

## HMM COMPOSITION OF SEGMENTAL UNIT INPUT HMM FOR NOISY SPEECH RECOGNITION

Kazumasa Yamamoto and Seiichi Nakagawa

Toyohashi University of Technology,  
Department of Information and Computer Sciences  
Tenpaku-cho, Toyohashi, 441, Japan  
{kyama,nakagawa}@slp.tutics.tut.ac.jp

### ABSTRACT

For robust speech recognition in noisy environments, various methods have been studied. In this paper, we apply parallel model combination (PMC) for segmental unit input HMM to recognize corrupted speech in additive noise. Since several successive frames are combined and treated as an input vector in segmental unit input modeling, the increased dimension of vector degrades the precision in estimating covariance matrices. Therefore Karhunen-Loeve expansion or LDA is used to reduce the dimension. Thus the inverse transformation of segmental statistics to cepstral domain is needed and correlations between frames have to be taken into account. We expanded the original PMC to segmental unit input HMM. Experimental results showed PMC for segmental unit input HMM proposed here gives better recognition performance than the original PMC.

### 1. INTRODUCTION

In practical speech recognition systems, the robustness for noisy speech is needed. Many researchers have been studied about this topic and many approaches have been proposed. These are roughly classified into several categories: robust feature parameters[1, 2], feature compensation[3, 4, 5] and model compensation methods[6, 7]. Among the model compensation methods, parallel model combination (PMC)[8] technique has been shown the effectiveness for noisy speech recognition.

On the other hands, we proposed a segmental unit input HMM for capturing the speech dynamics[9] and applied it to noisy speech recognition[10]. By using a segment, robustness in noisy speech recognition is expected because the correlation between frames is considered and dynamic features are also incorporated to feature parameters. Experimental results showed that the segmental unit input HMM trained with noisy speech gave better recognition performance in a noisy environment than frame-based HMM[10]. From this

result, we thought that HMM composition of segmental unit input HMM might be effective if it could be implemented.

This paper describes about implementation of PMC of segmental unit input HMM and shows experimental results compared with the original PMC technique.

### 2. PARALLEL MODEL COMBINATION

The procedure of original PMC is described as follows briefly[8]:

1. Gaussian cepstral means and covariances of speech HMMs and noise HMM are transformed to log spectral domain by DCT:

$$\mu_{(\log\_S)} = C\mu_{(cep\_S)}, \quad (1)$$

$$\Sigma_{(\log\_S)} = C\Sigma_{(cep\_S)}C^T, \quad (2)$$

where  $C$  is DCT matrix,  $\mu_{(cep\_S)}, \Sigma_{(cep\_S)}$  are mean and covariance matrix of speech HMMs in cepstral domain and  $\mu_{(\log\_S)}, \Sigma_{(\log\_S)}$  are mean and covariance matrix in log-spectral domain, respectively.

2. Transformed parameters are transformed again to linear spectral domain:

$$\mu_{(lin\_S),i} = \exp \left\{ \mu_{(\log\_S),i} + \frac{\sigma_{(\log\_S),ii}^2}{2} \right\}, \quad (3)$$

$$\sigma_{(lin\_S),ij}^2 = \mu_{(lin\_S),i}\mu_{(lin\_S),j} \{ \exp(\sigma_{(\log\_S),ij}^2) - 1 \}, \quad (4)$$

where  $\mu_{(lin\_S)}, \Sigma_{(lin\_S)}$  are mean and covariance matrix in linear-spectral domain, respectively.

3. Composition of speech HMMs and noise HMM is carried out:

$$\mu_{(lin\_SN)} = \mu_{(lin\_S)} + k\mu_{(lin\_N)}, \quad (5)$$

$$\Sigma_{(lin\_SN)} = \Sigma_{(lin\_S)} + k^2\Sigma_{(lin\_N)}, \quad (6)$$

where  $\mu_{(lin\_N)}, \Sigma_{(lin\_N)}, \mu_{(lin\_SN)}, \Sigma_{(lin\_SN)}$  are mean and covariance matrix of noise or compensated HMMs in linear-spectral domain, respectively, and  $k$  is a gain factor.

