Frame-Correlated HMM을 이용한 음성인식

김남수*, 은종관**
* 삼성종합기술원, 기반기술 연구소, 음성처리팀
** 한국과학기술원, 통신연구실

On the Use of a Frame-Correlated HMM for Speech Recognition

Nam-Soo Kim*, Chong-Kwan Un**
* Samsung Advanced Institute of Technology
** Dept. of Electrical Engineering, KAIST

ABSTRACT

We propose a novel method to incorporate temporal correlations into a speech recognition system based on the conventional hidden Markov model (HMM). With the proposed method using the extended logarithmic pool (ELP), we approximate a joint conditional PD by separate conditional PD's associated with respective components of conditions. We provide a constrained optimization algorithm with which we can find the optimal value for the pooling weights. The results in the experiments of speaker-independent continuous speech recognition with frame correlations show error reduction by 13.7 % with the proposed methods as compared to that without frame correlations.

I. INTRODUCTION

The success or failure of a hidden Markov model(HMM) system relies on how well the models can characterize the nature of real speech. Various approaches have been tried to take account of frame correlations for more realistic speech modeling. Some of them adopt the stochastic segment model or the dynamic system model in order to directly express speech feature trajectories[1]. In the case of continuous HMM, some approaches to change the topology of a conventional model have been developed[2][3]. Paliwal incorporated temporal correlation into the discrete HMM by conditioning the probability of current observation on the current state as well as on the previous observation[4]. With this approach, an output probability distribution(PD) is constructed for each possible pair of state and observation symbol. Even though this full parametrization is the most natural way to express the behavior of temporal correlations, the number of parameters to be estimated may increase excessively to get reliable estimates for the output PD's. As an alternative to this, a bigram-constrained(BC) HMM was proposed[5]. In the BC HMM, the spectral shape of an output PD in a state is restricted according to the observation symbol on the previous frame.

In this paper, we address the problem of efficient incorporation of frame correlations into the conventional HMM. Key issues in this approach are how to precisely express true temporal correlations, how to obtain robust estimates, and how to easily combine with the conventional HMM scheme. The BC HMM serves a good starting point to begin with for the reason that it can easily be combined with the traditional HMM recognition steps. But, as for the way to express temporal correlations, we consider it to be inadequate, though the intuition of restricting the spectral shape is quite appealing. We focus on the way to combine separate conditional PD's in order to precisely approximate the joint conditional PD.

For this purpose, several candidate strategies can be found in the field of statistics where the problem of aggregating a number of expert opinions is addressed under the name of group interaction, consensus belief emergence, or managerial expert use[6]. Among many pooling operators, the logarithmic opinion pool(LOP) attracts most, because it appears similar to the BC HMM while possessing a more flexible modeling capability. By adopting a schematic form of the LOP, we propose a new method to incorporate frame correlations based on the conceptual analogy in which we treat separate conditional PD's as if they were the opinions with which we can determine the aggregated opinion, i.e., the joint conditional PD. We will call the proposed method the extended logarithmic pool(ELP) where the word "extended" means that we expand the allowed region for pooling weights which lie on a positive simplex in the original LOP. With this ELP, we approximate a true joint conditional PD by means of separate conditional PD's in which the pooling weights are estimated so as to minimize approximation errors. To evaluate approximation error, we use the discrimination information which

indicates to what extent a PD deviates from a given reference PD[7]. The objective function expressed in terms of this discrimination information measure can be minimized by using a feasible direction method applicable to a wide range of constrained optimization problems[8].

In addition, we consider several related issues which are shown to be indispensable for enhancing the recognition performance when we apply ELP to HMM-based speech recognition. First, we suggest a method to derive phoneme-independent frame correlation PD's with maximum entropy. In spite of the fact that temporal correlation highly depends on phoneme identity, the use of phoneme-independent correlation PD's is preferred to that of phoneme-dependent ones for a moderate or small sized training data due to the requirement for robust parameter estimation. Next, we present a technique to combine two kinds of PD's through some exponents which are estimated according to the maximum mutual information(MMI) criterion[9]. Practically, the restriction of a state specific output PD usually yields too excessive concentration over only a small region of the observation space. Therefore, it is desirable for robust recognition to diffuse this concentration while maintaining most useful discriminability information.

