A Unified Way in Incorporating Segmental Feature and Segmental Model into HMM

Jun He Henri Leich

Lab T.C.T.S. Faculté Polytechnique de Mons, Boulevard Dolez 31, Mons, Belgium

ABSTRACT

There are two major approaches to speech recognition: frame-based and segment-based approach. Frame-based approach, e.g. HMM, assumes statistical independence and identical distribution of observation in each state. In addition it incorporates weak duration constrains. Segment-based approach is computational expensive and rough modelling easily occurs if not much 'templates' are stored. This paper presents a new framework to incorporate segmental feature and segmental model in a unified way into frame-based HMM to exploit the advantage of both methods. In the modified Viterbi algorithm, frame-based information prunes out the most probable path at each segment level to which segmental model can be applied with dramatically reduced computational load; at the same time, segmental score refines the score obtained by frame-based model at each level. In this way, the best path found in the end of Viterbi algorithm is optimal in both sense

1. INTRODUCTION

There are two major approaches to speech recognition: frame-based and segment-based. Frame-based approach, e.g. HMM modeling, assumes that a sequence of observations are statistically independent of each other so that elegant mathemathical tools can be utilized. This approach is computationally efficient and has been successfully applied in many different applications. However, there are several weakness inherent in HMM. For instance, it can not deal with the correlations and non-stationârity in a state. Besides that, HMM incorporate weak duration constraints in the phonemes they generate. In the past few years, many usefull strategies have been proposed, including explicit modelling of state occupancy by using duration density function [10]; incorporating duration information in a postprocessor [12], etc.

However it is comparatively difficult to handle the correlation and non-stationarity problem inherent in HMM. To improve the recognition rate, some researchers [1][4][11][5] have adopted the segment-based approach which is able to model the temporal variations in the underlying phonological units well. Ostendorf et al.[11] proposed to represent phones as sequence of observations, or stocastic segment. Ghitza et Sondhi[2] used templates as non-stationary state in HMM. Deng[1] dealt with non-stationarity by using parametric function to model the trajectory of mean in each state in HMM. We proposed an algorithm[5] to estimate the parameter of the mean function in the state of HMM so that several mean trajectories could exist in a state rather than one. However the segment-based approaches suffered from two weakness: One results from time warping a variable-length signal to a fixed-length segment model, which is a concatination of mean vectors. Conventionally, a sequence of observations is divided linearly into groups to compare with the mean assigning to it. Rough modelling easily occurs because the score of the observation sequence soly and highly relys on how the signal is mapped to different groups. If non-linear sampling is applied [2], it increases the flexibilty which make the extreme warping possible, and this may also result in a decrease in recognition rate.

The second weakness is that segment-based approaches are computational expensive unless accurate segmentation could be obtained. This is because every possible begining and ending time should be considered for each segment model in dynamic programming. So with each possible begining and ending time pair, the same set of manipulations should be done: time warp the sequence of speech, either linearly[11] or non-linearly[2], to the segment model; compare the distance between the time-warped sequence and the stocked segment and then calculate the probability. No recursion formula could be used even if there is only one observation difference .

To circumvent this weakness inherent in segment-based approach, BBN BYBLOS system[7] uses a segmental neural network (realized by MLP) in the postprocessor to rescore the result obtained from HMM. Since erroneous segmentation might occur when segmenting words into phonemes, this segment-based score might not be so reliable especially if the generalization ability of MLP is not well garanteed by training. Nevertheless, it does make up for, to some extent, the defect inherent in HMM.

Although this postprocessor approaches, has had some success [7][12], the questions of optimality of the estimate, robustness of the solution, and other criteria for successful utilization of segmental information remain unanswered [6].

0-7803-2431-5/95 $4.00 © 1995 IEEE

In this paper we propose a new framework to consider the segmental feature and segmental model in a unified way into the frame-based model. We use conventional HMM to calculate the probability matrix frame by frame, state by state in Viterbi algorithm. In searching for the most probable state sequence through the model via backtracking, we incorporate segmental score state by state, or phone by phone, as we wish to, so that the state sequence obtained in backtracking is "optimal" in terms of joint maximization of observations in the sequence both at the frame and the segment level rather than only at the frame level as in the conventional Viterbi algorithm nor soly at the segment level as in the conventional segment-based approach. Temporarily backtracking is needed in Viterbi process at each segment level. This will be explained in section 2, where the whole framework will be given. Then we will describe our experiments and show some results on the recognition of isolated words over telephone line.

