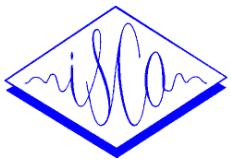

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Speech Recognition Using HMM with Decreased Intra-group Variation in the Temporal Structure

Nobuaki Minematsu and *Keikichi Hirose*
mine@gavo.t.u-tokyo.ac.jp hirose@gavo.t.u-tokyo.ac.jp

Dept. of Electronic Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 Japan

ABSTRACT

A new clustering scheme was proposed for the improvement of HMM-based phoneme recognition with temporal modeling. A precise observation of the temporal correspondences between the training data and their corresponding phoneme HMMs indicated that there were two extreme cases, one with several types of correspondences in a phoneme group that were completely different one from another, and the other with only one type. Although temporal modeling technique was commonly used to incorporate the temporal information in the HMMs, good modeling was not obtained for the former case. To cope with this problem, a new scheme was proposed where the training data for the phoneme of the former case were clustered into several smaller groups. The clustering was conducted so as to reduce the variation in the temporal correspondence in a group. After the clustering, a new HMM was constructed for each divided group. Using the proposed method, speaker dependent recognition experiments were conducted for the phonemes segmented from isolated words. A few-percent increase was observed in the recognition rate, indicating the validity of the proposed method.

1 INTRODUCTION

In terms of temporal structure of speech, HMMs with left-to-right configuration are commonly used for speech recognition. These basic HMMs, however, have the major problem that they cannot, only by themselves, represent adequately the temporal structure of speech. The probability that the transition stays at a state decreases exponentially with time. In order to cope with this problem, duration models have already been introduced where the duration distribution in each state was estimated from the training data. A duration model is usually constructed by estimating the probability (density) function $Pr_i(t_i)$, indicating the probability of the transition staying at state i for the period of t_i. Although this probability function can be obtained in the learning process by the Baum-Welch algorithm[1], it requires a lot of extra-time in computation as compared to just training basic HMMs. Moreover, the probability can be estimated only for discrete points of t_i, viz., integers indicating the number of loops for the self-loop transition at state i. If the training data are limited, as is usually the case, the probability function of discrete variable t_i may include large gaps and can only give us a rough idea about the duration distribution of HMM. Based on these considerations, the probability function was estimated in this paper by the "rematching" method, where the Viterbi paths (best matching paths) between the training data and their corresponding basic HMMs were traced to find out the temporal correspondences. Using the Gaussian or Gamma distribution function as the probability density function (henceforth, pdf), t_i can be treated as a continuous variable. In the recognition process of input speech, following the conventional Viterbi search, the likelihood score was modified by taking $Pr_i(t_i)$ into account[2]. The correlation between the temporal durations of the i-th and j-th states, usually ignored in other methods, can be easily incorporated into the duration model by the rematching method. Therefore, in the present paper, the duration model is represented in the form of joint pdf $Pr(t_1, t_2, .., t_i, .., t_n)$, where n is the number of states with self-loop transitions in an HMM.

Observation of the Viterbi paths for training samples of each phoneme indicated that there were cases with large variations in the temporal structure among samples in terms of the state transition. Certainly, there were many cases with small variations, and, for these cases, $Pr_i(t_i)$ may represent the temporal structure adequately. In the cases of large variations, however, since $Pr_i(t_i)$ is estimated as the average over the samples of a phoneme, it will not give the adequate modeling. To reduce the variations and to realize a better modeling of the temporal structure, a new clustering method was proposed where a phoneme group of training data with large variation was divided into several sub-groups. By constructing HMMs for the sub-groups and introducing duration models, the recognition rate may increase. Phoneme recognition experiments were conducted to prove this prediction.

