Journal of Information Hiding and Multimedia Signal Processing ©2012 ISSN 2073-4212
Ubiquitous International Volume **3**, Number **4**, October **2012**

Robust Speech Recognition by DHMM with A Codebook Trained by Genetic Algorithm

Shing-Tai Pan[1], and Tzung-Pei Hong[2]

Department of Computer Science and Information Engineering,
National University of Kaohsiung,
Kaohsiung, Taiwan 811.
[1]stpan@nuk.edu.tw, [2]tphong@nuk.edu.tw

Received July, 2011; revised June 2012

Abstract. *This paper uses genetic algorithms to train a codebook for the modeling of Discrete Hidden Markov Model (DHMM) applied to speech recognition. The GA-trained DHMM is then used to increase the recognition rate for Mandarin speeches. Vector quantization based on a codebook is a fundamental process to recognize the speech signal by DHMM. A codebook will be first trained by genetic algorithms through Mandarin speech features. The speech features are then quantized based on the trained codebook. Subsequently, the quantized speech features are statistically used to train the model of DHMM for speech recognition. All the speech features to be recognized should go through the codebook before being fed into the DHMM model for recognition. Moreover, the Empirical Mode Decomposition (EMD) process is applied here for the noises separation from speech signals. Experimental results show that the speech recognition rate can be improved by using genetic algorithms to train the model of DHMM and EMD process.*
Keywords: speech recognition, genetic algorithm, Discrete Hidden Markov Model, Empirical Mode Decomposition

1. **Introduction.** Recently, human life depends on electronic products more due to the development in IT technology. The interface between these products and user is quite important. Since the computation ability of CPU is enormously enhanced, the speech control for a electronic product becomes more realizable. It has been a long time for the development of speech recognition. Recently, the topic on the process of audio signal attracts more attention [1-3]. There are many researches about speech recognition [4-6] because speech recognition becomes more and more important and will be a standard interface between human and electronic products in the future.

As for the progress of the speech recognition technique, a former recognition technique is Dynamic Time Warping (DTW) [4] which used dynamic programming [7] to calculate the difference between the target speech and testing speech to recognize the testing speech. Then, Artificial Neural Network (ANN) was proposed to replace DTW for speech recognition. Because that the structure of ANN will be fixed after it is determined, the recognition rate cant be improved by online learning with more additive speech signals. Recently, Hidden Markov Model (HMM) [8] was widely applied on speech recognition [9-10]. It can solve the problem arises from variant speech speed and be constructed layer by layer to achieve automatic speech recognition (ASR). Instead of fixed input node as in ANN, the input number in HMM is variable. Hence, HMM is more appropriate for the speech recognition on the speech with non-constant speed. Besides, the speech

recognition rates by using HMM-based algorithms are always better than those by using ANN-based algorithms. Accordingly, HMM is viewed as the most effective approach on the speech recognition [11]. There are various HMM models can be applied to this application. These models can be categorized as Continuous Hidden Markov Model (CHMM) and Discrete Hidden Markov Model (DHMM) according to the type of the probability distributions used in HMM. Roughly speaking, if there is enough training data, the CHMM will perform better than the DHMM. However, the complexity of CHMM is much higher than DHMM, as well, the computing time and model scale of CHMM is much larger than DHMM. In addition, the DHMM provides more stable recognition results and faster training with the recognition accuracy that is no less than CHMM. This feature of DHMM makes it more suitable than CHMM to be implemented on a hardware for real-time applications. This is the reason why this paper adopts DHMM to recognize speech. Before speech recognition, speech signal have to be pre-processed. The pre-process of speech signal includes speech sampling, point detection, pre-emphasis, Hamming window and features capture. After these processes, we can evaluate the probabilities of every HMM model corresponding to each speech and find the model which has highest probability to be the result of recognition. The feature of speech signal which was used in this paper is obtained by Mel-Frequency Cepstrum Coefficient (MFCC) [8].

Mode Decomposition is applied here for separating noises from a speech signal. This allows us to recognize the speeches subject to various environmental noises. Moreover, speech feature quantization plays an important role on the training of DHMM model. All the features used for the training phase or for the testing phase of DHMM recognition systems must be quantized based on a trained codebook. It is very important to obtain a well-trained codebook, since the codebook affect enormously the training of DHMM model and hence the recognition rate. Consequently, in this paper, the GA is applied to train a better codebook to improve the recognition rate. GA is based on Darwins theory of evolution: Survival of The Fittest. It is an evolutionary algorithm which is most widely applied on solving optimization problems. The algorithm claimed that the nature of biological evolution is in the genes. Biological characteristics of each species are passed down through gene sequencing from previous generations. GA solves problems through gene encoding using a set of parameters while simulating the natural evolution process: selection, crossover, and mutation to find an optimal solution. Recently, some literatures used GA to enhance the performance of speech recognition and obtained impressed results [12-16]. Consequently, GA is used here to train an optimal codebook for DHMM. Besides, the EMD process is applied in this paper for the robust speech recognition problems.

