IMPROVED PARALLEL MODEL COMBINATION BASED ON BETTER DOMAIN TRANSFORMATION FOR SPEECH RECOGNITION UNDER NOISY ENVIRONMENTS

Jeih-weih Hung, Jia-lin Shen and Lin-shan Lee
Institute of Information Science, Academia Sinica
Nankang, Taipei, Taiwan, Republic of China
Email : {jwhung, jlshen}@iis.sinica.edu.tw

Abstract

The parallel model combination (PMC) technique has been shown to achieve very good performance for speech recognition under noisy conditions. However, there still exist some problems based on the PMC formula. In this paper, we first investigated these problems and some modifications on the transformation process of PMC were proposed. Experimental results show that this modified PMC can provide significant improvements over the original PMC in the recognition accuracies. Error rate reduction on the order of 12.92% was achieved.

1. Introduction

It is well known that the performance of a speech recognition system often degrades seriously while it is applied in the real environment. The performance degradation is mainly due to the mismatch between the training and test conditions. In order to achieve robust recognition in various noise conditions, many approaches were proposed and can be roughly classified into several categories: noise resistant speech feature representation, feature compensation methods and model compensation methods. Among the model compensation methods, the Parallel Model Combination (PMC) [1] technique has been shown to achieve very good performance for speech recognition in the presence of additive noise. With the PMC method the approximated noisy speech HMM's can be derived by properly combining the noise HMM and the pre-trained clean speech HMM's. Due to the fact that the speech signal is affected by the additive noise in the linear-spectral domain, in the PMC method it's suggested that the cepstral-based parameters of the clean speech HMM's and the noise HMM must be transformed to the log-spectral domain or the linear-spectral domain in order to perform the combination, and then inversely transformed back to the cepstral domain for normal recognition process.

Nevertheless, the PMC method doesn't necessarily provide good results in many cases as expected. The possible correlation between the speech signal and noise is one reason for this degradation, which has been discussed previously [2]. In this paper, we will discuss the other possible reason, which is the possible inaccuracy occurred while the log-spectral model parameters are transformed to the linear spectral ones. This inaccuracy is found to seriously influence the exactness of the following combination process in the linear spectral domain. As a result, it is expected that if more exact model transformation can be achieved, the accuracy of the resulting HMM's may be improved. In this paper, the preliminary experimental results indicate that this may be a correct direction.

The remainder of the paper is organized into 3 sections. In section 2, the transformation problem of the PMC method mentioned above will be discussed, and its counter measure will also be proposed. Section 3 then presents some preliminary experimental results using this modified PMC

method. Finally, a short conclusion is given in section 4.

2. The Problem of PMC and Its Counter Measure

Although under most noisy environments the PMC technique improves the recognition accuracy significantly as compared with the HMM's trained with clean speech, it still doesn't perform as well as the HMM's trained by noisy speech under matched conditions, and especially the performance gap becomes larger while the SNR becomes worse. There may be many reasons for this. One of them is the possible correlation between the speech signal and noise [2], and the problem occurred in the transformation process may be another, which will be discussed here.

In the transformation process of PMC, the original cepstral clean speech HMM's are first transformed into the log-spectral domain, and then transformed into the linear-spectral domain in order to perform the combination with noise model parameters. However, by comparing the values of the parameters for models transformed to he linear-spectral domain during the PMC processes with those for models trained directly in the linear-spectral domain with the same database, it's found that there exists a significant discrepancy between them. This discrepancy shows that the transformation between the log-spectral and linear-spectral domains used in PMC method was not performed very well. This may be one possible reason for which the performance of PMC techniques degrades in some cases. When examining the above transformation used in original PMC techniques, it is assumed that the log-spectrum x^l of the speech signal x is approximately normal-distributed as $N(\mu_{x^l},\sigma_{x^l}^2)$. Therefore the mean value of the speech signals in linear-spectral domain can be obtained below:

$$\begin{aligned}\mu_x = E(e^{x^l}) &= \frac{1}{K}\int_A e^{x^l}\exp\left(-\frac{(x^l-\mu_{x^l})^2}{2\sigma_{x^l}^2}\right)dx^l \\ &= \exp\left(\mu_{x^l}+0.5\sigma_{x^l}^2\right)\left[\frac{1}{K}\int_A \exp\left(-\frac{(x^l-\mu_{x^l}-\sigma_{x^l}^2)^2}{2\sigma_{x^l}^2}\right)dx^l\right],\end{aligned} \tag{1}$$

where $K=\int_A \exp\left(-\frac{(x^l-\mu_{x^l})^2}{2\sigma_{x^l}^2}\right)dx^l$.

