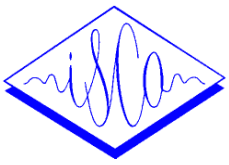

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SPOKEN WORD RECOGNITION USING THE ARTIFICIAL EVOLUTION OF A SET OF VOCABULARY

Tomio Takara and Eiji Nagaki

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

ABSTRACT

Hidden Markov models (HMMs) are widely used for automatic speech recognition. However, there is a problem still unresolved, i.e. how to design the optimal structure of the HMM. As an 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. In these applications, the evolutions occurred at each word class independently. However, many isolated word recognition systems are performed using a set of vocabulary. Therefore, the artificial evolution using the vocabulary set is thought to be more effective. In this paper, we propose the spoken word recognition using the artificial evolution of a set of vocabulary.

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 applications of a genetic algorithm (GA) [2] to search out such an optimal structure, and we showed these methods to be effective for isolated word recognition [3][4].

In these applications, the evolutions occurred at each word class independently. However, many isolated word recognition systems are performed using a set of vocabulary. Therefore, the artificial evolution using the vocabulary set is thought to be more effective.

In this paper, we propose the spoken word recognition using the artificial evolution of a set of vocabulary.

The GA was introduced on the basis of Darwin's principle of biological evolution (natural selection and mutation) and has been used for search, training and optimization. When we apply the GA to the selection of the HMM 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.

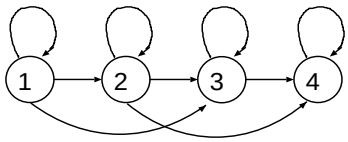

Figure 1: An example of HMM structure.

In our previous report, the HMM structure was represented as a matrix and then coded into a one-dimensional string [3]. This coding method is also adopted in this study. In addition to this method, we restricted the place of crossover at the boundaries of transition's group which are formed by grouping transitions with the same starting state. The reason for doing so is that the sum of probabilities of the transitions with the same starting state is equal to the unity, and this feature should not be destroyed by a new structure. To measure the fitness of the individual word represented by a string, we adopted the sum of log-likelihoods of the HMMs, whereas in the conventional method, the fitness is calculated independently for each category.

2. SPEECH RECOGNITION USING 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. 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

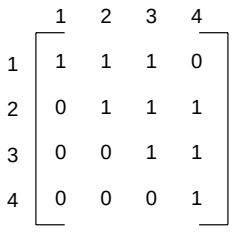

Figure 2: Matrix expression of Figure 1.

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:

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

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. In order to prevent being affected by a word length, in our method, the log-likelihood of a word is divided by the word length T.

3. SELECTION OF AN HMM STRUCTURE USING THE GENETIC ALGORITHM

3.1. 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.

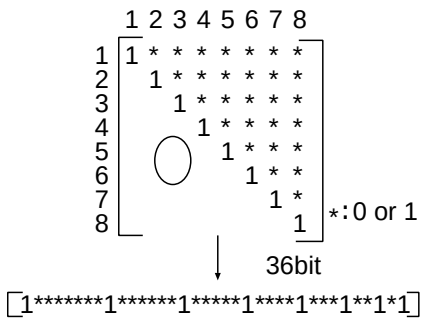

Figure 3: The coding for the HMM structure.

- Fitness evaluation
- Selection
- Crossover
- Mutation

3.2. Selection of an HMM Structure

We apply the GA to search for the optimal HMM structure. We adopt the left-right (L-R) type HMM structure because with it, we can associate time with the model states in a straightforward manner[1]. For simplicity, we set the number of states to be constant in all generations. The L-R HMM structure represented by the matrix form is coded into the genotype string as shown in Figure 3.

We adopt log-likelihood as the fitness. The fitness evaluation in our method is done as follows: First, the model parameters of an HMM 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, in which the log-likelihood is divided by the word length as described above. The log-likelihoods of each datum are averaged in the word class. The averaged log-likelihood is used as the fitness of the HMM structure.

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.

