Segmental Corrective Training for HMM Parameter Estimation in Speech Recognition

음성인식 시스템의 HMM 파라메터 추정을 위한 분절단위 교정 학습

Hoi Rin Kim*, Hwang Soo Lee**
김 회 린*, 이 황 수**

ABSTRACT

We present a modified corrective training method using state segment information in the hidden Markov model (HMM). The modified corrective training method corrects the HMM parameters using the segmental K-means algorithm instead of the forward-backward algorithm used in the conventional corrective training methods. This is motivated from the fact that the segmental K-means algorithm has more emphasis on the model state segment information which possesses common stochastic characteristics in speech signal. In a speaker-dependent phoneme and word recognition experiment, we show that the proposed algorithm results in higher recognition rate than the conventional corrective training method and requires less amount of computation. This shows the importance of state segment information in corrective training.

요 약

본 논문에서는 HMM 파라메터 추정을 위해 분절단위 정보를 이용하는 수정된 교정학습 방법을 제안한다. 수정된 교정학습 방법은 기존의 교정학습 방법에서 사용하는 전향-후향 알고리즘 대신에 분절단위 K-means 알고리즘을 사용하여 HMM 파라메터를 교정한다. 이 방식은 분절단위 K-means 알고리즘이 음성신호내의 공통의 통계적 특성을 가지는 상태단위 정보를 강조한다는 사실을 이용하였다. 화자종속 음소 및 단어인식 실험에서 제안된 알고리즘이 기존의 교정학습 방법보다 적은 계산량으로도 향상된 인식률을 보여주었다. 이것은 HMM 교정학습에서 상태단위 정보가 중요함을 보여준다.

I. INTRODUCTION

It is well known that the use of the HMM is very effective in automatic speech recognition [1]. Most HMM-based systems use the forward-backward algorithm for parameter estimation, which obtains an approximation to the maximum likelihood estimates(MLE) of the HMM parameters. MLE is based on the assumption that the underlying models are correct. In reality, however, this is extremely an inappropriate assumption about

*ETRI
**KAIST
접수일자 : 1992. 8. 9

the speech production process. To overcome the defects from the incorrect assumption, the corrective training method has been proposed[2][3][4]. The criterion in the corrective training method is not to maximize the likelihood for training data, but to minimize the error rate on training data. That is, when there is a recognition error on training data, or even when a wrong candidate gets too close to the right one, the initial model parameters are adjusted so as to lower the probability of the label responsible for the mistake or the near-miss. In this paper, we present a modified corrective training method using the state segment information in HMM, which improves the discriminant ability of the models.

Ⅱ. CORRECTIVE TRAINING

The corrective training(CT) algorithm has the following training procedure for a set of Markov models.

1. Using some training data and the forward-backward algorithm, obtain an estimate, λ of the parameters.
2. Perform speech recognition on the training data using the statistics λ.
3. If any utterance w is misrecognized as ω, adjust λ so as to make w more probable and ω less probable.
4. If any adjustments to λ were made, return to Step 2.

The corrective training is similar to an error-correction training procedure for the linear classifier. While the latter can be shown to converge [5], convergence for the corrective training has not been proven. Nevertheless, in practice the method does appear to converge.

Leaving aside the convergence problem, the corrective training is appealing from a pragmatic point of view. It is not assumed that the models are correct. For any set of models, the corrective training attempts to find statistics which make the models work. Instead of seeking statistics which maximize the likelihood, it acts directly on the error rate. Since the corrective training begins with the conventional forward backward parameter estimates, and since any parameter values that increase the error rate can be discarded, the final estimates after corrective training cannot have a higher error rate on the training data than MLE.

Of course, statistics which minimize the error rate on training data, do not necessarily minimize the error rate on test data. For this reason, it is important that the training data be representative of the intended application, and that as much data as practicable be made available for training purposes.

Let us now consider the implementation of the corrective training in more detail. Any adjustment to λ may introduce new errors. This is especially like where the probability of the correct model is only slightly greater than the probability of some other (incorrect) model. Therefore, λ is adjusted whenever the probability of an incorrect model is relatively close to the probability of the correct model. This helps to prevent new errors being introduced, and accelerates convergence. It also increases the difference in the probabilities of the correct and incorrect models, and hence the robustness of the recognizer.