II. EXTENDED LOGARITHMIC POOL

Before introducing ELP, we briefly discuss how LOP works for aggregating a number of expert opinions. Assume that two experts provide their opinions in terms of a PD over an observation set. Let $p(X \mid A)$ and $p(X \mid B)$ be the provided opinions where X is a random variable representing an observation. Then, by LOP

$$\hat{p}(X=x \mid A, B) = \frac{p(X=x \mid A)^{\lambda_A}\, p(X=x \mid B)^{\lambda_B}}{\sum_{y \in C} p(X=y \mid A)^{\lambda_A}\, p(X=y \mid B)^{\lambda_B}} \qquad (1)$$

where $\hat{p}(X \mid A, B)$ is the aggregated opinion and C is the whole observation set. λ_A and λ_B are the pooling weights in aggregation such that $\lambda_A, \lambda_B \geq 0$ and $\lambda_A + \lambda_B = 1$. ELP is motivated by the fact that we can treat $p(X \mid A)$ and $p(X \mid B)$ in (1) as if they were separate conditional PD's and $\hat{p}(X \mid A, B)$ as the approximated joint conditional PD. In addition, we expand the allowed region for pooling weights such that $\lambda_A, \lambda_B \geq 0$.

Now, for more general formulation, we assume that there are N conditions, $\eta^1, \eta^2, \cdots, \eta^N$. Let $Z^1, Z^2, \cdots, Z^N$ be the random variables representing each kind of conditions. By the ELP

$$\hat{p}(X=x \mid Z^1=\eta^1, Z^2=\eta^2, \cdots, Z^N=\eta^N; \lambda_1, \lambda_2, \cdots, \lambda_N)$$

$$= \frac{\prod_{i=1}^{N} p(X=x \mid Z^i=\eta^i)^{\lambda_i}}{\sum_{y \in C} \prod_{i=1}^{N} p(X=y \mid Z^i=\eta^i)^{\lambda_i}} \qquad (2)$$

where λ_i is a positive number indicating the pooling weight of the i th condition. For notational brevity, we use vector-like(row vector form) notations given as follows.

$$\mathbf{Z} = (Z^1, Z^2, \cdots, Z^N),$$

$$\boldsymbol{\eta} = (\eta^1, \eta^2, \cdots, \eta^N), \text{ and}$$

$$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_N).$$

Here, $\boldsymbol{\lambda}$ should be obtained such that the difference between the true joint conditional PD and that approximated via ELP is as small as possible. Therefore, we need an appropriate measure with which we can determine distance between two PD's. One of most natural measures is the discrimination information which indicates how far a PD deviates from a reference PD. If we let $p(X)$ and $q(X)$ be two PD's defined over C where the former is a reference, the discrimination information between them is defined by

$$D(p(X); q(X)) = \sum_{x \in C} p(X=x) \log \frac{p(X=x)}{q(X=x)}. \qquad (3)$$

Let

$$L(\boldsymbol{\lambda}) = D(p(X \mid \mathbf{Z}=\boldsymbol{\eta}); \hat{p}(X \mid \mathbf{Z}=\boldsymbol{\eta}; \boldsymbol{\lambda})) \qquad (4)$$

where $p(X \mid \mathbf{Z}=\boldsymbol{\eta})$ and $\hat{p}(X \mid \mathbf{Z}=\boldsymbol{\eta}; \boldsymbol{\lambda})$ denote true and approximated joint conditional PD's, respectively. The optimal value $\boldsymbol{\lambda}_{opt}$ is obtained under the criterion that

$$\boldsymbol{\lambda}_{opt} = \underset{\boldsymbol{\lambda} \in \Gamma}{argmin}\, L(\boldsymbol{\lambda}) \qquad (5)$$

where Γ is the allowed region for $\boldsymbol{\lambda}$ in R^N(N-dimensional Euclidean space). Here, we take Γ as the whole region where $\lambda_i \geq 0$ for all i. If, however, some λ_i is very large, attention will be paid only to small parts of the components in the approximated joint conditional PD. For that reason, we take Γ as a restricted region

$$\Gamma = \{\boldsymbol{\lambda} \mid 0 \leq \lambda_i \leq \Lambda_{\max}, i = 1, 2, \cdots, N\}. \qquad (6)$$