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 selection, we adopt two strategies and combine them. One is the fitness-ordered strategy in which 29 candidate individuals for the next generation are selected in the following probability:

$$P_s \propto N - i, \qquad (2)$$

where N is the population of a generation and i is the order of fitness. The other strategy is the elite-preservation strategy, in which an individual with the highest fitness always survives to be an individual of the next generation. 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 two points in a genotype string, and two strings are generated. 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.

3.3. Method of a Set of Vocabulary

In spoken word recognition systems with limited vocabulary, a word class with the largest likelihood is selected from reference word set of vocabulary. Therefore, it is reasonable for the artificial evolution to perform using a set of vocabulary.

In a conventional system which doesn't use the set of vocabulary, fitness is evaluated at each word class independently. Structures with the highest fitnesses are collected, and a recognition test is performed using these structures. Result is adopted as a recognition score at a generation. In the proposed method with a set of vocabulary, fitness is evaluated as a sum of the averaged log-likelihoods of each word class of a vocabulary. The recognition score at a generation is evaluated using this set of word classes' structures with the largest sum of averaged log-likelihoods. The total fitness is obtained as Eq. (2) using a order of the sum of log-likelihoods.

When we apply the genetic algorithm to the HMM, useless HMM structures may be generated in which there is no path from the initial state to the final state. We call such structures as dead structures because we can not use these structures for HMMs. In the conventional method, the dead structures are generated in the probability of one 30th, and they are weeded out in the selection. However, in the proposed method with a set of vocabulary, a set of structures of the vocabulary can not be used even when one of the structures is dead. The amount of dead structures becomes 11 (= 1/30 * 11) out of 30 individuals, and they are not weeded out in the selection. Therefore, we need to delete the dead structures.

The dead structure may be generated in crossover and mutation operations. Therefore, after these operations, the system checks a generated structure. When the structure is not dead, this structure is adopted as the structure of the next generation. On the other hand, if the structure is dead, the clone of one of its parents' structures is adopted for the next generation.

3.4. Genetic Operations at a State

Generally speaking for HMM, sum of the transition probabilities on the arcs that start from the same state should be equal to unity. Therefore, there is some relationship among probabilities whose origins are the same. In the conventional GA applications to the HMM, however, crossovers and mutations occurred at arbitrary arcs of transitions. In these methods, even if a new state with high performance is generated, its good characteristics may be destroyed by the crossover and the mutation operations. Therefore, we propose new GA operations for the HMM which is performed for a transition group with the same starting state.

In the crossover, we exchange parts of two chromosomes in a manner of state's unit, i.e. all transitions with the same starting sate are grouped to be a unit, and the unit is treated as the exchanged part. In a new method, the mutation also occurs at a state's unit, i.e. when one state is selected, the transitions that started from this state are all inverted of their existence. If a transition existed, it is deleted, and if there was no transition, a new transitions are generated.

4. RECOGNITION EXPERIMENT

4.1. Experimental Conditions

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[5]. 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[6], 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. Three dimensional vectors, whose components are second to fourth components of the FMS, are used as the feature vectors. For code-book generation, we use the clustering algorithm[7] in which the FMS-space is repeatedly divided into two sub-spaces, obtaining cluster centers which minimize the estimation error at each sub-space. The code-book size is 64.

In order to compare to the proposed method, we performed two recognition tests. One is a 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.

4.2. Experiment on the Set of Vocabulary

We show here results of recognition experiments which were done to study the effectiveness of the evolution for the set of vocabulary. The experiments were performed three times by changing the seeds of random numbers. The average of these results at each generation is shown in Fig. 4 where the horizontal axis represents the generation and the vertical axis represents the recognition score. The graph shows the highest score of the vocabulary set at each generation. The solid line shows the result of the pro-

