Effect of HMM Parameter on Robustness of Chinese Isolated Words Recognition

Liu De, Wang Mingjiang, Wu Zejun[1]

Abstract. The characteristics of speech signals are stable in short-term but unstable in long-term, so speech sequence modeling method based on Markov Chain can more effectively represent the feature of speech signals. According to some basic modeling unit, this method can also constructs sentence model of continuous speech, and the accuracy and flexibility of this method are relatively high. This paper firstly constructs a Chinese isolated word speech system whose vocabulary is small, and made the recognition rate achieve 100%. Secondly, this paper changes the number of HMM (Hidden Markov Model) states and the number of HMM observation symbols to test the effect that these two parameters have on the recognition robustness. The experimental results show that increasing the number of HMM states and the number of observation symbols can improve the robustness of isolated word speech recognition. When the values of these two parameters are greater than some certain value, continuing to increase these two parameters has no obvious effect on the recognition robustness.

Keywords: Isolated Words Recognition, Number of HMM States, Number of HMM Observation Symbols, Speech Recognition Robustness

[1] Liu De (🖂)
Harbin Institute of Technology Shenzhen Graduate School, 518055, Shenzhen, China
e-mail: liude19832006@126.com

Wang Mingjiang
Key Laboratory of The Technology of The Internet of Things, 518055, Shenzhen, China

Wu Zejun
Harbin Institute of Technology Shenzhen Graduate School, 518055, Shenzhen, China

*Funded by: The Key Laboratory Projects of Shenzhen (CXB201104220018A)

1 Introduction

Speech recognition technology has made tremendous progress since 1960s and is applied in wide range nowadays, from personal computer to mobile devices such as tablet PC, smart phone. So its market prospect is very broad. Speech recognition can be classified as 3 types: speaker recognition, continuous speech recognition and isolated words recognition [1][2]. This paper is mainly concerned about isolated words recognition (IWR). The process of isolated words recognition is illustrated in **Fig.1**.

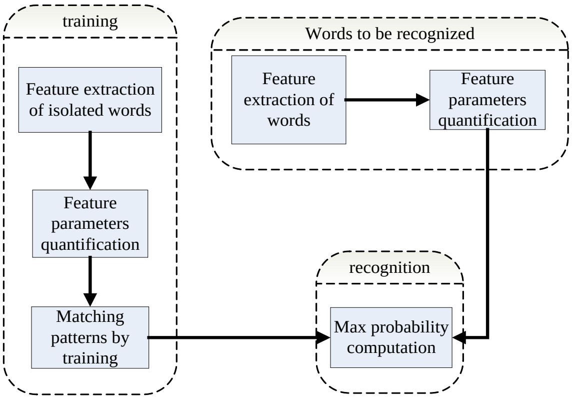
training
Feature extraction of isolated words
Feature parameters quantification
Matching patterns by training
Words to be recognized
Feature extraction of words
Feature parameters quantification
recognition
Max probability computation

Figure 1. The process of IWR

In Fig.1, the generation of matching patterns is the most complex step. The matching pattern is just a kind of acoustic model. The method of generating acoustic model is to use training algorithm to process the extracted features of speech signals. Hidden Markov Model (HMM) [1] which evolved from Markov Chain is a probability model represented by parameters, and can be used to describe the statistical properties of the random process. HMM model has always been a research focus and its application range has extended to various fields of speech processing. HMM model is a statistical probability model, so the training of HMM needs a large amount of speech data to generate a convergent probability model. The recognition process uses this model to compute the maximum likelihood probability to output the optimal state sequence.

For speech recognition based on HMM, each word to be recognized has one corresponding HMM model. Generally, we use HMM training algorithm to process the feature parameters of speech signals to obtain a set of HMM models. The feature parameters can be Linear Predicted Coefficients (LPCC) [3] or Mel-Frequency Cepstral Coefficients (MFCC). In this paper, we use MFCC [4] as the feature parameters. The recognition procedure combines the HMM models and feature parameters of words to be recognized, computes the maximum likelihood probability to output the recognition results.