The following algorithm incorporates the above considerations. It has two parameters, β and δ, which affect the rate of convergence and the extent to which the correct and incorrect probabilities are driven apart.

1. Using some labeled training data, apply the forward-backward algorithm iteratively to obtain an estimate λ of the parameters, and to compute the approximate frequencies $\hat{c}(k, a)$ for each label k in the label alphabet and each arc a in the arc inventory.
2. For each utterance u in the training data, use the statistics λ to compute the probability $P\{u|w\}$ for the correct model w, and the prob-

ability $P\{u|\omega_i\}$ for each incorrect model ω_i on the model list. Perform Step 3 for every triple u, w, ω_i, satisfying $\log P\{u|\omega_i\} > \log P\{u|w\} - \delta$.

3. Using the correct model w with the current statistics λ, and using only those labels corresponding to utterance u, apply one iteration of the forward-backward algorithm to obtain an estimate $\hat{c}_w^u(k, a)$ of the number of times each label k was produced by each arc a. Similarly, using the incorrect model ω_i, compute the approximate frequencies $\hat{c}_{\omega_i}^u(k, a)$. For each label k and arc a, replace $\hat{c}(k, a)$ by $\hat{c}(k, a) + \gamma(\hat{c}_w^u(k, a) - \hat{c}_{\omega_i}^u(k, a))$, where $\gamma = \beta$ if $(\log P\{u|w\} - \log P\{u|\omega_i\} \leq 0$, and γ decreases linearly from β to 0 as $(\log P\{u|w\} - \log P\{u|\omega_i\})$ increases from 0 to δ.
4. If any adjustments were made in Step 3, replace any negative frequencies $\hat{c}(k, a)$ by 0, recompute the parameters λ, and return to Step 2.

The parameter δ in Step 2 defines a near-miss and the parameter γ in Step 3 defines the degree of scaling. The error-correcting adjustments are made in Step 3. The frequencies $\hat{c}_w^u(k, a)$ for the correct model w are added to the existing frequencies, and the frequencies $\hat{c}_{\omega_i}^u(k, a)$ for the incorrect model ω_i are subtracted. This biases the frequencies in favor of the correct model and against the incorrect model, thereby increasing $P\{u|w\}$ and decreasing $P\{u|\omega_i\}$ as required.

Recently, an alternative error-correction strategy has been proposed by T. H. Applebaum and B. A. Hanson[3]. This method adjusts the frequencies of both the correct model and the incorrect model. That is, the error-correcting adjustments in Step 3 are modified as follows.

$$\hat{c}_w(k, a) = \hat{c}_w(k, a) + \gamma \cdot \hat{c}_w^u(k, a) \quad (1)$$

$$\hat{c}_{\omega_i}(k, a) = \hat{c}_{\omega_i}(k, a) - \gamma \cdot \hat{c}_{\omega_i}^u(k, a), \quad (2)$$

where $\hat{c}_w(k, a)$ is the approximate frequency for the correct model w, and $\hat{c}_{\omega_i}(k, a)$ is the approximate frequency for the incorrect model ω_i.

The experiments for the corrective training algorithm have been performed only on acoustically confused words since adjustments to the parameters will only be made for errors and near-misses. The results have shown that the performance of corrective training is better than either MLE via the forward-backward algorithm or MMIE, although the corrective traning algorithm is sensitive to selection of parameters.

Ⅲ. SEGMENTAL CORRECTIVE TRAINING

The coventional corrective training is performed using the forward-backward algorithm. That is, when the likelihood for each model in the vocabulary given an utterance in the training data is computed, the forward algorithm is used. Also, the estimate of the frequency count for each codeword in an output probability distribution is obtained using forward-backward algorithm. Therefore, this method biases the frequency counts in favor of the correct model and against the incorrect model in the sence of maximum likelihood.