2 TEMPORAL STRUCTURE IN HMM

2.1 Tracing of Viterbi Path

As mentioned above, an HMM can explicitly model the temporal information of speech by introducing a duration model. When constructing the duration model by the rematching method, the mean value and the variance of t_i for the training data of each phoneme are utilized. Since these parameters are calculated by tracing the

Viterbi paths, the path distributions in a phoneme group will indicate the features of the duration model directly. As already stated, large variations in the path distribution were observed for some phoneme groups. Since these variations cannot be well represented by a single distribution function like a Gaussian function, some preprocessing should be necessary for these phoneme groups. After these considerations, a quantitative analysis was conducted on the Viterbi paths by introducing the looping rate of a state, which was defined as the rate of self-loop transitions for the state to the total number of transitions in a Viterbi path.

2.2 Conditions of Experiment

The **ATR** speech database with labeling information were partly used as the speech material of the current experiment. As listed in **Tab. 1**, it was 5420 words for each of 5 adult male speakers. Sampled at 10 kHz with 16 bit accuracy, the material was segmented into phonemes by trained labelers. Speaker dependent experiments were conducted using the even numbered words for the training and the odd numbered words for the recognition (in section 4). For each speech sample, 16 LPC mel-cepstrum coefficients were first calculated as shown in **Tab. 2**. Then, for phonemes listed in **Tab. 3**, continuous HMMs were constructed by the Baum-Welch method. The number of speech segments used for the training of each phoneme HMM was 300 in almost all the cases, but was less than 300 for several HMMs. Structure of a HMM is schematically shown in the upper-left hand side of **Fig. 1**. After the calculation of these basic HMMs, each training data was rematched with the corresponding basic HMM to obtain the Viterbi path.

2.3 Results and Discussions

Figs. 1 and **2** show the results of analysis for several phonemes respectively for utterances of **MAU** and **MMY**. In these figures, each bar represents the "single occupancy rate (SOR)," which is defined as the rate of speech segments satisfying the following condition to the total number of segments used for the training of a phoneme HMM: self-loop transitions in one of three states occupying more than 70 % of the total transitions in a Viterbi path. Each bar consists of 3 blocks at most, which respectively represent the rates of self-loop transitions at the states 1, 2 and 3. While there are cases where a bar includes 2 or 3 blocks rather evenly, like /u,h/ of **MMY** for the case of two blocks and /g,h/ of **MAU** for the case of three blocks, there are also cases where a bar is almost occupied by one block, like /s/ of **MAU** and /sh/ of **MMY**.

When the SOR is low or when it is high and the dominant self-loop transitions occur only at a definite state, the variation should be small in the temporal correspondence between the training data and their corresponding HMM. In these cases, adequate modeling of the temporal structure is possible for a phoneme HMM by the duration model. When the SOR is high and the dominant self-loop transitions occur at one of three states rather randomly for each sample, however, the duration model cannot well represent the temporal structure. In these cases, a phoneme HMM is considered to be a composition of two or three distinct sub-models, and the duration model only represents their averaged features. **Fig. 3** schematically indicates the case of three sub-models, in each of which the transitions in the Viterbi path mostly occur in a definite state.

Tab. 1 Speech material used in the experiment.

Speaker	5 male adults. (**MAU,MHT,MTK,MMY,MMS**)
For training	300 or less segments for each phoneme from even numbered words in 5240 words.
For recognition	300 or less segments for each phoneme from odd numbered words in 5240 words.

Tab. 2 Condition of the experiment.

Acoustic feature	16 LPC mel-cepstrum coefficients.
Condition of analysis	16 LPC analysis by 25.6 msec Hamming window and 5 msec frame shift after 10 kHz sampling with 16 bit accuracy.
HMM	4 states left-to-right HMM with a single Gaussian distribution using a full covariance matrix for each transition.

Tab. 3 26 phonemes prepared for the experiment.
/a,i,u,e,o,p,t,k,ch,ts,b,d,g,s,sh,h,z, dj,m,n,r,w,y,N,Q,j/

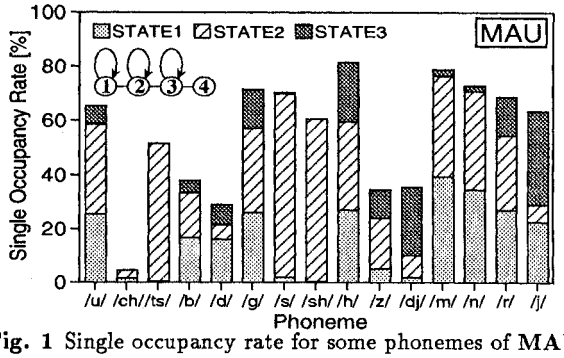

Fig. 1 Single occupancy rate for some phonemes of **MAU**.