This paper is organized as follows. In section II, GA is used to train a codebook for DHMM modeling. The modeling of DHMM for various speeches is presented in Section III. Section IV introduces the EMD process for some speeches subject to noises. The experiment of the proposed speech recognition system is then presented in Section V where the speech recognition rates for the speeches in the database AURORA II arerevealed. Finally, some conclusions are made in Section VI.

2. TRAIN CODEBOOK BY USING GA.

In this section, a codebook for the quantization of speech signal features is generated by using GA. Thereafter, a DHMM is modeled for speech recognition through the codebook in next section.

It is well known that GA solves a optimization problem through the evolution process: selection, crossover, and mutation [17]. It is important to define the chromosome in GA to meet the problem in hand. Each chromosome has multiple genes, and the number of the parameters in a problem will determine the number of genes. Encoding of genes can be divided into the following three ways: binary encoding, real number encoding, and

symbol encoding according to the parameter type in a problem and are shown in Figure 1 [18].

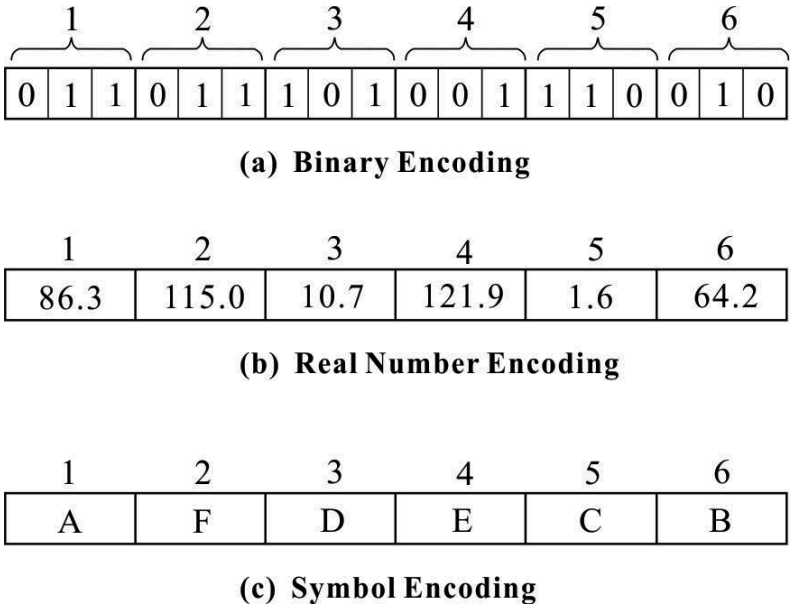


FIGURE 1. Three chromosome encodings: (a) binary encoding (b) real number encoding (c) symbol encoding [18].

In this paper, the GA is applied to find a better initial codebook for DHMM. A chromosome is synonymous for a codebook which contains some numbers of classifications of speech feature vectors. The real number encoding is used here, since the features are all real. Each generation of chromosomes is encoded here as a set of 3D matrix with the dimensions are equal to number of classification × number of features × number of chromosomes. According to the process of GA, we start by randomizing the first generation of chromosomes. Then the chromosomes of codebook are tested with speech signals by using DHMM. The recognition rate is identified as the fitness of the chromosomes. We select the best chromosomes based on natural selection. According to Survival of the "Fittest", the next generations chromosomes are produced by this generations best chromosomes. The selection method in this paper is based on the Roulette Wheel Selection. The crossover used here is linear crossover. After crossover, the mutation process will begin. If the mutation probability is larger than a preset mutation rate, we will randomly generate the new chromosome.

The codebook trained by GA repeats these steps until the fitness reaches the target recognition rate (Figure 2). The best codebook is then used to model the DHMM. During the training of weights by GA, the fitness function is defined as the reciprocal of the speech recognition error for the training speeches, which is described in eq. (1) and (2).

$$fitness = \frac{1}{MSE + 1} \tag{1}$$

$$MSE = \frac{1}{k_u \times i_u \times t_u} \sum_{k=1}^{k_u} \sum_{i=0}^{i_u} \sum_{t=1}^{t_u} |E_{k,i,t}|^2 \tag{2}$$

in which $E_{k,i,t}$ means the output error for t^{th} record of i^{th} literal by k^{th} person.

