An HMM/MLP Hybrid Approach for Improving Discrimination in Speech Recognition

Kyungmin Na and Soo-Ik Chae
School of Electrical Engineering, Seoul National University
San 56-1, Shinlim-dong, Kwanak-gu, Seoul 151-742, Korea
nkm@cheongon.snu.ac.kr, chae@sdgroup.snu.ac.kr

Abstract

In this paper, we propose an HMM/MLP hybrid scheme for achieving high discrimination in speech recognition. In the conventional hybrid approaches, an MLP is trained as a distribution estimator or as a VQ labeler, and the HMMs perform recognition using the output of the MLP. In the proposed method, to the contrary, HMMs generate a new feature vector of a fixed dimension by concatenating their state log-likelihoods, and an MLP discriminator performs recognition by using this new feature vector as an input. The proposed method was tested on the nine American E-set letters from the ISOLET database of the OGI. For comparison, a weighted HMM (WHMM) algorithm and GPD-based WHMM algorithm which use an adaptively-trained linear discriminator were also tested. In most cases, the recognition rates on the closed-test and open-test sets of the proposed method were higher than those of the conventional methods.

1. Introduction

Over the past few decades, the hidden Markov modeling (HMM) of a speech signal has become the prevailing approach in speech recognition because of its capability of modeling spectral and temporal variations in the speech signal [1]. However, the conventional *maximum likelihood* (ML) estimation method (*Baum-Welch algorithm*) often leads to poor discrimination for some tasks such as the recognition of the nine American E-set letters [2]-[7]. This is because there exists a mismatch between a true speech signal distribution and the chosen HMM, and the amount of available training data is always limited. Therefore, it is desirable to incorporate discrimination-based approaches into the conventional HMM frameworks [3].

Recently, there has been considerable effort in combining HMM with multilayer perceptron (MLP) [9]-[11] because it is proven that a properly-trained MLP becomes a Bayes optimal classifier, and thus has a powerful pattern discrimination ability. In these previous MLP/HMM hybrid systems, an MLP is trained as the *a posteriori* probability estimator for calculating state observation probability values [9], or as a vector quantization (VQ) labeler for generating an observation symbol sequence [10], and the HMMs perform recognition based on the output of the MLP as shown in Fig. 1. However, they still suffer from poor discrimination because the HMM model parameters of the hybrid system are estimated in the ML sense.

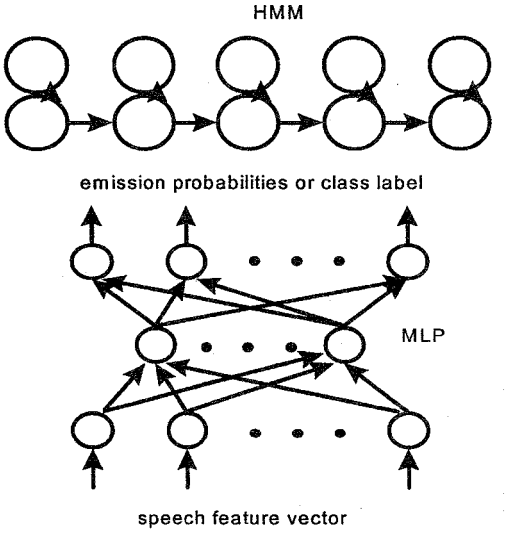

Figure 1. The conventional MLP/HMM hybrid scheme.

Therefore, we propose an HMM/MLP hybrid method where an MLP discriminator performs recognition based on a new feature vector obtained by concatenating the state log-likelihoods of HMMs. The structure of the proposed scheme is depicted in Fig. 2. Su and Lee showed that there is rich discrimination information in the

0-7803-4859-1/98 $10.00©1998 IEEE

concatenation of the state log-likelihoods of HMMs [3]. They proposed a modified adaptive learning procedure to obtain a set of linear discriminant functions. Afterwards, several authors proposed state-weighting approaches based on the generalized probabilistic descent (GPD) algorithm [12]-[13], or the use of an MLP [14]. In [14], an MLP is trained for phoneme classification, and then the output values of the MLP are used as the state-weights, which is different from the proposed scheme. The proposed method can utilize discrimination power of an MLP and rich discrimination information in the concatenation of the state log-likelihoods.

