An Efficient Model Parameter Compensation Method for Robust Speech Recognition

정용주(계명대)

<차 례>

<Abstract>

An Efficient Model Parameter Compensation Method for Robust Speech Recognition

Yong-Joo Chung

An efficient method that compensates the HMM parameters for the noisy speech recognition is proposed. Instead of assuming some analytical approximations as in the PMC, the proposed method directly re-estimates the HMM parameters by the segmental k-means algorithm. The proposed method has shown improved results compared with the conventional PMC method at reduced computational cost.

* Keywords: speech recognition, noise-robust speech recognition, HMM

1. Introduction

Model parameter compensation methods have shown successful results in noisy speech recognition based on hidden Markov models (HMMs) [1][2]. Among them, the parallel model combination (PMC) has been quite effective. Compared to other adaptation methods, its computational burden is relatively small and it does not require any additional adaptation data except the noise samples in the testing speech. However, the PMC assumes some analytical approximations for the convenience of analysis, which may introduce some errors in the model parameter adaptation process [3]. Another source of the errors may come from the inverse discrete cosine transformation (DCT) which the PMC utilizes to obtain the log-spectrum mean vector from the cepstrum mean vector as follows.

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

However, since the two vector spaces have different dimensions, the accurate inverse DCT can not be performed. That is, different vectors from the log-spectrum domain can be transformed to the same cepstrum vector so the inverse DCT will fail to recover the original values.

In our proposed method, we apply the segmental k-means algorithm [4] to find the statistics that are used to compensate the HMM parameter vectors in combination with the noise in the testing speech. Since the values of the statistics are obtained during the training, the computational cost in the testing is relatively small. Also, since the method does not assume the analytical approximations required in the conventional PMC, it can take into account the acoustical variations due to the noise more directly.

2. Compensation using the segmental k-means algorithm

2.1. Noise Assumption

We assume that the noise-corrupted speech in the cepstral domain is characterized by the following nonlinear equation.

$$\begin{aligned} Y^c &= C \log (X+N) \\ &= C \log X + C \log(i+\exp(\log N - \log X)) \end{aligned} \tag{2}$$

where X represents the linear spectral vector of the clean speech and N is the linear spectral vector of the additive noise signal. And i represents a unit vector while C, the matrix representing the DCT [5].

2.2. Mean Vector Adaptation

In HMM-based speech recognition, the parameter estimation is usually done by using the segmental k-means algorithm. Instead of using the clean speech, the noise-corrupted speech Y^c may be applied to the algorithm to obtain the updated mean vector as follows.

$$\begin{aligned} \hat{\mu}_{jk}^{c} &= \frac{\sum_{t=1}^{T} \gamma_t(j,k)(C \log X_t + C \log(i+\exp(\log N - \log X_t)))}{\sum_{t=1}^{T} \gamma_t(j,k)} \\ &= \mu_{jk}^{c} + \frac{\sum_{t=1}^{T} \gamma_t(j,k)(C \log(i+\exp(\log N - \log X_t)))}{\sum_{t=1}^{T} \gamma_t(j,k)} \\ &= \mu_{jk}^{c} + E(C \log(i+\exp(\log N - \log X_t))) \\ &= \mu_{jk}^{c} + \mu_{jk}^{n} \end{aligned} \tag{3}$$

Here, $\gamma_t(j,k)$ is the probability of being in state j at time t with k-th mixture component accounting for Y^c. In the above equation, we need to find the value of the noise N. We assume the noise to be deterministic and then, μ_{jk}^{n} is easily obtained as follows by ignoring the effect of the variance of X_t [2].

$$\begin{aligned} \mu_{jk}^{x} &= E(X_t) \\ \mu_{jk}^{n} &= C \log(i+\exp(\log N - \log \mu_{jk}^{x})) \end{aligned} \tag{4}$$

In our proposed method, we first obtain the mean vectors of X_t during the

training session and use it in the testing session to find μ_{jk}^{n} which is added to the original mean vector μ_{jk}^{x}. We can reliably estimate μ_{jk}^{x} during the training by using a large amount of clean speech and efficiently compensate the mean vectors using the noise N in the testing speech without performing the inverse DCT.

