A HIDDEN MARKOV MODEL WITH OPTIMIZED INTER-FRAME DEPENDENCE

F. J. Smith, J. Ming, P. O'Boyle, and A. D. Irvine
School of Electrical Engineering and Computer Science
The Queen's University
Belfast BT7 1NN, Northern Ireland, UK

ABSTRACT

An optimized hidden Markov model (HMM) with two kinds of inter-frame dependent observation structures, both built on the observation densities of a first-order dependent form, is presented to account for the statistical dependence between successive frames. In the first model, the dependence relation among the frames is determined optimally by maximizing the likelihood of the observations in both training and testing. In the second model, the dependence structure associated with each frame is described by a weighted sum of the conditional densities of the frame given individual previous frames. The segmental K-means and the forward-backward algorithms are implemented, respectively, for the estimation of the parameters of the two models. Experimental comparisons for an isolated word recognition task show that these models achieve better performance than both the standard continuous HMM and the bigram-constrained HMM.

1. INTRODUCTION

It has been well known that a major limitation to the standard HMM's in the modelling of speech signals is the state-conditioned independence assumption, which results in a loss of information about the temporal correlation between successive frames. The improvement of the conventional HMM approaches by incorporating some kind of modelling of the dynamic features has aroused much interest in recent years. The best-known approach is the addition of the 1st and 2nd time derivatives of the frame vectors. In [1] Kenny et al proposed a linear predictive HMM in which the frame sequence is modelled by a vector-valued AR process; Wellekens [2] and Paliwal [3] described, respectively, the HMM's with bigram constrained observations. More recently, direct scoring for segments rather than for individual frames has also been studied in the hybrid neural net / HMM [4] and two-dimensional (i.e. time-frequency) cepstral HMM's [5]. Some other approaches dealing with the same problem can also be found, e.g., in [6] where the recurrent neural networks were used to capture the sequential constraint and in [7] where explicit use of templates was suggested to represent the states.

In this paper we propose a new approach for incorporating the statistical dependence between successive frames in the HMM framework. The basic principle of this approach which differs from the previous ones is that the temporal correlation structure in an acoustic sequence is determined optimally together with the HMM parameters under the same criterion for model estimation. This is opposed to first specifying some temporal structure (e.g., the frame-lag [1-3] or frame-segment [4][5] structures) and then estimating the model parameters given these specifications. The joint optimization, with respect to both the temporal structure and the model parameters associated with it is, therefore, the major characteristic of our models. In this paper we will focus on the application of this principle to the HMM's with first-dependent observations. Two forms of the model based on the maximum likelihood criterion are investigated.

2. HMM'S WITH INTER-FRAME DEPENDENCE

Denote by $x = (x_1, \cdots, x_T)$ a sequence of observed frame vectors and λ the parameter set of an HMM.

Model 1. For the first model, the density function of x, given the state sequence $s = (s_0, \cdots, s_T)$, where $s_t \in \{1, \cdots, M\}$ and M is the number of the states, is defined as

$$p_\lambda(x|s,\tau) = \prod_{t=1}^{T} b_{s_t}(x_t|x_{\tau(t)}) \qquad (1)$$

where $b_{s_t}(x_t|x_{\tau(t)})$ is the observation density of the frame vector x_t from state s_t, which is assumed to be dependent on some previous frame $x_{\tau(t)}$, $\tau(t) < t$. The time-lag sequence $\tau = (\tau(1), \cdots, \tau(T))$ then characterizes a temporal dependence structure in the observed sequence, which is optimized together with the spectral structure (characterized by the state sequence) in both training and testing.

Eqn. (1) can be viewed as a general representation of the bigram-constrained HMM's. Although in each individual sequence the occurrence of one frame is most likely to depend on its immediate previous frame, the same dependent event-pairs, occurring in different sequences, may have different time intervals due to variations in speaking rate. The model (1), then, provides an explicit

0-7803-2431-5/95 $4.00

way of handling this variability. This was found to provide an enhanced robustness in our earlier experiments using a simpler version of this model [8]. For convenience, we call this model the *dependence-optimized model.*

Let $\lambda = (\pi, A, B)$ be the model parameter set, where $A = [a_{ij}]$ is the $M \times M$ state transition probability matrix, $\pi = [\pi_i]$ is the $M \times 1$ initial state probability vector, and $B = \{b_i: 1 \le i \le M\}$ is the observation density set where each b_i is defined on $R^K \times R^K$ (K is the dimension of the frame vector). For the model given by (1), we can write the joint density function of x and s, given λ and τ, as

$$p(x, s|\tau, \lambda) = p_\lambda(x|s, \tau)p(s|\lambda) = \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(x_t|x_{\tau(t)}) \quad (2)$$

where we assume that the probability of occurrence of s is independent of τ. For a given observation sequence, (2) gives the joint likelihood of the observation and state sequences given the dependence structure associated with the model parameters. This likelihood is used for maximization in training for estimation of the parameter set and in recognition for use as a score of the given observation sequence.

