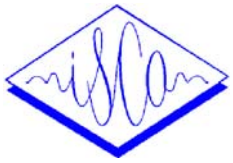

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

MANDARIN TELEPHONE SPEECH RECOGNITION USING MCE/GPD-BASED SPEAKER CLUSTER HMM

Sen-Chia Chang, Shih-Chieh Chien and Woei-Chyang Shieh
E000/CCL Industrial Technology Research Institute
Chutung, Hsinchu, Taiwan 310
E-mail:chang@ccl.itri.org.tw

10.21437/Eurospeech.1999-595

ABSTRACT

In this paper we successfully apply the MCE/GPD method to train speaker cluster HMM. The essential concept of our approach is to adjust all the parameters of the speaker cluster HMM simultaneously using each utterance of the whole training set. In other words, the parameters of each cluster-dependent HMM are no longer independently estimated by using only the training data of the speakers who belong to its corresponding cluster. To achieve this purpose, the discriminant function used in the MCE/GPD method need to be defined by the parameter set of the entire speaker cluster HMM. In our implementation, we define it as a function of the log-likelihood scores given the cluster-dependent HMMs. The proposed discriminative training procedure would increase the cluster separability and then improve the recognition rate.

Keywords: speaker cluster HMM, MCE/GPD

1. INTRODUCTION

It is well known that between-speaker variability will cause the recognition errors of speech recognition systems to increase. For example, a well-trained speaker-dependent speech recognition system always outperforms a speaker-independent (SI) speech recognition system. Using more detailed acoustic models is a promising way to achieve a higher recognition rate in SI speech recognition. The speaker cluster HMM (SC-HMM), which is able to accommodate between-speaker variations, have been shown to produce good performance. For example, the gender-dependent HMM (GD-HMM) [1,2] is a simple but successful implementation of the SC-HMM.

Clustering all speakers in the training set into clusters, and then training cluster-dependent HMMs for each speaker cluster independently by maximum likelihood (ML) estimation is a usually employed procedure for SC-HMM design. Here, we call this training procedure as the conventional approach. In pass years, most researches were interested in the first step of the conventional approach, that is, pre-clustering training speakers into clusters. An essential problem related to that is to find similarities across different speakers. A tree-structured speaker clustering method based on similarities across speakers defined by acoustic distances was proposed in [3]. In [4], a simple cluster tree was created according to three classes of speaking rate – fast, medium and slow. Natio et al. [5] used the vocal-tract-size related articulatory parameters associated with individual speakers to cluster speakers.

In this paper, we integrate the speaker clustering process and the model estimation process into a unified framework. It is realized by using the MCE/GPD (**M**inimum **C**lassification **E**rror / **G**eneralized **P**robabilistic **D**escent) [6][7] method with an objective function that is designed to minimize classification error. The essential concept of our approach is to train all of the adjustable parameters of the SC-HMM simultaneously using the whole training data. In other words, the parameters of each cluster-dependent HMM are no longer independently estimated by using only the training data of the speakers who belong to its corresponding cluster. To achieve this purpose, the discriminant function used in the MCE/GPD method need to be defined by the parameter set of the entire SC-HMM. In our implementation, we define it as a function of the log-likelihood scores given the cluster-dependent HMMs.

This paper is organised as follows. In section 2, training speaker cluster HMM based on the MCE/GPD method is described in detailed. In section 3, a Mandarin polysyllabic word recognition task is performed to evaluate the proposed SC-HMM training method. Finally, a brief summary and the future work are given in section 4.

2. MCE/GPD-BASED SPEAKER CLUSTER HMM

Before integrating the speaker clustering process and the model estimation process into a unified framework through the MCE/GPD method, the SC-HMM trained by the conventional approach is used as

the initialization model for the following discriminative training procedure. A hierarchical tree-structured speaker clustering algorithm is employed [3]. The hierarchical tree cluster organization is illustrated in Figure 1. The first level (root) of this tree is speaker independent. In the second level, speakers are clustered by gender. A Gaussian log-likelihood is used as a distance measure for clustering speakers within each gender group [8]. There are four detailed clusters at the leaves of the tree. For each speaker cluster, a cluster-dependent HMM is trained based on ML estimation using speech data of the speakers who belong to the corresponding cluster.

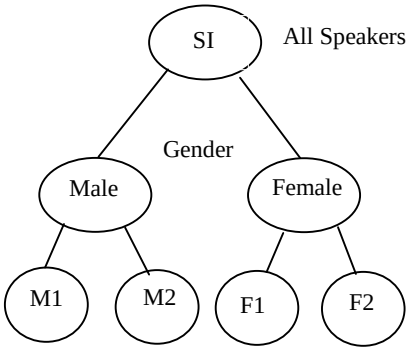

Figure 1. Hierarchical Cluster Tree Organization

2.1 Discriminant Function

When applying the MCE/GPD method to train an HMM-based speech recognizer, it is necessary to define a set of discriminant functions. If the parameters of each cluster-dependent HMM are adjusted independently by the MCE/GPD method, the discriminant function can be defined as the log-likelihood function. Since we hope to adjust all the parameters of the entire SC-HMM simultaneously using the whole training data, the discriminant functions have to be defined by the parameter set of the entire SC-HMM.