The HMM model has two very important parameters, the number of HMM states, denoted as *N*, and the number of HMM observation symbols, denoted as *M*. This paper optimized the above two parameters in the recognition of Chinese isolated words. Based on the recognition rate of 100%, by changing the value of *N* and *M*, we observe how these two parameters effects the robustness of recognition. Firstly, we determined the basic topology of the HMM models, including the number of HMM states, the length of sequence of observation. Secondly, we adjusted the parameters of HMMs, so as to precisely imitate each word's corresponding acoustic features.

2 The Establishment of HMM Model

2.1 The Basic Theory and Algorithm of HMM

The HMM model used in this paper is discrete HMM. A finite state machine in discrete time field is a simple HMM model. At any discrete time, the finite state machine stays at one state, and can jump from the current state to any state with some probability. The simplest HMM is illustrated in Fig.2 [5].

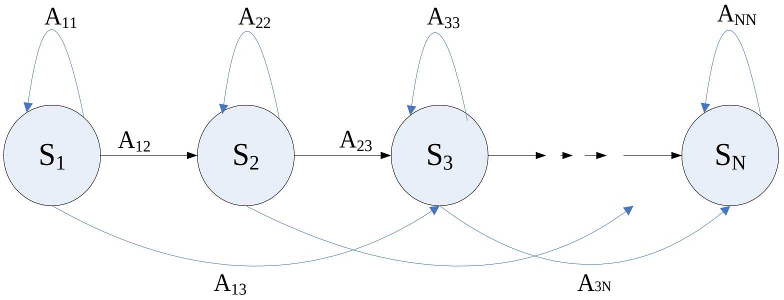

Figure 2. The discrete Markov process

The feature parameters of HMM are defined as following [1]:

(a) *N*, the number of HMM states. Although the states are hidden in HMM model, they have definite physical meanings in practical applications. In the later equations, each state is marked as {1, 2, 3, … , *N*-1, *N*}, the state at time *t* is marked as q_t.

(b) *M*, the number of observable symbols in each state. Each observation symbol is marked as $V=\{ v_1, v_2, v_3, \dots , v_M \}$, the sequence of observation is marked as $O=\{ o_1, o_2, o_3, \dots, o_T \}$, o_t is a observation symbol in set *V*, *T* is the length of sequence of observation.

(c) Probability distribution of state transition , $A=[a_{ij}]$,

$$a_{ij} = p[q_{t+1} = j \mid q_t = i] \quad 1 \le i \le N, 1 \le j \le N \tag{1}$$

(d) Probability distribution of sequence of observation, $B=[b_j(k)]$

$$b_j(k) = p[o_t = v_k \mid q_t = j] \quad 1 \le k \le M, 1 \le j \le N \tag{2}$$

(e) Probability distribution of initial state, $\pi=[\pi_i]$,

$$\pi_i = P[q_1 = i] \quad 1 \le i \le N \tag{3}$$

Based on these feature parameters, the sequence of observation $O=\{ o_1, o_2, o_3, \dots, o_T \}$ can be described as following:

(a) According to the probability distribution π, let the initial state $q_1=i$;

(b) Let the observation time $t=1$;

(c) According to the probability distribution *B* of observation symbols in current state, let $o_t=v_k$;

(d) According to the probability distribution *A*, let the state jump from current state $q_t=i$ to the next state$q_{t+1}=j$;

(e) Let $t=t+1$, if $t<T$ (the time sequence of observation is $t=1, 2, \dots , T$), then go back to (c) step, otherwise ends this step.

In summary, the HMM model of any isolated word can be represented by 2 parameters and 3 probability distribution matrix π, *A*, *B*. The HMM model is usually defined as $\lambda=(A, B, \pi)$.