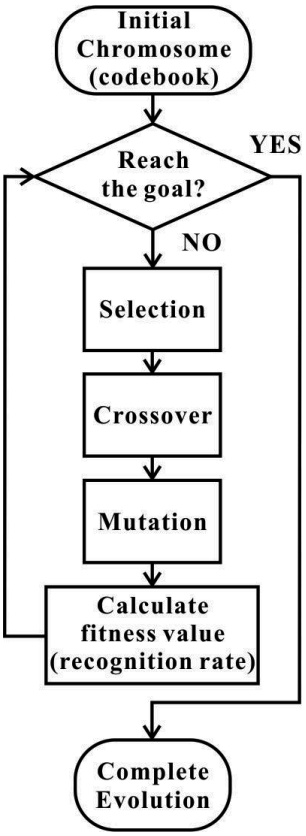


FIGURE 2. The GA flow chart for training codebook

3. **MODEL DHMM.** In a DHMM, the states are unobservable (hidden), but the outputs in each state are observable. Each state has a probability distribution over the possible output tokens (i.e., the observation). Therefore, the sequence of output tokens generated by the DHMM gives some information about the sequence of states. The relation between the features of speech, the observation and the hidden states of DHMM is depicted in Fig. 3. This section introduces the definition and training of the parameters for DHMM [19]. First, the definition of parameters in DHMM is introduced as follows.

λ : DHMM model,$\lambda = (A, B, \pi)$

$A : A = [a_{ij}]$,a_{ij} is the probability of state x_i transferring to state x_j, $a_{ij} = P(q_t = x_j \,|q_{t-1} = x_i)$

$B : B = [b_j(k)], b_j(k)$ is the probability of kth observation which is observed from state ,

i.e., $b_j(k) = P(o_t = v_k \,|q_t = x_j)$

$\pi : \pi = \pi[\pi_i], \pi_i$ is the probability of the case where the initial state is $x_i, \pi_i = P(q_1 = x_i)$

X : the state vectors of DHMM, $X = (x_1, x_2, \cdots x_N)$

V : the observation event vector of DHMM, $V = (v_1, v_2, \cdots v_M)$

O : the observation results of DHMM, $O = (o_1, o_2, \cdots o_T)$

Q : the resulting states of DHMM, $Q = (q_1, q_2, \cdots q_T)$

In DHMM, the probability of the observations according to the model $\lambda = (A, B, \pi)$ is calculated by the following equation (3) [19]:

$$P(O \,|\lambda) = \sum_{Q} P(O \,|Q, \lambda) P(Q \,|\lambda) = \sum_{q_1 \cdots q_T} \pi_{q_1} b_{q_1}(o_1) \cdot a_{q_1 q_2} b_{q_2}(o_2) \cdots a_{q_{T-1} q_T} b_{q_T}(o_T) \tag{3}$$

This equation enables us to evaluate the probability of the observations O ased on the DHMM model $\lambda = (A, B, \pi)$. The procedure for recognizing a speech is then depicted in Fig. 4. The DHMM models corresponding to each speech are first trained by using the training speech through a trained codebook. In the test phase, the feature for a test speech frame will be derived. Through the trained codebook, this feature is then quantized and becomes an observation of the DHMM. For each observation, the probabilities for all DHMM models are calculated according to eq. (3). The speech corresponding to the DHMM with largest probability is then assigned as the recognized speech.

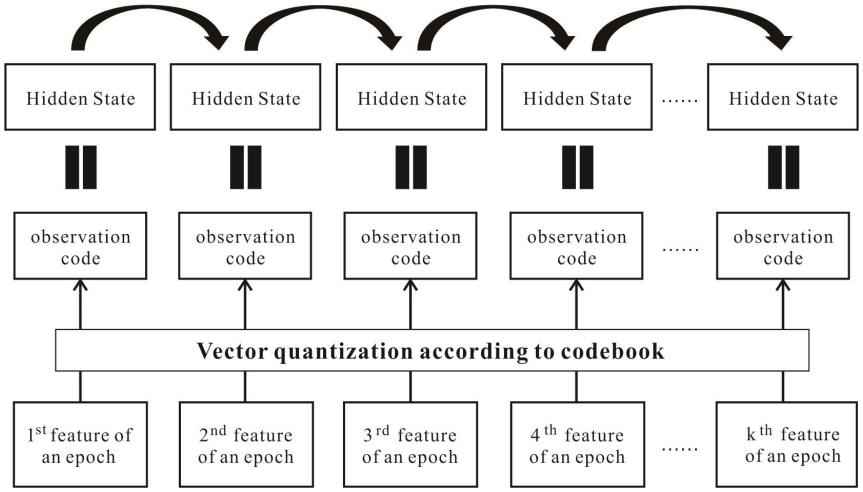


FIGURE 3. Relation between the features of speech, the observation and the hidden states of DHMM