Model 2. For the second model, the density function of x, given the state sequence s, is defined as

$$p_\lambda(x|s) = \prod_{t=1}^{T} b_{s_t}(x_t) = \prod_{t=1}^{T} \left[\sum_{\tau=1}^{N} w_{s_t\tau} f_{s_t\tau}(x_t|x_{t-\tau}) \right] \quad (3)$$

where the observation density of the frame vector x_t from state s_t, $b_{s_t}(x_t)$, is expressed as a weighted sum of the conditional densities of x_t given N individual previous frames; N is a predefined constant, $w_{s_t\tau}$ is the weight to the frame $x_{t-\tau}$ on which x_t is dependent in state s_t, and $f(\cdot|\cdot)$ is the component conditional density function constituting $b_{s_t}(x_t)$.

Unlike model 1, the dependence structure among the frames is characterized here by the weight sequence $\{w_{s_t\tau_t}: t = 1, \cdots, T\}$ associated with each time-lag sequence $(\tau_1, \cdots, \tau_T)$, where $1 \le \tau_t \le N$. We therefore call this model the *dependence-weighted model.* A similar weighted-sum representation principle was applied previously to a language-model with N-gram constraints [9]. Note that (3) is reduced to (1) if we assume that at each time t only one weight, $w_{s_t t-\tau(t)}$, exists while the others are zero.

From (3) we can write the density function of x given λ as

$$p(x|\lambda) = \sum_s p_\lambda(x|s)p(s|\lambda) = \sum_s \pi_{s_0} \prod_{t=1}^{T} \left[a_{s_{t-1}s_t} \sum_{\tau=1}^{N} w_{s_t\tau} f_{s_t\tau}(x_t|x_{t-\tau}) \right] = \sum_s \sum_{\tau_1=1}^{N} \cdots \sum_{\tau_T=1}^{N} \left[\pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} w_{s_t\tau_t} f_{s_t\tau_t}(x_t|x_{t-\tau_t}) \right] \quad (4)$$

where the summation for s is taken over all possible state sequences. Given an observation sequence, (4) gives the likelihood of the observation associated with the model parameter set. This likelihood is then maximized for model estimation and for recognition.

In both the models, the observation densities, i.e., the b_i's in (1) and the $f_{i\tau}$'s in (3), are defined as the multivariate Gaussian conditional density function of the form $g(x|x') = N(z, m_z, V_z)/N(x', m_{x'}, V_{x'})$, where $N(\cdot)$ denotes a normal function, $z = (x, x')$ is the joint vector of x and x', and m's and V's are the mean vectors and covariance matrices correspondingly, which are either state (for model 1) or state and time-lag (for model 2) dependent. Two kinds of covariance structure are considered for this density. For the first case, it is assumed that x and x' have diagonal covariance matrices and are uncorrelated except for their corresponding elements (i.e. spectral components). This results in a joint covariance matrix V_z in which each of the block covariance matrices has a diagonal form. For convenience therefore, we call this density the *diagonal-block covariance matrix density.* For the second case, full covariance matrices for x, x' and z are assumed.

Further, a mixture density, taking the density $g(x|x')$ defined above as the components, is implemented. This mixture density has a form $b_i(x|x') = \sum_k c_{ik} g_{ik}(x|x')$ and is applied particularly to the dependence-optimized model as defined by (1).

3. ALGORITHMS FOR MODEL ESTIMATION

Model 1. Given a training sequence x, the parameter set λ of the dependence-optimized model is estimated by maximizing the log likelihood $\max_{s,\tau} \log p(x, s|\tau, \lambda)$, where $p(x, s|\tau, \lambda)$ is given by (2). This maximization, obviously, includes the conventional bigram-constrained HMM's as a special case and is accomplished by using the segmental K-means procedure. The algorithm involves an iteration of alternate maximization of $\log p(x, s|\tau, \lambda)$, once over s and τ for a given λ, and then over λ for the resulting estimates of s and τ, $\bar{s}$ and $\bar{\tau}$. In particular, in

the estimation of λ, the optimization over b_i is equivalent to estimation of its parameter set given the data set $\{z_t\} = \{x_t,\ x_{\bar{\tau}(t)} : \bar{s}_t = i\}_{t=1,T}$. An efficient solution to the simultaneous maximization of s and τ given λ is implemented using a joint state-sequence decoding and dependence searching algorithm. More specifically, define $\delta_t(i)$ as the log likelihood of the best state and time-lag sequences ending in state i and for the observations $\{x_n\}_{n=1,t}$, then by induction we have the recursion formula

$$\delta_t(j) = \max_{1\le i\le M} [\delta_{t-1}(i) + \log a_{ij}] + \max_{\tau(t)} \log b_j(x_t|x_{\tau(t)}) \quad 1\le j\le M \tag{5}$$

