WEIGHTED PARALLEL MODEL COMBINATION FOR NOISY SPEECH RECOGNITION

Tai-Hwei Hwang and Hsiao-Chuan Wang
Department of Electrical Engineering, National Tsing-Hua University,
Hsinchu, Taiwan, ROC, 30043.
hcwang@ee.nthu.edu.tw

ABSTRACT

This paper proposes a modified parameter mapping scheme for parallel model combination (PMC) method. The modification aims to improve the discriminative capabilities of the compensated models. It is achieved by the rearrangement of the distributions of state models in order to emphasize the contribution of the mean in the following process. Both distributions of speech model and noise model are shaped in cepstral domain through a covariance contracting procedure. After the compensation steps, an expanding procedure of the adapted covariance is necessary to release the emphasis. Using this process, the discriminative capability is increased so that the recognition accuracy is improved. In this paper, the recognition of Chinese names demonstrates the improvement to the original PMC method, especially when SNR is low.

1. INTRODUCTION

A different environmental noise, which does not appear in the training data for reference models, is known as a critical factor which degrades the recognition accuracy seriously [1]. Many studies have been conducted to diminish the effect caused by the additive noise [2][3]. Among the studies, the parallel model combination (PMC) technique has been successfully developed to adapt the models trained by clean speech with the reference of environmental noise [4]. The statistical models of speech are expressed in cepstrum domain, while the effect of environmental noise is additive in linear spectral domain. To incorporate the noise statistics into speech models, a mapping for model parameters between cepstral domain and linear spectral domain is necessary. In the literature [5], a closed-form formulation was derived for the model transformation based on a log normal assumption.

In this paper, the transformation scheme is modified to obtain a better discriminative capability for pattern classification. The modification is done by shaping signal models; including speech models and noise models. The statistics models are rearranged in cepstral domain so that the vicinity of cepstral means are emphasized in the mapping. The shaping process is accomplished by contracting the covariance of a state model in cepstral domain using a scalar factor. Both covariance matrices of state models of speech and noise are divided by a same factor for the contraction. After the models are adapted in linear spectral domain, they are transformed back into cepstral domain and the covariance terms are expanded by the same factor to release the emphasis. The effect resulted from the procedure can be observed in the improvement of the discriminative capability of the adapted models. The improvement is quite significant to the test condition at low SNR, and is useful to the noisy speech recognition. This modified PMC method is termed weighted PMC (W-PMC).

This paper is organized as follows. In section 2, the original PMC method is briefly reviewed and a modification in the model mapping is introduced. In section 3, A discriminative test on selected vowels shows the effect by the modification. In section 4, a recognition task of Chinese names under the corruption of additive noise is conducted to evaluate the W-PMC method. Finally, a conclusion is given in section 5.

2. WEIGHTED PARALLEL MODEL COMBINATION

2.1. Review of Parallel Model Combination

Parallel model combination (PMC) method adapts speech models to the test condition with additive noise as follows. Assume that the observation function of speech model is Gaussian, which can be characterized by a mean and a covariance $\{\mu^c, \Sigma^c\}$. In this paper, the super script c means a parameter in cepstral domain and l means that in log spectral domain. The speech to be recognized is often represented in cepstral domain, while the noisy speech is the addition of background noise and the speech in linear spectral domain. To match the test condition, the model parameters of clean speech have to be transformed into linear spectral domain for incorporating the statistics of background noise $\{\widetilde{\mu}, \widetilde{\Sigma}\}$. The transformation can be divided into two sequential stages. At the first stage, the model parameters in log spectral domain are derived from the cepstral domain by the inverse discrete cosine

This research has been supported by the National Science Council, Taiwan, ROC under contract NSC-86-2213-E-007-020.

transform (IDCT, denoted by $\mathbf{C}^{-1}$),

$$\mu^l = \mathbf{C}^{-1}\mu^c \text{ and } \Sigma^l = \mathbf{C}^{-1}\Sigma^c(\mathbf{C}^{-1})^T,$$

where a super script T means a transpose of a matrix. At the second stage, based on a log normal assumption [5], the i-th component of the mean vector μ and the covariance in the linear spectral domain can be computed by

$$\mu_i = \exp(\mu_i^l + \sigma_{ii}^l / 2) \text{ and } \sigma_{ij} = \mu_i\mu_j[\exp(\sigma_{ij}^l) - 1].$$