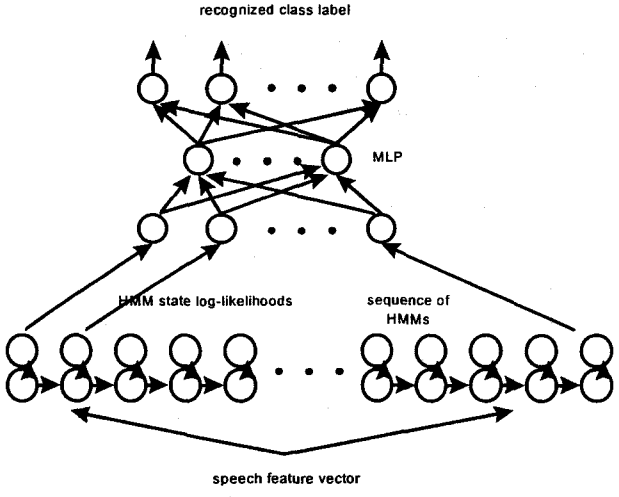

Figure 2. The proposed HMM/MLP hybrid scheme.

In the following Section, we describe the features of the proposed HMM/MLP hybrid scheme. In Section 3, experimental results on the nine American E-set letters from the ISOLET database [15] of OGI (Oregon Graduate Institute) are summarized. Finally, conclusions are presented in Section 4.

2. Proposed Hybrid Scheme

The proposed HMM/MLP hybrid system consists of 5-state left-to-right discrete HMMs and an MLP discriminator. For the HMMs, a VQ codebook with 128 entries is generated using the LBG algorithm, and each HMM is trained by the conventional Baum-Welch algorithm. For the MLP discriminator, a typical 3-layer perceptron is adopted because it is empirically granted that the use of 1 hidden layer is enough in most real applications though the number of hidden layers should be determined according to the complexity of a given problem.

In this 3-layer perceptron, the number of input neurons becomes the product of the number of states in an HMM and the number of classes to be classified, and thus becomes 45. The number of output neurons is the same with the number of classes to be classified, and thus is 9.

However, there is no analytical method in selecting the optimal number of hidden neurons up to date. Thus, in many previous applications, the number of hidden neurons has been empirically chosen to be large enough by rule of thumb after several trial-and-errors. We also determine it as 45 after several trials over different training conditions.

To train this 3-layer perceptron, we adopt the on-line error backpropagation algorithm where the error gradient with respect to the connection weight vector is estimated pattern by pattern. To reduce the possibility of getting stuck in a local minimum and to accelerate the convergence speed, a random pattern presentation scheme is used. The connection weights are initialized with small random values between -0.5 and +0.5.

3. Experimental Results

We evaluated this system on the nine American E-set letters {b, c, d, e, g, p, t, v, z} extracted from the ISOLET spoken letter database of the OGI [15]. The ISOLET database contains the letters of the English alphabet spoken in isolation by 150 speakers. Each speaker uttered each of the letters twice. Among them, the first recordings from 120 speakers (TR set) were used for training, and the second recordings from the same 120 speakers (CT set) were used for closed-test. The remaining recordings from 30 speakers (OT set) were used for open-test. The baseline performances of the discrete HMM-based system trained by the Baum-Welch algorithm were 76.9 % for the TR set, 62.5 % for the CT set, and 60.4 % for the OT set.