The arguments i and $\tau(t)$ which maximize (5) for each t and j are stored to retrieve the best state and time-lag sequences for the complete observation. The solution at the end of an observation, time T, is given by $\max_i \delta_T(i)$, which is used as the score of that observation in recognition. (5) is similar to the conventional Viterbi algorithm, except for the extra search for the $\tau(t)$'s. The complexity of the search is proportional to the range of the dependence being searched.

Model 2. For the dependence-weighted model, the likelihood function defined by (4) is maximized for estimation of the parameter set that includes π, A, $\{f_{i\tau}\}$ and $\{w_{i\tau}\}$ for $1\le i\le M$ and $1\le \tau\le N$. This maximization is implemented using the forward-backward recursion algorithm which iteratively maximizes the likelihood with respect to a new estimate $\bar{\lambda}$ given the previous estimate λ. The main reason for taking the forward-backward recursion instead of the Viterbi algorithm is to prevent the dependence weights, i.e. $w_{s_t\tau}$'s, from taking only two values: one or zero. Define the forward probabilities

$$\alpha_t(i) = \sum_{j=1}^{M} \alpha_{t-1}(j) a_{ji} b_i(x_t) \quad 1\le i\le M,\ 1\le t\le T \tag{6}$$

with $\alpha_0(i) = \pi_i$, where $b_i(x_t)$ is defined in (3), and the backward probabilities

$$\beta_t(i) = \sum_{j=1}^{M} \beta_{t+1}(j) a_{ij} b_j(x_{t+1}) \quad 1\le i\le M,\ T-1\ge t\ge 0 \tag{7}$$

with $\beta_T(i) = 1$, we have

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T} \alpha_{t-1}(i) a_{ij} \left[\sum_{\tau=1}^{N} w_{i\tau} f_{i\tau}(x_t|x_{t-\tau})\right] \beta_t(j)}{\sum_{t=1}^{T} \alpha_{t-1}(i)\beta_{t-1}(i)} \tag{8}$$

$$\bar{w}_{i\tau} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{M} \alpha_{t-1}(j) a_{ji} f_{i\tau}(x_t|x_{t-\tau}) w_{i\tau} \beta_t(i)}{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)} \tag{9}$$

$$\bar{m}_{z,i\tau} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{M} \alpha_{t-1}(j) a_{ji} w_{i\tau} f_{i\tau}(x_t, x_{t-\tau}) \beta_t(i) \cdot z_t}{\sum_{t=1}^{T}\sum_{j=1}^{M} \alpha_{t-1}(i) a_{ji} w_{i\tau} f_{i\tau}(x_t, x_{t-\tau}) \beta_t(i)} \tag{10}$$

$$\bar{V}_{z,i\tau} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{M} \alpha_{t-1}(j) a_{ji} w_{i\tau} f_{i\tau}(x_t, x_{t-\tau}) \beta_t(i) \cdot \tilde{z}_t \tilde{z}_t^*}{\sum_{t=1}^{T}\sum_{j=1}^{M} \alpha_{t-1}(i) a_{ji} w_{i\tau} f_{i\tau}(x_t, x_{t-\tau}) \beta_t(i)} \tag{11}$$

where $\tilde{z}_t = z_t - m_{z,i\tau}$, $z_t = (x_t, x_{t-\tau})$ is the joint vector with time lag τ and $f_{i\tau}(x_t, x_{t-\tau})$ is the density of this joint vector. $m_{z,i\tau}$ and $V_{z,i\tau}$ contain the mean vector and covariance matrix of the component vector $x_{t-\tau}$ that are required for the computation of the conditional density.

4. EXPERIMENTS

Speaker-dependent recognition experiments are performed based on a database consisting of the E-set (b, c, d, e, g, p, t, v, z) collected from 4 (two male and two female) speakers. The systems we chose for comparison are a standard continuous HMM based recognizer (HTK) [10] and the bigram-constrained HMM ([3], continuous version) in which the dependence of each frame is fixed to the immediate previous frame. All the systems assume a left-to-right Markov chain with 5 states. For the dependence-optimized and -weighted models, the searching and weighting ranges for the dependence are set to be six frames. The speech is sampled at 10 kHz and each frame has a span of 25.6 ms with an overlap of 10.6 ms. The 10th-order LPC cepstral coefficients together with 10 delta cepstral coefficients and a delta power are calculated for each frame as the feature vector. 20 utterances of each word are used to train a model for each speaker, and another 30 utterances of each word / speaker are used for testing. The recognition results are shown in Table 1. In all cases, except where otherwise indicated, the covariance matrices used are diagonal or diagonal-block forms. The digits in parentheses in the mixture cases are the numbers of mixtures which were found to be the best from a maximum of 5 mixtures per state.

