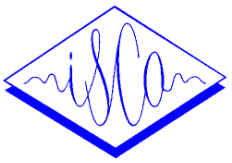

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SPEECH MODEL COMPENSATION WITH DIRECT ADAPTATION OF CEPSTRAL VARIANCE TO NOISY ENVIRONMENT

Tai-Hwei Hwang, Kuo-Hwei Yuo, and Hsiao-Chuan Wang*

*E000/Computer & Communication Labs, Industrial Technology Research Institute,
Chutong, Hsinchu, Taiwan 310

Department of Electrical Engineering, Tsing Hua University,
Hsinchu, Taiwan 300
Email: hthwei@atc.ccl.itri.org.tw; hcwang@ee.nthu.edu.tw

ABSTRACT

A modified parallel model combination (PMC) for noisy speech recognition is proposed such that both speech cepstral mean and variance are adapted without the mapping of variance between cepstral and log-spectral domains. By investigating an adapted scalar random variable of log-energy in the way of PMC, we observe that the adapted variance of log-energy can be roughly predicted by the energy ratio of source signals. Based on the observation, we propose that the cepstral variance of the adapted model can be approximated according to the local signal-to-noise ratio (SNR) of a state. The combined cepstral variance is then assigned to be the variance of clean speech, the variance of noise, or the average variance of clean speech and noise. The performance of using this approximation method is compared with the original PMC. Our experiment shows that the degradation of the performance is small, but the proposed method has greatly reduced the computational cost as comparing with the PMC method.

1. INTRODUCTION

Ambient noises often degrade the performance of automatic speech recognition seriously [1]. The noise adds some spectral components to the speech signal and makes the reference models of clean speech failed to match the noisy speech. One method to improve the performance is to retrain the reference models with a matched noisy speech database [2]. However, the retrain process is time-consuming and is unfeasible to a practical system. Instead of retraining the models, the parallel model combination (PMC) method can adapt the models to the noisy environment [3][4] during the recognition phase. In this approach, it assumes that the alignment of noisy speech models is invariant from the clean speech, and only the feature distributions of model are altered by the additive noise. If the feature distribution is expressed as a mixture of Gaussian densities, the adaptation of feature distribution can be performed by compensating the mean and variance for each Gaussian density. Since the noise is additive in the linear spectral domain, the cepstral mean and variance have to be transformed into the linear spectral domain. Once the combination for mean and variance are done in the linear spectral domain, the mean and variance are transformed back into the cepstral domain for speech recognition. The PMC method that has been developed from a log-normal assumption can provide efficient formulations for the transformation and make the adaptation feasible [3]. However, its computational cost may become very large as the number of speech models increases. Some efforts have been made to speed up the model adaptation, such as the data driven PMC [4] and the method to reduce the number of adaptations [5].

In this paper, we propose a modified formulation to adapt the cepstral variance in the PMC method. The modification is inducted by an adapted scalar random variable of log-energy in the way of PMC. In such simple case, we observe that the adapted variance of log-energy can be grossly predicted by the energy ratio of source signals. Accordingly, we generalize the case to multivariate speech models in order to simplify the model adaptation procedure in PMC. In the proposed method, there is no need to perform the mapping of covariance matrix from the log-spectral domain to the cepstral domain. Thus, it can effectively reduce the computation in the PMC method. Though the adapted variance is roughly approximated, our experiments show that there is a small degradation on the recognition rate as comparing with the standard PMC method.

This paper is organized as follows. In section 2, we make a brief review of the standard PMC method that is based on the log-normal assumption. In section 3, we conducted an example of PMC on two scalar random variables to figure out the effect on the adapted variance. In section 4, a novel adaptation scheme for the cepstral variance is proposed. In section 5, the experimental setting and its results are discussed. Finally, a conclusion of this work is given.