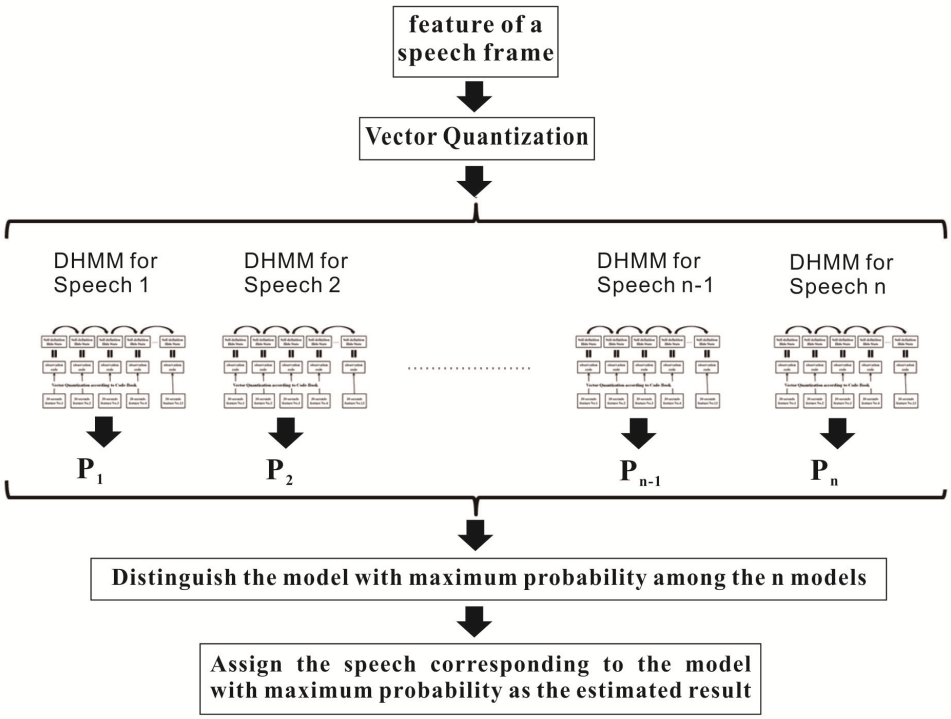


FIGURE 4. Procedure for the speech recognition via DHMM

4. **EMPIRICAL MODE DECOMPOSITION (EMD).** This study applies EMD to decompose the environmental noises and clean speech signal from a contaminated speech signal. In this section, the procedure for performing EMD is introduced. The purpose of EMD is to shift the original non-stationary data series until the final data series are stationary. The main step to perform EMD operation is to divide a speech signal into several intrinsic mode functions (IMFs). The condition for the data series to be an IMF can be described as follows [20] .

1. In the whole data series, the number of local extremes and the number of zero crossings must either equal or differ at most by one.
2. At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

The procedure of EMD for a data series or a signal is then introduced as follows. The Cubic Spline [21] is used to generate the upper envelop and lower envelop of the signals during the process of finding IMFs. Let the original signal is $X(t)$ and $Temp(t) = X(t)$.
step1 :
Find the upper envelop $U(t)$ and lower envelop $L(t)$ of the signal $Temp(t)$. Calculate the mean of the two envelops $m(t) = [U(t) + L(t)]/2$. The component of $Temp(t)$ is obtained by the equation:

$$h(t) = Temp(t) - m(t) \tag{4}$$

step2 :
Check whether the signal $h(t)$ satisfies the conditions of IMF or not. If it is, then the first IMF is obtained as $imf_1(t) = h(t)$ and go to next step, else assign the signal $h(t)$ as $Temp(t)$ and go to Step 1.
step3 :
Calculate the residue $r_1(t)$ as

$$r_1(t) = Temp(t) - imf_1(t) \tag{5}$$

Assign the signal $r_1(t)$ as $X(t)$ and repeat Step 1 and Step 2 to find $imf_2(t)$.
step4 :
Repeat Step 3 to find the subsequent IMFs as follows.

$$r_n(t) = r_{n-1}(t) - imf_n(t), n = 2, 3, 4, \cdots \tag{6}$$

This step is end when the signal $r_n(t)$ is constant or a monotone function.

After the EMD procedure Step 1 Step 4 is finished, the following decomposition of $X(t)$ is obtained.

$$X(t) = \sum_{i=1}^{n} imf_i(t) + r_n(t) \tag{7}$$

The EMD procedure can be illustrated by the flowchart in Fig. 5.

However, it is hard to satisfy the second condition of IMF in practical application, since zero mean value of the envelopes for all time t is almost impossible. Hence, a looser condition is lunched to replace the second condition of IMF. An index for the mean value of the envelopes and a threshold is used to construct this looser condition. In general, the threshold is assigned in the range 0.2 to 0.3. Moreover, the index can be calculated through the following equation [20].

$$SD_{ik} = \frac{\sum_{t=0}^{T} \left| h_{i(k-1)}(t) - h_{i(k)}(t) \right|}{\sum_{t=0}^{T} h_{i(k-1)}^2(t)} \tag{8}$$

in which $h_{i(k)}(t)$ is the kth iteration for ith IMF. It is noted that the function $h_{i(k-1)}(t) - h_{i(k)}(t)$ in numerator of eq. (8) is equal to the mean $m_{i(k)}(t)$, i.e., $h_{i(k)}(t) = h_{i(k-1)}(t) - m_{i(k)}(t)$.