Recently, Juang, B. H. has proposed another parameter estimation method for the HMM-based system, called the segmental K-means algorithm[6]. This method focuses on the probability of the most likely state sequence as opposed to summing the probabilities over all possible state sequences in each model. This is motivated from the fact that the modeling by the segmental K-means algorithm may be reasonable in the sence that modeling and decoding must be performed on a same criterion, since the Viterbi scoring algorithm with backtrace search is efficient for decoding words (or phonemes) in continuous speech recognition. Moreover, the segmental K-means algorithm has more emphasis on the model state segment information which possesses common stochastic characteristics in speech signal, and can avoid numerical difficulties associated with the forward-backward algor-

ithm, such as probability calculation and scaling [7]. On the other hand, this algorithm tends to over-estimate the HMM parameters to the training data, and consequently may result in poor performance on the test data. However, a proper use of the state segment information will improve the disciminant property between the models.

From these considerations, we can expect that the modified corrective training with segmental correction will enhance the recognition performance of the HMM-based system. The procedure of the modified algorithm which we call as segmental corrective training is as follows.

1. Using some labeled training data, apply the forward-backward algorithm iteratively to obtain an estimate λ of the model parameters, and apply the segmental K-means algorithm to compute the approximate counts $\hat{c}_w(k, a)$ for each label k and each arc a in each phoneme model w.
2. For each phoneme utterance u in the training data, use the statistics λ to compute the state-optimized likelihood $Pr\{u, s^*|w\}$ for the correct phonemen model w, and the state-optimized likelihood $Pr\{u, s^*|w_i\}$ for each incorrect phoneme model w_i on the phoneme list. Perform step 3 for every triple u, w, w_i, satisfying $\log Pr\{u, s^*|w_i\} > \log Pr\{u, s^*|w\} - \delta$.
3. Using the Markov model for the correct phoneme w the current statistics λ, and using only those labels corresponding to the utterance u, apply one iteration of the segmental K-means algorithm to obtain an estimate $\hat{c}^u_w(k, a)$ of the number of times each label k was produced by each arc a. Similarly, using the model for the incorrect phoneme w_i, compute the approximate counts $\hat{c}^u_{w_i}(k, a)$. Then, update both $\hat{c}_w(k, a)$ and $\hat{c}_{w_i}(k, a)$ as the followings.

$$\hat{c}_w(k, a) = \hat{c}_w(k, a) + \gamma \cdot \hat{c}^u_w(k, a) \tag{3}$$

$$\hat{c}_{w_i}(k, a) = \hat{c}_{w_i}(k, a) - \gamma \cdot \hat{c}^u_{w_i}(k, a), \tag{4}$$

where

$$\gamma = \begin{cases} \beta, \text{ if } (\log Pr\{u, s^*|w\} - \log Pr\{u, s^*|w_i\}) \le 0 \\ \beta\left(1 - \dfrac{\log Pr\{u, s^*|w\} - \log Pr\{u, s^*|w_i\}}{\delta}\right), \\ \text{if } 0 < (\log Pr\{u, s^*|w\} - \log Pr\{u, s^*|w_i\}) \le \delta. \end{cases}$$

4. If any adjustments were made in step 3, replace any negative counts $\hat{c}_w(k, a)$ by 0, ce compute the parameters in λ, and return to step 2.

The state-optimized likelihood $Pr\{u, s^*|w\}$ in step 2 is obtained by the Viterbi scoring algorithm, and the best state sequence s^* in the model w for the utterance u is also obtained by backtrackin the Viterbi decoding path. The near-miss factor δ affects the extent to which to which the correct and incorrect probabilities are driven apart, and the learning rate β affects the rate of convergence.

Ⅳ. SIMULATION RESULT

The performance of the proposed segmental corrective training method was obtained for a speaker dependent phoneme and word recognition system whose recognition units are phonemes. We used 43 Korean context-independent phoneme models, and each word was modeled by concatenating the corresponding phoneme models. Each phoneme HMM is a simple left-to-right model with 3 states and 7 transitions as shown in Fig1. The transitions are tied into three groups for robust estimation of output probabilities. Transitions in the same group share the same output

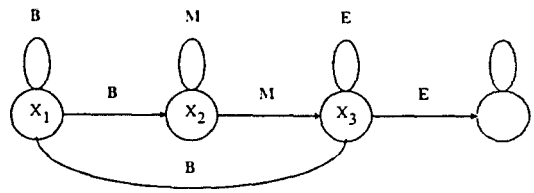

Fig. 1. The phoneme model used in this system.

probabilities represented by B, M, and E. This model assumes that there are at most three steady states for a phoneme, which are indicated by the self-loops. Furthermore, the lower transition explicitly models short durations. The vocabulary consists of phonetically balanced 100 Korean words and one male speaker uttered 100 words with 5 repetitions : utterances obtained from 4 repetitions were used in the training phase and the remaining was used in the test phase.