4. Their parameters are inverse-transformed to log-spectral domain:

$$\begin{aligned} \mu_{(\log\_SN),i} &= \log \mu_{(\text{lin}\_SN),i} \\ &- \frac{1}{2} \log \left\{ \frac{\sigma_{(\text{lin}\_SN),ii}^2}{\mu_{(\text{lin}\_SN),i} \mu_{(\text{lin}\_SN),i}} + 1 \right\}, \end{aligned} \quad (7)$$

$$\sigma_{(\log\_SN),ij}^2 = \log \left\{ \frac{\sigma_{(\text{lin}\_SN),ij}^2}{\mu_{(\text{lin}\_SN),i} \mu_{(\text{lin}\_SN),j}} + 1 \right\}, \quad (8)$$

5. And inverse-transformed to cepstral domain.

$$\mu_{(\text{cep}\_SN)} = C^{-1} \mu_{(\log\_SN)}, \quad (9)$$

$$\Sigma_{(\text{cep}\_SN)} = C^{-1} \Sigma_{(\log\_SN)} (C^{-1})^T, \quad (10)$$

6. Recognition is performed.

### 3. PMC OF SEGMENTAL UNIT INPUT HMM

#### 3.1. Segmental Unit Input HMM

Given an input time sequence  $y = y_1 y_2 \cdots y_T$  ( $T$  is input length) and a state sequence  $x = x_1 x_2 \cdots x_T$ , for the probability of  $y$  from output probability computational formulation of HMM, following expressions are derived[9]:

$$\begin{aligned} &P(y_1 \cdots y_T) \\ &= \sum_x P(y_1 y_2 \cdots y_T, x_1 x_2 \cdots x_T) \\ &= \sum_x P(y_1 y_2 \cdots y_T | x_1 x_2 \cdots x_T) P(x_1 x_2 \cdots x_T) \\ &= \sum_x \prod_i P(y_i | y_1 y_2 \cdots y_{i-2} y_{i-1}, x_1 x_2 \cdots x_{i-1} x_i) \\ &\quad \times P(x_i | x_1 x_2 \cdots x_{i-1}) \\ &\approx \sum_x \prod_i P(y_i | y_{i-3} y_{i-2} y_{i-1}, x_{i-1} x_i) P(x_i | x_{i-1}) \end{aligned} \quad (11)$$

$$= \sum_x \prod_i \frac{P(y_{i-3} y_{i-2} y_{i-1} y_i | x_{i-1} x_i)}{P(y_{i-3} y_{i-2} y_{i-1} | x_{i-1} x_i)} P(x_i | x_{i-1}) \quad (12)$$

$$\approx \sum_x \prod_i \frac{P(y_{i-1} y_i | x_{i-1} x_i)}{P(y_{i-1} | x_{i-1} x_i)} P(x_i | x_{i-1}) \quad (13)$$

$$= \sum_x \prod_i P(y_i | y_{i-1}, x_{i-1} x_i) P(x_i | x_{i-1}) \quad (14)$$

$$\approx \sum_x \prod_i P(y_i | x_{i-1} x_i) P(x_i | x_{i-1}), \quad (15)$$

where Eqs.(11)(12) define conditional density HMM of 4-frame width, and Eqs.(13)(14) define conditional density HMM of 2-frame width. Eq.(15) is traditional standard HMM.

On the other hand, Eqs.(12) can be approximately re-written as:

$$\begin{aligned} &\text{Eq.(12)} \\ &\approx \sum_x \prod_i P(y_{i-3} y_{i-2} y_{i-1} y_i | x_{i-1} x_i) P(x_i | x_{i-1}). \end{aligned}$$

This equation defines the segmental unit input HMM of 4-frame width[9]. By using segment of several combined frames, the segmental unit input HMMs can treat both static and dynamic features.

