RELAX FRAME INDEPENDENCE ASSUMPTION FOR STANDARD HMMS BY STATE DEPENDENT AUTO-REGRESSIVE FEATURE MODELS

Ying Jia, Jinyu Li

Intel China Research Center
#601, North Office Tower, Beijing Kerry Center, #1 GuangHua Road, Beijing 100020, PRC
Email: jeffel_jia@yahoo.com, Tel: (86-10) 8529,8800 ext 1677, Fax: (86-10) 85298717

ABSTRACT

In this paper, we propose a new type of frame-based hidden Markov models (HMMs), in which a sequence of observations are generated using state-dependent auto-regressive feature models. Based on this correlation model, it can be proved that expressing the probability of a sequence of observations as a product of probabilities of decorrelated individual observations doesn't require the assumption of frame independence. Under the maximum likelihood (ML) criteria, we also derived re-estimation formulae for the parameters (mean vectors, covariance matrix, and diagonal regression matrice) of the new HMMs using an Expectation Maximization (EM) algorithm. From the formulae, it's interesting to see that the new HMMs have extended the standard HMMs by relaxing the frame independence limitation. Initial experiment conducted on WSJ20K task shows an encouraging performance improvement with only 117 additional parameters in all.

1. INTRODUTION

The advent of hidden Markov model (HMM) has brought about a considerable progress in speech recognition technology over the last two decades, and nowhere has this progress been more evident than in the area of Large Vocabulary Continuous Speech Recognition (LVCSR). However a number of unrealistic assumptions with HMMs are still regarded as obstacles for its potential effectiveness. A major one is the inherent assumption that successive observations are independent and identically distribution (IID) within a state. It follows from the mechanics of the speech generation process that in reality the observations are highly dependent and correlated. Furthermore, under maximum likelihood (ML) criteria, the performance of a HMM-based system relies on how well the model can characterize the nature of real speech.

2. FRAME INDEPENDENCE ASSUMPTION OF STANDARD HMMS

In the statistical approach to automatic speech recognition, the mathematical optimal solution dictates that the recognizer follows the maximum a posteriori (MAP) decision rule

$$\hat{W} = \arg\max p(W|O) = \arg\max p(O|W)p(W)$$

where W is a word string hypothesis for a given acoustic observation O. $p(O|W)$ is the acoustic model, and

$$p(W) = \prod_{i=1}^{L} p(w_i | w_{i-1}, \cdots, w_{i-N})$$

is the N-gram language model. When deriving the acoustic model score $p(O|W)$, a hidden state sequence $q_1^T \in \Gamma$ is usually introduced as

$$\begin{aligned} p(O|W) &= \sum_{\Gamma} p(o_1^T, q_1^T | W) \\ &= \sum_{\Gamma} p(o_1^T | q_1^T, W) \cdot p(q_1^T | W) \end{aligned}$$

in which it is assumed the hidden process can fully account for the conditional probability of the acoustic signal.

In the frame-based HMM approach, the state sequence probability $p(q_1^T|W)$ can be rewritten by applying the Markov first order assumption as

$$\begin{aligned} p(q_1^T|W) &= p(q_0) \prod_{t=1}^{T} p(q_t | q_{t-1}, W) \\ &= \pi_{q_0} a_{q_0 q_1} a_{q_1 q_2} \cdots a_{q_{T-1} q_T} \end{aligned}$$

Given a hidden state sequence q_1^T, the joint observation probability along the state sequence $p(o_1^T | q_1^T, W)$ can be written as a product of probabilities of individual observation vector o_t conditioned on previous observations o_1^t and state partial sequence q_1^t, namely

$$p(o_1^T | q_1^T) = \prod_{t=1}^{T} p(o_t | o_1^t, q_t, q_1^{t-1})$$

0-7803-7041-4/01/$10.00 ©2001 IEEE

To make the above equation computationally tractable for standard HMM, it's necessary for us to make the frame independence assumption, which implies that all observations are statistically dependent on the state that generate them, not on the previous observations, i.e., $p(o_t|o_1^t, q_t, q_1^{t-1}) = p(o_t|q_t)$. According to this frame independence assumption, the joint observation probability can be rewritten as

$$p(o_1^T|q_1^T) = \prod_{t=1}^{T} p(o_t|o_1^t, q_t, q_1^{t-1}) \cong \prod_{t=1}^{T} p(o_t|q_t)$$

Although the frame independence assumption is clearly inappropriate for speech sounds, the standard HMM in practice has worked extremely well for various types of speech recognition tasks.

