Temporal Hidden Markov Models

Dat Tran

School of Information Sciences and Engineering
University of Canberra, ACT 2601, Australia

ABSTRACT

Hidden Markov model (HMM) is a doubly stochastic process. The observable process produces a sequence of observations and the hidden process is a Markov process. The HMM assumes that the occurrence of one observation is statistically independent of the occurrence of the others. To avoid this limitation, the temporal HMM is proposed. The hidden process in the temporal HMM is the same but the observable process is now a Markov process. Each observation in the training sequence is assumed to be statistically dependent on its predecessor and codewords or Gaussian components are used as states in the observable Markov process. Speaker identification experiments performed on 138 Gaussian mixture speaker models in the YOHO database shows a better performance for the temporal HMM compared to the standard HMM.

1. INTRODUCTION

The hidden Markov model (HMM) approach is the well-known and widely used statistical method of characterizing the spectral properties of the time frames of a speech pattern [6]. There are two assumptions in the first-order HMM. The first is the Markov assumption, i.e. a new state is entered at each time t based on the transition probability, which only depends on the previous state. It is used to characterise the sequence of the time frames of a speech pattern. The second is the output-independence assumption, i.e. the output probability depends only on the state at that time regardless of when and how the state is entered [4]. A process satisfying the Markov assumption is called a Markov model. An observable Markov model is a process where the output is a set of states at each instant of time and each state corresponds to an observable event. The HMM is a doubly stochastic process with an underlying Markov process which is not directly observable (hidden) but which can be observed through another set of stochastic processes that produce observable events in each of the states [6].

The HMM-based training methods have become widely applied in speech recognition, voice authentication, on-line (dynamic) handwriting recognition, signature authentication, and face recognition systems. However, there is a limitation of this approach. The HMM assumes that the occurrence of one feature in the training data is statistically *independent* of the occurrence of the others. This assumption is not appropriate for speech or handwriting recognition because a spoken or written word is represented as a time series of features and hence the features are correlated in time. The proposed temporal models can avoid this limitation of the HMM.

In order to represent that correlation, the use of codewords in a codebook obtained by vector quantization (VQ) modeling as states of a Markov chain was developed [1] for isolated word recognition. The proposed research extends this idea to a general framework and hence it can apply to HMMs, Gaussian mixture models (GMMs) and their fuzzy versions. In he proposed approach, each codeword in the codebook or each Gaussian component in the Gaussian mixture model is a specific state of the Markov chain. The state-transition probabilities of the Markov chain are used to represent the dependence between acoustic features. For example, if codeword v_k comes after codeword v_i, it is considered that there is a probability of transition from state v_i to state v_k..

The temporal HMM is the standard HMM using the above-mentioned Markov chain approach. Each feature vector in the training sequence is assumed to be statistically *dependent* on its predecessor in the proposed temporal model. In the HMM, observations in the sequence O are assumed to be *independent*. Therefore the temporal model has avoided the limitation of the HMM. A simpler version of the temporal HMM is the temporal GMM, which is also presented in this paper.

Speaker identification experiments performed on 138 Gaussian mixture speaker models in the YOHO database shows a better performance for the temporal HMM compared to the standard HMM.

2. HIDDEN MARKOV MODEL

Let $S=\{s_1, s_2, ..., s_T\}$, $O=\{o_1, o_2, ..., o_T\}$, and $V=\{v_1, v_2, ..., v_M\}$ be a state sequence, an observation sequence and a discrete set of observation symbols,

respectively. The compact notation $\Lambda = \{\pi, A, B\}$ indicates the complete parameter set of the HMM where $\pi = \{\pi_i\}$, $\pi_i = P(s_1 = i)$ is the initial state distribution; $A = \{a_{ij}\}$, $a_{ij} = P(s_t = j \mid s_{t-1} = i)$ is the state transition probability distribution, and $B = \{b_j(k)\}$, $b_j(k) = P(o_t = v_k \mid s_t = j)$ is the observation symbol probability distribution, denoting the probability that a symbol $o_t = v_k$ is generated in state j. There are three basic problems for HMMs. Here we concentrate on the evaluation and reestimation problems. The decoding problem is not considered in this time.

