mismatch tends to require a larger expansion factor. Table 2 shows the appropriate values of a_0 and the actual variance in dB with respect to the geometric mean level for each noise. From this table, the optimal value of a_0 for each noise seems to be inversely proportional to the magnitude of the actual noise variance. Thus, the appropriate expansion factor is such that it results in the zeroth variance of 15dB to 25dB depending on the extent of SNR mismatch.

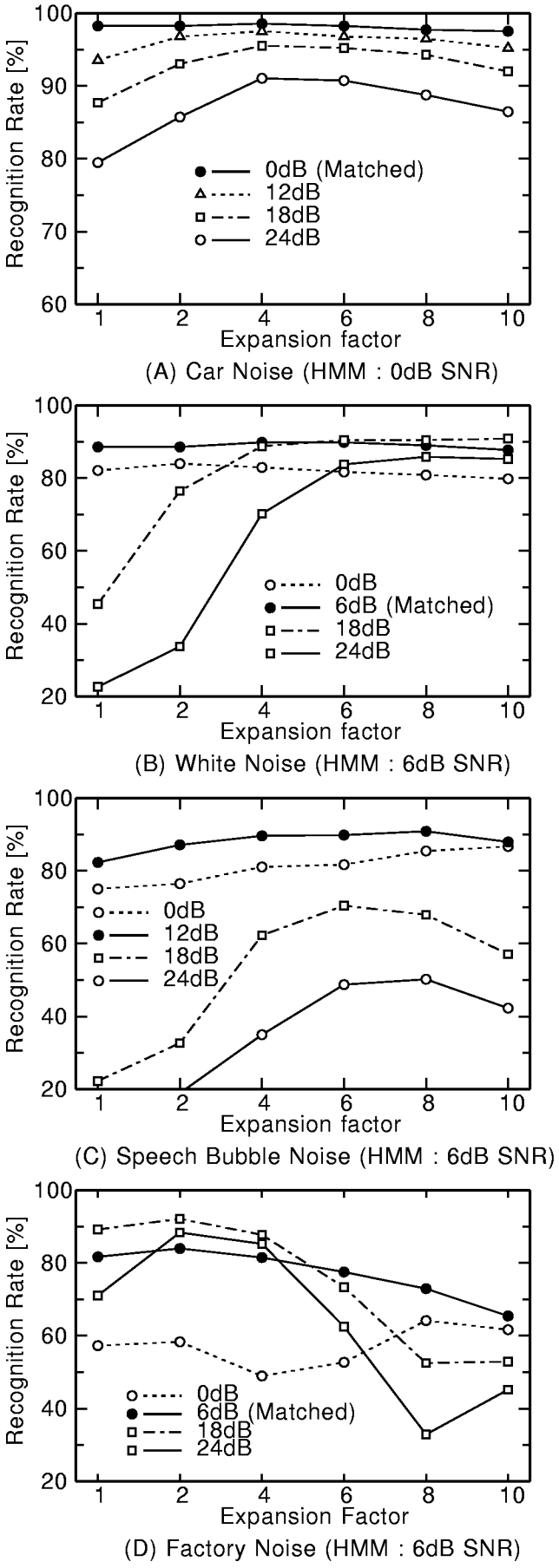

Figure 2. Effect of variance expansion on recognition accuracy at several test speech SNRs

Table 2. Variances and their optimum expansion factors of the zeroth cepstrum for the four types of noises

Noise	White	Speech	Car	Factory
σ_0 [dB]	2.4	2.5	6.2	9.8
a_0	6∼10	6∼8	4∼6	2∼3

3.3. Effect of Higher Order Variances

The additional expansion of the higher oder cepstral variances is examined. First, all the variances σ_0 to σ_{16} are equally expanded by the factors of $\alpha = 1 \sim 10$. The experimental results shows that the optimal value of α for each noise is mostly the same as that in the zeroth variance expansion. Figures 3 (A) to (D) compare the recognition performances as a function of test speech SNR obtained by PMC ($\alpha = 1$) and two types of variance expansion, $a_0 = \alpha$ and $a_0 \sim a_{16} = \alpha$, for four types of noises. In these figures the SNRs of HMMs for car, white, speech bubble and factory noises are set to -6 dB, 6 dB, 12 dB, and 18 dB, respectively, and the expansion coefficients for these noise are set to 6, 6, 6, 2, respectively. As shown these figures, at higher SNR than the matched SNR, the recognition performance for car and factory noises is further improved over that obtained by expansion of only σ_0. In particular, the recognition rate for car noise increases from 60% to 90% at 30dB SNR, which is 36dB higher than the matched SNR. On the other hand, at lower SNR than the matched SNR, the recognition performance is degraded except factory noise.

Then, the expansion of σ_0 to σ_5 is examined as well. These results are included in figures 3 (A) and (B). This variance expansion of lower cepstra maintains the advantage at higher SNR at the expense of slight degradation in recognition performance at lower SNR. As a result of looking into the mean vectors and variances in composite HMMs, it was found that cepstral variances are greatly affected by extending variance expansion to higher order cepstra. Figure 4 compares average standard deviation of each Gaussian in the HMM of /zero/ computed by three type of variance expansion for car noise of 0dB SNR. The first and the last states correspond to low SNR speech segments, and have small variances. Thus, the mean variances of low SNR states increase with the number of variances expanded. These large variances might make HMMs insensitive to the mismatch in level and global spectrum. Consequently, it seems to be appropriate to apply the variance expansion to several lower order cepstra.