3. REVIEW OF RESEARCH EFFORTS ON FRAME CORRELATION MODELING

Under maximum likelihood (ML) criteria, the performance of a HMM-based system relies on how well the HMMs can characterize the nature of real speech. For this reason, various approaches have been tried to take account of frame correlation for more realistic modeling. These efforts are generally known by the name of "frame correlation modeling".

The family of segment models tries to directly express speech feature trajectories. The basic modeling unit is not a frame but a phonetic unit. This family of models relaxes both the stationarity and the independence assumptions within a standard HMM state. While they seem to be successful in extracting dynamic cues for speech recognition under a suitable trajectory assumption, they are not based on widely availiable HMM technology.

Deng et al. [6] used a regression polynomial function of time to model the trajectory of the mean in each state. A similar model was suggested by Gish and Ng [7] for a keyword spotting task. Russell and Holmes, and Gales and Young [8] extended the model suggested by Deng, by assuming a parametric segmental model with random coefficients, that are sampled once per segment realization. Therefore, the mean trajectory is a stochastic process instead of a fixed parameter. Digalakis [9] proposed a dynamical system model which generalizes the Gauss-Markov model to a Kalman filter framework, by assuming noisy observations.

Several authors have proposed nonparameteric segment models. A major advantage of nonparametric models is that they are not sensitive to the shape of the feature trajectory that needs to be approximated. Consequently, they are also not sensitive to the segment partition problem. On the other hand, nonparameteric models might require more data to train the model on, since they are less constrained that parametric models. The first nonparametric approach to a nonstationary state HMM was the stochastic segment model (SSM) suggested by Ostendorf and Roukos [10] in 1989. The SSM assigns a Gaussian distribution to the entire segment which is resampled to a fixed length. A nonparametric approach to a nonstationary state HMM with an additional step of time wrapping was suggested by Ghitza and Sondhi, in which the trajectory of the mean in a given state is set equal to that state realization in the training set whose dynamic time warping (DTW) distance from all other sequences in the ensemble is minimal. More recently, Kimball et al. [10] suggested a nonparametric approach that models each segment by a discrete mixture of nonparametric mean trajectors.

The most recent progress was made by Hsiao-Wuen Hon [5]. In his method the segment-based and frame-based HMM are combined together by a unreliable conditional probability decomposition assumption.

In the case of continuous HMM's, a Gaussian probability density function (PDF) assumption is made between adjacent feature vectors in C.J.Wellekens [1]. In P. Kenny [2], a linear prediction technigue is used to parameterize frame correlation. Paliwal [3] incorporated temporal correlation into discrete HMM's by conditioning the probability of the current observation on the current state as well as on the previous observation. S. Takahashi [4] propose a bigram-constrained (BC) HMM in which the probability of the current observation dependents on the current state as well as on the previous observation. But a BC HMM is obtained by combining a VQ-code bigram and the traditional HMM. So the number of parameters to be estimated in BC HMM is less than the number of the full parameterization method proposed by Paliwal. A remarkable point of BC HMM is that it has provided a method to combine the joint conditional PD by two separate conditional PD. All these efforts have been devoted to a decomposition of the probability $p(o_t|o_1^{t-1}, q_j, \lambda)$.

4. STATE DEPENDENT AUTO-REGRESSIVE FEATURE MODEL

Here we use a state dependent auto-regressive (AR) model to characterize the frame correlation between successive observation vectors, i.e., the observation vectors within a state are generated according to

$$o_t = \sum_{i=1}^{N} a_i o_{t-i} + e_t + n_t$$

where a_i are diagonal matrices, so that a auto-regressive model applies to each component of the vector o_t. e_t is a mean vector of this Gaussian component. n_t is the noise signal between the actual observation o_t and the predicted observation $\hat{o}_t$ with zero mean.

The reasons for us to use state dependent auto-regressive model to characterize frame correlation stem from the speech generation model and its application in speech coding. In the time domain, the speech waveforms are generated directly by the excitation source and vocal tract, and the vocal tract can be reasonably well parameterized by time-varying auto-regressive filter models. Based on this modeling framework, which is known as linear predictive coding, speech coding has made a great progress from a 32 kpbs to 4.8kbps. In the cepstral domain, the rationality comes from the fact that each cepstral frame is extracted from a window of speech samples.