Assuming that the SC-HMM is composed of R cluster-dependent HMMs, and the parameter set for each of the cluster-dependent HMMs is denoted as Λ_r, r = 1, 2, … , R. The discriminant function, which is defined by the entire parameter set $\Gamma = \{\Lambda_1, \Lambda_2, \ldots, \Lambda_R\}$, has the following form

$$g_i(X;\Gamma) = \log\left[\frac{1}{R}\sum_{r=1}^{R} w_r(X)\exp\left[h_i(X;\Lambda_r)\eta\right]\right]^{\frac{1}{\eta}}, \quad i = 1, 2, \ldots, M \qquad (1)$$

where X is an observation sequence and M is the number of classes (words) to be classified. $h_i(X;\Lambda_r)$ is the log-likelihood function defined only on the parameter set Λ_r. η is a positive weighting number. $w_r(X)$ is a cluster weighting function that indicates the degree to which X belongs to the rth speaker cluster. In this paper, we simply set $w_r(\bullet)$ to be a zero-one function and use it to indicate which Λ_r will be updated when X is given. For example, if we want to update all the parameters of the entire SC-HMM using the whole training data, all w_r will be assigned to be 1. On the other hand, if we only want to adjust the parameters within male sub-tree when X is uttered by a male speaker, w_r will be set to be 1 for male part and to be 0 for the others.

2.2 Generalized Score Function

The discriminant function defined by equation (1) could be regarded as a generalized score function of the SC-HMM based speech recognition system. With this definition, the speech recognition system is operating under the following decision rule:

$$C(X) = C_i \text{ if } g_i(X;\Gamma) = \max_j g_j(X;\Gamma) \qquad (2)$$

There are two usually employed decision strategies in the case of using SC-HMM for SI speech recognition. The first one is based on pre-selecting the cluster where the testing speaker belongs. Only the corresponding Λ_r of the selected cluster is adopted for performing recognition process. In the second strategy, one candidate is obtained independently for each speaker cluster at first, and then the candidate with the best score is chosen as the final recognition result. If we set $w_r(X)$ to be 1 and $w_{s,s\neq r}$ to be 0, the decision rule will be

$$C(X) = C_i \quad \text{if} \quad g_i(X;\Gamma) = \max_j h_j(X;\Lambda_r) \qquad (3)$$

It is a mathematical form of the first decision strategy. When $\eta \to \infty$ and w_r is set to be 1 for all r, $g_i(X;\Gamma) = \max_r h_i(X;\Lambda_r)$. The decision rule will have the form:

$$C(X) = C_i \text{ if } \quad g_i(X;\Gamma) = \max_{r,j} h_j(X;\Lambda_r) \qquad (4)$$

In this case, our decision strategy is equivalent to the second decision strategy mentioned above. Because the decision rule defined by equation (2) is a general form of the conventional decision strategies, we hope better recognition rates may be achieved when some different η and w_r are used.

3. EXPERIMENTS

Effectiveness of the proposed SC-HMM training method is examined by simulations on a Mandarin polysyllabic word recognition task. MAT speech database is used in our experiments.

3.1 Database and Features

MAT (**M**andarin speech data **A**cross **T**aiwan) speech database [9] was collected at eight recording stations in Taiwan through telephone networks. The speakers were chosen to reflect the population of the gender, the dialect, the educational level, and the residence in Taiwan. A subset of MAT speech database, MATDB-4, is used in the following experiments. The vocabulary size is 1062. The length of each word is ranging from two to four syllables. 14886 utterances spoken by 560 speakers (292 male speakers and 268 female speakers) are assigned for training. The testing data contains 3697 utterances spoken by 140 speakers (70 male speakers and 70 female speakers). All speech signals were sampled at a rate of 8 kHz and preemphasized with a digital filter, $1-0.95z^{-1}$. It was then analyzed for each Hamming-windowed frame of 20 ms with 10 ms frame shift. The recognition features consist of 12 mel-cepstral coefficients, 12 delta mel-cepstral coefficients, the delta energy, and the delta-delta energy. Cepstral mean normalization [10] is employed to remove the telephone channel effects on an utterance-by-utterance basis.

3.2 Sub-syllable Models

We employ 138 sub-syllable models, including 100 3-state right-context-dependent INITIAL HMMs and 38 5-state context-independent FINAL HMMs, as basic recognition units for each cluster-dependent system. The observation distribution for each state of the HMM is modeled by a multivariate Gaussian mixture distribution. The number of mixture components in each state varies from 1 to 16 depending on the amount of training data, and each of the mixture components has a diagonal covariance matrix. For silence, a single-state model with 16 mixtures is used.

3.3 Experimental Results

In our experiments the decision rule is based on equation (2), that is, the discriminant function defined by equation (1) is used as the score function. In recognition phase, the value of w_r is simply set to be 1 for each employed speaker cluster. Three different η values ($\eta=1$, $\eta=2$ and $\eta\to\infty$) are used in all experiments.