This means that SD_{ik} is the ratio of the energy of $m_{i(k)}(t)$ to that of $h_{i(k-1)}(t)$. In practical applications, the second condition of IMF is then replaced by the looser condition that the SD_{ik} should be smaller than the assigned threshold. That is, if $SD_{ik} < 0.3$, for example, and the first condition of the IMF is satisfied, then the iterations for ith IMF stops and we got a new IMF.

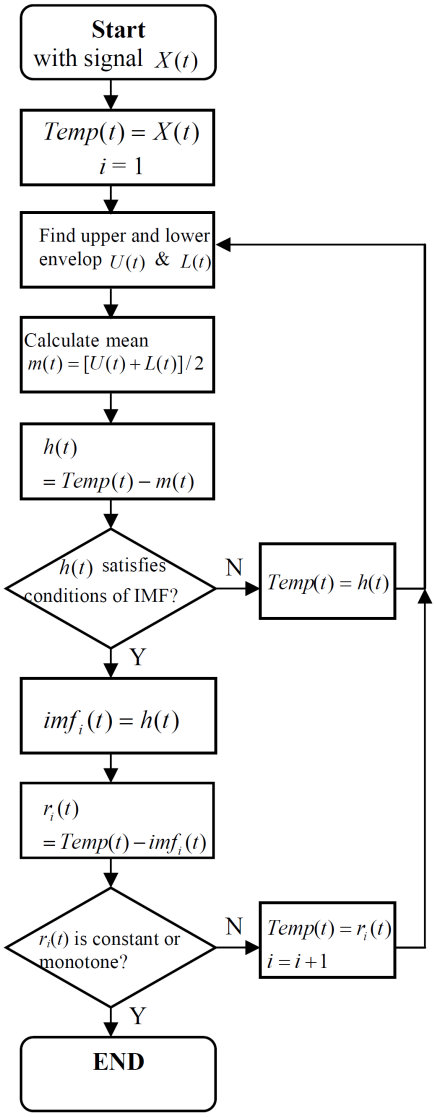


FIGURE 5. Flowchart of EMD

5. **EXPERIMENTAL RESULTS.** In this experiment, two sets of database are used for recognition: one is the database with Mandarin language recorded in our laboratory, another is the AURORA II in which the speeches are contaminated by various environmental noises. About the DHMM model, 7 hidden states are used and 64 observations are adopted for each hidden state. Hence, the size of codebook used in this experiment is 64×13. The parameter matrix in DHMM are of the size 7×7, 7×64 and 1×7 for the matrix A, B, and π, respectively. The specs for the GA and DHMM in this experiment are listed in Table 1 and Table 2, respectively.

5.1. **Codebook trained by GA.** In this experiment, the speeches 0 9, which are commonly used in the speech recognition research for Mandarin language, are recognized. One hundred data for each single speech are recorded. This means that there are totally one thousand data for this experiment. For the hold-out experiment, 50% of the data are used for training while the remaining for testing. Each speech signal is divided into 20 frames with variable overlap rate. Each frame has 13 features derived from 256 sampling points. The evolution of GA on training the codebook for speech recognition is depicted in Fig. 6 in which totally 100 iterations are shown. The evolution of GA converges to its final results about at 30 iterations and gets a very high recognition rate. The numerical results in this experiment are listed in Table 3, in which the recognition rates for the codebook trained by GA and by K-mean algorithm are compared. It is obvious that the speech recognition rates for all the speech in the table as well as the average speech recognition rate are all improved by using GA to train the codebook. This verifies the performance of GA on training the codebook for DHMM modeling. Besides, the speech recognition rates for the speech subject to two noised environments, a supermarket environment and a road environment, which are the environments that many peoples will visit, are listed in Table 4 and Table 5, respectively. According to these tables, the speech recognition rates are improved by using GA to train codebook. This reveals that the proposed method performs well even if some noised speech is considered.

TABLE 1. Specs for the GA in the experiments.

Items	Specs
Chromosome number in each generation	30 (with real number coding)
Selection	Roulette Wheel Selection
Mutation rate	0.4
Crossover	linear
Fitness function	The mean of recognition rate (M_{SSRR})

TABLE 2. Specs of the DHMM in the experiments.