This paper uses Baum-Welch algorithm to compute the above 3 matrixes. The *N* and *M* are fixed, *A*, *B* and π are computed as (4), (5) and (6).

$$\overline{\pi}_i = \frac{\alpha_1(i)\beta_1(i)}{\sum_{j=1}^{N} \alpha_1(i)\beta_1(i)} \tag{4}$$

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(o_t) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)} \tag{5}$$

$$\overline{b_j(k)} = \frac{\sum_{t=1, o_t=k}^{T} \alpha_t(i)\beta_t(i)}{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)} \tag{6}$$

In the above three equations, $\overline{a}_{ij}$, $\overline{b_j(k)}$ and $\overline{\pi}_i$ are the normalized values to replace a_{ij} , $b_j(k)$ and π_i .

In (5) and (6), $\alpha_t(i)$ is the forward probability at time *t* in state *i*; $\beta_t(i)$ is the backward probability at time *t* in state *i*, both of whose calculating method are iterative.

The forward probability is the probability computed in such condition that the λ is given, the state is *i* at time *t*, and the sequence of observation is {*o*1, *o*2, *o*3, …, *ot*} at time *t.* (7) and (8) are the computing formulas.

$$\alpha_1(i) = \pi_i b_i(o_1) \quad 1 \le i \le N \tag{7}$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] \cdot b_j(o_{t+1}) \;\; 1 \le t \le T-1, 1 \le i \le N, 1 \le j \le M \tag{8}$$

The backward probability is the probability computed in such condition that λ is given, the state is *i* at time *t*, and the sequence of observation is {*o*t+1, *o*t+2, …, *o*T} from time *t*+1 to time *T*. (9) and (10) are the computing formulas.

$$\beta_T(i) = 1 \quad 1 \le i \le N \tag{9}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \;\; t = T-1, T-2, \ldots, 1; 1 \le i \le N; 1 \le j \le M \tag{10}$$

Please be noted, because the computing process is iteration, the values of a_{ij} and $b_j(k)$ are the values computed in previous step of computing $\alpha_t(i)$ and $\beta_t(i)$. When to compute $\alpha_t(i)$ and $\beta_t(i)$ in the first iteration, a_{ij} =1, $b_j(k)$ =1.

2.2 The Implementation of HMM Parameters

The structure of HMM model includes *N*---the number of HMM states and *M*---the number of observation symbols. For isolated words recognition, the number of HMM states can be decided by the length of the speech. For discrete HMM, the number of observation symbols is decided by sample space, but limited by the computation complexity, we take 16~32 as the values. We take the left-to-right structure model as the shape of Markov chain.

In the process of generating HMM parameters by Baum-Welch algorithm, an important question is how to select the initial model. When $P(O \mid \lambda)$ reaches its maximum value, the values of HMM parameters are optimal. Generally speaking, the initial values of π and *A* have nearly no effect on the final HMM models, so we can use random values as their initial values, only satisfies the properties of probability. But the initial values of *B* have an effect on the final HMM models, so we use a relatively complex method to select the initial values of *B*.

In this paper, we take N=20 as the initial value of the number of HMM states, T=32 as the value of the length of observation sequence. The number T is just the number of codebook in the process of feature parameters quantification. And at the same time, let M=20 be the initial value of the number of observation symbols. On the above condition, we can generate a probability distribution matrix A for the state transition with the size of 20x20; a probability distribution matrix B for the observation sequence with the size of 20x32; an initial probability distribution π with the size of 1x20.

The sequence of observation in this paper is computed as following. For example, the Chinese word “ZhongGuo”, suppose there are T useful frames after feature extraction, and each frame is a 39 dimensional vector. After vector quantification, each frame will be numbered with a numeric. The T numeral numbers construct a sequence of observation.

3 Experimental Results and Discussion

This paper uses 32 words for training to establish the matching patterns (HMM model). And each word has 40 speech signal files—20 males and 20 females reading the same words. So totally 1280 speech files are used to extract the MFCC, establish the codebook and model the matching patterns (HMM). To verify the recognition rate and observe the robustness, 12 words are used. Each test word is read by 6 males and 6 females, so the test set has totally 384 speech files.