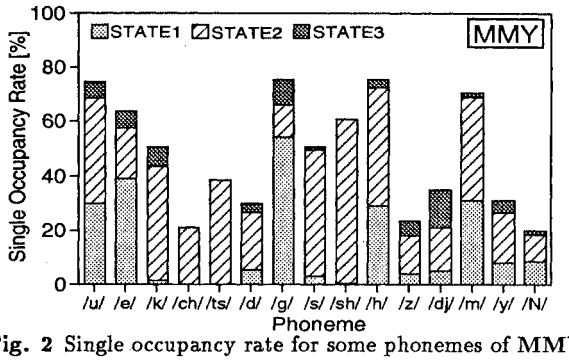

Fig. 2 Single occupancy rate for some phonemes of **MMY**.

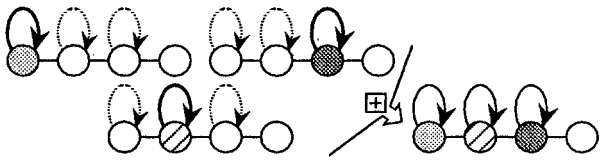

Fig. 3 HMM as the composition of 3 sub-models.

These results indicate that the temporal correspondences between the training data and their corresponding HMMs may sometimes include large variations even for a single phoneme. For these cases, the duration models will not correctly represent the temporal structure of speech. This incorrect modeling may cause the recognition errors and an additional processing is necessary to cope with the variations in the temporal correspondences. From this point of view, a new method was proposed to reduce the temporal variations and, therefore, to increase the recognition performance, as is discussed in the following sections.

3 DECREASE OF INTRA-GROUP TEMPORAL VARIATION

Although the variation among training data of a phoneme group was already discussed in several works, most of them were basically related to the distribution of the spectral features but not the temporal features of speech. In these works, the training data of a phoneme group were often clustered into those of several smaller groups[3], but the clustering was conducted so as to reduce the variation in spectral domain. Therefore, the training data of the divided group may still include variation in the temporal correspondence. The variation in temporal domain should be taken into account especially when introducing the duration model to the basic HMM. Based on these considerations, the issue of temporal variation was dealt with directly in the proposed method. A clustering scheme was also adopted in the method to divide a phoneme group into sub-groups, but the criterion of the clustering was the reduction of variation in temporal domain. Concretely, the basic HMMs were first constructed for phonemes and were then utilized for the clustering of the training data.

Appropriate indices or thresholds should be set up for the good clustering. In the current method, the rematching method was adopted to calculate indices representing the degree of temporal distortion among the training data. As schematically shown in **Fig. 4**, the clustering was conducted with the following procedures:

1. Construct basic HMMs by the Baum-Welch method.
2. Calculate looping rate $\theta(i)$ for each state of each training sample by the rematching method. The rate $\theta(i)$ represents the rate of self-loop transitions at state i to the total number of transitions in a Viterbi path.
3. Classify the samples satisfying the condition of $\theta(i)>\varepsilon$ into a group candidate, which will be referred to as the state i group, henceforth. The threshold ε should be larger than 0.5 to avoid a sample belonging to two state groups. With this procedure, a phoneme group will be divided into $n+1$ group candidates, viz., n state groups and another group with samples not satisfying the condition of $\theta(i)>\varepsilon$ for any i.
4. Calculate $\lambda(i)$ for i from 1 to n, which is the rate of the training data classified to the state i group to the total number of data of the original phoneme group.
5. For each phoneme group, select the "sub-groups" satisfying the condition of $\alpha<\lambda(i)<\beta$ from $n+1$ group candidates obtained by the procedure 3.