In the original PMC technique, the integral in equation (1) is assumed operated over an infinite interval A=[-∞, ∞] such that the integral term

$$\frac{1}{K}\int_{-\infty}^{\infty}\exp\left(-\frac{(x^l-\mu_{x^l}-\sigma_{x^l}^2)^2}{2\sigma_{x^l}^2}\right)dx^l = 1 \tag{2}$$

such that equation (1) becomes

$$\mu_x = \exp\left(\mu_{x^l}+0.5\sigma_{x^l}^2\right). \tag{3}$$

However, the above assumption that the distribution range of speech data is two-sided infinite seems not very reasonable. One of the reasons is that since the amount of the available speech data is limited, it distributes indeed within a finite interval. The other unreasonable part is that this distribution implies finite probability exists even for the log-spectrum of the speech signals extending to infinity.

From the above observation, if we replace the infinite integral range [-∞, ∞] with a more reasonable one, say $[-\infty,\mu_{x^l}+a\sigma_{x^l}]$, where a is a parameter depending on the models, then equation (1) becomes

$$\mu_x = \exp\left(\mu_{x^l}+0.5\sigma_{x^l}^2\right)f\left(\sigma_{x^l}^2\right) \tag{4},$$

where

$$\begin{aligned}f(\sigma_{x^l}) &= \frac{1}{K}\int_{\infty}^{\mu_{x^l}+a\sigma_{x^l}}\exp\left(-\frac{(x^l-\mu_{x^l}-\sigma_{x^l}^2)^2}{2\sigma_{x^l}^2}\right)dx \\ &= \frac{\sigma_{x^l}}{K}\int_{-\infty}^{a-\sigma_{x^l}} e^{-\frac{y^2}{2}}dy\end{aligned}. \tag{5}$$

By comparing equation (3) and (4), we found that the modified formula of mean value is the PMC-derived formula multiplying a weighting factor $f\left(\sigma_{x^l}^2\right)$. One also can see that the weighting factor will obviously decrease with the increase of $\sigma_{x^l}^2$.

Besides the mean value, the covariance of clean speech signal in the linear spectral domain can also be obtained in similar

manner:

$$
\begin{aligned}
(\Sigma_x)_{ij} &= E(x_i x_j) - E(x_i)E(x_j) \\
&= E\left(e^{x_i^l + x_j^l}\right) - E\left(e^{x_i^l}\right)E\left(e^{x_j^l}\right) \\
&= \exp\left(\mu_{x_i^l} + \mu_{x_j^l} + 0.5\Sigma_{x_i^l + x_j^l}\right) f\left(\Sigma_{x_i^l + x_j^l}\right) \\
&\quad - \exp\left(\mu_{x_i^l} + 0.5(\Sigma_{x^l})_{ii}\right) f\left((\Sigma_{x^l})_{ii}\right) \exp\left(\mu_{x_j^l} + 0.5(\Sigma_{x^l})_{jj}\right) f\left((\Sigma_{x^l})_{jj}\right) \\
\text{where} \quad \Sigma_{x_i^l + x_j^l} &= (\Sigma_{x^l})_{ii} + (\Sigma_{x^l})_{jj} + 2(\Sigma_{x^l})_{ij}
\end{aligned}
\tag{6}
$$

After the mean and covariance $[\mu_x, \Sigma_x]$ of clean speech signal in the linear-spectral domain are obtained by equation (4) and (6), they combine with noise parameters $[\tilde{\mu}, \tilde{\Sigma}]$ to form the noisy speech mean and covariance $[\hat{\mu}_x, \hat{\Sigma}_x]$ as follows:

$$\boldsymbol{\mu}_{\hat{x}} = \boldsymbol{\mu}_x + g\tilde{\boldsymbol{\mu}}, \tag{7}$$

$$\boldsymbol{\Sigma}_{\hat{x}} = \boldsymbol{\Sigma}_x + g^2\tilde{\boldsymbol{\Sigma}}, \tag{8}$$

where g is a gain factor. In the next step, the noisy speech mean and covariance $[\mu_{\hat{x}}, \Sigma_{\hat{x}}]$ need to be transformed back to the log-spectral domain by help of the inverse functions of equation (4) and (6), which are quite complicated to obtain, however. By observing equation (4), we see that the inverse function of equation (4) is

$$\mu_{\hat{x}_i^l} = \log \frac{\mu_{\hat{x}_i}}{f\left((\Sigma_{\hat{x}^l})_{ii}\right)} - \frac{1}{2}(\Sigma_{\hat{x}^l})_{ii}, \tag{9}$$

that is, the value of the noisy log-spectral mean depends on the diagonal term of the log-spectral covariance matrix, which is still unknown. To reduce the complexity of computation, we use the following equation to gain the noisy log-spectral covariance matrix:

$$(\Sigma_{\hat{x}})_{ij} = \frac{\exp\left(\mu_{x_i^l} + \mu_{x_j^l}\right)(\Sigma_{x^l})_{ij} + \exp\left(\tilde{\mu}_i^l + \tilde{\mu}_j^l\right)(\tilde{\Sigma}^l)_{ij}}{\left(\exp\left(\mu_{x_i^l}\right) + \exp\left(\tilde{\mu}_i^l\right)\right)\left(\exp\left(\mu_{x_j^l}\right) + \exp\left(\tilde{\mu}_j^l\right)\right)} \tag{10}$$

which is derived from Taylor expansion approximation [3]. By using equation (10) and then equation (9) both the log-spectral mean and covariance of the noisy speech signal can be acquired, and next they are transformed back to the cepstral domain for the recognition process.