Assuming that the speech and noise are independent and additive in linear power spectral domain, the adapted mean and covariance can be obtained by

$$\bar{\mu}_i = g\mu_i + \tilde{\mu}_i \text{ and } \bar{\sigma}_{ij} = g^2\sigma_{ij} + \tilde{\sigma}_{ij},$$

where factor g is a gain term providing the match of signal power to test condition. Assuming that the combined distribution in linear spectral domain is log normal, the above mapping process can be straightforwardly inverted. Therefore, the linear domain parameters are transformed back to the log spectral domain by

$$\bar{\mu}_i^l = \log(\bar{\mu}_i) - 0.5\log\left(\frac{\bar{\sigma}_{ii}}{\bar{\mu}_i^2} + 1\right)$$

$$\text{and } \bar{\sigma}_{ij}^l = \log\left(\frac{\bar{\sigma}_{ij}}{\bar{\mu}_i\bar{\mu}_j} + 1\right),$$

and secondly, back to the cepstral domain by

$$\bar{\mu}^c = \mathbf{C}\bar{\mu}^l \text{ and } \bar{\Sigma}^c = \mathbf{C}\bar{\Sigma}^l\mathbf{C}^T.$$

2.2 Modified Parameters Mapping Scheme

In the above model adaptation scheme, the model parameters in linear spectral domain are the consequence of the mapping from log spectral domain. As the Gaussian distribution is concerned, the neighborhood of the mean is the portion with the highest density and that keeps the most discriminative information from the other distributions. If a distribution in log spectral domain is shaped so as to emphasize the vicinity of mean, the corresponding distribution in linear spectral domain will account for the change. Using this concept, if the observation function of model states are properly shaped in log spectral domain, the combination of parameters in linear spectral domain will be the consequence that keeps the dominant information, and the adapted models will be more efficient to pattern classification. Intuitively, the emphasis of the mean can be achieved through a contraction of the covariance. This procedure can be performed before the mapping of parameters by dividing the covariance of all signal models in log spectral domain with a pre-defined factor.

Once the models are adapted, the covariance in log spectral domain are expanded by multiplying the same factor for the succeeding pattern classification. The expanding process is necessary because it keeps the modified mapping process as an unitary transformation. On the other hand, the previous emphasis of the mean destroys the strategy of the training model in the sense of maximum likelihood, and results in a degradation of the performance of pattern matching. The expanding process could be a remedy to the problem. The contract-and-expand of covariance in log spectral domain also can be performed in the cepstral domain because of the linearity of IDCT. The adaptation method with the contract-and-expand procedure is termed weighted parallel model combination (W-PMC) method in the study, since the highlight of the mean can be considered as a weighting operation to performing original Gaussian integration. The effect by the W-PMC will be demonstrated in the following discriminative test and a recognition task.

3. DISCRIMINATIVE TEST

The benefit gained from the covariance contract-and-expand operation is primarily demonstrated by a discriminative test of vowels. Five vowels, /a/, /e/, /i/, /o/, and /u/ were uttered by a male speaker in a quiet environment. The speech wave forms were digitized in 8 kHz and the mel-cepstrum analysis is applied for each 0.032 second to obtain 12 mel-frequency cepstrum coefficients (MFCCs) as a feature vector [7]. Assume that the distribution of feature vector of a vowel is Gaussian and its model parameters can be obtained by

$$\mathbf{m}_v = \frac{1}{N_v}\sum_{t=1}^{Nv}\mathbf{c}_{v,t} \text{ and } \Sigma_v = \frac{1}{N_v}\sum_{t=1}^{N_v}(\mathbf{c}_{v,t} - \mathbf{m}_v)(\mathbf{c}_{v,t} - \mathbf{m}_v)^T,$$

where v indicates one of the five vowels, t is the frame index, and N_v is the total number of frames of an observation of v. A confusing model of v, denoted by $\bar{v}$, is defined for which one is not v but gives the largest likelihood for an observation of v, *i.e.*,

$$\bar{v} = \arg\max_{u \neq v} \log(N(\mathbf{o}_v; \mathbf{m}_u, \Sigma_u)).$$

A discriminative scoring of an observation of v is defined by the likelihood ratio

$$Sc_v \overset{def.}{=} \log(N(\mathbf{o}_v; \mathbf{m}_v, \Sigma_v)) - \log(N(\mathbf{o}_v; \mathbf{m}_{\bar{v}}, \Sigma_{\bar{v}})).$$

At first, the discriminative scoring is applied to the clean utterances with respect to the clean speech models. Resulting scores and correspondent confusing vowels are tabulated in Table 1.1. The clean speech of five vowels are artificially added with Gaussian white noise and babble noise, extracted from NOISEX-92 database, in SNR 20dB, 10dB, and 0dB to generate the noisy speech. Corresponding to the test SNR and the noise type, the

model parameters of clean speech are adapted using W-PMC method parameterized by four contract-and-expand factors, 1, 2, 5, and 10. The scoring procedures are applied to the noisy speech with respect to the adapted models and the results are listed in Table 1.2 and Table 1.3 for both noises. For each case, it is observed that the average score becomes smaller as the noise power increasing. This phenomenon may explain the degradation of noisy speech recognition, that the additive noise blurs the difference among templates even a compensation scheme was applied to them. In the case of additive babble noise, the discriminative scores increase when incorporating a bigger contract-and-expand factor. The trend is still kept for the case of white noise in 10dB and 0dB. For the case of white noise in 20db, the scores get smaller as applying a bigger contract-and-expand factor. However, the score is still high enough for effective pattern classification. From the results, it implies that the modified mapping process of model parameters is more beneficial for the speech recognition in low SNR.