Although the repetition of above clustering is theoretically possible, it was not conducted for the current experiment taking the limited size of the training data into account. After the clustering, an HMM was constructed for each sub-group.

The temporal variation of a phoneme group can be represented quantitatively by the rate of the training data included in the state groups selected as sub-groups, to the total number of data in the original group, which shall be called "temporal variation rate (TVR)." The TVR can also be defined for each sub-group as an index for its temporal variation by repeating the above procedures again. **Fig. 5** shows TVRs before and after the clustering, averaged over all the phoneme groups for each speaker. The left-hand side bars indicate the case of $(\varepsilon,\alpha,\beta)=(0.7,0.3,0.7)$, while the right-hand side bars indicate the case of $(0.7,0.25,0.75)$. As shown in the figure, the clustering can effectively reduce the temporal variations in phoneme groups. The speaker variation in TVR was found to be reduced also by the clustering.

4 RECOGNITION EXPERIMENTS

The discussion in section **2.3** implies that the reduced intra-group temporal variation increases the validity of the duration model for the HMM-based recognition. Therefore, a prediction comes up from the results in section **3** that the clustering based on the variation in temporal domain may increase the recognition rate. To prove this prediction, recognition experiments were conducted for the phoneme segments from the odd numbered word utterances, shown in **Tab. 1**.

4.1 Duration Model

As already mentioned in section **1**, the duration model was represented as joint probability density using a multivariate pdf to treat t_i as a continuous variable and to incorporate the correlation between t_i and t_j into the model. For the univariate case, comparison was already conducted between Gaussian and Gamma

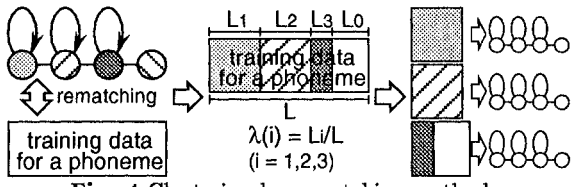

Fig. 4 Clustering by rematching method.

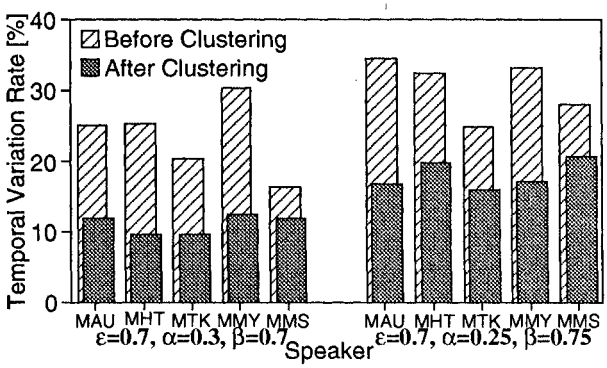

Fig. 5 Temporal variation rate before and after clustering.

distribution functions as pdf, and a better result was obtained for the latter function[4]. For the current study, however, taking the following two points into account, the multivariate Gaussian pdf was adopted. One is the complexity in the estimation of population parameters of Gamma distribution function[5] and the other is the aim of the study being just to indicate the advantage of duration model with a joint pdf. In the rematching process, a Viterbi path was traced to produce a duration vector of order n, whose i-th element was t_i. The mean vector and the covariance matrix were then calculated for a phoneme group or a sub-group using duration vectors of all the samples of the group.

4.2 Conditions of Experiment

The above experiments were performed under almost the same conditions as indicated in section **2**. Recognition experiments were conducted for the following 5 cases using phoneme segments from the odd numbered word utterances (see **Tab. 1**):

CASE 1 Using the basic HMM.
CASE 2 Using the basic HMM and its duration model with a *diagonal* covariance matrix.
CASE 3 Using the basic HMM and its duration model with a *full* covariance matrix.
CASE 4 Using the HMM after clustering.
CASE 5 Using the HMM after clustering and its duration model with a full covariance matrix.

The experiments were conducted with the following assumption: If the decrease of intra-group temporal variation has some positive effects on the duration model, the improvement of the recognition rate from CASEs 4 to 5 should be larger than that from CASEs 1 to 3.