The speech signal is processed by using an eighth order LPC (linear predictive coding) analysis, with a 45 ms Hamming window and a 15 ms shift. A 24-dimensional feature vector, consisting of 12 LPC-driven liftered cepstral coefficients and 12 delta cepstral coefficients [3], is applied to an HMM-based system at every frame, and the HMM-system generates a 45-dimensional input vector at the end of frames through Viterbi decoding. Note that the 45-dimensional vector is obtained by concatenating the state log-likelihoods from 9 HMMs with 5-states [3]. The input vector to the 3 layer perceptron is divided by a proper positive constant (100.0 in this work) for limiting its dynamic range.

In training the 3-layer perceptron, the magnitude of the learning rate should be chosen carefully after several trials since the convergence speed and the final performance depends on it. In these experiments, the learning rates of 0.01, 0.05, 0.1, 0.2, and 0.3 were tried, and we found that the performances were similar. The

maximum number of iterations was limited to 300. Since the on-line error backpropagation algorithm is used in this simulation, there is a fluctuation in the recognition rates on the TR, CT and OT sets as shown in Fig. 3. The training criterion is the conventional mean square error (MSE) criterion.

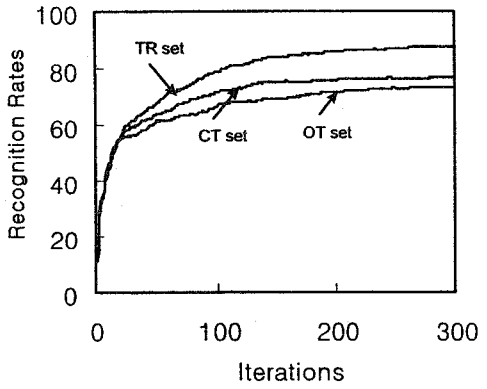

Figure 3. The recognition rate as a function of the number of iterations in the proposed method (learning rate = 0.05).

For comparison, a weighted HMM (WHMM)-based system using a modified adaptive learning procedure was implemented and tested. Since the GPD algorithm is an extended version of the adaptive learning procedure, we also implemented the GPD-WHMM system, and tested it. Note that the GPD-WHMM approach is a little bit different from those in [12] and [13] where only state-weighting functions are incorporated into the HMM-based system. The GPD-WHMM system results in a set of linear discrimination functions instead of the state-weighting functions.

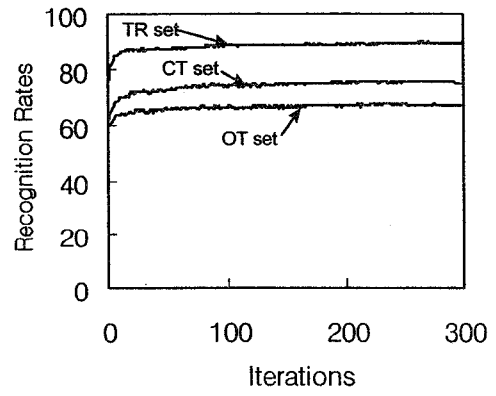

Figure 4. The recognition rate as a function of the number of iterations in the WHMM approach (Su and Lee's).

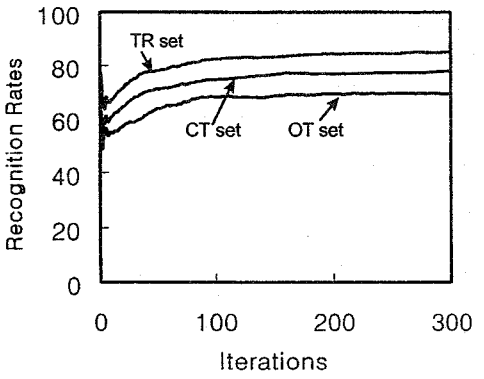

Figure 5. The recognition rate as a function of the number of iterations in the GPD-WHMM approach.

A series of experiments for these two approaches were performed under various different training conditions. Typical curves on the recognition rate evolution of these two methods were given in Fig. 4 and Fig. 5, respectively.