2. THE FRAMEWORK

We present the new framework within the context of recognition of isolated words which are represented as a sequence of subword units, such as phonemes, modeled by HMM with a few states.

The question is that given the observation sequence $O=O_1O_2 \cdots O_T$, which represents the spectrum of the speech signal at frame level, and the model λ, how do we choose a state sequence $Q=q_1q_2 \cdots\ q_T$ which is optimal in maximizing $P(O,Q/\lambda)$ with segmental constraints? In another word, how can we incorporate segmental feature $V=V_1(t_1),V_2(t_2),\cdots V_k(t_k),\cdots,V_M(t_M)$ before the segmentation $t_1t_2 \cdots t_k \cdots t_M$ is finally known by Viterbi algorithm? Here T is the number of observation and M is the number of segment. t_k is the end boundary of the segment V_k. $V_k(t_k)$ represents the segmental feature or the transformed segmental vector up to time t_k. The end of a segment should be the end of a state but the end of a state is not necessarily the end of a segment. It depends on how the segment is defined. If we define a segment as a phoneme, it is modeled by several states of HMM. However if a segment is supposed to be a microsegment, it corresponds to each state in HMM. The following algorithm can be applied to any definition.

Since Viterbi algorithm is a dynamic programming algorithm, it has the property that the remaining decisions must constitute an optimal policy with regard to the state resulting from the decision made before. So at each state there already exists optimal path towards it and the segmentation is known in this sense. In another word $V_1(t_1)$, $V_2(t_2), \ldots V_k(t_k)$ is state dependent and time dependent segmental feature or vectors.

The criterion used to find the best state sequence here is to maximize $P(O,Q \mid \lambda)$, which is usually the summation of score of frames along the path, with the segmental constraint produced by $V_1(t_1)$, $V_2(t_2), \ldots V_k(t_k)$. We use the Viterbi algorithm but put in a state synchronous manner for the sake of simplicity although it can be used frame synchronously as in [9].

2.1. Modified Viterbi Algorithm.

Suppose that

$$\delta_{j+1}(t) = \max_{q_1 \cdots q_{t-1}} P\left[q_1 \cdots q_t = j+1, O_1 \cdots O_t \middle| \lambda\right]$$

is the best score along a single path, at time t, with the first t observations ends in state $j+1$.

We also need to define two probability variables, $P_{j+1}\left(O_{u+1}^t\right)$, $W_j(V_j(u))$, one from the point of frame-based model and the other from the point of segment-based model.

First, $P_{j+1}\left(O_{u+1}^t\right)$ is defined as observation sequence probability from time $u+1$ to time t at state $j+1$.

$$P_{j+1}\left(O_{u+1}^t\right) = b_{j+1}(O_{u+1}) b_{j+1}(O_{u+2}) \cdots b_{j+1}(O_t)$$

Here we have assumed statistical independence of observations. $b_{j+1}(\cdot)$ is the observation probability distribution in state $j+1$.

Second, $W_j(V_j(u))$ is defined as the probability of a segment represented by a feature, like duration, or by a transformed segmental vector, like in SSM[11]. We will explain how to use it later.

The main idea used in this new framework for incorporating segmental feature or segmental model in the frame-based Viterbi algorithm is to replace the transition coefficient in conventional HMM with $W_j(V_j(u))$ at each transition from one state to another because it is known that duration probability density inherent in HMM by using transition coefficient has a exponential form which is inappropriate for most physical signals.

The modified Viterbi algorithm is as follows.