Items	Specs of DHMM
Size of codebook	**64×13**
Hidden state	**7**
Observation code	**13**
$\lambda = (A, B, \pi)$	
A	**matrix dimensions equal to Hidden state × Hidden state**
B	**matrix dimensions equal to Observation code × Hidden state**
π	**matrix dimensions equal to Hidden state × 1**

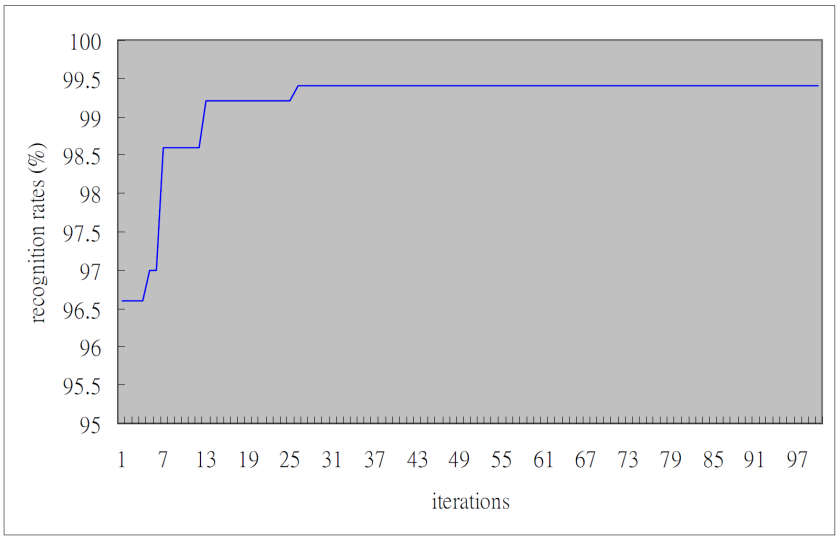


FIGURE 6. Iterations of GA for training the codebook

TABLE 3. Experimental results for the codebook trained by GA and k-mean algorithm

Speeches	Recognition rate for the codebook trained by GA	Recognition rate for the codebook trained by k-mean algorithm
0	***1***	1
1	***1***	1
2	***1***	1
3	***1***	0.96
4	***1***	0.92
5	***1***	0.94
6	***0.94***	0.92
7	***1***	0.94
8	***1***	0.92
9	***1***	0.96
Avg.	***0.994***	0.956

TABLE 4. Experimental results for the codebook trained by GA and k-mean algorithm in the noised environment of supermarket

Supermarket		
speech	Recognition rate for the codebook trained by GA	Recognition rate for the codebook trained by k-mean algorithm
0	***0.56***	0.4
1	***0.3***	0.16
2	***0.76***	0.74
3	***0.92***	0.24
4	***0.76***	0.84
5	***0.46***	0.28
6	***0.92***	0.8
7	***0.68***	0.48
8	***0.84***	0.76
9	***0.32***	0.24
Avg.	***0.652***	0.494

TABLE 5. Experimental results for the codebook trained by GA and k-mean algorithm in the noised environment of road of supermarket

Road		
speech	Recognition rate for the codebook trained by GA	Recognition rate for the codebook trained by k-mean algorithm
0	***0.94***	0.86
1	***0.7***	0.54
2	***0.82***	0.72
3	***0.88***	0.86
4	***0.84***	0.84
5	***0.72***	0.7
6	***0.74***	0.7
7	***0.7***	0.6
8	***0.76***	0.66
9	***0.72***	0.58
Avg.	***0.782***	0.706

5.2. **EMD on noised speech from AURORA II.** In this experiment, speech recognition on the AURORA II database is investigated through the proposed methods. The proposed approaches were then evaluated under the AURORA II [22] testing environment with an English connected digit-string corpus. The clean speech signals from TIDigits database, in which the speeches form 55 male and 55 female with sampling rate of 8 KHz

are included, are used to train the proposed speech recognition system. Then, three testing sets (sets A, B, and C) defined by AURORA II are used for testing of the proposed method. The testing set A includes four different types of environmental noise, which is subway, babble, car, and exhibition. The testing set B includes another four different types of environmental noise which is restaurant, street, airport, and train station. The testing set C then includes two environmental noise types, subway and street, respectively from sets A and B, plus additional convolutional noises [23]. In all the three sets A, B, and C, the signal-to-noise ratio (SNR) is tested ranged from 20 dB to 0 dB with steps of 5 dB.

Table 6-8 reveal the results by using DHMM alone for the speech recognition while the tables Table 9-11 show the results by applying EMD to the speech signals before the DHMM modeling. In these tables, different environmental noises from AURORA II database are explored. The average recognition rates from the speech signals with SNR 0 dB to 20 dB are also revealed. From these tables, it is obvious that the recognition rates for the speech subject to different environmental noises are mainly improved by applying EMD process. Figure 7 shows the comparison of average recognition rates between the results with and without applying EMD on the speech signals.