The overall performance comparison is summarized in Table 1. The results in Table 1 are the best ones on the OT set for each algorithm, and other results under various different conditions varied slightly in comparison with these ones. In most cases, the recognition rates on the CT and OT sets of the proposed method were higher than those of the WHMM and GPD-WHMM approaches. In addition, the results indicated that the performance on the OT set of the proposed scheme was better than those of the GPD-WHMM algorithm while the performances on the CT set of both methods were nearly same.

Table 1. Performance comparison.

Algorithm	CT set	OT set
Baum-Welch	62.5 %	60.4 %
WHMM	75.9 %	69.8 %
GPD-WHMM	77.2 %	70.2 %
Proposed	77.1 %	73.7 %

4. Conclusions

In this paper, an HMM/MLP hybrid scheme was proposed to improve the discriminatory power of the conventional HMM-based speech recognizer. The proposed method uses an MLP discriminator whose input vector becomes the concatenation of the state log-likelihoods of HMMs. Because of the MLP's discrimination power and rich discriminative information in the HMM state log-likelihoods, the proposed scheme performs better than other similar approaches such as the WHMM and the GPD-WHMM algorithms. However,

compared to these two similar approaches, its computational burden increases since it requires larger number of weights and additional nonlinear activation function calculations.

Currently, the proposed method is applied to an isolated speech recognition. However, in near future, we will develop the way of applying this method to connected word and continuous speech recognition.

References

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 72, no. 2, pp. 257-286, 1989.

[2] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 3, pp. 257-265, 1997.

[3] K.-Y. Su and C.-H. Lee, "Speech recognition using weighted HMM and subspace projection approaches," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 1, Part I, pp. 69-79, Jan. 1994.

[4] S. Katagiri, C.-H. Lee, and B.-H. Juang, "New discriminative training algorithm based on the generalized probabilistic descent method," *Proc. 1991 IEEE Workshop Neural Networks for Signal Processing*, pp. 299-308, 1991.

[5] P.-C. Chang and B.-H. Juang, "Discriminative training of dynamic programming based speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 2, pp. 135-143, 1993.

[6] W. Chou, B.-H. Juang, and C.-H. Lee, "Segmental GPD training of HMM based speech recognizer," *Proc. IEEE Internat. Conf. Acoustic. Speech Signal Process.*, pp. 473-476, 1992.

[7] W. Chou, B.-H. Juang, and C.-H. Lee, "Minimum error rate training based on N-best string models," *Proc. IEEE Internat. Conf. Acoustic. Speech Signal Process.*, pp. II-652-II-655, 1993.

[8] S. Amari, "A theory of adaptive pattern classifiers," *IEEE Trans. Electron. Comput.*, vol. 16, no. 3, pp. 299-307, 1967.

[9] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, "Connectionist probability estimators in HMM speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 1, Part II, pp. 161-174, Jan. 1994.

[10] P. L. Cerf, W. Ma, and D. V. Compernolle, "Multilayer perceptrons as labelers for hidden Markov models," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 1, Part II, pp. 185-193, Jan. 1994.

[11] G. Rigoll, "Maximum mutual information neural networks for hybrid connectionist-HMM speech recognition system," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 1, Part II, pp. 175-184, Jan. 1994.

[12] F. Wolfertstetter and G. Ruske, "Discriminative state-weighting in hidden Markov models," in *Proc. ICSLP'94*, pp. 219-222, 1994.

[13] O. W. Kwon and C. K. Un, "Discriminative weighting of HMM state-likelihoods using the GPD method," *IEEE Signal Processing Letters*, vol. 3, no. 9, pp. 257-259, Sep. 1996.

[14] Y. J. Chung and C. K. Un, "Multilayer perceptrons for state-dependent weightings of HMM likelihoods," *Speech Communication*, vol. 18, pp. 79-89, Jan. 1996.

[15] R. Cole, M. Fanty, Y. Nuthusamy, and M. Gopalakrishnan, "Speaker-independent recognition of spoken English letters," in *Proc. IJCNN'90*, vol. II, pp. 45-51, 1990.

[16] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Tech. J.*, vol. 62, no. 4, pp. 1035-1075, 1983.