As shown in table 1, both the dependence-optimized and the dependence-weighted models improve the performance in comparison to the standard HMM. The

Table 1. Recognition Results of the Systems for 4-Speaker's E-set

Model	Dependence-Optimized			Dependence-Weighted	Standard HMM			Bigram HMM
Speaker	Single	Mixture	Full cov	Single	Single	Mixture	Full cov	Single
A	83.70	90.00 (3)	80.37	85.93	79.63	86.66 (4)	78.14	79.25
I	93.89	97.41 (2)	90.37	94.81	89.62	93.70 (3)	87.40	91.48
L	91.85	94.81 (2)	82.59	94.44	84.07	89.25 (3)	81.11	86.29
P	85.93	92.59 (2)	79.62	87.40	80.74	88.14 (4)	78.14	81.85
Average	88.84	93.70	83.23	90.65	83.51	89.43	81.19	84.71

improvement is more significant for the diagonal covariance cases but less significant for the full covariance cases due to the increased number of parameters. Also, both the models outperform the bigram-constrained HMM. Particularly, the histogram of the maximum likelihood time-lags (i.e. $\bar{\tau}(t)$'s) in the dependence-optimized model, which is accumulated over all the utterances being correctly recognized in the single mixture case, is shown in Fig. 1. As expected, the dependence-weighted model performs better than the single-Gaussian dependence-optimized model because of the use of more information. We didn't try multiple mixtures for the static spectral representation in the model because of the parameter size.

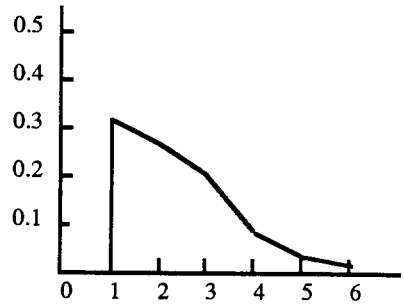

Fig. 1 Histogram of the maximum likelihood time lag in the dependence-optimized model

5. CONCLUSION

In this paper we studied the problem of improving the HMM's ability in capturing the temporal correlations in acoustic sequences. Unlike the conventional approaches, we proposed a model in which a joint optimization is performed over both the temporal structure and the model parameters. We particularly investigated two kinds of observation structures built upon the first-order dependent densities and presented the algorithms that implement the joint optimization for the model estimation. The experimental results show clearly the advantage of this optimization over some conventional approaches.

REFERENCES

[1] P. Kenny, M. Lennig, and P. Mermelstein, "A linear predictive HMM for vector-valued observations with applications to speech recognition," IEEE Trans. ASSP, vol. ASSP-38, pp. 220-225. 1990.

[2] C. J. Wellekens, "Explicit correlation in hidden Markov models for speech recognition," ICASSP-87, pp. 384-387, 1987.

[3] K. K. Paliwal, "Use of temporal correlation between successive frames in a hidden Markov model based speech recognizer," ICASSP-93, pp. 215-218, 1993.

[4] G. Zavaliagkos et al, "A hybrid segmental neural net / hidden Markov system for continuous speech recognition," IEEE Trans. SAP, vol. 2, pp. 151-159, 1994.

[5] B. P. Milner and S. V. Vaseghi, "Speech modelling using cepstral-time feature matrices and hidden Markov models," ICASSP-94, pp. 601-604, 1994.

[6] T. Robinson, "A real-time recurrent error propagation network word recognition system," ICASSP-92, pp. 617-620, 1992.

[7] O. Ghitza and M. M. Sondhi, "Hidden Markov models with templates as non-stationary states: an application to speech recognition," Computer Speech and Language, vol. 2, pp. 101-119, 1993.

[8] J. Ming and F. J. Smith, "Inter-frame dependent hidden Markov model for speech recognition," IEE Electronics Letters, vol. 30, pp. 188-189, 1994.

[9] P. O'Boyle, M. Owens, and F. J. Smith, "A weighted average n-gram model of natural language," Computer Speech and Language, vol. 8, pp. 337-349, 1994.

[10] S. J. Young, HTK–Hidden Markov Model Toolkit V1.5, CUED, Sept. 1993.