USE OF TEMPORAL CORRELATION BETWEEN SUCCESSIVE FRAMES IN A HIDDEN MARKOV MODEL BASED SPEECH RECOGNIZER

K.K. Paliwal

Computer Systems and Communications Group
Tata Institute of Fundamental Research
Homi Bhabha Road, Bombay 400005, India

ABSTRACT

Though the hidden Markov modeling (HMM) technique has been successfully applied to various speech recognition applications, it has one major limitation. It assumes state-conditioned stationarity of the observation vectors, implying that the occurrence of one observation vector is independent of others if these vectors are generated by the same state. In most of the situations, this assumption of stationarity is not valid as the time sequence of the observation vectors is highly correlated. In the present paper, we try to use this temporal correlation by conditioning the probability of the current observation vector on the current state as well as on the previous observation vectors. Results from an isolated word recognition experiment using discrete HMMs are reported to illustrate the point.

1. INTRODUCTION

Though the hidden Markov modeling (HMM) technique has been successfully applied to various speech recognition applications, it has one major limitation. It assumes state-conditioned stationarity of the observation vectors, implying that each state is a stationary sou'rce generating independent, identically distributed (IID) observation vectors. This means that observation vectors within a state are identically distributed and the occurrence of one observation vector is independent of others if these vectors are generated by the same state. In certain situations (e.g. steady-state vowels), this assumption of stationarity is reasonable. But, in most of the cases (e.g. vowel-consonant or consonant-vowel transitions, glides and diphthongs), this assumption of stationarity is not valid as the time sequence of the observation vectors is highly correlated. Thus, there is a strong need for incorporating the temporal correlation between successive observation vectors in the HMM framework.

In the literature [2-8], some studies have been reported where the assumption of state-conditioned stationarity is somewhat relaxed. For example, Ostendorf and Roukos [2] have used a stochastic segment model which can, in principle, avoid the IID assumption completely. However, in their implementation, they have assumed the observation vectors to be independent within a state, because the computational complexity becomes exorbitantly high otherwise. Explicit use of templates to represent states in the stochastic segment model has the advantage that it does not have to make the assumption of identical distribution of observation vectors within a state. A similar study has been reported by Ghitza and Sondhi [3]. Deng [4] has used a parametric model to represent the trend within a state, thus avoiding the assumption of identical distribution of observation vectors. Kenny et al. [5] have used a state-conditioned linear prediction model to remove correlation between successive observation vectors, and treated the resulting residual vectors as independent and identically distributed. Nonlinear predictors have been used recently by a number of authors for removing this temporal correlation between successive observation vectors [6-8].

In the present paper, we try to use this temporal correlation by conditioning the probability of the current observation vector on the current state as well as on the previous observation vectors. This is done by introducing the state-conditioned transition probabilities between successive observation symbols (or, vector quantizer labels) for the discrete HMMs. We use these HMMs in a speaker-independent isolated word recognition experiment, and provide results which illustrate the usefulness of incorporating temporal correlation between successive observation vectors. Though we have not studied here the use of this type of explicit incorporation of temporal correlation for continuous HMMs, a theory has been developed in [9] for this purpose.

0-7803-0946-4/93 $3.00 © 1993 IEEE

2. THEORY

Consider a discrete HMM $\lambda = [N, M, \pi, A, B]$, where N = the number of states in the model, M = the number of output symbols in the discrete alphabet of the model, $\pi = \{\pi_i, 1 \leq i \leq N\}$, the initial state probability vector (π_i is the probability that the model is in state i initially), $A = \{a_{ij}, 1 \leq i,j \leq N\}$, the transition matrix of underlying Markov chain (a_{ij} is the probability of transition from state i to state j), and $B = \{b_j(k), 1 \leq j \leq N, 1 \leq k \leq M\}$, the model output symbol probability matrix ($b_j(k)$ is the probability of outputting the symbol k when the model is in state j). Consider an input utterance represented by a sequence of observation symbols, $X_1^T = \{X_1, X_2, \ldots, X_T\}$, where T is the number of frames in the input utterance. In order to compute the probability, $P(X_1^T \mid \lambda)$, of the observation sequence X_1^T being generated by the model λ, define the probability of partial observation sequence X_1^t and state j at time t (denoted by q_j^t) as

$$\alpha_j(t) = P(X_1^t, q_j^t \mid \lambda). \tag{1}$$

Then, by definition

$$\alpha_j(1) = \pi_j P(X_1 \mid q_j^1, \lambda), \tag{2}$$

and

$$P(X_1^T \mid \lambda) = \sum_{j=1}^{N} \alpha_j(T). \tag{3}$$

A forward recursion relation for the computation of probability $\alpha_j(t)$ can be derived as follows:

$$\begin{aligned} \alpha_j(t) &= P(X_1^t, q_j^t \mid \lambda) \\ &= \sum_{i=1}^{N} [P(X_1^{t-1}, q_i^{t-1} \mid \lambda) \\ &\qquad P(X_t, q_j^t \mid X_1^{t-1}, q_i^{t-1}, \lambda)] \\ &= \sum_{i=1}^{N} [\alpha_i(t-1) P(q_j^t \mid X_1^{t-1}, q_i^{t-1}, \lambda) \\ &\qquad P(X_t \mid X_1^{t-1}, q_i^{t-1}, q_j^t, \lambda)] \,. \end{aligned} \tag{4}$$

Since the state of the model at a given frame depends only on the state of the model at the preceding frame and is independent of the preceding observation symbols, it follows that

$$\begin{aligned} P(q_j^t \mid X_1^{t-1}, q_i^{t-1}, \lambda) &= P(q_j^t \mid q_i^{t-1}, \lambda) \\ &= a_{ij}. \end{aligned} \tag{5}$$

Note that the last equality makes use of the time invariance property of transition probabilities. Substituting Eq. (5) in Eq. (4) and using the assumption that the probability of the observation symbol at given frame is independent of the state of the model in the preceding frame, it follows

$$\alpha_j(t) = \sum_{i=1}^{N} \alpha_i(t-1) a_{ij} P(X_t \mid X_1^{t-1}, q_j^t, \lambda). \tag{6}$$

As mentioned earlier, the standard HMM approach assumes the state-conditioned stationarity of the observation vectors (or symbols). This means that

$$P(X_t \mid X_1^{t-1}, q_j^t, \lambda) = P(X_t \mid q_j^t, \lambda) \tag{7}$$

Substituting Eq. (7) in Eq. (6), it follows

$$\begin{aligned} \alpha_j(t) &= \sum_{i=1}^{N} \alpha_i(t-1) a_{ij} P(X_t \mid q_j^t, \lambda) \\ &= \sum_{i=1}^{N} \alpha_i(t-1) a_{ij} b_j(X_t). \end{aligned} \tag{8}$$

This is the famous recursion relation used in the standard HMM approach for computing the forward probability [10].

As mentioned earlier, the assumption of state conditioned stationarity (given by Eq. (7)) is the cause of the major limitation in the standard HMM approach. In order to avoid it, we start from Eq. (6). This equation provides an effective procedure to incorporate the temporal correlation between successive observation symbols. The temporal correlation can be extended to as many frames as required. For example, suppose it is required to incorporate temporal correlation between observation vectors of two successive frames. In this case, Eq. (6) can be rewritten as

$$\begin{aligned} \alpha_j(t) &= \sum_{i=1}^{N} \alpha_i(t-1) a_{ij} P(X_t \mid X_{t-1}, q_j^t, \lambda) \\ &= \sum_{i=1}^{N} \alpha_i(t-1) a_{ij} b_{jX_{t-1}}(X_t), \end{aligned} \tag{9}$$

where $b_{jk}(l)$ is the probability of outputting the symbol l given that model is in state j and has outputted the symbol k in the previous frame.

Comparison of Eq. (9) with Eq. (8) reveals that this approach is comparable to the standard HMM approach in terms of computation cost. However, this approach has the problem that it requires M^2N parameters $\{b_{jk}(l)\}$, which is much larger in number than the MN parameters $\{b_j(l)\}$ needed in a standard HMM. However, this problem of large number of parameters can be solved by judiciously increasing the size of training data and applying some smoothing technique (e.g., deleted interpolation [11]).

Note that this approach is developed here for discrete HMMs. However, it can be easily extended to continuous HMMs [9].

3. EXPERIMENTAL RESULTS

In this section, we conduct speech recognition experiments where we study the use of discrete HMMs with and without temporal correlation. Results of these experiments are described in this section.

In our experiments, we have used a speaker independent, isolated word speech recognition system with 5-state left-to-right HMMs. Each frame is represented here by a 12-dimensional observation vector (consisting of 12 cepstral coefficients derived through linear prediction analysis). The vocabulary consists of 9 English E-set alphabets (i.e., B, C, D, E, G, P, T, V and Z). The data base has two sets of data, each consisting of one utterance of each of the nine words by each of 100 speakers (50 men and 50 women). One set of data is used for training and another set for testing. The training and testing tokens were recorded over local dialed-up telephone lines, bandpass filtered to 200–3200 Hz, and digitized at a sampling rate of 6.67 kHz. An eighth-order linear prediction analysis was performed every 15 ms with a frame width of 45 ms using the autocorrelation method (with Hamming window and preemphasis), and 12 cepstral coefficients were computed from the 8 linear prediction coefficients. These 12 cepstral coefficients are weighted by a cepstral window (or, lifter) [12], and are treated as the 12 components of an observation vector. Endpoints of each utterance were manually determined.