Utterances were low-pass filtered with the cut-off frequency of 4.5KHz and sampled at 10KHz. End points of the utterances were detected manually to obtain the exact performance of the proposed method. Smoothed power spectra were obtained in every 10msec interval by using the homomorphic processing method. As the input feature vectors of HMM, sixteen frequency-band energies in a mel-frequency domain were computed from the spectra and vector-quantized. The size of the VQ codebook was 256.

The HMM parameters for each phoneme were estimated by five different methods : first, the forward-backward estimation (FB), second, the segmental K-means re estimation (only for the output probabilities) of the HMM parameters initialized by the forward-backward algorithm (FB+SKM), third, the conventional corrective training of the HMM parameters estimated by the first method (FB+CT), fourth, the segmental corrective training of the HMM parameters estimated by the second method (FB+SKM+SCT), and last, the segmental corrective training of the HMM parameters estimated by the first method (FB+SCT). In these experiments, we did not adjust the transition probabilities, which are relatively unimportant in our system. We set the learning rate β to 1 from the previous works [2][3], and the near-miss factor δ to $\delta_0 \cdot \log Pr\{u \mid w\}$ for the conventional method and to $\delta_0 \cdot \log Pr\{u, s^* \mid w\}$ for the segmental method, where δ_0 was varied from 0 to 0.03. Also, five iterations of corrective training were performed.

Fig.2 and Fig.3 show the convergence behaviors for the three HMM parameter estimation methods with the corrective training procedure. In the figures, the number of errors mean the number of times an incorrect phoneme was found to be more probable than the correct phoneme in the training script and the number of adjustments mean the total number of errors and near-misses. As shown in Fig.2, the number of errors decrease fast until the 3rd iteration and no appreciable reduction in error counts results after that. The error rates on the training data were lowest for the FB+SKM+SCT method. The recognition accuracies evaluated on test data by the Viterbi scoring for the five parameter estimation methods are given in Table 1. Phoneme accuracies were obtained on 375 phoneme test data in the phonetically balanced 100 words. Word accuracies were obtained on the 100 word test data. Best performances for the three methods with corrective training procedure were obtained at $\delta_0 = 0.02$ and interation = 3. The FB+SCT method yields the highest recognition rate. The performance of the FB+SKM+SCT method was ranked 2nd on the test data while it yields the best performance on the training data. This is due to the fact that the FB+SKM+SCT method over-estimates the model parameters to the training data by using the state segment information too muc.

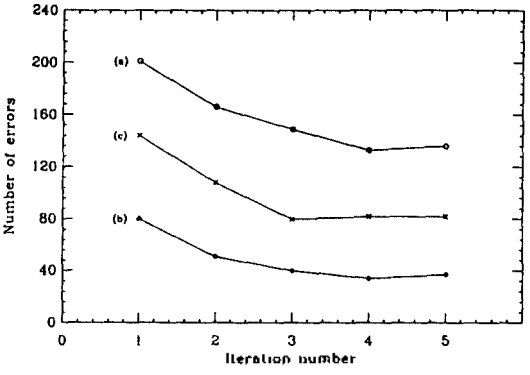

Fig. 2. Number of errors versus iteration number for three estimation methods ($\beta = 1$, $\delta_0 = 0.02$, and training data size = 1505).
(a)FB+CT (b)FB+SKM+SCT
(c)FB+SCT