In this paper, we use self adaptive method to change the values of N and M. By observing the recognition rate and the robustness, we can find the optimal values of N and M for each word. Specifically, for each word, in the process of establishing HMM models, increase its N from 12 to 24, the step size is 2, at the same time the value of M is fixed. Then fix the value of N, increase the value of M from 16 to 32, and the step size is 2.

The experimental results show that the recognition rate is 100% when the value of N varies from 14 to 24, and the value of M varies from 22 to 32. The key indicator of evaluating the recognition robustness is the difference (denoted as $\triangle P=P_1-P_2$) between the first recognition probability and the second recognition probability of each word. The greater the difference, the better the recognition robustness. In such case, when there are other interference factors, such as noise or the pronunciation may be not clear, the recognition result is still right. Conversely, if the difference ($\triangle P$) is relatively small and there is interference, the difference may become negative. That is to say, the second recognition probability becomes the first recognition probability, the recognition result is wrong.

3.1 The Effect of the Number of HMM States On Recognition Robustness

When the number of HMM states---N changes, the difference (△P) between the first recognition probability (P_1) and the second recognition probability (P_2) is illustrated in **Fig.3**. The horizontal axis represents the serial number of each word, in Fig.3, the vertical axis represents the value of △P; in **Fig.4**, the vertical axis represents the average value ($\overline{P}_1$) of the first recognition probability P_1; in **Fig.5**, the vertical axis represents the variance value of the first recognition probability P_1. The black curve with rectangle point on it means that the value of N equals to 14 and the red curve with circle point on it means that the value of N equals to 16 and so on, see Fig.3, Fig.4 and Fig.5.

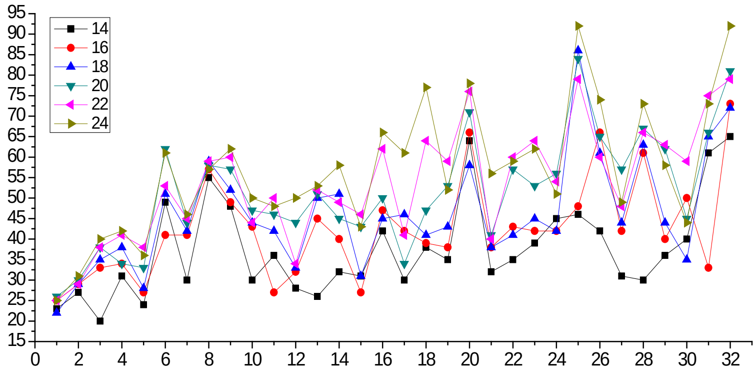

Figure 3. △P with different N

From Fig.3, we can see that the △P becomes greater as the number of HMM states (N) increases, which means that the recognition robustness has been improved. For example, the pink curve is above the green curve almost everywhere. But for some individual words, the △P becomes less as N increases, which means that each word has its own optimal N.

From Fig.4, we can see that the first recognition probability of nearly every word becomes greater as N increases. Cause the value of recognition probability is negative, the less the absolute value, the greater the probability. For example, the curve with yellow forward triangle point on it is below the curve with pink backward triangle point on it, which means that the $\overline{P}_1$ of each word becomes greater as N increases.

From Fig.5, we can see that the variance of P_1 does not always become greater as N increases. The main reason is that as N increases, the denominator of (5) becomes greater, which makes the backward probability decrease. But this does not imply the recognition robustness gets worse, for only △P is the representative of recognition robustness. In fact, we know that the recognition robustness becomes better as N increases, see Fig.3.