Moreover, restricting the region for $\boldsymbol{\lambda}$ as in (6) is much helpful in seeking $\boldsymbol{\lambda}_{opt}$. Seeking $\boldsymbol{\lambda}_{opt}$ can efficiently be accomplished by a feasible direction method or an active set method, which is suitable for solving constrained optimization problems[8].

III. FRAME-CORRELATED HMM BASED ON ELP

Let S be a set of all states and C be a set of all observation symbols. By ELP, the likelihood of current observation in each state is evaluated as follows.

$$\hat{p}(X=x_t \mid Z=x_{t-1}, s) = \frac{p(X=x_t \mid s)^{\lambda_s}\, p(X=x_t \mid Z=x_{t-1})^{\lambda_c}}{\sum_{y \in C} p(X=y \mid s)^{\lambda_s}\, p(X=y \mid Z=x_{t-1})^{\lambda_c}} \tag{7}$$

where x_t is an observation symbol at time t and s indicates a state in S. The pooling weights, λ_s and λ_c are estimated with given training data by following the approach in Section II. As for λ_s and λ_c, we can estimate them for all possible pairs of state and observation symbol. But, this may give many unseen pairs for which we can not estimate pooling weights. Therefore, it is robust to cluster all the pooling weights into a smaller number of groups.

In (7), $p(X=y \mid s)$ is a component of the traditional output PD in s which has been estimated during the conventional HMM training phase. On the other hand, $p(X=y \mid Z=x_{t-1})$ represents an element of the frame correlation PD. It is generally agreed that the correlation PD highly depends on each phoneme. On this ground, the use of phoneme-dependent correlation PD's is known to be more beneficial than using phoneme-independent ones. If, however, the amount of training data is not so sufficient for supporting well estimated phoneme-dependent correlation PD's, one may have undesirable results during the recognition phase. Moreover, since the number of parameters has a close relationship with robust recognition, it is desirable to use the same correlation PD for all phonemes.

Assume that there are N_p phonemes, $q_1, q_2, \cdots, q_{N_p}$, used as units for recognition. Let $c(x, y \mid q)$ indicate counts of the event in which current observation is y given that the just previous output is x and the current phoneme identity is q. Then, the phoneme specific co-occurrence probabilities for the symbol pairs can be calculated as

$$p(X=x, Z=y \mid q_i) = \frac{c(x, y \mid q_i)}{\sum_{u, v \in C} c(u, v \mid q_i)} \quad \text{for} \quad 1 \le i \le N_p . \tag{8}$$

For deriving phoneme-independent correlation PD's, it is necessary to sum these phoneme specific co-occurrence probabilities with appropriate weights as the following:

$$p(X=x, Z=y; \omega) = \sum_{i=1}^{N_p} \omega_i\, p(X=x, Z=y \mid q_i) \tag{9}$$

where

$$\sum_{i=1}^{N_p} \omega_i = 1 \quad \text{and} \quad \omega_i \ge 0 \quad \text{for} \quad 1 \le i \le N_p . \tag{10}$$

Once (9) has been evaluated for each symbol pair, the correlation PD's are obtained by the normalizing operation.