2.3. Covariance Adaptation

The covariance matrix for the noisy speech can be similarly obtained. However, for the compensation of the covariance, we need to know the noise N in advance during the training session to find the necessary statistics. Therefore, the effect of the mismatch between the assumed noise N during the training and the observed noise $\hat{N}$ in the testing should be compensated using the Taylor series approximation as follows.

$$\begin{aligned} &C\log(i + \exp(\log \hat{N} - \log X_t)) \approx \\ &C\log(i + \exp(\log N - \log X_t)) \\ &\quad + \left[\frac{\partial C\log(i+\exp(\log N - \log X_t))}{\partial N}\right](\hat{N} - N) \end{aligned} \tag{5}$$

By using the above approximation in (5), the updated covariance matrix is obtained as follows.

$$\begin{aligned} \hat{\Sigma}_{jk}^{c} = E((Y_t^c - \hat{\mu}_{jk}^{c})(Y_t^c - \hat{\mu}_{jk}^{c})^T) &= E((Y_t^c Y_t^{cT}) + \hat{\mu}_{jk}^{c}\hat{\mu}_{jk}^{cT} \\ &\quad - 2E(Y_t^c)\hat{\mu}_{jk}^{cT} \end{aligned} \tag{6}$$

Also,

$$\begin{aligned} E(Y_t^c) = &E(C\log X_t) + E(C\log(i+\exp(\log N - \log X_t))) \\ &+ E\left(\left[\frac{\partial C\log(i+\exp(\log N - \log X_t))}{\partial N}\right]\right)(\hat{N} - N) \end{aligned}$$

The use of vector Taylor series have been done in the previous works [2][6], but our approach differs in that the derivatives are estimated during the training phase and

later used for testing. The proposed method can be also applied to the dynamic parameters as well. This is another advantage over the PMC where the delta and the delta-delta parameters cannot be compensated directly if linear regression coefficients are used for the dynamic parameters.

3. Experimental Results

3.1. Data Preparation

In this section, the performance of the proposed method of adapting the HMM parameter vectors is evaluated on speaker-independent isolated word recognition experiments. The vocabulary consists of 75 phoneme-balanced Korean words. And, the basic recognition unit is the set of 32 phoneme-like units that are modeled by the left-to-right continuous density HMM. The baseline HMM is trained by the segmental k-means algorithm using 4,500 utterances from 60 speakers. For the testing, the noise-corrupted 6,000 utterances from 80 speakers are used. The noisy speech was obtained by adding a car noise to the clean speech at various signal-to-noise ratios (SNRs). 13-th order mel-frequency cepstral coefficients (MFCCs) and their time derivatives (delta-MFCCs) using the regression coefficients are considered as the feature vectors.

3.2. Results

In <Table 1>, we compare the recognition rates of the proposed method with other approaches when static 13-th order MFCCs are used as the feature vectors. First, the recognition rates of the baseline recognizer with clean speech HMMs are shown. As there was no effort for compensation in the baseline recognizer, the recognition rate dropped severely at 20dB or below. We also show the recognition results when the baseline recognizer was retrained at the same SNRs as in the testing (matched conditions). The recognition results were improved considerably compared with the clean speech HMMs because the acoustic variation due to noise can be more successfully reflected in the HMM during the training process.

In <Table 1>, we compare the results of the proposed method with the conventional PMC methods. We could see that better performance is obtained when we adapt only mean vectors. Compensating the diagonal covariance matrix in addition to

the mean vectors did not lead to improved results in the proposed method as well as in the PMC methods. This degraded performance with the variance adaptation have also been reported in some previous works [6]. The reduced values of the estimated covariances may be the reason for the weak robustness against noise and speaker variability resulting in poor performances. And, generally, for the variance adaptation, we will need more adaptation data than for the mean vector. The noise samples in the testing speech which we used for the compensation may be not enough for the reliable estimation of the covariance matrix. The results from the log-normal PMC and log-add PMC are both presented.

<Table 1> Comparison in the word recognition rates (%) of the proposed method with other approaches when static 13-th order MFCCs are used as the feature vectors.