The Evaluation Problem given the observation sequence O, and the model λ, the problem is how to choose the model which best matches the observations for the purpose of classification or recognition. For solving this problem, we obtain [6]

$$P(O \mid \lambda) = \sum_{all\ S} P(O \mid S, \lambda) P(S \mid \lambda) \tag{1}$$

or

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i) \tag{2}$$

where $\alpha_t(i) = P(o_1, o_2, ..., o_t, s_t = i \mid \lambda)$ and $\beta_t(i) = P(o_{t+1}, ..., o_T \mid s_t = i, \lambda)$ are the forward and backward variables, respectively [6].

The estimation problem given the observation sequence O, how do we adjust the model parameters Λ to maximize $P(O|\Lambda)$? This problem is solved by applying the Baum-Welch reestimation algorithm as follows [6]

$$\bar{\pi}_i = \gamma_1(i), \qquad \bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)},$$

$$\bar{b}_j(k) = \frac{\sum_{t \in o_t = v_k} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{3}$$

where $\gamma_t(i) = P(s_t = i \mid O, \lambda) = \sum_{j=1}^{N} \xi_t(i,j)$ (4)

and $\xi_t(i,j) = P(s_t = i, s_{t+1} = j \mid O, \lambda)$

$$= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)} \tag{5}$$

Note that a scaling procedure is required for implementing (3) since the dynamic range of the $\alpha_t(i)$ computation will exceed the precision range of any machine for sufficiently large t.

3. TEMPORAL HIDDEN MARKOV MODEL

The use of codewords in a codebook as states of a Markov chain was developed for isolated word recognition [1]. The proposed research extends this idea to a general framework and hence it can apply to HMMs, Gaussian mixture models (GMMs) and their fuzzy versions.

HMMs have been successfully applied to speech recognition, however there is a limitation of this approach. The HMM assumes that the occurrence of one feature is statistically *independent* of the occurrence of the others. This assumption is not appropriate for speech or handwriting recognition because a spoken or written word is represented as a time series of features and hence the features are correlated in time. Let $O = o_1, o_2, ..., o_T$ denote a stochastic process in discrete time. The probability that the t-th variable o_t takes the value w_t depends on the values taken by all the previous variables. Using the Markov assumption, the probability that the t-th variable o_t takes the value w_t depends on the immediately preceding value o_{t-1} as follows:

$$P(o_t = w_t \mid o_{t-1} = w_{t-1}, o_{t-2} = w_{t-2}, ..., o_1 = w_1) = P(o_t = w_t \mid o_{t-1} = w_{t-1}) \tag{6}$$

The stochastic processes based on this assumption are termed Markov processes. Markov chains are Markov processes for which state variables are restricted to have a finite number of values and the probability $P(o_t = w_t \mid o_{t-1} = w_{t-1})$ is assumed to be invariant in time. The sequence $W = w_1, w_2, ..., w_T$ represents a sequence of states. In order to apply Markov chains theory to temporal models, the feature vectors are considered as outputs of Markov chains. Let $X = x_1, x_2, ..., x_T$ be a sequence of feature vectors which represents a spoken or written word, a feature vector x_t can be mapped to either a member of the set of codewords $V = \{v_1, v_2, ..., v_K\}$ obtained by vector quantization (VQ) modeling or a member of the set of Gaussian components $G = \{g_1, g_2, ..., g_K\}$ by GMM. The state sequence W may be either a codeword sequence $w_1 = v_i, w_2 = v_j, ..., w_T = v_m$ or a Gaussian sequence $w_1 = g_i, w_2 = g_j, ..., w_T = g_m$, where $1 \le i, j, m \le K$. Therefore, each codeword in V or each Gaussian in G is a specific state of the Markov chain. The state-transition probabilities of the Markov chain are used to represent the dependence between acoustic features. For example, if codeword v_k comes after codeword v_i in W, it is,

considered that there is a probability of transition from state v_i to state v_k.

It should be noted that each observation in the sequence O is assumed to be statistically *dependent* on its predecessor in the proposed temporal model. In the HMM, observations in the sequence O are assumed to be *independent*. Therefore the temporal model has avoided the limitation of the HMM.