2. BRIEF REVIEW OF LOG-NORMAL PMC

Assume that the cepstral distribution of speech signal can be expressed as a mixture of Gaussian densities. For simplicity, one Gaussian density is assumed for the cepstral distribution. The cepstral mean vector and covariance matrix of speech signal are expressed by $\{m^c, \Sigma^c\}$, where the super script c means a parameter in cepstral domain. To obtain the spectral mean and variance, $\{m^c, \Sigma^c\}$ are first transformed into log-spectral domain by taking inverse discrete cosine transformation (IDCT)

$$m^l = \mathbf{C}^{-1} m^c \tag{1}$$

$$\text{and } \Sigma^l = \mathbf{C}^{-1} \Sigma^c (\mathbf{C}^{-1})^T, \tag{2}$$

where the super script *l* indicates the expression in log-spectral domain, $\mathbf{C}^{-1}$ is the transformation matrix of IDCT, and a super script *T* means the transpose of a matrix. As a signal feature is assumed Gaussian in cepstral domain, it is also Gaussian in log-spectral domain because of the linearity of DCT. The spectral mean can be computed from the Gaussian integration of taking exponential on the log-spectrum. As derived in [2], the *i*-th component of spectral mean and variance can be efficiently computed as follows,

$$m_i = \exp(m_i^l + s_{ii}^l / 2) \tag{3}$$

$$\text{and } s_{ij} = m_i m_j [\exp(s_{ij}^l) - 1]. \tag{4}$$

Assuming that the speech and the noise are independent, the spectral mean and covariance of noisy speech can be obtained by

$$\hat{m}_i = g m_i + \tilde{m}_i \tag{5}$$

$$\text{and } \hat{s}_{ij} = g^2 s_{ij} + \tilde{s}_{ij}, \tag{6}$$

where *g* is a gain term providing the match of signal power to test condition, $\tilde{m}_i$ is the *i*-th component of spectral mean of noise, and $\tilde{s}_{ij}$ is the element of spectral covariance of noise indexed by *ij*. Assuming that the distribution of combined signal in log-spectral is normal, the above mapping process can be straightforwardly inverted. Therefore, the linear domain parameters are transformed back to the log-spectral domain by

$$\hat{m}_i^l = \log(\hat{m}_i) - 0.5 \hat{s}_{ii}^l \tag{7}$$

$$\text{and } \hat{s}_{ij}^l = \log\left(\frac{\hat{s}_{ij}}{\hat{m}_i \hat{m}_j} + 1 \right), \tag{8}$$

and secondly, back to the cepstral domain by

$$\hat{m}^c = \mathbf{C} \hat{m}^l \tag{9}$$

$$\text{and } \hat{\Sigma}^c = \mathbf{C} \hat{\Sigma}^l \mathbf{C}^T. \tag{10}$$

3. RELATIONSHIP BETWEEN ADAPTED CEPSTRAL VARIANCE AND ENERGY RATIO

Consider two random variables of signals *a* and *b* which are characterized by log-energy. Let these random variables be of normal distribution and their means and variances are $\{m_a^l, s_a^l\}$ and $\{m_b^l, s_b^l\}$. These two scalar random variables are the source signals to be combined in the linear energy domain. The mean and variance of the combined signal can be computed as follows according to the formulation of PMC.

$$m_a = \exp(m_a^l + s_a^l / 2), \tag{11}$$

$$s_a = m_a^2 [\exp(s_a^l) - 1], \tag{12}$$

$$m_c = m_a + m_b \tag{13}$$

$$s_c = s_a + s_b \tag{14}$$

$$m_c^l = \log(m_c) - 0.5 s_c^l \tag{15}$$

$$s_c^l = \log\left(\frac{s_c}{m_c^2} + 1 \right) \tag{16}$$

Using equations (13) and (14), we obtain

$$\frac{s_c}{m_c^2} = \frac{s_a + s_b}{(m_a + m_b)^2} = \frac{m_a^2 (\exp(s_a^l) - 1)}{(m_a + m_b)^2} + \frac{m_b^2 (\exp(s_b^l) - 1)}{(m_a + m_b)^2}. \tag{17}$$

If the energy of *a* is much larger than that of *b*, i.e., $m_a >> m_b$, then

$$\frac{m_a^2}{(m_a + m_b)^2} \cong 1 \text{ and } \frac{m_b^2}{(m_a + m_b)^2} \cong 0. \tag{18}$$