Table 1: Confusing vowels and their discriminative log likelihood scores with model adaptation using W-PMC.

α	Reference Models of Clean Vowels					Average
	/a/	/e/	/i/	/o/	/u/	
NA	o/772.3	u/404.5	u/721.2	u/286.9	o/43.8	435.7

Table 1.1: Under quiet environment (NA: No model adaptation in this case)

α	20dB					Average
	/a/	/e/	/i/	/o/	/u/	
1	o/510.6	u/230.1	u/294.2	u/297.9	e/30.3	272.6
2	o/494.8	u/220.3	u/299.0	u/290.4	e/28.24	266.5
5	o/486.5	u/218.1	u/302.3	u/283.3	e/25.9	263.2
10	o/484.0	u/218.4	u/303.5	u/280.1	e/25.2	262.2

α	10dB					Average
	/a/	/e/	/i/	/o/	/u/	
1	o/104.4	u/98.3	u/119.4	u/138.0	e/21.9	96.4
2	o/118.1	u/106.8	u/123.4	u/131.6	e/17.3	99.4
5	o/133.8	u/112.4	u/126.6	u/129.6	e/16.7	103.8
10	o/140.8	u/114.3	u/127.8	u/129.2	e/17.0	105.8

α	0dB					Average
	/a/	/e/	/i/	/o/	/u/	
1	o/38.6	u/35.4	u/48.3	a/42.6	i/6.4	34.2
2	o/43.9	u/40.9	u/55.0	a/43.7	e/10.7	38.8
5	o/45.7	u/45.6	u/59.2	a/44.5	e/11.4	41.3
10	o/46.2	u/47.4	u/60.7	a/44.7	e/12.1	42.2

Table 1.2: Contaminated by Gaussian *white* noise in 20dB, 10dB, and 0dB. (α , contract-and-expand factor for W-PMC)

α	/a/	/e/	/i/	/o/	/u/	Average
1	o/555.2	u/285.6	u/138.5	u/138.5	u/111.6	226.6
2	o/582.4	u/299.6	u/146.4	u/122.5	e/43.6	238.9
5	o/591.2	u/304.7	u/153.2	u/131.4	e/44.8	245.1
10	o/593.2	u/306.1	u/155.8	u/134.8	e/45.4	247.1

α	10dB					Average
	/a/	/e/	/i/	/o/	/u/	
1	o/141.5	u/72.2	u/71.0	u/45.9	o/27.6	71.7
2	o/159.8	u/84.0	u/73.4	u/50.3	o/26.6	78.8
5	o/172.9	u/93.3	u/74.9	u/53.8	o/27.1	84.4
10	o/177.6	u/96.9	u/75.5	u/55.1	o/28.4	86.7

α	0dB					Average
	/a/	/e/	/i/	/o/	/u/	
1	o/35.7	u/15.8	u/28.7	u/10.8	o/10.7	20.3
2	o/38.3	u/16.7	u/30.8	u/10.4	o/11.0	21.4
5	u/40.2	u/17.4	u/32.3	u/10.0	o/18.6	23.7
10	u/40.8	u/17.7	u/32.9	u/9.9	o/22.3	24.7

Table 1.3: Contaminated by *babble* noise in 20dB, 10dB, and 0dB. (α , contract-and-expand factor for W-PMC)

4. RESULTS ON NOISY SPEECH RECOGNITION

A Chinese name recognition task was conducted to evaluate the performance of using W-PMC method. The database of Chinese names were collected from 18 males and 11 females in a quiet environment. Each speaker pronounced a name list once, in which one list was consisted of 120 Chinese names. There are about three or four Mandarin syllables for each piece of name. The speech data from 12 male and 7 female speakers were collected as the training data (about 2/3 of the database) and the remains were the test data. The speech waveform was digitized in 8 kHz and segmented into frames of 32ms with 50% overlap. Speech features were extracted frame by frame from a mel-frequency analysis using a 20-filter bank. The feature vector compromised 13 mel-frequency cepstrum coefficients (MFCCs), include a zeroth term required in the model adaptation only, and 12 delta MFCCs.

