HMM speech recognition with reduced training

Foo, Say Wei; Yap, Timothy

1997

Foo, S. W. & Yap, T. (1997). HMM speech recognition with reduced training. Proceedings of the International Conference on Information, Communications and Signal Processing 1997: (pp. 1016-1019).

https://hdl.handle.net/10356/91395

https://doi.org/10.1109/ICICS.1997.652134

International Conference on
Information, Communications and Signal Processing
ICICS '97
Singapore, 9-12 September 1997

HMM Speech Recognition with Reduced Training

Say Wei Foo, Timothy Yap
Department of Electrical Engineering
National University of Singapore
10 Kent Ridge Crescent, Singapore 119260

Abstract

One of the problems faced in automatic speech recognition is the amount of training required to adapt the machine to the speaker way of pronunciation. To a certain extent, the accuracy of correct recognition is proportional to the amount of training and adaptation carried out. This is especially true when a large vocabulary is involved. For certain applications, it is desirable that the training requirement be reduced to the bare minimum without sacrificing the accuracy of recognition. In this paper, the minimum number of training required to achieve an acceptable degree of accuracy for a speaker dependent speech recognition system based on the Hidden Markov Model (HMM) is investigated. A method is also proposed which retains the same degree of accuracy of recognition with much reduced training.

1. Introduction

In speech recognition, a distinction is generally made between the recognition of utterances from a speaker who has previously 'enrolled' his voice (speaker dependent recognition) and a speaker whose voice the recognizer has never 'heard' previously (speaker independent recognition). Speaker dependent recognition finds use in situations where the system is designed for known speakers. This kind of system can be used for applications such as computer voice dictation and voice activated dialing for telephone. Speaker independent systems are used in situations where a large number of people are likely to have access to the system. Generally, the recognition error rate of speaker dependent systems is lower than that of speaker independent ones. However, most speaker dependent systems require the speakers to spend considerable time to train the systems in order to achieve high accuracy. For certain applications, it is desirable to minimize this task of enrollment.

Hidden Markov Modeling (HMM) is one of the most widely used methods for the automatic recognition of spoken utterances [1,2]. The purpose of our study is to determine within a practical framework how a speaker dependent system may be realized with minimal training without sacrificing the accuracy.

In Section 2, a quick review of HMM for isolated word recognizer is presented to establish the notation and terminology for subsequent discussions. In Section 3, an HMM based speech recognizer system is described. The proposed technique of reduced training is discussed in Section 4 followed by concluding remarks in Section 5.

2. HMM for Isolated Word Recognition

The type of HMM that is used in this paper is known as the left-right Bakis model with only single transitions allowed. This model is shown in Fig. 1.

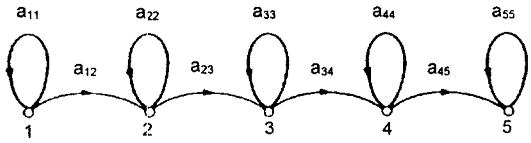

Fig. 1 Left-right model with only single transitions allowed

The Markov model is characterized by the following parameters:

i. N = the number of states in the model.

ii. M = the number of output symbols in the discrete alphabet of the model.

iii. A = $\{a_{ij}\}$, the transition matrix of the underlying Markov chain. Here, a_{ij} is the probability of making a transition to State j, given that the model is in State i.

iv. B = $\{B_{jk}\}$ = $\{b_j(k)\}$, the model output symbol probability matrix, where $b_j(k)$ is the probability of outputting symbol k, given that the model is in State j.

v. $\pi = \{\pi_i\}$, i = 1,2,...,N, the initial state probability vector. For the left-right model shown in Fig. 1, we assume that the system always begins in State 1, i.e., $\pi_1 = 1$, $\pi_i = 0$, $i \neq 1$.

0-7803-3676-3/97/$10.00 © 1997 IEEE

Isolated word recognition using HMM consists of two phases: training and recognition (or classification). In the training phase, the training set of observations is used to derive a set of reference models of the above type, one for each word in the vocabulary. In the classification phase, the probability of generating the test observation is computed for each reference model. The test is classified as the word whose model gives the highest probability.

2.1 Recognition Algorithm

Given the observation sequence **O** = {o1, o2, ..., oT} and the model λ (i.e., N, M, A, B, π), the probability of **O** having been generated by model λ is

$$P(\mathbf{O}|\lambda) = \sum_{all\, q} P(\mathbf{O}|q,\lambda) P(g|\lambda)$$
$$= \sum_{q_1, q_2, K, q_T} \pi_{q_q} b_{q_1}(\mathrm{o}1) a_{q_1 q_2} b_{q_2}(\mathrm{o}2) a_{q_2 q_3} \Lambda\ a_{q_{T-1} q_T} b_{q_T}(\mathrm{o}_T) \tag{1}$$

The summation in Eq. (1) is more readily calculated by introducing a forward partial probability variable $\alpha_t(i)$, as

$$\alpha_t(i) = P(o_1\ o_2\ \ldots\ o_t,\ q_t = i|\lambda) \tag{2}$$

and applying the iteration procedure below.