	0dB	10dB	20dB
Clean speech HMM	32.5	71.3	88.7
Matched Conditions	84.4	91.8	94.3
log-normal PMC (mean only)	82.6	90.8	93.8
log-normal PMC (mean + variance)	73.9	89.8	93.1
PMC log-add	82.7	90.6	93.7
Proposed method (mean only)	83.6	91.5	93.9
Proposed method (mean + variance)	83.4	91.2	93.9

The results of the log-add PMC were comparable to the log-normal PMC although the log-add PMC is a simplified version. This may be due to the fact the log-normal PMC introduces some analytical approximations. We also note that the inverse DCT used in the PMC methods is another reason for the possible errors occurring in the domain transformation. We could see that the proposed method outperformed the PMC methods entirely. In particular, at 0dB, the recognition rate of the proposed method with only mean vector adaptation was 83.6% while 82.7% was obtained for the log-add PMC. This means that we could reduce by half the difference in the performance between the log-add PMC and the matched condition HMMs. This is remarkable considering the fact that the retraining method may give us the benchmark

performance attainable.

In <Table 2>, we show the results when the delta-MFCCs are added to form 26-th order feature vectors and only mean vectors are compensated. As expected, the recognition rates were increased considerably for all the approaches compared with the results in <Table 1>. But, for the PMC methods, compensating the delta- MFCC mean vectors resulted in some performance degradation. There seems to be some approximation errors in (5) for updating the delta-MFCC mean vectors. However, in our proposed algorithm, we can see far better recognition results when we update the delta-MFCC mean vectors in addition to the static mean vectors, because the proposed method adapts the delta-MFCCs mean vectors directly using the statistics obtained during the training.

<Table 2> Comparison in the word recognition rates (%) of the proposed method with other approaches when the delta-MFCCs are added to the feature vectors. The scores in the parenthesis are for the cases when only static means are updated.

	0dB	10dB	20dB
Clean speech HMMs	55.7	89.9	94.6
Matched conditions	89.1	95.6	97.5
log-normal PMC (mean only)	86.7 (86.9)	94.1 (93.7)	96.8 (93.8)
log-add PMC	86.1 (87.1)	93.9 (93.8)	96.7 (96.7)
Proposed method (mean only)	90.4	95.9	97.5

To see the effect of the proposed method when detailed acoustic phonetic models are used, we show in <Table 3>, the results of the compensation when 255 tri-phone models are used instead of the 32 phoneme like units. 26-th order MFCCs are used as the eature vectors and delta-MFCC mean vectors are not compensated for the log-add PMC. Although, the baseline recognition rates are quite high even at 0dB, we can see that the proposed method improves recognition results compared with the log-add PMC method as expected.

<Table 3> Comparison in the word recognition rates (%) of the proposed method with other approaches when 255 tri-phone models are used.

	0dB	10dB	20dB
Clean speech HMMs	82.8	95.8	98.3
Matched Conditions	96.9	98.7	98.8
Log-add PMC	94.7	97.4	98.3
Proposed Method	95.9	97.9	98.6

4. Conclusions

In this paper, we proposed an efficient method for the HMM parameter compensation in noisy speech recognition. As the method utilizes the statistics obtained during the segmental k-mean training process, simple adaptation is performed using the noise samples in the testing speech without requiring much computational cost in the testing. It also has the merit that dynamic parameters are easily compensated without requiring analytical approximations used in the previous approaches. From the experimental results, we could see that it outperformed the conventional PMC methods where some approximations are assumed for the convenience of analysis.

References

[1] M. J. F. Gales, "Model Based Techniques for Noise-Robust Speech Recognition", Ph.D. Dissertation, University of Cambridge, 1995.

[2] P. J. Moreno, "Speech Recognition in Noisy Environments", Ph.D. Dissertation, Carnegie Mellon University, 1996.

[3] J-W. Hung, J-L. Shen, L-S. Lee, "New approaches for domain transformation and parameter combination for improved accuracy in parallel model combination (PMC) techniques", *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 8, pp.842-855, 2001.

[4] L. R. Rabiner, B. H. Juang et al., "Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities", *AT&T Tech. J.*, Vol. 64, No. 6, pp.1211-1234, 1985.

[5] S. B. Davis, P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 28, pp.357-366, 1980.

[6] S. Sagayama, Y. Yamaguchi et al., "Jacobian approach to fast acoustic model adaptation", *Proc. ICASSP*, 1997.

접수일자: 2003년 2월 5일
수정일자: 2003년 3월 7일
게재결정: 2003년 3월 8일

▶ 정용주(Yong-Joo Chung)
주소: 704-701 대구시 달서구 신당동 1000번지
소속: 계명대학교 전자공학과
전화: 053) 580-5925
FAX: 053) 580-5165
E-mail: yjjung@kmu.ac.kr