In the experiment, each Mandarin syllable is represented by a concatenation of context-dependent sub-word models. These sub-word models can be classified into two categories, one is a set of initials which include their transition portions and the other is a set of finals. The initial model and the final model are consisted of three states and four states, respectively. The observation distribution of each state is a mixture of four Gaussian probability densities. Full covariance matrices are used in the compensation process, while only the diagonal components are adopted to compute the likelihood scores for simplicity.

The noisy speech is generated by artificially adding three types of noises, Gaussian *white*, *babble* and *factory* noises, extracted from NOISEX-92 database, to the speech waveform in five SNR's. The noise model is trained from the noise data of 2 seconds, which is modeled by one state with a mixture of two Gaussian densities. A baseline system is the one without any noise compensation scheme. Four contract-and-expand factors, 1, 2, 5, and 10 are experimentally applied in W-PMC method to compare their effects. In case of using 1 as the factor, W-PMC is equivalent to the original PMC. In the experiments, only the MFCC portion of speech models are adapted, leaving the delta portion unchanged. The gain term g is set to 1 for all cases without loss of generality. The results in terms of recognition error rates are listed in Table 2. The improvement by using W-PMC method is obvious for the selected additive noises, especially when SNR is lower than 10dB. Furthermore, the increasing of α tends to decrease the error rates when the SNR is low. However, the results are not consistent in some cases where the SNR is high. For example, in the case of 20dB white noise, the error rate increases from 6.3% to 6.8% when contracting factor changes from 1 to 10. The results are consistent with the discriminative test, where the discriminative scores are increased at low SNR and decreased at high SNR when incorporating with a larger contract-and-expand factor.

White	*20dB*	*15dB*	*10dB*	*5dB*	*0dB*
No adaptation	7.1	14.4	31.6	61.8	87.5
PMC	6.3	9.3	17.9	37.2	64.8
W-PMC(2)	6.5	9.5	16.6	29.9	55.9
W-PMC(5)	6.6	9.8	16.4	29.6	53.8
W-PMC(10)	6.8	9.9	16.7	29.9	53.4

Babble	*20dB*	*15dB*	*10dB*	*5dB*	*0dB*
No adaptation	5.4	9.7	22.4	57.5	87.8
PMC	4.0	5.2	9.1	23.1	58.5
W-PMC(2)	3.8	5.2	8.8	20.1	52.1
W-PMC(5)	3.9	5.1	8.9	19.6	49.6
W-PMC(10)	4.1	5.1	9.2	19.2	50.2

Factory	*20dB*	*15dB*	*10dB*	*5dB*	*0dB*
No adaptation	5.2	10.0	25.6	59.0	89.0
PMC	4.5	6.8	11.5	26.8	64.8
W-PMC(2)	4.4	6.6	10.8	22.7	55.8
W-PMC(5)	4.6	6.5	11.2	21.8	53.3
W-PMC(10)	4.4	6.3	11.3	22.2	52.8

Table 2: Chinese names recognition error rate compensated by PMC and W-PMC(α) for the contamination of *white*, *babble*, and *factory* noise, respectively.

5. CONCLUSION

In this study, we introduce a modified scheme for the mapping of the model parameters in the parallel model combination method. Using the modification, a contract-and-expand procedure of the covariance, the discriminative capabilities of the adapted models are improved. The effect will be more significant by assigning a bigger contract-and-expand factor in low SNR. Therefore, the proposed method is useful to improve the recognition accuracy of noisy speech, especially at low SNR.

6. REFERENCES

[1] Juang, B. H. "Speech recognition in adverse environments." *Computer Speech and Language* 5, 1991, pp. 275-294.

[2] Gong, Y. "Speech recognition in noisy environments: A survey," *Speech Communication,* Vol. 16, 1995, pp. 261-291.

[3] Vaseghi, S. V. & Milner, B. P. "Noise compensation methods for hidden Markov model speech recognition in adverse environments," *IEEE Trans. on Speech and Audio Processing,* Vol. 5, 1997, No. 1.

[4] Gales, M. J. F. & Young, S. J. "Robust continuous speech recognition using parallel model combination," *IEEE Trans. on Speech and Audio Processing*, 1996, vol. 4, No. 5.

[5] Gales, M. J. F. & Young, S. J. "Cepstral parameter compensation for HMM recognition in noise," *Speech Communication* 12, 1993, pp. 231-239.

[6] Varga, A. & Steeneken, H. J. M. "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication 12*, 1993, pp. 247-251.

[7] Davis, S. B. & Mermelstein, P. "Comparison of parametric representations for monosyllabic word recognition in continusly spoken sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 28, No. 4, 1980, pp. 357-366.