When using the segmental unit input HMM where several successive frames are given as one input vector, since the dimension of vector increases, it results in decreasing the precision in the estimation of covariance matrices. In order to solve this problem, Karhunen-Loeve expansion (K-L expansion) or linear discriminant analysis (LDA) is used to reduce the dimension. We used K-L expansion.

#### 3.2. PMC of Segmental Unit Input HMM

In HMM composition of segmental unit input HMM, since K-L expansion or LDA is used, the inverse K-L transformation of segmental statistics to cepstral domain is needed. Fig.1 illustrates an example of 4-frame segment. In this case,  $\mu_{(\text{cep}\_S)}$ , that is inverse transformation of  $\mu_{(\text{kl}\_S)}$ , and  $\Sigma_{(\text{cep}\_S)}$ , that is also inverse transformation of  $\Sigma_{(\text{kl}\_S)}$ , can be treated in the same way of the original PMC (because diagonal blocks are explicitly intra-frame covariance matrices), but there is a problem that how to treat the other blocks, that is inter-frame covariances, of  $\Sigma_{(\text{cep}\_S)}$ . Here, inverse K-L transformed covariance matrices are divided into frame level blocks and DCT is carried out to the each block like as diagonal blocks. If a segment consists of  $N$ -frame vectors and the frame vector consists of  $d$  coefficients, then the mean has  $N$  cepstral vectors (each vector has  $d$  coefficients) and the covariance matrix consists of  $N \times N$  cepstral covariance blocks (each block is  $d \times d$  matrix). So we modified the procedure as follows:

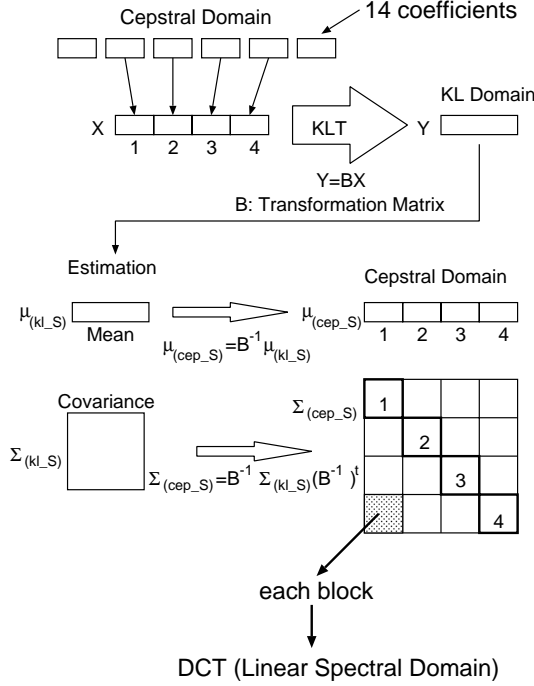
0. Gaussian means and covariances of speech HMMs and noise HMM are transformed by inverse K-L transformation:

$$\mu_{(\text{cep}\_S)} = B^{-1} \mu_{(\text{kl}\_S)}, \quad (16)$$

$$\Sigma_{(\text{cep}\_S)} = B^{-1} \Sigma_{(\text{kl}\_S)} (B^{-1})^T, \quad (17)$$

where  $\mu_{(\text{kl}\_S)}$ ,  $\Sigma_{(\text{kl}\_S)}$  are mean and covariance matrix of speech HMMs, respectively, and  $B$  is K-L transformation matrix.