In this case, we can obtain the approximation $\frac{s_c}{m_c^2} \cong \exp(s_a^l) - 1$, and $s_c^l \cong \log(\exp(s_a^l) - 1 + 1) = s_a^l$. Conversely, if $m_a << m_b$, then $s_c^l \cong s_b^l$. Thus, it shows that the combined variance in the extreme case will be either s_a^l or s_b^l. Let's consider the case of equal energy, $m_a \cong m_b$. Assume that the variances of the source signals are similar, $s_a^l \cong s_b^l$, we can approximate both variances with their average $\hat{s}_a^l \cong \bar{s}^l$ and $\hat{s}_b^l \cong \bar{s}^l$, where $\bar{s}^l = 0.5(s_a^l + s_b^l)$. Bring these approximates into equation (17), we can obtain $\frac{s_c}{m_c^2} \cong 0.5(\exp(\bar{s}^l) - 1)$, and

$$s_c^l \cong \log\left(0.5(\exp(\bar{s}^l) + 1)\right) = \log(\exp(\bar{s}^l) + 1) - \log(2).$$

Assume $\exp(\bar{s}^l) >> 1$, we can obtain an approximate of combined variance

$$s_c^l \cong \bar{s}^l - \log(2) = 0.5(s_a^l + s_b^l) - \log(2). \tag{19}$$

4. APPROXIMATED VARIANCE ADAPTATION ON SPEECH MODELS

The above relationship between the combined variance and the energy ratio is generalized to the compensation of speech model by using PMC method. In this case, the energy of a multivariate Gaussian density is defined by the summation of spectral means. As revealed in the above, if the energy of speech is much greater than the energy of noise, the variance of combined signal will approach to the variance of speech. Conversely, it will approach to the variance of noise. Both cases may appear in most situations of model adaptation. In other words, the former case is generally held for the vowel parts, and the latter is for the consonant parts in an utterance. Besides the extreme cases, the remainders are considered as equal energy, and equation (19) should be applied. In transferring to cepstrl domain, log(2) can be dropped since we consider only the cepstral coefficients with quefrency index greater or equal to 1. Therefore, we can approximate the adapted variance with the average variances of speech and noise. The adaptation scheme of variance can be concluded as follow.

$$\hat{\sigma}_{ii}^{c} = \begin{cases} \sigma_{ii}^{c}, & \text{if } r > \rho \\ \tilde{\sigma}_{ii}^{c}, & \text{if } r < 1/\rho \\ 0.5(\sigma_{ii}^{c} + \tilde{\sigma}_{ii}^{c}), & \text{elsewhere} \end{cases}, \quad \text{for } 1 \le i \le p, \tag{20}$$

where the energy ratio is defined by $r = \sum_{j=0}^{N-1} \mu_j \Big/ \sum_{j=0}^{N-1} \tilde{\mu}_j$, N is the dimension of the spectral mean, p is the order of truncated cepstrum, and ρ is a preset threshold. Thus far, a simplified version of PMC method can be obtained by replacing equation (10) with equation (20). Because of the replacement, it can significantly simplify the computation.

5. EXPERIMENTS ON NOISY SPEECH RECOGNITION

5.1 Database and Feature Extraction

A Chinese name recognition is conducted to verify the proposed method. The recognition targets are 120 human names and there are 3480 test utterances in the experiment. The database of Chinese names, which was generated by Computer & Communication Labs (CCL) of Industrial Technology Research Institute, were collected from 18 males and 11 females through a microphone in a quiet environment. Each piece of name is consisted of three or four Mandarin syllables. The speech data is divided into two equal parts, the training set and the testing set. These two sets are exchanged for each experiment condition. In addition, a database of Chinese words and phrases, which was generated by Telecommunication Laboratories of Chunghwa Telecom Corp., is used as part of the training set. This database was collected from 51 males and 50 females through another microphone in a quiet environment. Each speaker pronounced 50 words or phrases of 4 Mandarin syllables.

These speech databases are sampled in 8K Hz and segmented into frames of 32ms with 50% overlap. Each segment of signal is multiplied by a Hamming window before performing the Fourier transform. Features are extracted from a mel-frequency analysis by using a 20-filter bank. Applying discrete cosine transform on the filtered energy vector, we obtain a truncated mel-frequency cepstrum with first 13 mel-frequency cepstral coefficients (MFCCs), including the zero-th MFCC that is needed in the PMC method. Besides, 12 delta MFCCs obtained by linear filtering 12 MFCCs are also involved to form a feature vector.