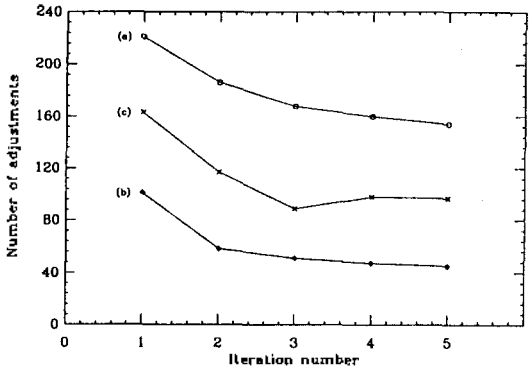

Fig. 3. Number of adjustments versus iteration number for three estimation methods ($\beta = 1$, $\delta_0 = 0.02$, and training data size = 1505).
(a)FB+CT (b)FB+SKM+SCT
(c)FB+SCT

Table 1. Performance comparison of the five parameter estimation methods in speaker-dependent phoneme and word recognition (at $\beta = 1$, $\delta_0 = 0.02$, and iteration = 3).

Estimation method	Phoneme accuracy(%)	Word accuracy(%)
FB	71.5	88
FB+SKM	72.5	88
FB+CT	72.5	89
FB+SKM+SCT	70.9	90
FB+SCT	74.9	93

V. CONCLUSION

We have studied a training procedure based on the modified corrective training which uses the segmental information in speech signals. While the conventional corrective training corrects the HMM parameters by the forward-backward algorithm, the modified algorithm corrects the HMM parameters by the segmental K-means algorithm. In the initialization procedure for the corrective training algorithms, two HMM parameter sets are used : (1) a set estimated by the forward-backward algorith, (2) a set re-estimated by the segmental K-means algorithm from the set obtained in (1).

We have experimented the proposed algorithm on a speaker-dependent phoneme and word recognition system using context-independent phoneme models. From the experimental results, the proposed segmental corrective training initialized by the forward-backward algorithm has been shown to have the best performance. The gain on the phoneme recognition accuracy against the conventional corrective training is 2.4%, and the gain on the word recognition accuracy is 4%. Therefore, we can conclude that the use of the segmental information in corrective training improves the discriminant ability of the phoneme models, and so improves the recognition performance for the HMM-based phoneme or word recognition systems. This reflects the importance of state segment information in corrective training.

ACKNOWLEDGEMENT

Thanks to Dr. Yong Ju Lee, Byung Nam Yoon, and Dr. Chul Hee Kang in ETRI for their thoughtful considerations.

REFERENCES

1. L. R. Rabiner, "An introduction to hidden Markov models," IEEE ASSP Mag., Jan. 1986.
2. L. R. Bahl, et al., "A new algorithm for the estimation of hidden Markov model parameters," Proc. ICASSP88, Paper S11.2, 1988.
3. T. H. Applebaum and B. a. Hanson, "Enhancing the discrimination of speaker independent HMMs with corrective training," Proc. ICASSP89, Paper S6.12, 1989.
4. K. F. Lee and S. Mahajan, "Corrective and reinforcement learning for speaker-independent continuous speech recognition," Computer Speech and Language, Vol.4, pp.231-245, 1990.
5. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," Parallel distributed processing : exploration in the microstructure of cognition, vol. 1, MIT Press, Cambridge, Massachusetts, 1986.
6. B. H. Jung and L. R. Rabiner, "The segmental K-means algorithm for estimating parameters of hidden Markov models," IEEE Trans. on ASSP,

Vol.38, pp.1639-1641, 1990.
7. N. Merhav and Y. Epharaim, "Hidden Markov modeling using the most likely state sequence," Proc. ICASSP91, Paper S7.16, 1991.

▲Hoi Rin Kim

Hoi Rin Kim was born in Seoul, Korea, on Mar. 9, 1961. He received the B.S. degree in Electronics Engineering from Hanyang University, Seoul, in 1984, and the M.S. and Ph.D. degrees in Electrical Engineering from Korea Advanced Institute of Science and Technology(KAIST), Seoul, in 1987 and 1992, respectively.

Since Oct. 1987, he has been with the Machine Translation Section, Electronics and Telecommunications Research Institute(ETRI), Daejeon, where he is currently a senior member of technical staff. His research interests include speech recognition, neural networks, speech coding, and speech synthesis.

▲Hwang Soo Lee

Hwang Soo Lee is Associate Professor, Information and Communication Engineering, KAIST. (Vol.6, No.3)