5. RELAX THE LIMITATION OF FRAME INDEPENDENCE ASSUMPTION

Based on the above state dependent auto-regressive feature models, we can see that given current state q_t and previous N frames $o_{t-N}\cdots o_{t-1}$, o_t has the same distribution with n_t. Namely

$$p\left(o_t\middle|o_1^{t-1},q_t\right)=p\left(n_t\middle|o_{t-N}^{t-1},q_t\right)$$

So the likelihood of a state sequence hypothesis can be written as

$$p\left(o_1^T\middle|q_1^T\right)=\prod_{t=1}^{T}p\left(o_t\middle|o_1^{t-1},q_t,q_1^{t-1}\right)$$

$$=\prod_{t=1}^{T}p\left(n_t\middle|o_{t-N}^{t-1},q_t\right)$$

Therefore without frame independence assumption, we can also express the joint probability of the observations o_1^T as a product of probabilities of noisy individual observations n_t.

6. EM-BASED REESTIMATION FORMULAE FOR HMM PARAMETERS

For HMM states modeled by Gaussian mixture, it has been proved that maximization of the likelihood $p(O|W)$ equals to maximizing Q

$$Q=\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_{q_t,m}(t)\ln p\left(n_t\middle|o_{t-N}^{t-1},q_t\right)$$

Applying the state dependent auto-regressive feature model, the above Q-function can be rewritten as

$$\begin{aligned}Q&=\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_{q_t,m}(t)\ln p\left(n_t\middle|o_{t-N}^{t-1},q_t\right)\\&=\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_{q_t,m}(t)\ln p\left(o_t-\sum_{i=1}^{N}a_{m,i}o_{t-i}-e_{t,m}\middle|o_{t-N}^{t-1},q_t\right)\\&=\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_m(t)\left(\ln 2\pi|W_m|+\left(o_t-\sum_{i=1}^{N}a_{m,i}o_{t-i}-e_{t,m}\right)^T W_m^{-1}\left(o_t-\sum_{i=1}^{N}a_{m,i}o_{t-i}-e_{t,m}\right)\right)\end{aligned}$$

To maximize the Q-function with respect to mixture parameters, an EM algorithm can be applied. For each utterance, the mixture occupancy is the missing data. So the following iterative EM algorithm can be derived.

Expectation Step: Given mean e_m, variance W_m, and correlation matrices $a_{m,i}$, the expected alignment $\gamma_m(t)$ can be given using forward-backward algorithm as

$$\gamma_m(t)=p\left(q_{s,m}\middle|e_{m,t},W_m,a_{m,i},o_{t-N}^{t-1}\right)=\alpha_m(t)\beta_m(t)$$

Maximization Step: Given expectation of the missing data, differentiating Q with respect to mixture parameters (mean, variance and correlation matrix) and setting them to zero gives the following estimation formulas.

$$e_{m,t}=\frac{\sum_{t=1}^{T}\gamma_m(t)\left(o_t-\sum_{i=1}^{N}a_{m,i}o_{t-i}\right)}{\sum_{t=1}^{T}\gamma_m(t)}$$

$$W_m=diag\left(\frac{\sum_{t=1}^{T}\gamma_m(t)\left(o_t-\sum_{i=1}^{N}a_{m,i}o_{t-i}-e_{m,t}\right)\left(o_t-\sum_{i=1}^{N}a_{m,i}o_{t-i}-e_{m,t}\right)^T}{\sum_{t=1}^{T}\gamma_m(t)}\right)$$

For diagonal matrix $a_{m,i}$ $(1\le i\le N)$, the vector formed by N k-th diagonal elements from diagonal matrices can be estimated as

$$\begin{bmatrix}a_{m,1}^{(k)}\\\vdots\\a_{m,N}^{(k)}\end{bmatrix}=\begin{bmatrix}\sum_t\gamma_m(t)o_{t-1}^{(k)}w_m^{(k)-1}o_{t-1}^{(k)} & \cdots & \sum_t\gamma_m(t)o_{t-N}^{T}w_m^{(k)-1}o_{t-1}^{(k)}\\\vdots & \ddots & \vdots\\\sum_t\gamma_m(t)o_{t-1}^{(k)}w_m^{(k)-1}o_{t-N}^{(k)} & \cdots & \sum_t\gamma_m(t)o_{t-1}^{(k)}w_m^{(k)-1}o_{t-N}^{(k)}\end{bmatrix}^{-1}\times\begin{bmatrix}\sum_t\gamma_m(t)\left(o_t^{(k)}-e_{m,t}^{(k)}\right)w_m^{(k)-1}o_{t-1}^{(k)}\\\vdots\\\sum_t\gamma_m(t)\left(o_t^{(k)}-e_{m,t}^{(k)}\right)w_m^{(k)-1}o_{t-N}^{(k)}\end{bmatrix}$$

Therefore the N diagonal correlation matrices can be simultaneously estimated using the above formula in an element by element fashion.