Temporal Gaussian Mixture Model

The parameter model set is denoted as $\lambda = (q, p)$ where $q = [q(i)]$ and $p = [p(i,j)]$.

$$q(i) = P(o_1 = g_i),$$

$$p(i,j) = P(o_t = g_j \mid o_{t-1} = g_i), \quad 1 \le i,j \le K \quad (7)$$

satisfying

$$\sum_{i=1}^{K} q(i) = 1, \qquad \sum_{j=1}^{K} p(i,j) = 1 \qquad 1 \le i, j \le K \quad (8)$$

Using the Lagrangian method and the maximum likelihood estimation method, the model parameters are calculated as follows

$$q(i) = \frac{1}{L}\sum_{l=1}^{L} P(g_i \mid x_1^{(l)}, \lambda)$$

$$p(i,j) = \frac{\sum_{l=1}^{L}\sum_{t=2}^{T_l} P(g_i \mid x_{t-1}^{(l)}, \lambda) P(g_j \mid x_t^{(l)}, \lambda)}{\sum_{l=1}^{L}\sum_{t=2}^{T_l} P(g_i \mid x_{t-1}^{(l)}, \lambda)} \quad (9)$$

where $1 \le i, j \le K$, $1 \le l \le L$, L: number of training sequences and $P(g_i \mid x_t^{(l)}, \lambda)$ denotes the posterior probability used in the GMM method to update mixture weights, mean vectors and covariance matrices.

Temporal Hidden Markov Model

The parameter model set is denoted as $\lambda = (A, B, \pi)$ where $B = (q, p)$, $q = [q(i)]$ and $p = [p(i,j)]$ are temporal model parameters, A and π are standard HMM parameters. The probability $P(O \mid \lambda)$ is calculated as follows

$$P(O \mid \lambda) = \sum_{S} P(O, S \mid \lambda) = \sum_{S} P(O \mid S, \lambda) P(S \mid \lambda) \quad (10)$$

where

$$P(O \mid S, \lambda) = q(o_1 = w_1 \mid S, \lambda) \prod_{t=2}^{T} p(o_t = w_t \mid o_{t-1} = w_{t-1}, S, \lambda) \quad (11)$$

$$P(S \mid \lambda) = P(s_1 \mid \lambda) \prod_{t=2}^{T} P(s_t \mid s_{t-1}, \lambda) = \pi_{s_1} \prod_{t=2}^{T} a_{s_{t-1}s_t} \quad (12)$$

where π_{s_1} and $a_{s_{t-1}s_t}$ are HMM parameters. If the state sequence W is represented by the codeword sequence V, we obtain the discrete temporal hidden Markov model (DTHMM). The continuous temporal hidden Markov model (CTHMM) is obtained if the Gaussian sequence is used to represent the state sequence W. The forward-backward algorithm is used to calculate both state transitions $p(i,j)$ and a_{ij}.

4. EXPERIMENTAL RESULTS

Database description

The YOHO corpus was designed for speaker verification systems in office environments with limited vocabulary. There are 138 speakers, 106 males and 32 females. The vocabulary consists of 56 two-digit numbers ranging from 21 to 97 pronounced as "twenty-one", "ninety-seven", and spoken continuously in sets of three, for example "36-45-89", in each utterance. There are four enrolment sessions per speaker, numbered 1 through 4, and each session contains 24 utterances. There are also ten verification sessions, numbered 1 through 10, and each session contains 4 utterances. All waveforms are low-pass filtered at 3.8 kHz and sampled at 8 kHz. Speech processing was performed using HTK V2.0, a toolkit [10] for building hidden Markov models (HMMs). The data were processed in 32 ms frames at a frame rate of 10 ms. Frames were Hamming windowed and pre-emphasized. The basic feature set consisted of 12th-order mel-frequency cepstral coefficients (MFCCs) and the normalized short-time energy, augmented by the corresponding delta MFCCs to form a final set of feature vector with a dimension of 26 for individual frames

Algorithmic Issues

GMMs are initialized as follows. Mixture weights, mean vectors, and covariance matrices were initialized with essentially random choices. Covariance matrices are diagonal, i.e. $[\sigma_k]_{ii} = \sigma_k^2$ and $[\sigma_k]_{ij} = 0$ if $i \ne j$, where σ_k^2, $1 \le k \le c$ are variances. A variance limiting constraint was applied to all GMMs using diagonal covariance matrices [7]. This constraint places a minimum variance value $\sigma_{\min}^2 = 10^{-2}$ on elements of all variance vectors in the GMM in our experiments. Each speaker was modelled by using 96 training utterances in four enrolment sessions without end-point detection. Error rates therefore were not too low to allow meaningful comparisons between the current and proposed methods. GMMs were trained in text-independent mode.