Although the simplest way to determine the weights in (9) is just using the normalized phoneme counts in the training data, it has a tendency to concentrate on frequently observed phonemes. To compensate for this excessive concentration, one may use uniform weights, which are shown to be better empirically than the normalized phoneme counts. Here, we suggest a new method to obtain phoneme-independent correlation PD's when we are given phoneme specific co-occurrence counts. Our method is based on the principle of maximum entropy, which is one of most frequently adopted strategies whenever one does not have any a priori information about the unseens. By the maximum entropy principle, we estimate the weights such that

$$\omega^o = \underset{\omega}{argmax}\, \mathrm{H}(p(X, Z; \omega)) \tag{11}$$

where

$$\mathrm{H}(p(X, Z; \omega)) = -\sum_{x \in C} \sum_{y \in C} p(X=x, Z=y; \omega) \log p(X=x, Z=y; \omega) . \tag{12}$$

We can follow the procedure similar to that in Section II for reaching ω^o except for the additional equality constraint that the sum of weights is 1. The feasible direction method involving equality constraints is also described in [8].

From Bayesian point of view, we can consider $\hat{p}(X \mid Z=x_{t-1}, s)$ as if it were the a posteriori PD when the a priori is $p(X \mid s)$ with observation x_{t-1} at time $t-1$. Generally, in speech signals, $p(X \mid Z=x_{t-1})$ is with low entropy, which means that it has the shape of high peaks and deep valleys. Thus, the a posteriori, $\hat{p}(X \mid Z=x_{t-1}, s)$ may also resemble this low entropy shape, which shows high possibilities of rejecting unseen data in the recognition phase. We propose a new method to control the contributions of the a priori and the a posteriori PD's, which requires only slight modification to the conventional HMM recognition scheme. Our method is based on the well-known codebook exponents[9]. Let $p_i(X)$ and $p_o(X)$ denote the a priori and the a posteriori PD's, respectively, defined over an observation set C. Then, the combination of the a priori and the a posteriori PD's becomes

$$\hat{p}(X=x; \mu) = p_i(X=x)^{\mu_i} \cdot p_o(X=x)^{\mu_o} \quad , x \in C \tag{13}$$

where $\mu = (\mu_i, \mu_o)$ denotes the exponent of each PD such that

$$\mu_i + \mu_o = 1 \quad , 0 \le \mu_i, \mu_o \le 1 . \tag{14}$$

Suppose that a word w is realized by an output symbol string, $y_1, y_2, \cdots, y_T$ and we concern only the most prominent state sequence when evaluating word probability. Then, given a word model M with the prior-posteriori combination shown in (13) and w, we can find the most dominant state sequence, $s_1, s_2, \cdots, s_T$. With the determined state sequence, the word probability is represented as

$$Pr(w \mid M;\mu) = D \prod_{t=1}^{T} p(X=y_t \mid s_t)^{\mu_1} p(X=y_t \mid Z=y_{t-1}, s_t)^{\mu_2} \quad (15)$$

where D is the state sequence probability depending on s_1, s_2, $\cdots$, s_T and y_0 denotes the initial null output.

The estimation of exponents is based on the MMI criterion which has been widely used for enhancing the discrimination ability of parameters for recognition[9]. Let w denote a word in the training set, U, and M_w be a model corresponding to w. Then, the objective function to be optimized under the MMI criterion is defined by

$$I(\mu) = \prod_{w \in U} \frac{Pr(w \mid M_w;\mu)}{\frac{1}{K}\sum_{i=1}^{K} Pr(w \mid M_{i(w)};\mu)} \quad (16)$$

where $M_{i(w)}$ is a model which gives the ith highest score for w. To maximize $I(\mu)$ with respect to μ, we take the approximated version of Baum-Welch algorithm, which is shown to be effective for optimization with a rational type of objective functions[10].

IV. CONTINUOUS SPEECH RECOGNITION

A. Baseline Recognition System

The vocabulary consists of 102 Korean words representing month, day, date and time. In the vocabulary, there are many confusable word groups in which a word is different from others by only a small number of phonemes. 90 speakers(43 males ar 47 females) uttered 20-30 sentences to construct the database used for training and evaluation. Utterances from 70 speakers(33 males and 37 females) constructed the training data in which there were 1631 sentences and 5122 words, and those from the other 20 speakers were used to form the test data containing 439 sentences and 1448 words. Each utterance was low-pass filtered with a cut-off frequency of 4.5 kHz and digitized with a sampling rate of 16 kHz. We used twelfth-order linear predictive coding(LPC) cepstral coefficients and differenced LPC cepstral coefficients as the feature vectors, and extracted them in every frame of 10 ms. Two separate codebooks were constructed such that the number of codewords is 256 for each codebook.

