SELECTION OF THE OPTIMAL STRUCTURE OF THE CONTINUOUS HMM USING THE GENETIC ALGORITHM

Tomio Takara, Yasushi Iha, and Itaru Nagayama

Department of Information Engineering, University of the Ryukyus
1 Senbaru, Nishihara, Okinawa 903-0213 JAPAN
takara@ie.u-ryukyu.ac.jp

ABSTRACT

The hidden Markov models (HMMs) are widely used for automatic speech recognition because they have a powerful algorithm used in estimating the model's parameters, and also achieve a high performance. Once a structure of the model is given, the model's parameters are obtained automatically by feeding training data. However, there is still an unresolved problem with the HMM, i.e. how to design an optimal HMM structure. In answer to this problem, we proposed the application of a genetic algorithm (GA) to search out such an optimal structure, and we showed this method to be effective for isolated word recognition. However, the test of this method was restricted to discrete HMMs. In this paper, we propose a new application of the GA to the continuous HMM (CHMM) which is thought to be more effective than the discrete HMM. We report the results of our experiment showing the effectiveness of the genetic algorithm in automatic speech recognition.

1. INTRODUCTION

The hidden Markov models (HMMs)[1] are widely used for automatic speech recognition because they have a powerful algorithm used in estimating the model's parameters, and also achieve a high performance. Once a structure of the model is given, the model's parameters are obtained automatically by feeding training data.

However, there is still an unresolved problem with the HMM, i.e. how to design an optimal HMM structure. In answer to this problem, we proposed the application of a genetic algorithm (GA)[2] to search out such an optimal structure, and we showed this method to be effective for isolated word recognition[3].

However, the test of this method was restricted to discrete HMMs.

In this paper, we propose a new application of the GA to the continuous HMM (CHMM) which is thought to be more effective than the discrete HMM. We also report the results of our recognition experiment showing the effectiveness of the GA in the CHMM.

In our previous report, the HMM structure was represented as a matrix and then coded into a one dimensional string. In addition to this coding, we newly code the initial output probability of each state at each generation. For the current selection, we use the roulette strategy and the elite-preservation strategy with two elite individuals instead one. The first elite is an individual which has the highest fitness, and the other is a newly added set of individuals with the highest recognition score. For the crossover, we adopt one point crossover of the string which is a coded form of the matrices with added output probabilities. For the mutation, we again generate or delete transitions between states, further, we randomly shift the value of the mean vectors of the normal density function.

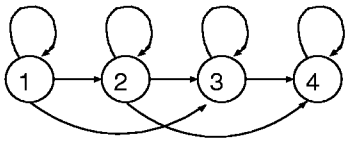

Figure 1: An example of HMM structure.

	1	2	3	4
1	1	1	1	0
2	0	1	1	1
3	0	0	1	1
4	0	0	0	1

Figure 2: Matrix expression of Figure 1.

2. SPEECH RECOGNITION USING CONTINUOUS HIDDEN MARKOV MODELS

A hidden Markov model (HMM) is understood as a generator of vector sequences, and has a number of states connected by arcs. Figure 1 illustrates an example of an HMM structure, in which the circles and the arrow arcs represent the states and the state-transitions, respectively. In each state, there is an output probability distribution of an acoustic vector, and each transition is associated with a state-transition probability. The output probability of the continuous HMM is represented in a multidimensional normal density function. These probabilities are called the model parameters and can be estimated effectively by using the Baum-Welch algorithm[1]. An HMM structure can be expressed in a matrix form $C = (c_{i,j})$. When $c_{i,j} = 1$, there exists a transition from state i to state j, and when

$c_{i,j} = 0$, the transition does not exist. For example, the matrix expression of the structure of Figure 1 is shown in Figure 2. The matrix expression of an HMM will be used for the coding of the genetic algorithm.