TABLE 6. Recognition rates (DHMM) for various SNR speech signals from Test A

Test A				
SNR	**Subway**	**Babble**	**Car**	**Exhibition**
clean	97.51	98.28	98.49	97.63
20db	92.22	84.82	88.78	89.49
15db	86.79	72.4	80.73	84.79
10db	76.92	50.6	58.88	70.66
5db	41.37	20.2	41.42	37.29
0db	23.6	11.47	18.06	21.9
Avg.	**64.18**	**47.898**	**57.574**	**60.826**

TABLE 7. Recognition rates (DHMM) for various SNR speech signals from Test B

Test B				
SNR	**Restaurant**	**Street**	**Airport**	**Train Station**
clean	98.1	97.78	97.9	99.2
20db	85.95	89.62	87.12	88.33
15db	74.33	80.36	77.63	79.59
10db	54.1	63.2	51.23	54.8
5db	20.43	40.45	23.62	21.94
0db	13.53	19.1	12.82	15.25
Avg.	**49.668**	**58.546**	**50.484**	**51.982**

TABLE 8. Recognition rates (DHMM) for various SNR speech signals from Test C

Test C		
SNR	**Subway**	**Street**
clean	98.55	98.63
20db	89.23	90.1
15db	81.54	80.9
10db	68.17	64.78
5db	32.72	34.97
0db	20.05	21.76
Avg.	**58.342**	**58.502**

TABLE 9. Recognition rates (DHMM+EMD) for various SNR speech signals from Test A

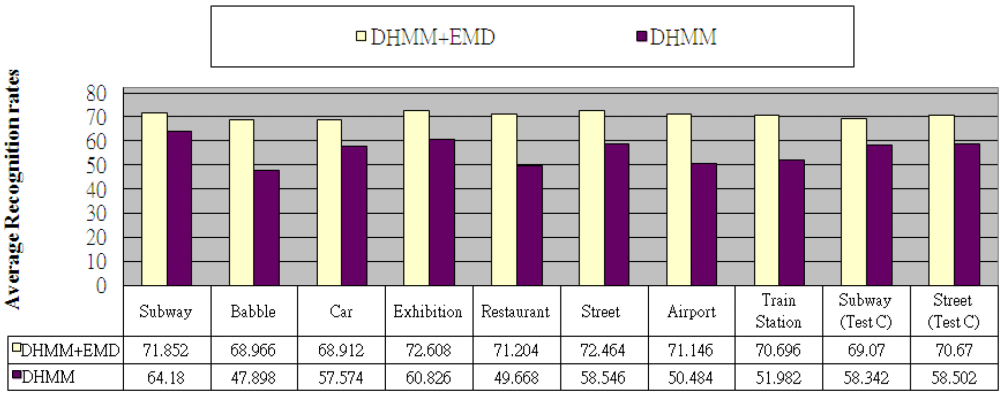
Test A				
SNR	**Subway**	**Babble**	**Car**	**Exhibition**
clean	97.3	97.42	97.31	97.01
20db	93.1	89.2	95.92	94.52
15db	88.9	86.72	86.62	89.32
10db	86.2	74.25	70.12	79.54
5db	60.26	59.26	55.1	62.94
0db	30.8	35.4	36.8	36.72
Avg.	**71.852**	**68.966**	**68.912**	**72.608**

TABLE 10. Recognition rates (DHMM+EMD) for various SNR speech signals from Test B

Test B				
SNR	**Restaurant**	**Street**	**Airport**	**Train Station**
clean	97.22	97.56	97.4	97.1
20db	93.2	93.2	91.33	90.64
15db	89.96	86.4	85.1	86.42
10db	80.1	83.63	77.1	77.42
5db	56.24	62.37	63.9	59.6
0db	36.52	36.72	38.3	39.4
Avg.	**71.204**	**72.464**	**71.146**	**70.696**

TABLE 11. Recognition rates (DHMM+EMD) for various SNR speech signals from Test C

Test C		
SNR	**Subway**	**Street**
clean	97.8	97.88
20db	91.21	95.2
15db	84.02	84.24
10db	77.42	77.8
5db	57.5	58.77
0db	35.2	37.34
Avg.	**69.07**	**70.67**



	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Train Station	Subway (Test C)	Street (Test C)
DHMM+EMD	71.852	68.966	68.912	72.608	71.204	72.464	71.146	70.696	69.07	70.67
DHMM	64.18	47.898	57.574	60.826	49.668	58.546	50.484	51.982	58.342	58.502

FIGURE 7. Comparison of average recognition rates between the results with and without applying EMD.

6. **CONCLUSIONS.** The GA had been used on training a codebook for modeling of Discrete Hidden Markov Model (DHMM) to improves the speech recognition rate for the Mandarin. First, a codebook was trained by genetic algorithms to obtain a better result. Based on the trained codebook, the speech features are quantized by the trained codebook. Subsequently, the quantized speech features are used to the modeling of DHMM for the speech recognition. All the speech features should go through the quantization procedure before being fed into the DHMM model for recognition. Besides, the EMD process is used to separate the speech signals and noises for the speeches with environmental noises in AURORA II. Experimental results reveal that the speech recognition rate can be improved, for clean or noised speech, by using genetic algorithm to train the codebook for the model of DHMM as well as EMD process.