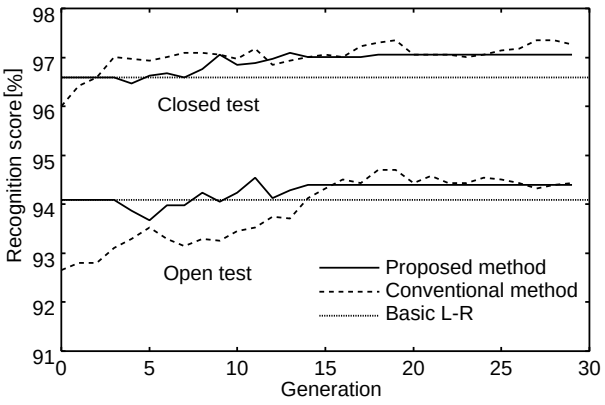

Figure 4: Evolution for a Set of Vocabulary.

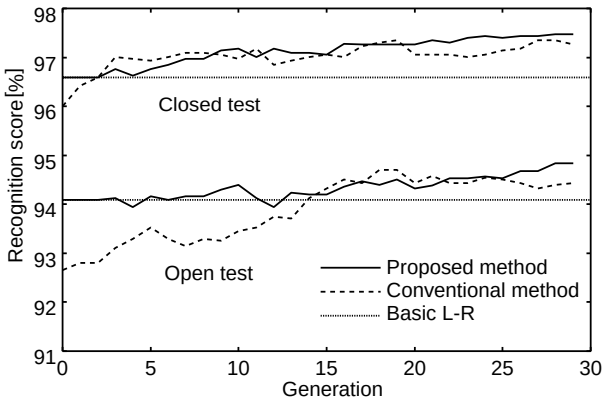

Figure 5: Evolution for a State's Unit.

posed method using the vocabulary set, and the dashed line shows that of the conventional method without the vocabulary set. For comparison, the recognition scores of B-L-R structure are shown in doted line. The figure is divided into two groups. The upper group shows the results of the closed test, and the lower group shows the results of the open test.

From this figure, we can see that the recognition scores increase as the generation proceeds until about the 15th generation both in the closed and the open tests. After the 15th generation, they do not increase and converge to some value. The reason of this can be thought that a number of dead structures increases because we adopted the set of vocabulary, and the probabilities of crossover and mutation become small in substance because the crossover or mutation were not performed substantially when the dead structures were generated. However, we will obtain better result if we make better process after the dead structures' generation because, in the open test, the recognition scores around at the beginning generations are higher than that of the conventional one.

4.3. Experiment on the State's Unit

We show next the result of the state's unit method in which the GA operations for a state's unit are expanded to the set of vocabulary. The experiments were performed three times (cases) by changing the seeds of random numbers. Fog. 5 shows the average of these three cases. The solid line shows the result of the proposed method of GA operations for a state's unit. The dashed line shows the result of the conventional one in which the categories are processed independently without the set of vocabulary and the GA operations are not performed for a state's unit.

From this figure, we can see that the recognition scores increase as the generation proceeds in the closed test and the open test. In all cases of the open test, the recognition scores at the last generation of the proposed method were greater than that of the conventional method. In two cases of the experiments, the recognition scores of the proposed method increased after the 20th generation and could be expected to increase even after the 30th generation, whereas that of the conventional ones converge to upper limits around at the 20th generation. Consequently, the proposed method that adopts the set of vocabulary and the GA operation for the state's unit is shown to be effective.

5. CONCLUSION

We applied the genetic algorithm to select the optimal HMM structure for isolated word recognition. Major features of this method are to evolve the set of vocabulary and to perform the GA operation at a state as a unit. We performed recognition experiments showing that the proposed method is effective for automatic speech recognition using the genetic algorithm.

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, 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, 1997.

4 Takara, T., Iha, Y., and Nagayama, I.: "Selection of the Optimal Structure of the Continuous HMM Using the Genetic Algorithm", ICSLP-98, Vol. 3, pp.751-75, 1998.

5 NIST: "TIDIGITS CD-ROM Set", NIST, 1991.

6 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, 1983.

7 Nakagawa, S.: "Speech Recognition Using Probability Model", (in Japanese) pp. 18-26, Corona Co., Tokyo, 1988.