3. Experimental Results

Some experiments were performed to verify the concept mentioned here. The training speech database used in the experiments contains 2 sets of 1345 isolated syllables in Mandarin Chinese produced by a speaker. It is used to train 113 right context-dependent (RCD) INITIAL HMM's and 41 context-independent (CI) FINAL HMM's. Another set of 1345 syllables produced by the same speaker is used as the test data to be recognized in speaker dependent mode. 14 order mel-frequency cepstral coefficients are used as the feature parameters. Noise HMM's for different levels of white noise to be added into the clean speech are also individually trained, composed of one state and one mixture per state. Furthermore, matched HMM's are also trained with the noise corrupted speech data.

We first show the transformation problem of PMC mentioned in section 2. Figure 1 shows the linear spectral envelopes (linear spectral mean) for 3 sets of models averaged over all the 113 INITIAL models and 41 FINAL models. The first set of models is obtained by transformation using (3), the original transformation formula of PMC. The second set of models is obtained by transformation using (4), our proposed modified formula. The third set of models is trained directly in the linear spectral domain with the same database.

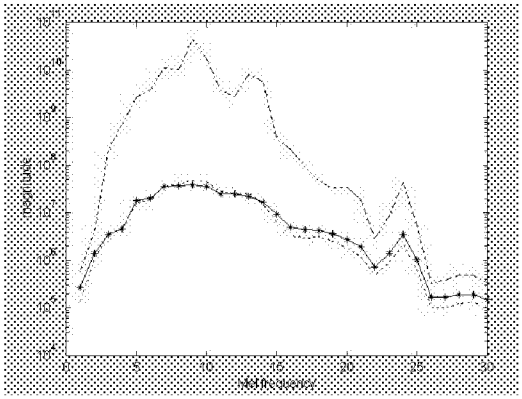

Fig 1. The spectral envelopes, where "-*-*-*-* " is the true linear spectral envelope, the solid line is the spectral envelope derived from the original PMC formula (3), and the dash line is

the spectral envelope derived from equation (4), the modified transformation formula.

From Figure 1 we see that, as previously mentioned, there exists a significant discrepancy between the PMC-derived linear spectral envelope and the true one. However, it is also shown that our proposed method makes the transformed linear spectral envelope approaches the true one a great deal.

Noise	SNR (dB)	①	②	③	Modified PMC HMM's	
					④	⑤
White	30	48.33	80.67	69.81	77.70	77.32
	20	14.42	71.15	44.98	62.53	64.98
	10	2.01	49.74	24.16	43.05	44.09
F16	30	66.25	84.54	78.74	81.93	81.71
	20	29.81	81.04	58.22	70.93	72.49
	10	8.55	69.22	25.87	48.85	50.93

Table 1. Recognition accuracies using different versions of models:① Clean HMM's, ② Matched HMM's ③ Original PMC HMM's ④ Better transformation ⑤ Better transformation and correlation

Next, the recognition accuracies using different versions of models on white noise and F16 noise are compared in Table 1. One can find that by using the clean speech HMM's, the recognition accuracies are seriously degraded, especially when SNR becomes worse. Secondly, it's also shown that the recognition rates can be significantly improved when the matched noisy speech HMM's are used. However, it's time-consuming and therefore impractical to retrain the noisy speech HMM's. In the original PMC models, only a short period of noise is used to train the noise models, thus is practically much feasible. As can be found in the fifth column of Table 1, the recognition performance of the original PMC method is improved significantly when SNR is high, but the improvements are reduced when SNR becomes lower. However, the modified PMC proposed here (listed in the last two columns) always provides better compensation compared to the original PMC, especially in the low SNR cases. In the sixth column, only the transformation problem mentioned in section 2 is considered and it uses equation (4) to perform the domain-transformation. Obviously the accuracies in the sixth column are actually much improved compared with the results of the original PMC-derived speech models. Finally, if both the transformation problem and the correlation problem are considered, the recognition rates will be further improved, as shown in the last column. From the experimental results, we see that the performance of the compensated HMM can be significantly enhanced by our proposed method.

4. Conclusion

In this paper, it is shown that proper modification of the domain-transformation in the original PMC method can produce much better acoustic models and accuracies for speech recognition under noisy conditions, especially when SNR is low. It is believed that the recognition performance can be further improved with more accurate domain-transformation process.

References

[1] M.J. Gales and S.J. Young, "Robust Speech Recognition in Additive and Convolutional Noise Using Parallel Model Combination", Computer, Speech and Language 9, pp. 289-307, 1995.

[2] J.W. Hung, J.L. Shen and L.S. Lee, "Improved Robustness for Speech Recognition Under Noisy Conditions Using Correlated Parallel Model Combination", 1998 IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '98).

[3] P.J. Moreno, B. Raj and R.M. Stern, "A Vector Taylor Series Approach for Environment-independent Speech Recognition", 1996 IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '96)