An HMM is a finite-state machine that changes state once every time unit. Each time, t, a state, j, is entered, an acoustic speech vector, $\mathbf{y}_t$, is generated with probability density $b_j(\mathbf{y}_t)$. The transition from state i to state j is governed by the probability $a_{i,j}$. The joint probability of a vector sequence $\mathbf{Y}$ and state sequence $\mathbf{X}$, given some model M is calculated as the product of the transition probabilities and the output probabilities:

$$p(\mathbf{Y}, \mathbf{X}|M) = a_{x(0),x(1)} \prod_{t=1}^{T} b_{x(t)}(\mathbf{y}_t) a_{x(t),x(t+1)} \quad (1)$$

where $x(0)$ is constrained to be the model entry state and $x(T+1)$ is constrained to be the model exit state. Eq.(1) can be rewritten in a logarithmic form $\log p(\mathbf{Y}, \mathbf{X}|M)$. Using the Viterbi algorithm, $\log p(\mathbf{Y}|M)$ can be approximated by finding the state sequence $\mathbf{X}$ that maximizes Eq.(1). We adopt the Viterbi algorithm to calculate the log-likelihood, $\log p(\mathbf{Y}|M)$.

In a spoken word recognition system using HMMs, the HMMs for each word class are previously prepared. When a spoken word is inputted, the log-likelihoods for each HMM of a word class are calculated and the word class maximizing this value is determined as the word class of the inputted word.

We also use the log-likelihood to evaluate the fitness of the genetic algorithm. In this case, the evaluations are done for each training datum and are averaged in the word class.

3. SELECTION OF A CHMM STRUCTURE USING THE GENETIC ALGORITHM

The genetic algorithm (GA) was introduced on the basis of the principle of biological evolution (natural selection and mutation) and has been used for search, training or optimization. In this algorithm, a candidate for the solution of a problem is represented by a one dimensional string of genotype on a chromosome. The string is decoded into a phenotype and its fitness is evaluated. Individuals with higher fitness survive and individuals with lower fitness die. The procedure of the GA is as follows:

1 Set initial generation.

2 Repeat following GA operations until the terminating condition is satisfied.

- Fitness evaluation
- Selection
- Crossover
- Mutation

When we apply the GA to the selection of the CHMM structure, we need to specify the coding method of the structure and the fitness measure of each individual HMM, as well as the selection, crossover and mutation operations.

In our previous report, the HMM structure was represented as a matrix and then coded into a one dimensional string[3]. The Left-to-Right (L-R) CHMM structure represented by the matrix form is coded into the genotype string as shown in Figure 3. In addition to this coding, we newly code the initial output probability of each state at each generation. Because the output probability of the CHMM is represented in a multidimensional normal density function, we put, in the order of time, the mean vectors and the covariance matrices of each states.

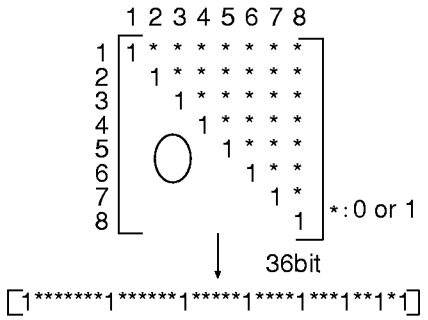

Figure 3: The coding for the HMM structure.

To measure the fitness of the individual represented by a string, we adopt the sum of log-likelihoods of the CHMM. The fitness evaluation in our method is done as follows: First, the model parameters of an CHMM structure are randomly initialized and estimated by the Baum-Welch algorithm using training data. Next, the log-likelihood is calculated for each training datum using the Viterbi algorithm. The log-likelihoods of each datum are summed in an individual. And the error rate, which is estimated in the closed test, is added to the log-likelihood as a penalty. The value f_i of an individual i is given as follows:

$$f_i = (\sum_{j=1}^{M} L_{ij}) - E_I \quad (2)$$

$$E_I = (\sum_{j=1}^{M} L_{ij}) \cdot (10 \cdot e_I) \quad (3)$$

where M is the number of the training data, L_{ij} is the log-likelihood of the individual i calculated by a training datum j, e_I is the error rate of the group I which includes the individual i, and E_I is the penalty according to the error rate. Individuals are ordered according to f_i in a generation, then the fitness F_i is given as:

$$F_i \propto 1/k_i \quad (4)$$

where k_i is the order of the individual i.