1. Initialization

$$\alpha_1(i) = \pi_i b_i(o_1), \qquad 1 \le i \le N.$$

2. Recursion

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(\mathbf{o}_{t+1}) \tag{3}$$

Eq. (1) can then be determined from

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{4}$$

The summation in Eq.(1) can also be calculated using backward partial probability variable $\beta_t(j)$

$$\beta_t(i) = P(o_{t+1}\ o_{t+2}\ \ldots\ o_T \mid q_t = i, \lambda) \tag{5}$$

and following the steps detailed below.

1. Initialization

$$\beta_T(i) = 1$$
$$i=1,..,N. \tag{6}$$

2. Induction

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{o}_{t+1}) \beta_{t+1}(j)$$

$$t = T - 1, T - 2, \ldots, 1, \text{ and } i=1,..,N \tag{7}$$

3. Termination

$$P(\mathbf{O}|\lambda) = \sum_{j=1}^{N} \beta_1(j) \pi_j b_j(\mathbf{o}_1) \tag{8}$$

2.2 Training Algorithm

The objective in the training phase is to optimize the probability P(**O**|λ), where **O** is the observation sequence and λ is the model to be trained. One of the algorithms commonly used is the Baum-Welch re-estimation algorithm [3]. In most cases, training observations come in sets, since a single training observation is insufficient for the training of HMMs. Thus, the batch training version of this algorithm is almost always used. The formula for the re-estimation is reproduced here,

$$a_{ij} = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i) a_{ij} b_j\left(\mathbf{o}_{t+1}^{(k)}\right) D_{t+!} \beta_{t+1}^k(j)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i) \beta_t^k(i)} \tag{9}$$

$$b_j(l) = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i) \beta_t^k(i)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i) \beta_t^k(i)} \tag{10}$$

3. The System

The block diagram of a HMM based speech recognition system is shown in Fig. 2.

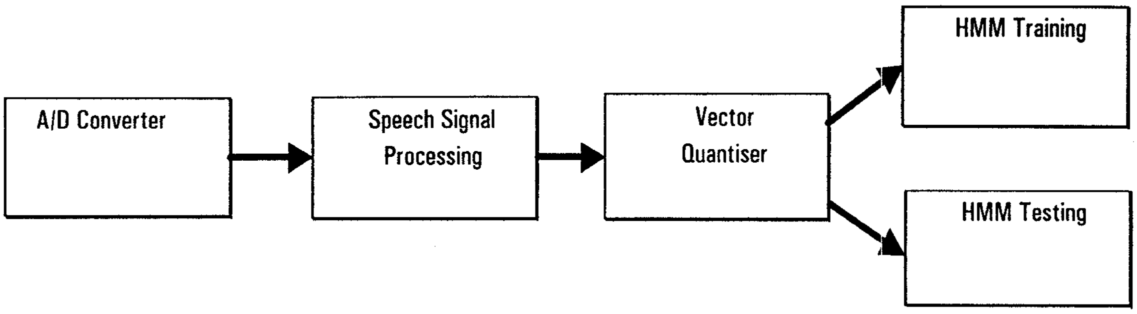

Fig. 2 Structure of HMM based recognizer.

The recording of a spoken word normally consists of a segment of speech that is buffered on both sides by comparatively long periods of silence and sometimes sound artifacts such as mouth clicks as shown in Fig. 3. It is thus desirable to isolate the actual spoken word from the silence periods. A simple two threshold forward–backward algorithm [4] was developed to determine the speech end-points.

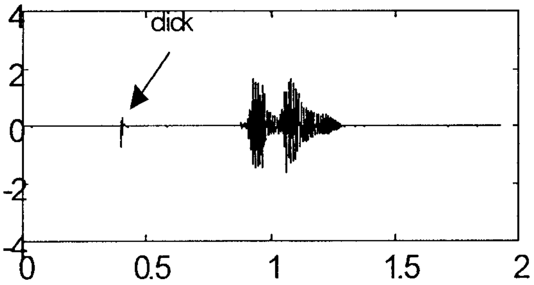

Fig 3. Example of a sound artifact in the recorded speech signal.

The next step in the signal processing stage is to window the signal into 20ms frames with an overlap of 10ms. An 8-pole LPC analysis is performed on each frame, and cepstral vector of length 12 is obtained. To simplify computational load, these cepstral vectors are coded using vector quantization. For the system proposed in this paper, the code book is obtained using the K-means algorithm [5]. It is found that the size of the codebook used in the vector quantization process plays an important part in the performance (i.e. recognition accuracy) of the system. A code book size of 306 is chosen as it provides a reasonable compromise between the magnitude of distortion and complexity.

Thus the variable M is 306 for the Hidden Markow Model. N, the number of states, is chosen to be 6. It is found that this number is sufficient for isolated word recognition.