At first, experiments using SI models trained with ML (label as "SI-ML") and MCE (label as "SI-MCE") are performed and taken as baseline results. The recognition error rates are displayed in Table 1. The effectiveness of the MCE/GPD method is shown. Since there is only one speaker cluster in the case of using SI-HMM, there is no difference on performance when different η being used.

To examine the performance of the SC-HMM, three experiments are run. They contain the use of: (1) GD-HMM (label as "GD"); (2) SI-HMM and GD-HMM (label as "SI+GD"); (3) SI-HMM, GD-HMM, and 4 detailed acoustic models (label as "SI+GD+leaf(4)-1"). Here, the parameters of each cluster-dependent HMM are updated with MCE independently by using only the training data from the speakers who belong to its cluster. Their recognition error rates are summarized in Table 1. It shows that the SC-HMM always outperforms the SI-HMM. We also find that the best recognition result is achieved when $\eta=2$. It means the generalized score function defined in equation (1) is more suitable for the SC-HMM based speech recognition system.

In the following experiments, the proposed MCE/GPD-based SC-HMM training algorithm is applied to further refine the models estimated with ML. Because the acoustic difference between genders is evident, we only adapt the parameters of cluster-dependent HMMs within a gender group simultaneously with the MCE/GPD method. The recognition result is displayed in Table 1 (label as "SI+GD+leaf(4)-2"). From Table 1, we can see that our proposed training method outperforms the conventional one. The recognition error rate drops to 6.7% in the case where 7 speaker clusters are used.

System	$\eta=1$	$\eta=2$	$\eta\to\infty$
SI-ML	10.2	10.2	10.2
SI-MCE	8.8	8.8	8.8
GD	7.8	7.6	7.7
SI+GD	7.4	7.3	7.5
SI+GD+leaf(4)-1	7.3	7.2	7.6
SI+GD+leaf(4)-2	7.1	6.7	7.4

Table 1: Recognition Error Rates using SC-HMMs (%)

4. SUMMARY

The SC-HMM that enables accommodate between-speaker variations is a promising way to achieve a higher recognition rate in SI speech recognition. In this paper, the MCE/GPD method is applied to train SC-HMM. In our approach, all the parameters of the SC-HMM are simultaneously adjusted by using the whole training set. In other words, the parameters of each cluster-dependent HMM are no longer independently estimated by using only the training data of the speakers who belong to its corresponding cluster. To achieve this purpose, the discriminant function used in the MCE/GPD training method need to be defined by the parameter set of the entire SC-HMM. In our implementation, we define it as a function of the log-likelihood scores given the cluster-dependent HMMs. The effectiveness of our porposed SC-HMM training method has been shown in Mandarin polysyllabic word recognition task. A recognition rate of 93.3% is achieved in the case where 7 clusters are used.

In this paper, we simply set $w_r(\bullet)$ to be a zero-one function. In the future, we will investgate its effect on speech recognition when a continuous function is used.

5. ACKNOWLEDGMENT

This paper is a partial result of the project No. 3P11100 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, R.O.C.

The authors would like to thank the Association for Computational Linguistics and Chinese Language Processing in Taiwan for kindly supplying the database.

6. REFERENCES

[1] P. C. Woodland, C. J. Leggetter, J. J. Odell, V. Valtchew, and S. J. Young, "The 1994 HTK large vocabulary speech recognition system", *Proc. of ICASSP'95*, vol. 1, pp. 73-76, 1995.

[2] X. Huang, K. F. Lee, H. W. Hon and M. Y. Hwang, "Improved acoustic modeling with the SPHINX speech recognition system", *Proc. of ICASSP'91*, pp.345-348, 1991.

[3] T. Kosada and S. Sagayama, "Tree-structured speaker clustering for fast speaker adaptation", *Proc. of ICASSP'94*, vol. 1, pp.245248, 1994.

[4] T. J. Hazen and J. R. Glass, "A comparison of novel techniques for instantaneous speaker adaptation", *Proc. of EUROSPEECH'97*, pp. 2047-2050, 1997.

[5] M. Naito, L. Deng and Y. Sagisaka, "Speaker clustering for speech recognition using the parameters characterizing vocal-tract dimensions", *Proc. of ICASSP'98*, pp. 981-984, 1998.

[6] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. on Signal Processing*, SP-40, no. 12, pp. 3043-3054, 1992.

[7] W. Chou, C.-H. Lee and B.-H. Juang, "Segmental GPD training of an hidden Markov model based speech recognizer," *Proc. of ICASSP-92*, pp. 473-476, 1992.

[8] Y. Gao, M. Padmanabhan and M. Picheny, "Speaker Adaptation based on pre-clustering training speakers", *Proc. EUROSPEECH'97*, pp. 2091-2094, 1997.

[9] H.-C. Wang, "***MAT*** – A project to collect Mandarin speech data through telephone networks in Taiwan", *Computational Linguistics and Chinese Language Processing*, vol. 2, no. 1, pp. 73-90, 1997.

[10] F.-H. Liu, R. M. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition", *Proc. of Human Laguage Technology*, pp. 69-74, 1993.