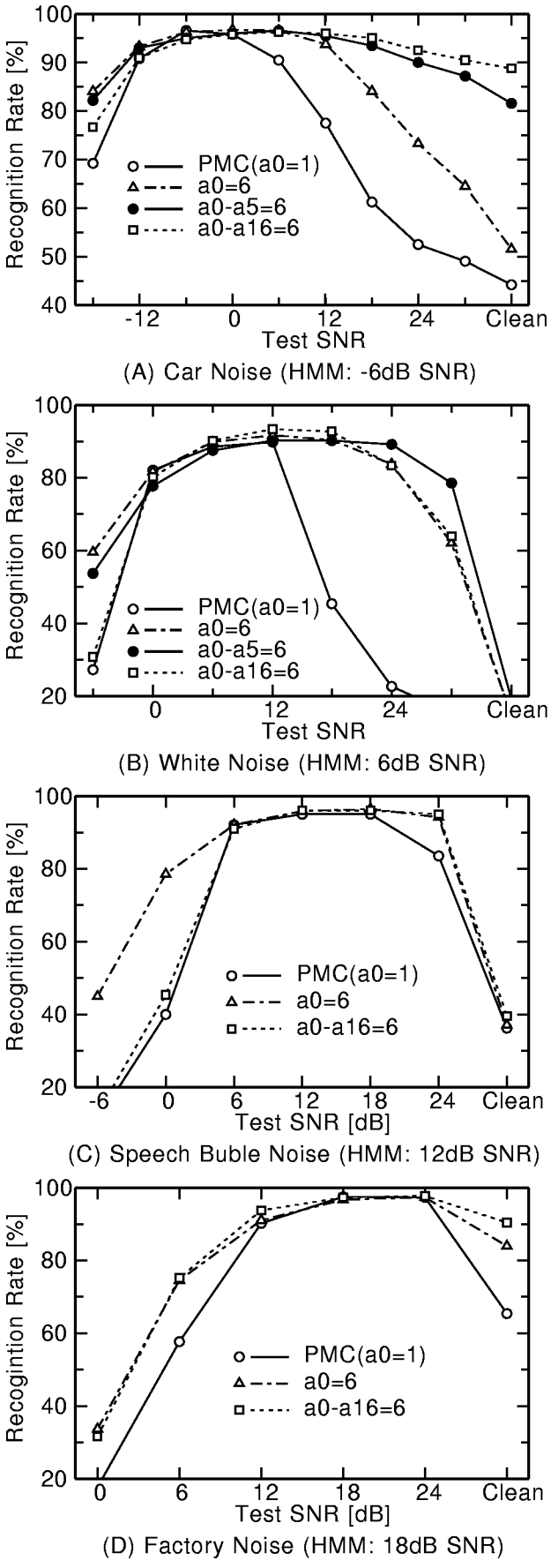

Figure 3. Comparison of the standard PMC and the proposed methods with three types of variance expansion.

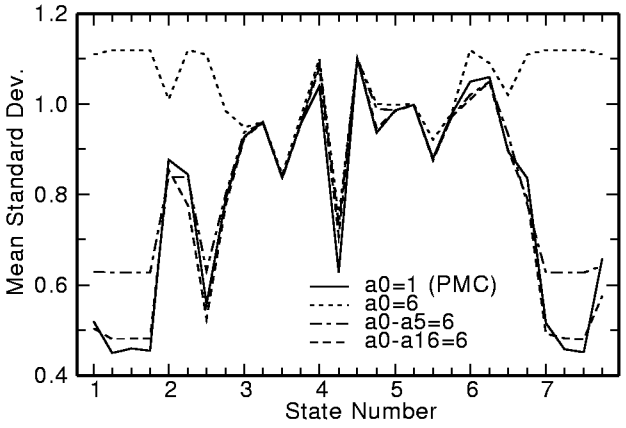

Figure 4. Effect of variance extension on average standard deviation for each Gaussian in the HMM, /zero/, of 0 dB SNR.

4. CONCLUSION

In this paper we have examined the variance expansion technique in PMC to cope with mismatch condition due to noise level and/or spectral variation. The proposed method have been shown to greatly improve the robustness to a wide range of mismatch conditions just by expanding variances of several lower cepstra in a noise model to around 20dB with respect to the geometric mean level. Presently, we are evaluating this technique on other various noise environment and are extending to the problem on spectral variation of noise.

REFERENCES

[1] Gales N.J. and Young S.J.,"An improved approach to the hidden markovl model decomposition of speech and noise," Proceedings of ICASSP, pp.232-236, 1992.

[2] Martine F., Shikano K., Minami Y., and Okabe Y., "Recognition of noisy speech by composition of hidden Markov models," IEICE Technical Report, SP92-96, pp.9-16, 1992.

[3] Davis S.B. and Mermelstein P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions ASSP, Vol.28, pp.357-366, 1980.

[4] Varga A.P., Steeneken H.J.M., Tomlinson M. and Jones D., "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Tech. Rep., DRA Speech Res. Unit, 1992.

[5] Itahashi S.,"Recent Speech Database Projects in Japan," Proceedings of ICSLP, pp.1081-1084, 1990.