1) Initialization:

$$\delta_1(t) = P_1\left(O_1^t\right) \qquad 1 \le t \le T_{1e}$$

$$\varphi_1(t) = 0$$

2) Recursion:

$$\delta_{j+1}(t) = \max_{T_{jb} \le u \le T_{je}} \left[\delta_j(u) W_j\left(V_j(u)\right)\right] P_{j+1}\left(O_{u+1}^t\right)$$

$$\varphi_{j+1}(t) = \arg\max_{T_{jb} \le u \le T_{je}} \left[\delta_j(u) W_j\left(V_j(u)\right)\right] P_{j+1}\left(O_{u+1}^t\right)$$

$$T_{(j+1)b} \le t \le T_{(j+1)e}$$

$$1 \le j < N$$

3) Termination:

$$P(O,Q|\lambda)$$

$$= \max_{T_{(N-1)b} \le u \le T_{(N-1)e}} \left[\delta_{N-1}(u) W_{N-1}\left(V_{N-1}(u)\right)\right] P_N\left(O_{u+1}^T\right)$$

Here T_{jb} ,T_{je} is the possible begining and ending boundary for state j.

Since $P_{j+1}\left(O_{u+1}^t\right)$ only requires information from the previous frame, this does not cost more computation than conventional HMM.

$$P_{j+1}\left(O_{u+1}^t\right) = P_{j+1}\left(O_{u+1}^{t-1}\right) b_{j+1}(O_t)$$

2.2. Combination of Two Sorts of Models.

The advantage of using $W_j(V_j(u))$ is that it can merge two separately considered ideas—segmental feature (e.g. duration) and segment-based model into one and apply them into frame-based model (HMM).

From the property of dynamic programming algorithm, it is clear that there already exits the best path with regard to the state up to present. If the definition of a segment is a phoneme and the state j is the final state of this phoneme, then the boundary of this segment can be found by temporarily backtracking to the begining state of this phoneme (Fig. 1). The way of incorporation can be:

1) Suppose that $V_j(u)$ represents the duration of the segment ending at time u in state j. After temporarily retrieving we find that this segment starts at time r according to the best path to the state j. Then,

$$V_j(u) = u - r;$$

$$W_j(V_j(u)) = P_d(u - r);$$

where $P_d(\cdot)$ is the duration probability density function at state j. It can be gamma function as used by Levinson[10] or simply be uniformly distributed fiunction bounded by minimun and maximun duration constraints as in [8][3].

2) Suppose that $V_j(u)$ represents a segmental vector,

$$V_j(u) = F(O_r, O_{r+1}, \cdots O_u)$$

$F(\cdot)$ is a resampling transformation used to transform a variable length observed segment to a fixed length.

$$W_j\left(V_j(u)\right) = P_s\left[F(O_r, O_{r+1}, \cdots O_u)\right]$$

$P_s(\cdot)$ can be a multivariate Gaussian function used for modelling the segment.

So for time u at state j there is only one most probable segment obtained and resampled to have a fixed length segmental vector. There is no need to consider all the possible begining and ending times to be segmentation boundaries as used in convensional segment-based model. On the other hand non-linear sampling is made possible because the observation sequence in this segment has already been grouped into several states. Resampling in each states to form the segmental vector could not only avoid scaling the variations in speaking rate uniformly by linear sampling but also extreme warping by non-linear resampling in a segment without any restriction. In fact it is the frame-basis information before state j which prunes all the possible boundaries to the most probable one and make an effective non-linear sampling possible by grouping the observation sequence into several states in advance according to their physical poperty.

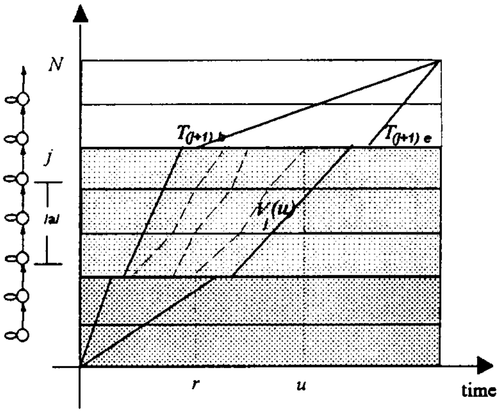

Fig.1 Temporarily backtracking to find the segment boundary for phoneme /a/.

3) It is also possible to apply segmental feature at different level (e.g. state duration, phoneme duration, global word duration) in the same Viterbi algorithm. Each source of information $W_j(V_j(u))$ is applied in Viterbi algorithm at the state which corresponds to the end of a segment before the transition from one state to another happened.

It is clear now that this new framework can incorporate two source of information together: the frame-based $P_j(\cdot)$ prunes the possible segmentations for $V_j(u)$ so that computational load is dramatically reduced; $W_j(V_j(u))$ refine the score $\delta_j(u)$ at each level (state level, phoneme segment level) so that the best path found in the end is not only optimal from the frame point of view but also from the segment point of view.