In discrete HMMs, the cepstral space is represented in terms of M codevectors. These M codevectors are obtained from the training set data using the k-means algorithm [13] using total squared error as the distortion measure. Recognition results are obtained with and without temporal correlation on the training set data. These are shown in Table 1 as a function of M. It can be seen from this table that the present approach provides significant improvement in recognition performance (with the use of temporal correlation between two successive vectors).

Effect of temporal correlation on the speech recognition performance is also studied on the test data set. Results are shown in Table 2 as a function of M. It can be seen from this table that speech recognition performance of the system improves by incorporating the temporal correlation, though the improvement is not much. This happens due to the fact that the training data set is small in size. It does permit reliable estimate of parameters $\{b_{jk}(l)\}$. However, this problem can be overcome by judiciously increasing the size of training data and applying some smoothing technique (e.g., deleted interpolation [11]).

Table 1: Speech recognition accuracy with and without temporal correlation on the training data set.

M	Recognition accuracy (in %)	
	without correlation	with correlation
8	41.44	59.11
16	45.44	77.56
32	57.33	95.00
64	62.78	99.11

Table 2: Speech recognition accuracy with and without temporal correlation on the test data set.

M	Recognition accuracy (in %)	
	without correlation	with correlation
8	38.56	40.00
16	40.78	42.67
32	45.22	44.78
64	45.67	46.56

4. CONCLUSIONS

In this paper, we have incorporated the temporal correlation between successive frames in an HMM-based speech recognizer. This is done by making the probability of the current observation vector dependent on the previous observation vectors. Our preliminary results show that this approach provides significant improvement in recognition performance (with the use of temporal correlation between two successive frames alone).

5. REFERENCES

[1] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE,* Vol. 77, No. 2, pp. 257-286, Feb. 1989.

[2] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition", *IEEE Trans. Acoustics, Speech and Signal Processing,* Vol. 37, No. 12, pp. 1857-1869, Dec. 1989.

[3] O. Ghitza and M.M. Sondhi, "Hidden Markov models with templates as states: An application to

speech recognition", *Proc. IEEE Workshop on Automatic Speech Recognition*, Arden House, Harriman, NY, pp. 27-28, Dec. 1991.

[4] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal", *Signal Processing*, Vol. 27, No. 1, pp. 65-78, Apr. 1992.

[5] P. Kenny, M. Lennig and P. Mermelstein, "A linear predictive HMM for vector-valued observations with applications to speech recognition", *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 38, No. 2, pp. 220-225, Feb. 1990.

[6] K. Iso and T. Watanabe, "Speaker-independent word recognition using a neural prediction model", *Proc. IEEE Conf. Acoust., Speech and Signal Processing*, Albuquerque, NM, pp. 441-444, Apr. 1991.

[7] E. Levin, "Word recognition using hidden control neural architecture" *Proc. IEEE Conf. Acoust., Speech and Signal Processing*, Albuquerque, NM, pp. 433-436, Apr. 1991.

[8] J. Tebelskis and A. Waibel, "Large vocabulary recognition using linked predictive neural networks" *Proc. IEEE Conf. Acoust., Speech and Signal Processing*, Albuquerque, NM, pp. 437-440, Apr. 1991.

[9] C.J. Wellekens, "Explicit correlation in hidden Markov models for speech recognition", *Proc. IEEE Conf. Acoust., Speech and Signal Processing*, pp. 384-387, 1987.

[10] L.R. Rabiner, S.E. Levinson and M.M. Sondhi, "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition", *Bell Syst. Tech. J.*, Vol. 62, No. 4, pp. 1075-1105, Apr. 1983.

[11] L.R. Bahl, P.F. Brown, P.V. de Souza, R.L. Mercer and D. Nahamoo, "A fast algorithm for deleted interpolation", *Proc. 2nd European Conf. Speech Communication and Technology*, Genoa, Italy, pp. 1209-1212, Sept. 1991.

[12] B.H. Juang, L.R. Rabiner and J.G. Wilpon, "On the use of bandpass liftering in speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-35, pp. 947-954, 1987.

[13] Y. Linde, A. Buzo and R.M. Gray, "An algorithm for vector quantizer design", *IEEE Trans. Commun.*, Vol. COM-28, pp. 84-95, Jan. 1980.