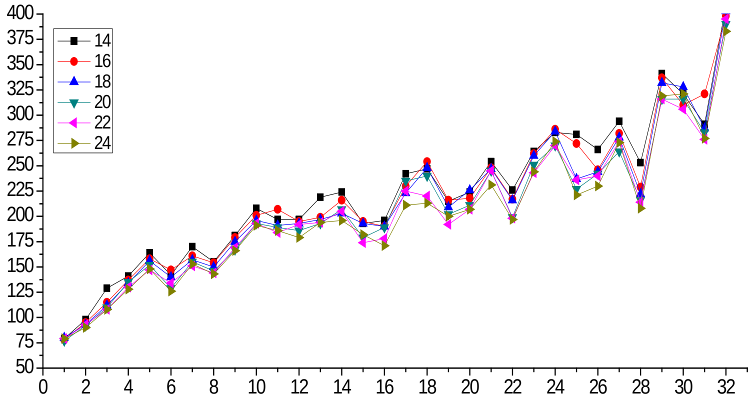

Figure 4. The means of P_1 with different N

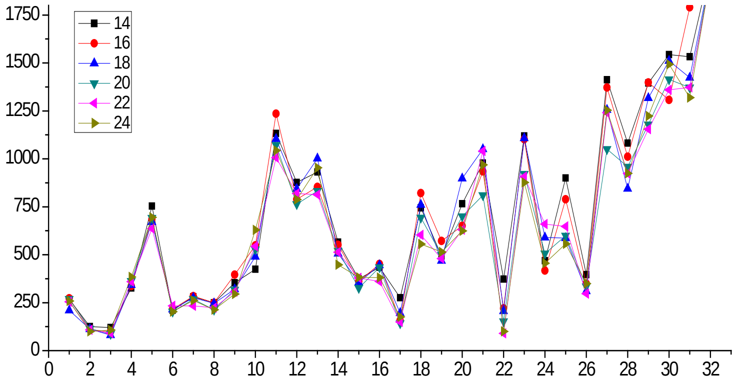

Figure 5. The variance of P_1 with different N

3.2 The Effect of the Number of Observation Symbols On Recognition Robustness

When the number of observation symbols---M changes, the difference ($\triangle P$) between the first recognition probability(P_1) and the second recognition probability(P_2) is illustrated in **Fig.6**. The horizontal axis represents the serial number of each word, in Fig.6, the vertical axis represents the value of △P; in **Fig.7**, the vertical axis represents the means ($\bar{P}_1$) of first recognition probability P_1; in **Fig.8**, the vertical axis represents the variance value of first recognition probability. The curve line with rectangle point on it means the value of M equals to 22 and the curve line with red circle point on it means the value of N equals to 24 and so on, see Fig.6, Fig.7 and Fig.8.

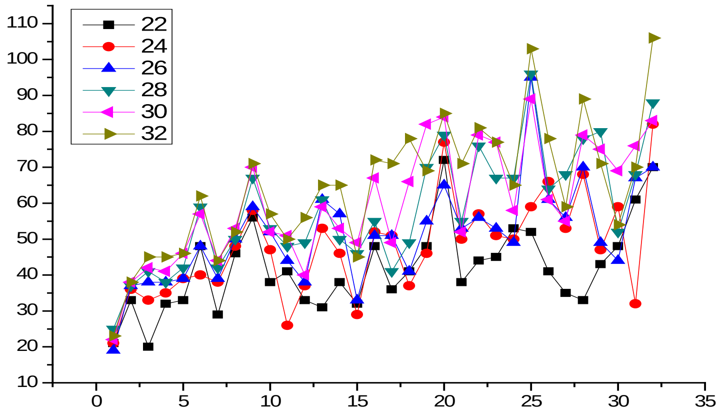

Figure 6. △P with different M

From Fig.6, we can see that as the number of observation symbols increases, the relationship between different states can be reflected more clearly, so the △P of each isolated word becomes greater. For example, the yellow curve with forward triangle on it is above the pink curve with backward triangle point on it. Even there is some unpredictable disturbance such as noise, the system can still correctly recognize each word. That is to say, the recognition robustness is improved, especially for some words, such as the 16th, 17th, 18th, 32th word.

From Fig.7, we can see that the first recognition probability (P_1) of each word also becomes great as M increases. For probability, its value is negative, so the less the absolute value, the grater the recognition probability. For example, the yellow curve with forward triangle on it is below the pink curve with backward triangle on it. For some individual word, the increase of P_1 can be 50~100, such as word 17, word 23 and word 29.

From Fig.8, we can see that the variance of P_1 becomes less as M increase. This implies that the first recognition probability becomes more stable.