1. Gaussian cepstral means and covariances of speech HMMs and noise HMM in each block are transformed



**Figure 1. Inverse K-L transformation of mean and covariance matrix**

by DCT:

$$\mu_{(\log\_S)n} = C\mu_{(cep\_S)n}, \quad (18)$$

$$\Sigma_{(\log\_S)nm} = C\Sigma_{(cep\_S)nm}C^T, \quad (19)$$

where  $\mu_{(cep\_S)} = \{\mu_{(cep\_S)n}^T\}^T$ ,  
 $\Sigma_{(cep\_S)} = [\Sigma_{(cep\_S)nm}]$  ( $n, m = 1, \dots, N$ ).

2. - 4. are same as the original PMC.
5. The parameters are transformed inverse to cepstral domain:

$$\mu_{(cep\_SN)n} = C^{-1}\mu_{(\log\_SN)n}, \quad (20)$$

$$\Sigma_{(cep\_SN)nm} = C^{-1}\Sigma_{(\log\_SN)nm}(C^{-1})^T \quad (21)$$

6. K-L transform is performed.

$$\mu_{(kl\_SN)} = B\mu_{(cep\_SN)}, \quad (22)$$

$$\Sigma_{(kl\_SN)} = B\Sigma_{(cep\_SN)}B^T. \quad (23)$$

7. Recognition is performed.

## 4. RECOGNITION EXPERIMENTS

### 4.1. Speech Database and HMM

In experiments, continuous density output distribution HMMs with 5 states (4 output with 16 mixture diagonal

**Table 1. Speech analysis condition**

sampling frequency	12kHz
preemphasis	$1 - 0.98z^{-1}$
Hamming window width	21.33ms(256 points)
frame period	8ms(96 points)

Gaussian distribution per state) having duration control are used. They were trained by syllable-segmented data from ATR speech database (6 male speakers). After that, they were retrained by MAP estimation[11] by using database of Acoustic Society of Japan (30 male speakers). These models are called ‘‘Clean Model’’. The number of syllables used as recognition unit was 113.

The acoustic analysis conditions are shown in Table 1. 14 dimensional FFT cepstrum coefficients are used as speech feature parameters for a frame. Segments consist of successive 4 frames and their dimension are reduced to 20.

As testing data, a noisy speech database was produced by adding noise to clean speech data at specified SNR(0,10,20dB). 104 sentences of ‘‘Sightseeing Guide Task around Mt. Fuji’’ uttered by each of other 6 male speakers and recorded in recording room were used as clean speech data. Noise of automobile cabin and exhibition booth noise are used as environmental noise[12] and white noise is also used.

Noise HMM for PMC is composed of one state and one mixture per state. Furthermore, noisy environment HMMs are also trained with noise corrupted speech data(adding noise to clean speech data described above at 10dB SNR only). These models are called ‘‘Noise Model’’.

Syllable recognition rate in continuous speech without language model is used as a performance measure.

### 4.2. Experimental Results

Continuous syllable recognition experiments were carried out on speaker-independent mode to compare original PMC and proposed method.

Tables 2,3,4 show the experimental results of white noise, noise of automobile cabin and noise of exhibition hall, respectively, where ‘‘Frame’’ denotes use of frame-based (traditional) HMM and ‘‘Seg’’ denotes use of segmental unit input HMM.

As can be seen from tables, the recognition performance is seriously degraded with clean speech HMMs when SNR becomes worse. In the case of segmental unit input HMM, the performance is better than frame-based model, but is absolutely worse. Using the models trained with corrupted speech, the recognition performance significantly improved.