We set the number of states to be eight because, in our preliminary experiments, the HMM structure with eight states achieved the best score for spoken word recognition. We set 30 for the number of individuals in a generation. The GA procedure is independently performed for each word class.

In the initial condition, one of the individuals is set as a basic L-R (B-L-R) HMM structure in which $a_{i,j} = a_{i,j+1} = 1$; others $= 0$, because the B-L-R structure achieved the

highest score in our preliminary recognition experiments. The other 29 individual genotype strings are randomly generated.

For the current selection, we use the roulette strategy and the elite-preservation strategy with two elite individuals instead one. The first elite is an individual which has the highest fitness, and the other is a newly added set of individuals with the highest recognition score.

For the crossover, we adopt one point crossover of the string which is a coded form of the matrices with added output probabilities. Before the crossover, the candidates are randomly selected and paired. Then the crossover operation is done for each pair. The crossover occurs in the probability 0.6 at one point in a genotype string, and two strings are generated.

For the mutation, we again generate or delete transitions between states, further, we randomly shift the value of the mean vectors of the normal density function. For the mutation, each bit of the string is inverted in the probability 0.03.

After the GA operations are repeated 30 times (or generations), the GA procedure is terminated.

4. RECOGNITION EXPERIMENT

In order to evaluate the proposed method, we performed recognition experiments. The speech data used in our recognition test are English numeral words from the database TIDIGITS[4]. For training, 11 numeral words "one" to "nine", "zero" and "oh" were uttered twice by 20 American males and 20 American females. In an open test, we used the same vocabulary of the above numeral words this time uttered by another group of 20 males and 20 females.

The speech sampling rate is 10kHz, and overlapping sections of 25.6ms of speech weighted by the Blackman window are analyzed every 10ms to give FFT power spectra. The power spectra are transformed to FMSs[5], which are the Fourier transforms of Mel Sone spectra whose frequency-axes are warped to be the mel scale and magnitude-axes are warped to be the sone scale. Five dimensional vectors, whose components are second to sixth components of the FMS, are used as the feature vectors. Different CHMM structures were used for each word category.

In order to compare to the proposed method, we performed two recognition tests. One is a conventional method without the GA using the B-L-R structure. Resulting recognition scores were 96.6% for the training data set in the closed test and 94.1% for the test data set in the open test. These recognition scores are cited in the following graphs.

In the proposed method, we set one of the initial individuals to be the B-L-R structure and monitor the recognition score whether it becomes higher or not than that of the B-L-R structure. Because we adopt the elite preservation strategy, we can certainly get better structures than the B-L-R structure whenever they exist. Result of the recognition test using the proposed method is shown in Figure 4.

From this figure, we can see that the recognition score becomes higher as each generation proceeds, and becomes higher than that of the B-L-R structure. This is true not only for a training set but also for a test set. We performed the experiments three times and the same results

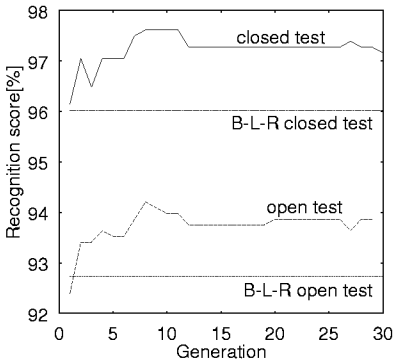

Figure 4: Proposed method.

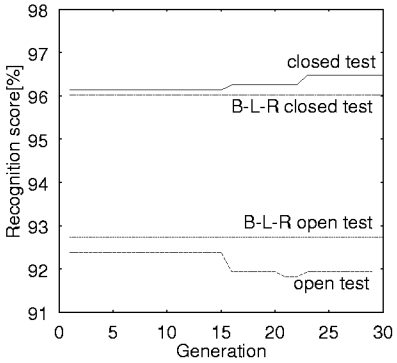

Figure 5: Without the coding of the output probability.

were obtained. These show that using the GA to select the CHMM structure is quite effective for automatic speech recognition.

Consequently, it is shown that the genetic algorithm is effective for spoken word recognition.