For isolated words speech recognition, based on the average recognition rate unchanged, we should as possible as to improve the recognition robustness. From the experimental results, we know that increasing the number of HMM states and the number of observation symbols is a relatively good way. But when the number of HMM states and the number of observation symbols exceed to some certain values, continuing to increase these two parameters will not help to improve the recognition robustness. So by the method of iterative computation, we can find the optimal values of the above two parameters for each isolated word.

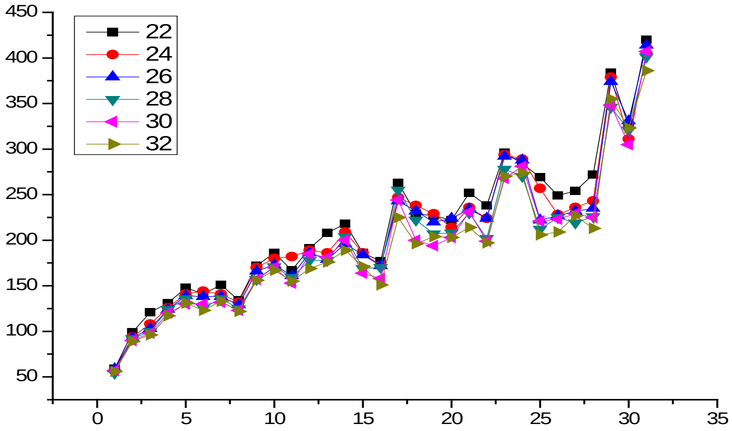

Figure 7. The means of P_1 with different M

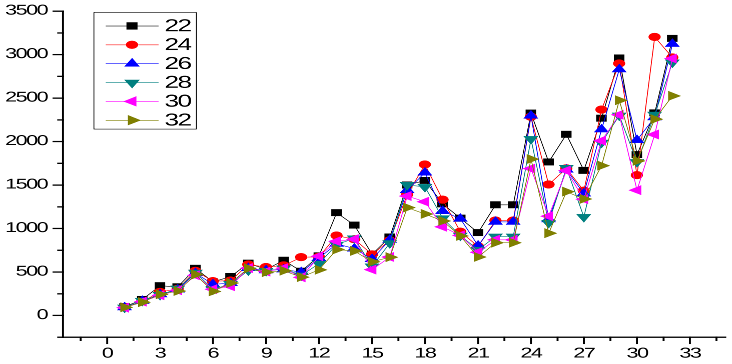

Figure 8. The variance of P_1 with different M

4 Conclusions

This paper uses HMM models as the matching patterns in Chinese isolated words speech recognition. By changing the values of the number of HMM states and the number of observation symbols, we researched the effect that these two parameters have on the recognition rate and recognition robustness. The experimental results show that the number of HMM states and the number of observation symbols have no effect on recognition rate within a certain range, but have obvious effect on recognition robustness. When the values of these two parameters are beyond a certain range, the recognition robustness does not improve anymore, the recognition rate even decreases. There is one another question we need to consider that the memory space for storing state transition matrix and observation probability matrix becomes large dramatically as the number of HMM states and the number of observation symbols increase. So we need to consider it comprehensively, to achieve the balance of storage space and accuracy.

References

1. Antonio M. Peinado, Jose C. Segura, *Speech Recognition Over Digital Channels—Robustness and Standards*, John Wiley & Sons, Ltd, 2006.
2. John Holmes, Wendy Homes, S*peech Synthesis and Recognition*, 2nd ed., London and New York, 2001.
3. Assaleh K T, Mammone R J. “New LP-derived features for speaker identification”. IEEE Trans. on Speech and Audio Proc., 1994, 2(4): 630-637.
4. Fatma zohra. Chelali, Amar. DJERADI, “MFCC and vector quantization for Arabic fricatives Speech/Speaker recogniton”, Multimedia Computing and Systems, 2012: 284-289
5. Jianing Dai, “Isolated Word Recognition Using Markov Chain Models”, IEEE Trans. on Speech and Audio Processing., Vol. 3, No. 6, 1995: 458-463.