27 phoneme models involving silence model were used as the basic units of recognition. Each unit was modeled by a three-state discrete HMM which is a simple left-to-right model without skipping. All the HMM parameters were trained according to the maximum likelihood(ML) criterion using the segmental approach. In order to avoid difficulties arising from zero probabilities, we interpolated the trained output PD's with uniform PD. For this, we divided the training data into two blocks so as to keep separate counts on each block, and then carried out deleted interpolation(DI) with five ranges of counts. Word recognition rate of the baseline system without frame correlation was shown to be 73.0 %.

B. Experimental Results for Frame-Correlated HMM

We compared the recognition accuracies of frame-correlated HMM with various types of phoneme-independent frame correlation PD's. First, we set the pooling weights such that $(\lambda_s, \lambda_c) = (1, 1)$ for all the pairs of state and output symbol, which truly is the case of BC HMM. The frame correlation PD's were obtained independently for each parameter set based on the training speech data.

Three different methods for weighting the phoneme specific counts were attempted: natural, uniform and maximum entropy-based weights. Natural weights indicate that we use normalized phoneme counts per frame observed in the training data while uniform weights mean that equal weights are given to all phonemes. Maximum entropy-based weights were derived with the feasible direction method accompanied by an equality constraint as well as by inequality constraints. For convenience, we note the phoneme-independent frame correlation PD's with these weights by $\mathbf{FC}_n$, $\mathbf{FC}_u$ and $\mathbf{FC}_e$, respectively.

Recognition results for the BC HMM are shown in Table I in which the result of baseline system without frame correlation is also shown for the purpose of comparison. Word recognition accuracy of the BC HMM with $\mathbf{FC}_n$ was 72.7 % which was slightly lower than that of the baseline in which frame correlation was not considered. On the other hand, BC HMM with $\mathbf{FC}_u$ and $\mathbf{FC}_e$ reduced the word error rate of the baseline by 3.0 % and 5.9 %, respectively. From these recognition results, we can see that $\mathbf{FC}_e$ outperforms both $\mathbf{FC}_n$ and $\mathbf{FC}_u$. This demonstrates that when incorporating frame correlation PD's, it is needed to flattening the shape while maintaining valuable information inherent in them.

Next, we used different pooling weights for each phoneme. All the experiments were conducted with $\mathbf{FC}_e$ which yielded the best result in the case of BC HMM. Values of the pooling weights were chosen such that they minimized approximation errors of the ELP within the given training data. During the optimization procedure, we set $\Lambda_{max} = 1$, which was later found suitable since optimal pooling weights fell inside the defined region in most of experiments. We denote the derived values of pooling weights as $\lambda(ELP)$.

Table II shows the recognition result with $\lambda(ELP)$. Word recognition rate with $\lambda(ELP)$ was 71.6 % which was lower than that of the BC HMM by 3.0 %. That the recognition performance

was even worse than the baseline without frame correlation is rather surprising. After careful examination, we have found that the performance degradation with $\lambda(ELP)$ are caused by two reasons. One is that derived values for λ_s is usually smaller than λ_c for each phoneme, which may cause losing in discriminating capability of state-specific output PD's. The other is that joint conditional PD's estimated in the training data are too sparse to get reliable pooling weights.