5.2 Experimental Setting

The noisy speech is artificially produced by adding three types of noises, Gaussian *white*, *babble* and *factory* noises, extracted from NOISEX-92 database, to the speech waveform in five SNR's [6]. A system without any noise compensation scheme, denoted by 'No Adapt.', is included to show the baseline performance. The result by using the standard PMC method is denoted by 'STD'. To demonstrate the performance gained from the adaptation of variance, the result by using the PMC but leaving the variance unaltered, denoted by 'NVA', is also reported. The proposed method, where the mean is adapted by the PMC but the variance is adapted directly in cepstral domain by equation (20), is denoted by 'DIR'. The threshold of energy ratio in equation 20, ρ, is set to 10 in the experiment. The model adaptation is performed for each test utterance and the noise model is updated during the speech inactive period. The power normalization of speech signal has been applied to the database, and hence the gain term g can be set to 1 in the combination step. For simplicity, the cepstral covariance matrices are assumed diagonal in the computation of likelihood scores.

5.3 Results

The recognition results, as listed in Table 1, show that 'NVA' can reduce most error rates from 'No Adapt.'. The improvement resulted from the adaptation of variance, i.e., 'STD', seems not so significant for the test utterance in high SNR. However, it is crucial to the cases in low SNR. For instance, an error reduction of 25.3% from 'NVA' is obtained when the test speech is corrupted by white noise in SNR = 0dB. It is also observed that there is no significant performance gained from the adaptation of variance in the case of babble noise. This case may be explained by the similarity between the test speech and the babble noise. The recognition rate of 'DIR' is very close to 'STD' in most test conditions. Though the adaptation strategy of using 'DIR' is coarse and simple, it is good enough in the case of SNR greater than 5dB.

Table 1, Error rates (%) of Chinese name recognition in various adaptation schemes ($r = 10$ in method 'DIR').

White	*20dB*	*15dB*	*10dB*	*5dB*	*0dB*
No Adapt.	2.6	6.3	19.4	52.1	87.8
STD	2.0	3.5	8.0	18.9	45.4
NVA	2.2	3.8	9.1	24.5	60.8
DIR	2.0	3.5	8.7	20.0	46.7

Babble	*20dB*	*15dB*	*10dB*	*5dB*	*0dB*
No Adapt.	1.6	4.6	16.8	50.6	85.8
STD	1.2	2.0	4.9	15.5	45.1
NVA	1.2	2.0	4.8	15.9	48.5
DIR	1.2	1.9	4.9	15.6	46.0

Factory	*20dB*	*15dB*	*10dB*	*5dB*	*0dB*
No Adapt.	1.7	3.2	11.6	41.7	87.9
STD	1.4	2.0	4.7	15.1	44.3
NVA	1.4	2.2	5.3	19.0	58.5
DIR	1.2	2.0	5.3	15.6	46.0

6. CONCLUSION

A modified PMC method is proposed by using a direct adaptation scheme on the cepstral variance. The experimental results show that its performance is comparable to the standard PMC in the cases of SNR greater than 5dB. Even though the adapted variance could be a coarse one, its performance degradation is small. In addition, the simplified approach removes the operation of the transformation of adapted variance from the log-spectral domain to the cepstral domain. It gains a lot of speed up in real time applications.

7. ACKNOWLEDGEMENTS

This research has been partially sponsored by the National Science Council of Taiwan, under contract number NSC-88-2614-E-007-002.

1. 8. REFERENCES

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

3. Gales, M. J. F. "Predictive model-based compensation schemes for robust speech recognition". *Speech Communication* 25, pp. 49-74, 1998.

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

5. Gales, M. J. F. and Young, S. J. "A fast and flexible implementation of parallel model combination". *Proceedings of ICASSP-95*, pp. 133-136, 1995.

6. Komori, Y., Kosaka, T., Yamamoto, H., & Yamada, M. " Fast parallel model combination noise adaptation processing". *Proceedings of. Eurospeech* 97, pp. 1523-1526, 1997.

7. 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*, pp. 247-251, 1993.