ACKNOWLEDGEMENT. This research work was supported by the National Science Council of the Republic of China under contract NSC 100-2221-E-390-025-MY2.

REFERENCES

[1] B. Milner and J. Darch, Robust acoustic speech feature prediction from noisy mel-frequency cepstral coefficients, *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 338-347, 2011.

[2] L. Buera, A. Miguel, O. Saz, A. Ortega, and E. Lleida, Unsupervised data-driven feature vector normalization with acoustic model adaptation for robust speech recognition, *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 296-309, 2010.

[3] J. W. Hung and W. H. Tu, Incorporating codebook and utterance information in cepstral statistics normalization techniques for robust speech recognition in additive noise environments, *IEEE Signal Processing Letters*, vol. 16 ,no. 6, pp. 473-476, 2009.

[4] C. Wan and L. Liu, Research and improvement on embedded system application of DTW-based speech recognition, *Proc. of International Conference on Anti-counterfeiting, Security and Identification*, pp. 401-404, 2008.

[5] J. T. Chien and M. S. Lin, Frame-synchronous noise compensation for hands-free speech recognition in car environments, *IEE proceedings. Vision, image and signal processing*, vol. 147, no. 6, pp. 508-515, 2000.

[6] Y. Zhan, L. H., K. C. Kwak and H. Yoon, Automated speaker recognition for home service robots using genetic algorithm and dempsterVshafer fusion technique, *IEEE Trans. Instrumentation and Measurement*, vol. 58, no. 9, pp. 3058-3068, 2009.

[7] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms 2nd Editon*, McGraw-Hill, 2002.

[8] X. Huang, A. Acero, and H. Wuenon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*,Prentice Hall PTR Upper Saddle River, USA, 2001.

[9] J. Tao, L. Xin and P. Yin, Realistic visual speech synthesis based on hybrid concatenation method, *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 469-477, 2009.

[10] S. Kwong and C. W. Chau, Analysis of parallel genetic algorithms on HMM based speech recognition system, *IEEE Trans. Consumer Electronics*, vol. 43, no. 4, pp. 1229-1233, 1997.

[11] Y. Tsao and C. H. Lee, An Ensemble Speaker and Speaking Environment Modeling Approach to Robust Speech Recognition, *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 1025-1037, 2009.

[12] C. W. Chau, S. Kwong, C. K. Diu, and W. R. Fahrner, Optimization of HMM by a genetic algorithm, *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1727-1730, 1997.

[13] F. Sun and G. Hu, Speech recognition based on genetic algorithm for training HMM, *Journal of Electronics Letters*, vol. 34, no. 16, pp. 1563-1564, 1998.

[14] K. J. Won, A. Prugel-Bennett and A. Krogh, Training HMM structure with genetic algorithm for biological sequence analysis, *Journal of Bioinformatics*, vol. 20, no. 18, pp. 3613-3619, 2004.

[15] Z. Linghua, Y. Zhen, and Z. Baoyu, A new method to train VQ codebook for HMM-based speaker identification, *Proc. of International Conference on Signal Processing*, pp. 651-654, 2004.

[16] X. Y. ZHANG, J. P. Wu, Y. W. ZHANG, and Q. S. ZHANG, Optimum vector quantization codebook design for speaker recognition, *Proc. of International Conference on Signal Processing*, pp. 1397-1402, 2004.

[17] R. L. Haupt and S. E. Haupt, *Practical Genetic Algorithms*,Wiley Interscience, USA, 2004.

[18] F. T. Lin, Evolutionary computation part 2: genetic algorithms and their three applications, *Journal of Taiwan Intelligent Technologies and Applied Statistics*, vol. 3, no. 1, pp. 29-56, 2005.

[19] P. Blunsom, *Hidden Markov Model*,The University of Melbourne, Department of Computer Science and Software Engineering, 2004.
`http://www.cs.mu.oz.au/460/2004/materials/hmm-tutorial.pdf`

[20] N. E. Huang, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, *Proc. of the Royal Society A: Mathematical, Physical and Engineering Sciences*, pp. 903-995, 1998.

[21] J. H. Mathews and K. D. Fink, *Numerical Methods Using MATLAB,* 4^{th} *Edition*, Prentice-Hall, 2004.

[22] H.G. Hirsch and D. Pearce, The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, *Proc. of The ISCA Tutorial and Research Workshop on Automatic Speech Recognition: Challenges for the new Millennium* , 2000.

[23] C. W. Hsu and L. S. Lee, Higher order cepstral moment normalization for improved robust speech recognition, *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 205-219, 2009.