Performance improvements of the frame-correlated HMM with phoneme-dependent pooling weights were achieved in three steps. In the first step, we set $\lambda_s = 1$ and only λ_c was searched for each phoneme. Values for these pooling weights are denoted by $\lambda(ELP1)$. With $\lambda(ELP1)$, the word accuracy was improved to 75.1 % which was higher than that of the BC HMM by 0.5 %. In the second step, we smoothed the estimated joint conditional PD's in the training data. For this, we interpolated each PD with its co-occurrence smoothed one by using the DI technique. Pooling weights were derived based on these smoothed joint conditional PD's, and they are denoted by $\lambda(ELP2)$. Word accuracy with $\lambda(ELP2)$ was 75.7 % which reduced the recognition error of BC HMM by 4.3 %. In the last step, we set the pooling weights by $\lambda(ELP2)$ and applied the priori-posteriori combination mentioned in Section III. Exponents were separately estimated for each output symbol based on the MMI criterion. When applying the MMI criterion, we took only three candidates of highest score into consideration with the help of the N-best search algorithm. $\lambda(ELP2)$ with priori-posteriori combination yielded the word recognition accuracy of 76.7 %, which resulted in reducing the error rate of the baseline without frame correlation by 13.7 %.

V. CONCLUSIONS

In this paper, we proposed a novel method to incorporate frame correlations into a conventional HMM-based recognition system. With the proposed ELP, a joint conditional PD can be expressed in terms of separate conditional PD's associated with respective components. The convexity property of the approximation error function enables to guarantee the existence of global optimum pooling weights, and the feasible direction method is applied to seek them. We also suggested two techniques for application of the ELP to practical word recognition. When constructing the phoneme-independent frame correlation PD's, we introduced a scheme of adding phoneme specific counts with weights that maximize the decided entropy. In addition, we presented a way to combine two kinds of PD's via the use of exponents, which are estimated under the MMI criterion for the purpose of improving the discrimination capability. We evaluated the performances of the frame-correlated HMM through speaker-independent continuous speech recognition experiments, and conclude that the proposed methods are efficient in practical applications.

REFERENCES

[1] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 1857-1869, Dec. 1989.

[2] P. Kenny, M. Lennig, and P. Mermelstein, "A linear predictive HMM for vector-valued observations with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-38, pp. 220-225, Feb. 1990.

[3] C. J. Wellekens, "Explicit correlation in hidden Markov model for speech recognition," in *Proc. of Int. Conf. Acoust., Speech, Signal Processing*, pp. 383-386, 1987.

[4] K. K. Paliwal, "Use of temporal correlation between successive frames in a hidden Markov model based speech recognizer," in *Proc. of Int. Conf. Acoust., Speech, Signal Processing*, pp. 215-218, 1993.

[5] S. Takahashi, T. Matsuoka and K. Shikano, "Phonemic HMM constrained by statistical VQ-code transition," in *Proc. of Int. Conf. Acoust., Speech, Signal Processing*, pp. 553-556, 1992.

[6] C. Genest and J. V. Zidek, "Combining probability distributions: A critique and an annotated bibliography," *Statistical Science*, vol. 1, pp. 114-148, 1986.

[7] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, Mass.: Addison-wesley, 1987.

[8] D. G. Luenberger, *Linear and Nonlinear Programming*. Reading, Mass.: Addison-Wesley, 1984.

[9] Y. Normandin, *Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem, Ph. D. Thesis, McGill University*, 1991.

[10] P. S. Gopalakrishnan et al., "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Trans. Inform. Theory*, vol. IT-37, pp. 107-113, Jan. 1991.

TABLE I
RECOGNITION RESULTS OF BC HMM WITH VARIOUS FRAME CORRELATION PD'S

correlation PD	error rate(%)			accuracy(%)
	INS	DEL	SUB	
FC_a	1.6	0.2	25.5	72.7
FC_b	1.0	0.3	24.9	73.8
FC_c	1.0	0.3	24.1	74.6
baseline	1.0	0.7	25.3	73.0

TABLE II
RECOGNITION RESULTS OF ELP
WITH PHONEME-DEPENDENT POOLING WEIGHTS

weights	error rate(%)			accuracy(%)
	INS	DEL	SUB	
$\lambda(ELP)$	1.0	0.5	26.9	71.6
$\lambda(ELP1)$	1.0	0.3	23.6	75.1
$\lambda(ELP2)$	1.0	0.4	22.9	75.7
$\lambda(ELP2)$ with pri-post combination	1.2	0.4	21.7	76.7