5. DISCUSSION

The major features of the proposed method are as Follows:

(1) One of the initial individuals is the B-L-R structure.

(2) The initial output probability of each state at each generation is coded .

(3) The elite-preservation strategy with two elite individuals is adopted.

(4) The fitness is given as the log-likelihood with a penalty based on the error rate.

We discuss here the effect of these features according to the recognition experiments.

Figure 5 shows the result of the recognition experiment in which the output probabilities are not coded. From this figure, we can see that the recognition score of the closed test is improved, however, that of the open test is not. Consequently, it is shown that the coding of the output probability is effective for the improvement of the recognition score of the open test.

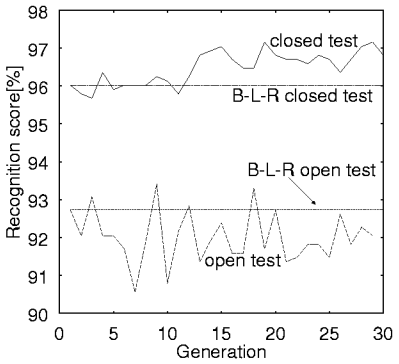

Figure 6: Without the elite-preservation strategy.

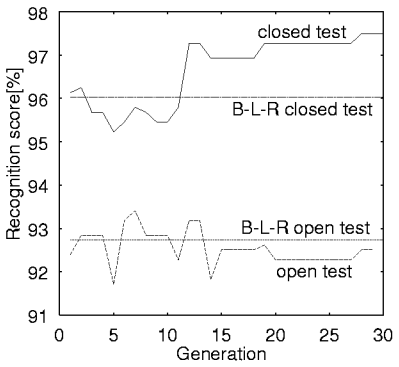

Figure 7: Only the elite by log-likelihood is preserved.

Figure 6 shows the result of the recognition experiment in which the elite-preservation strategy is not used. Figure 7 shows the result of the recognition experiment in which only the elite by log-likelihood is preserved. These show the effectiveness of the elite-preservation strategy of the proposed method. From Figure 7, it is also shown that the elite by recognition score is effective in the open test.

Figure 8 shows the result of the recognition experiment in which the B-L-R structure is not set in the initial individuals. From this figure, we can see that the recognition score of the closed test becomes around that of the B-L-R structure, and the recognition score of the open test improves slowly. This shows that the optimal structure is near to the B-L-R structure.

6. CONCLUSION

We applied the genetic algorithm to select the optimal HMM structure for isolated word recognition. Major features of this method are the coding and the GA operation for the output probability and the adoption of two kinds of elite preservation strategy. We performed recognition experiments showing that the GA is effective for automatic speech recognition.

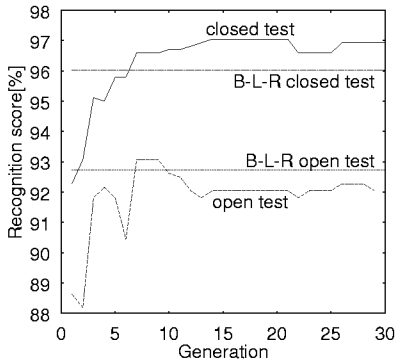

Figure 8: Without the B-L-R structure in the initial individuals.

REFERENCES

[1] Rabinar, L. R. : "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, 77, 2, pp. 257-286 (Feb. 1989).

[2] Goldberg, D. E.: "Genetic Algorithms in Search, Optimization, and Machine Learning", Addison-Wesley Publishing Co., Inc., Reading, Massachusetts (1989).

[3] Takara, T., Higa, K., and Nagayama, I.: "Isolated Word Recognition Using the HMM Structure Selected by the Genetic Algorithm", ICASSP-97, Vol. 2, pp. 967-970 (Apr. 1997).

[4] NIST: "TIDIGITS CD-ROM Set", NIST (Feb. 1991).

[5] Takara, T. and Imai S.: "Isolated Word Recognition Using DP-Matching and Maharanobis' Distance", (in Japanese) Trans. IECE Japan, J66-A, 1, pp. 64-70 (Jan. 1983).