4. Reduced Training

Experiments were carried out to assess the amount of training required for accurate recognition. A vocabulary of 20 words was chosen. 50 utterances of each word were recorded from a single speaker. The codebook was established using 400 utterances: 20 for each of the 20 words.

The HMMs were then batch trained (Eq. (1)) with different numbers of utterances from 1 to 40 for each word and tested with the remaining 10 utterances. It was found that the recognition accuracy increases with the number of utterances used for training. A recognition accuracy of 99.5 % is achieved when all 40 utterances were used for training.

To achieve an accuracy of 99.5%, a training set of 40 utterances per word is required. From a practical standpoint, this number is far too large. It would be very inconvenient for users to provide such a large number of training utterances before the recognizer can be used with confidence.

There are two possible solutions to this problem. Firstly, a pre-trained speaker independent recognizer could be used. Alternatively, a new technique to reduce the amount of training data required would be ideal.

However, the performance of a speaker independent recognizer is deemed to be poorer than a speaker dependent system. This is especially true when comparison is made between a system that is optimized for the speaker and one that is not. The challenge would be to devise a system that performs like a speaker dependent system, yet requires minimal training to optimize for a particular speaker. In this paper, a technique attempting to meet this challenge is proposed for HMM based system. The technique adopts a two-tier approach: a base line speaker independent system is first developed which is further optimized by an individual speaker.

The formulae used for training, Equations (5) and (6), to obtain a_{ij} and $b_j(l)$, consist merely of running sums taken from k = 1 to k = K, where k is the observation number and K is the total number of observation sequences in the batch. By continuing to add to these running sum, the system is optimized for the additional training sequences.

The training procedure the two tier approach is then as follows. First, generic models of each word in the vocabulary are developed by batch training with speech samples from a variety of sources and in sufficient quantity to achieve high recognition. The running totals are retained as bases. When the system is to be adapted to a particular speaker, training utterances are recorded and these are used to update the running totals and the relevant parameters re-estimated. As the training samples from the users are obviously more important than the samples obtained for the generic model, more weights will be accorded to the training samples obtained from the user. This is achieved by multiplying the sums obtained from the user data by a weighing factor, and add them to the previous totals to get the new estimates for a_{ij} and $b_j(l)$.

Experiments were conducted to test the effectiveness of this approach and to assess the additional amount of training required to achieve a given degree of accuracy. For the experiment, the initial estimates used were the models that were trained earlier (let's call this speaker - Speaker 1). A single utterance of each word was taken from another speaker (Speaker 2), and used for the adaptive training process. Subsequently, another three samples of each word were collected from Speaker 2 for testing purposes. Out of the 30 tests, 23 were correctly recognized.

Additional utterances were then collected from Speaker 2 for the purpose of adaptive training. The same 3 test utterances were maintained for future tests. It was found that when the adaptive training set was increased to a size of 4, 100% accuracy was achieved. A graph of the number of errors against the adaptive training size is shown in Fig. 4. This size of training set is far less (90% less) than the 40 that was required to obtain a similar accuracy when we built the models from scratch.

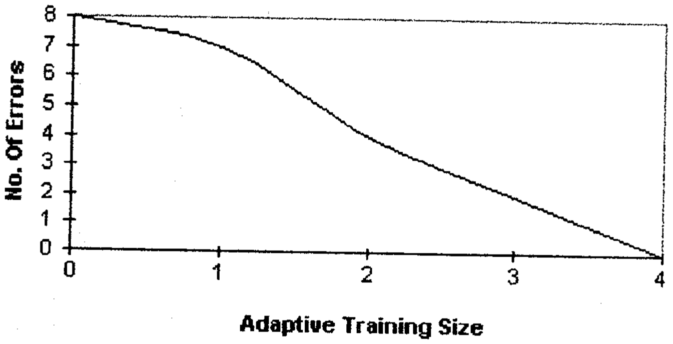

Fig.4 Number of recognition errors in 30 tests as the adaptive training size is increased.

5. Conclusion

A HMM based speaker dependent speech recognizer with reduced training is proposed. It is known that accurate speech recognizer based on HMM could be obtained if a sufficiently large training set could be obtained. To reduce the amount of training while maintaining the accuracy, a two-tier approach is proposed. In this approach, the model trained with a large generic training set is used as a base, speaker dependent training samples are added to this base with more weights given to these samples.

Experiments show that using the two-tier approach, the number of training sets required can be significantly reduced without compromising the accuracy of recognition.

References

[1] S.J. Cox, "Hidden Markov Models for Automatic Speech Recognition: Theory and Application, Computer Recognition of Speech 209-229.

[2] L.R. Rabiner, "A Tutorial in Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE (Feb 1989), pp. 257-86

[3] L.E.Baum, "An Inequality and Associated Maximisation Technique in Statistical Estimation for Probabilistic Functions of a Markov Process", Inequalities,3 (1972),pp.1-8

[4] S.H.Yap,"Hidden Markov Model Appioach to Speech recognition", Thesis 1997.

[5] Robert M.Gray, "Vector Quantization," IEEE ASSP Mag. (April 1984), pp. 4-26.