**Table 2. Continuous syllable recognition rate in white noise: no language model, average of 6 speakers, %Correct**

METHOD		0dB	10dB	20dB	clean
Clean Model	Frame	3.2	7.1	17.8	44.8
	Seg	8.6	11.0	20.9	55.7
Noise Model	Frame	12.3	28.6	23.6	32.3
	Seg	16.3	36.1	28.5	47.2
PMC	Frame	18.4	28.9	40.6	—
	Seg	22.4	32.6	45.8	—

**Table 3. Continuous syllable recognition rate in noise of automobile cabin: no language model, average of 6 speakers, %Correct**

METHOD		0dB	10dB	20dB	clean
Clean Model	Frame	25.8	37.6	43.4	44.8
	Seg	28.9	43.6	53.3	55.7
Noise Model	Frame	35.1	41.8	43.6	43.1
	Seg	46.3	56.7	57.3	55.4
PMC	Frame	39.9	43.7	44.6	—
	Seg	40.6	54.6	56.3	—

However these models are impractical because of training cost. Also in this case, segmental unit input HMMs are better recognition performance than frame-based HMMs. And experimental results showed that the proposed method is outperformed the original PMC in almost of various conditions. In a case of the noise of exhibition hall, however, the recognition performance was degraded with the proposed method. We think this degradation was caused because many human speech included in the noise of exhibition hall was enhanced by using segment.

## 5. CONCLUSION

In order to improve the accuracy of speech recognition in noisy environments, we proposed and implemented the PMC of segmental unit input HMM to noisy speech recognition. Experimental results showed in almost all cases better recognition rates for the PMC of segmental unit input HMM than for the original PMC method in various noise at various SNR.

## REFERENCES

- [1] J.Koehler, N.Morgan, H.Hermansky, H.Gunter-Hirsh and G.Tong, "Integrating RASTA-PLP into speech recognition," Proc.ICASSP-94, pp.I-421-I-424, Apr. 1994.
- [2] K.Aikawa, H.Singer, H.Kawahara and Y.Tohkura, "Cepstral representation of speech motivated by time-frequency masking: An application to speech recogni-

**Table 4. Continuous syllable recognition rate in noise of exhibition hall: no language model, average of 6 speakers, %Correct**

METHOD		0dB	10dB	20dB	clean
Clean Model	Frame	8.8	21.0	36.8	44.8
	Seg	14.1	28.7	46.6	55.7
Noise Model	Frame	23.3	36.9	34.1	29.0
	Seg	26.8	47.8	45.8	37.6
PMC	Frame	28.7	39.2	42.8	—
	Seg	9.7	38.4	53.0	—

tion," J. Acoust. Soc. Am., Vol.100, no.1, pp.603-614, July 1996.

- [3] S.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech & Signal Process., Vol.27, no.2, pp.113-120, Feb. 1979.
- [4] M.Rahim and B.-H.Juang, "Signal bias removal for robust telephone based speech recognition in adverse environments," Proc.ICASSP-94, pp.I-445-I-448, Apr. 1994.
- [5] A.Acero and R.M.Stern, "Environmental robustness in automatic speech recognition," Proc.ICASSP-90, pp.849-852, Apr. 1990.
- [6] A.P.Varga and P.K.Moore, "Hidden Markov model decomposition of speech and noise," Proc.ICASSP-90, pp.845-848, Apr. 1990.
- [7] K.Takagi, H.Hattori and T.Watanabe, "Rapid environment adaptation for speech recognition," J. Acoust. Soc. Jpn.(E), Vol.16, no.5, pp.273-281, Sep. 1995.
- [8] M.J.F.Gales and S.J.Young: Cepstral parameter compensation for HMM recognition in noise, Speech Communication, vol.12, no.3, pp.231-239(1993)
- [9] S.Nakagawa and K.Yamamoto, "Evaluation of segmental unit input HMM," ICASSP-96, pp.439-442, 1996.
- [10] K.Yamamoto and S.Nakagawa, "Evaluation of segmental unit input HMM in noisy environments," Proc. ICSP'97, pp.643-648, Aug. 1997.
- [11] Y.Tsurumi and S.Nakagawa, "An unsupervised speaker adaptation method for continuous parameter HMM by maximum a posteriori probability estimation," ICSLP-94, pp.431-434, Sept. 1994.
- [12] S.Itahashi, "Recent speech database projects in Japan," ICSLP-90, pp.1081-1084, 1990.