3. EXPERIMENTS.

We use a database of a vocabulary of 53 isolated words. Every word is spoken by 87 speakers once and it was collected over the american telephone line. 47 speakers are used for training and 40 speakers are used for testing.

Each frame uses 30ms window of signal. Consecutive frames are spaced 10 ms apart. 12 LPC cepstrum coefficients and 12 delta-cepstral coeffients are computed for each frame. Two groups of experiments have been conducted, one without using cepstral mean substruction (**CMS**) [13] and the other with in order to compensate the channel influence by telephone line.

Table 1. Recogniton Rate with Different Methods

	no CMS	with CMS
Baseline	73.7%	80.5%
Mode I	80.06%	86.54%
Mode I+Mode II	82.14%	87.4%
Mode III	80.06%	86.4%

The baseline recognition system uses 3 mixtures of Gaussian density function at each state in hidden Markov model .

According to the last section there are at least 3 sorts of incorporation mode of the segment-based information into the frame-based Viterbi algorithm. The first mode (**Mode I**)— to use state duration constraints into each state, has been realized by [3][8]. In our experiment, after several iterations of training of HMM, duration information is reliable and is extracted for use in the re-training of the HMM model so that the segmentation of the signal is more accurate, which hopes to result in a more accurate model. This method proves to be very effective by the experiment because it changes .the convergence point from the local minimum given by the standard HMM reestimation formula.

The second mode(**Mode II**) is to apply segmental model to the optimal path pruned by Viterbi algorithm at the end of segment state. To insure multi-trajectory in each segment, the probability density function for the segmental model is represented as a mixture of the Gaussian function. Experimental results reveal that it improves a little from the **Mode I**.

The third mode (**Mode III**) is to incorporate different duration information at different level while calculating the probability matrix in Viterbi algorithm: state duration at the end of each state; phoneme duration at the end of phoneme state; word duration at the end of whole word. It seems that no more improvement from **Mode I** meaning that state duration constraint is sufficiently well done.

REFERENCES.

1. Deng , L. " A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Processing*, 27(1),65-78.
2. Ghitza,O, Sondhi,M.M, "Hidden Markov models with templates as non-stationary states:an application to speech recognition", *Computer Speech and language* (1993)2, pp101-119.
3. Gu, H.Y, Tseng, C.H, and Lee, L.S, "Isolated-utterance speech recognition using hidden Markov modles with bounded state durations", *IEEE Trans. on Signal Processing*, Vol.39, No. 8, 1991.
4. Gong,Y.F, Haton,J.P., "Stocastic trajectory modelling for speech recognition", pp I-57—I-60, *ICASSP'94*.
5. He, J, Leich, H, "Combining stocastic trajectory model and discriminative feature in speech recognizer", vol 2, pp681-684, *ICASSP'94*.
6. Juang,B.H, "Issues in using hidden Markov models for speech recognition", *Advances in Speech Signal Processing*, pp509-554
7. Kubala,F, Anastasakos,A, Makhoul,J, Nguyen,L, Schwartz,R, Zavaliagkos,G,"Comparative experiments on large vocabulary speech recognition", *ICASSP'94*, Vol 1, pp561-564, 1994.
8. Lee,C.H, "On the use of some robust modeling techniques for speech recognition", *Computer Speech and Language*, pp35-52, No.3, 1989.
9. Lee, C.H, "A frame-synchronous network Search Algorithm for Connected Word Recognition", *IEEE Trans. on ASSP*, Vol.37, No. 11, 1989.
10. Levinson, S.E, "Continuously variable duration hidden Markov models for speech analysis", *ICASSP'86*, pp1241-1244, 1986.
11. Ostendorf, M, Roukos, S, "A stocastic segment model for phoneme-based continuous speech recognition", *IEEE Trans. ASSP*, Vol 37, No.12, pp1857-pp1869, 1989.
12. Rabiner,L.R, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, Vol 77, No 2, pp257-286, 1989.
13. C, Monné,J , Jouvet,D "On-line adaptation of speech recognizer to variations in telephone line conditions", *Proceeding of EUROSPEECH'93*, pp1247-1250.