4.3 Results and Discussions

Fig. 6 shows the recognition rates of 5 cases for each speaker. The clustering was conducted with the condition of $(\varepsilon,\alpha,\beta)$=(0.7,0.3,0.7), and, after the clustering, around 40 HMMs were constructed for 26 phonemes of each speaker. In the figure, the parenthesized number below each speaker's label indicates the number of HMMs constructed after the clustering. For phonemes with plural HMMs, recognition was conducted with the similar process as in the case of multi-template recognition. The score of CASE 3 exceeds that of CASE 2 for every speaker, indicating the validity of the introduction of a multivariate pdf to the duration model. Increase in the recognition rate from CASE 1 to CASE 4 indicates that the proposed clustering method is valid for the improvement of recognition performance even without duration models. The recognition rate was further increased by the introduction of the duration model as indicated in the score for CASE 5. The numbers on the top of the bars for CASE 3 and for CASE 5 of each speaker respectively indicate the increase of recognition rate by the introduction of duration models to the basic HMMs (CASE 3−CASE 1) and that to the HMMs after the clustering (CASE 5−CASE 4). As shown in the figure, a larger increase was obtained for the HMMs after the clustering for each speaker. **Fig. 7** shows the error reduction rates

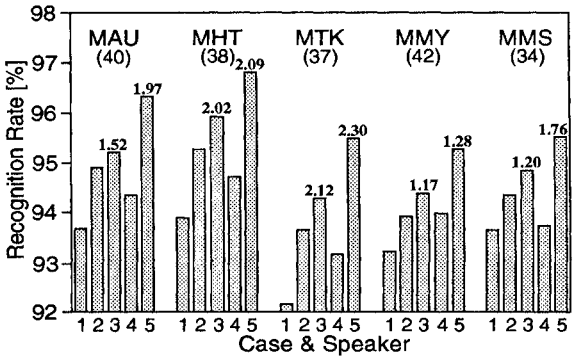

Fig. 6 Recognition rates in 5 cases for each speaker.

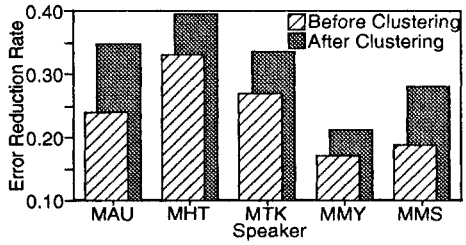

Fig. 7 Error reduction rate before and after clustering.

by the duration model before and after the clustering. These results coincide with the assumption, indicating the validity of the clustering scheme based on the intra-group temporal variation.

5 CONCLUSIONS

Variations in the temporal correspondences between the training data and their basic HMMs were discussed based on the observation of the Viterbi paths. For several phonemes, it was found that there were large temporal variations even among the training data for a single phoneme. Consequently, the duration model from these data could not model their temporal structure adequately. In order to reduce the variation, a new method was proposed, where the training data with large intra-group variation were clustered into sub-groups. The results of phoneme recognition experiments showed the validity of the method. Optimization of the method, including the introduction of Gamma pdf, will increase the performance. This is now under the experiment.

REFERENCES

[1] Y.Hashimoto, *et al.*, "*Investigation on Segmentation of Continuous Speech Using Hidden Markov Model*," Reports of Spring Meet. Acoust. Soc. Jpn., 3-P-2, pp.231-232, 1988.

[2] L.R.Rabiner, *et al.*, "*Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities*," AT&T. Tech. J., 64, 6, pp.1211-1231, 1985.

[3] S.Sagayama, "*Phonemic Environment Clustering, Principle and Algorithm*," IEICE Technical Report, SP87-86, pp.1-6, 1987.

[4] Y.Takada, *et al.*, "*A Comparative Study on Duration Models*," Reports of Autumn Meet. Acoust. Soc. Jpn., 2-5-14, pp.75-76, 1991.

[5] S.E.Levinson, "*Continuous Variable Duration Hidden Markov Models for Automatic Speech Recognition*," Computer Speech and Language, No.1, pp.29-45, 1986.