Speaker Identification

Let λ_k, $k = 1,.., M$, denote speaker models of M speakers. Given a feature vector sequence X, a classifier is designed to classify X into M speaker models by using M discriminant functions $f_k(X)$, computing the similarities between the unknown X and each speaker model λ_k and selecting the model λ_{k^*} if

$$k^* = \arg\max_{1 \le k \le M} f_k(X) \tag{13}$$

where $f_k(X) = P(X \mid \lambda_k)$ (14)

$P(X \mid \lambda_k)$ is the likelihood function for the standard GMM method and

$$f_k(X) = q(x_1 = g_1 \mid S, \lambda) \prod_{t=2}^{T} p(x_t = g_j \mid x_{t-1} = g_i, S, \lambda_k) \tag{15}$$

$q(i)$ and $p(i, j)$ are calculated as shown in (9) for the proposed temporal GMM method.

Experimental Results

Figure 1 shows the speaker identification error rates averaged on the YOHO 138 speakers. Speaker models consist of 16, 32 and 64 Gaussian mixtures, respectively. The identification error rate obtained by using the temporal GMM method is lower than that obtained by using the standard GMM method in all of the three different model sizes.

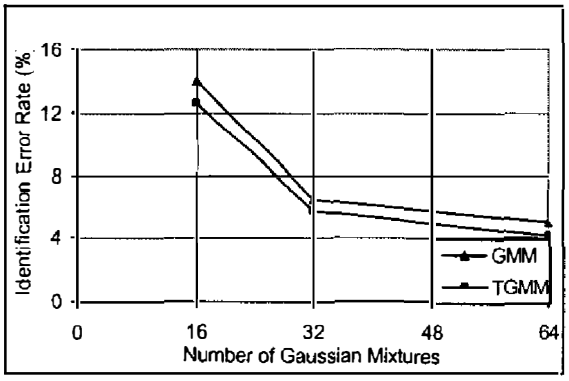

Figure 1: Speaker identification error rate (in %) averaged on 138 speakers for speaker models consisting of 16, 32 and 64 Gaussian mixtures using the standard GMM and the temporal GMM methods.

5. CONCLUSION

A new approach to hidden Markov modeling has been proposed in this paper. The proposed temporal hidden Markov model employs the Markov process for both the observable and the hidden processes. Each feature vector in the training sequence is assumed to be statistically dependent on its predecessor in the proposed temporal model. In the HMM, observations in the sequence O are assumed to be independent. Therefore the temporal model has avoided the limitation of the HMM. Speaker identification experiments performed on 138 Gaussian mixture speaker models in the YOHO database shows a better performance for the temporal GMM compared to the standard GMM. More experiments on the temporal HMM are investigating for speaker recognition.

6. ACKNOWLEDEMENT

The author would like to acknowledge the support of the Divisional Research Institute Grant, University of Canberra, Australia.

7. REFERENCES

[1] J. Dai, *"Isolated word recognition using Markov chain models"*, IEEE Transactions on Speech and Audio Processing, vol. 3, no. 6, 1995.

[2] R.O. Duda and P.E. Hart, *"Pattern classification and scene analysis"*, John Wiley & Sons, 1973.

[3] S. Furui, *"Recent advances in speaker recognition"*, Patter Recognition Letters, 18, pp. 859-872, 1997.

[4] X.D. Huang, Y. Ariki, and M.A. Jack, *"Hidden Markov models for speech recognition"*, Edinburgh University Press, 1990.

[5] J. B. Millar, J. P. Vonwiller, J. M. Harrington, and P. J. Dermody, " The Australian National Database of Spoken Language", in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 97-100, 1994.

[6] L. R. Rabiner and B. H. Juang. *Fundamentals of speech recognition*. Prentice Hall PTR, 1993.

[7] D.A. Reynolds, *"Robust text-independent speaker identification using Gaussian mixture models"*, IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, January 1995.

[8] D.A. Reynolds, *"A Gaussian mixture modeling approach to text-independent Speaker Identification"*, PhD thesis, 1993.

[9] D. Tran and M. Wagner, "Fuzzy Gaussian Mixture Models for Speaker Recognition", Australian Journal of Intelligent Information Processing Systems (AJIIPS), vol. 5, no. 4, pp. 293-300, 1998

[10] P. C. Woodland et. al., *"Broadcast news transcription using HTK"*, in Proceedings of ICASSP, USA, 1997.

[11] N. Kambhatla, *"Local models and Gaussian mixture models for statistical data processing"*, PhD thesis, Oregon Graduate Institute of Science & Technology, pp. 175-177, 1996.