From the above formulae, we see that the standard HMM is a special case of the new HMM if we assume the observations are independent from each other, i.e. correlation matrixes are zero matrix.

7. EXPERIMENT RESULTS

An initial investigation of the use of new models was carried out on a large-vocabulary speaker independent continuous speech recognition task. Experiments were conducted on Wall Street Journal 20k English task. The baseline system was a gender-independent within-word-triphone mixture-Gaussian tied state HMM system. In this model set, all the speech models had a three emitting state, left-to-right topology. Two silence models were used. The first silence model, a short pause model, had a single emitting state which may be skipped. The other silence model was a fully connected tree emitting state model used to represent longer period of silence. The speech was parameterized into 12 MFCC's, along with normalized log-energy and the first and second differentials of these parameters. This yielded a 39-dimensional feature vector, to which cepstral mean normalization was applied. The acoustic training data consisted of 36696 utterances from the SI-284 WSJ0 and WSJ1 sets. The ICRC LVCSR system was trained using decision-tree-based state clustering to define 6617 triphone states. A 24k word list and dictionary was used with the trigram language model. All decoding used a dynamic-network decoder.

For the particular implementation of the new models considered here, all states of all context-dependent phones associated with all monophone were assigned to the same set of diagonal correlation matrices. The order of the auto-regressive feature model is 3. Therefore this resulted in only 117 additional parameters. The process of building the correlation matrices was first to mix-up the final number of components. A conversion from standard models to new HMMs were made by setting the 117 additional correlation parameters to zero. Finally 5 iterations of embeded forward-backward reestimation were performed.

The experiment results were compared in table 1. It's really encouraging to see that the additional 117 parameters drop the word error rate from 11.8 (baseline) to 11.4. It should be emphasized that the WER for most speakers were cut down.

Table 1: Performance of a standard system (S) and a frame correlated system (N) on 333 testing utterances.

	Avg	440	441	442	443	444	445	446	447
S	11.8	9.9	21.0	12.4	14.6	11.9	6.1	10.3	8.4
N	11.4	9.5	20.8	11.8	13.0	12.0	5.9	9.7	9.6

8. CONCLUSIONS

We have extended the standard HMMs to a new type of HMMs by removing the frame independence assumption. In our new models, mean, variance, and a set of diagonal correlation matrices are parameters of each Gaussian component. These parameters can be re-estimated using the formulae derived from an EM algorithm. Actually the standard HMM is a special case of the new model when we assume the frames are independent. Without frame independence assumption, it has been proved that we can also express the probability of a sequence of observations as a product of probabilities of noisy individual observations if a reasonable state dependent auto-regressive feature model is used. Initial experiment conducted on WSJ20K task shows an encouraging performance improvement with only 117 additional parameters in all. Therefore the new models convince us of some interesting research directions opened to follow.

REFERENCE

1. C.J.Wellekens, Explicit correlation in hidden Markov model for speech recognition, Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp.383-386,1987.
2. P.Kenny, M.Lenning and P.Mermelstein, A Linear predictive HMM for vector-valued observation with applications to speech recognition. IEEE trans. On Acoustics, Speech and Signal Processing, pp. 220-225, 1993.
3. K.K. Paliwal, Use of temporal correlation between successive frames in hidden Markov model based speech recognizer. Proc. Int, Conf. Acoustics, Speech, and Signal Processing, pp.553-556, 1992.
4. S.Takahashi, Phoneme HMM's constrained by frame correlations. Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp. 219-222, 1993.
5. Hsiao-Wuen Hon and Kuansam Wang, "Unified Frame and Segment Models for Automatic Speech Recognition", ICASSP2000, Turkey.
6. L. Deng, M. Aksmanovic, D. Sun, and J. Wu, "Speech Recognition using hidden Markov models with polynomial regression functions as nonstationary states," IEEE trans. Speech, Audio Processing, Vol.2, pp 507-520, 1994.
7. H. Gish and K. Ng, "A segmnetal speech model with applications to nonstationary states: An application to speech recognition," in Proc. Int. Conf. Acoustics, Speech, Signal Processing, 1993, pp.447-450.
8. M.Gales and S. J. Young, "Segmental hidden Markov models," in Proc. Eurospeech, 1995, pp. 1579-1582.
9. V. V. Digalakis, "Segment-based stochastic models of spectral dynamics for continuous speech recognition," Ph.D. dissertation, Boston Univ., Boston, MA, 1992.
10. M. Ostendorf and V. Digalakis, "A stochastic segment model for phoneme-based continuous speech recognition," IEEE trans. Acoust., Speech, Signal Processing, vol. 37, pp. 1857-1869, 1989.