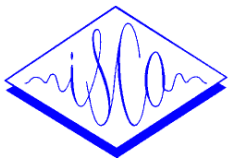

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

NONLINEAR TIME ALIGNMENT IN STOCHASTIC TRAJECTORY MODELS FOR SPEECH RECOGNITION

Mohamed Afify[1,2] *Yifan Gong*[1] *Jean-Paul Haton*[1]

[1]CRIN–CNRS & INRIA Lorraine, BP 239, 54506 Vandoeuvre–lès–Nancy, France
[2]National Telecommunication Institute, Cairo, Egypt

ABSTRACT

A nonlinear time alignment technique is presented in the framework of stochastic trajectory models (STM). We show how to obtain maximum likelihood (ML) estimates of model parameters, and how to use the technique during recognition with a slight additional computational overhead. Experimental results for a French 850 word continuous speech task are given. For a 10 speaker population, we test with various degrees of nonlinearity, and the introduced technique provides a slight improvement (about 1%) in the average word recognition rate.

1. INTRODUCTION

Hidden Markov models (HMM) have proven to be efficient for continuous speech recognition [1]. Yet their performance –being frame oriented– is limited by the independence assumption given an underlying state sequence, and by the stationarity assumption on state statistics. A new promising technique for speech recognition based on segment modelling has recently been used [2]-[4], and recognition improvements have been reported using this approach.

Recently Vinics, a continuous speech recognition system has been successfully implemented using stochastic trajectory models (STM) [3]. A basic assumption about the trajectory models used in Vinics is linear time alignment. In this work we try to relax this assumption by extending the linear alignment function to a general warping function.

Originally linear warping was motivated by its simplicity, and by the assumption that at the phoneme level it preserves some information in discrimination of some sounds. On the otherhand it could be argued that linear warping is not optimal in the sense that the model likelihood could be further optimized with respect to the alignment function, further nonlinear alignment provides a new degree of freedom which may result in a better representation of the data in the states of the trajectory, and finally we are encouraged by HMMs which owe a part of their success to the powerful nonlinear alignment technique inherent in their structure.

We are not aware of any attempts to use nonlinear warping in segment based models except in [4], this was accomplished by using classical DTW throughout training and testing. Here we apply the nonlinear warping as a postprocessor during training, this provides a kind of smoothing on the final trajectories after establishing their structure, and also facilitates testing various degrees of nonlinearity starting from the original structure. Also we point out how to use nonlinear alignment during testing with a slight additional computational overhead.

In Section 2 we briefly review the original trajectory modelling in Vinics. The nonlinear alignment technique is discussed in detail in Section 3. In Section 4 multi-speaker sentence recognition results are given for a 850 word application. The experiments compare the nonlinear alignment approach to the original formulation. Finally conclusions are given in Section 5.

2. THE BASELINE SYSTEM

2.1. Basic Formulation

We assume that each phoneme s is represented by K trajectories in a parameter space. Each trajectory is assumed to consist of Q points. The probability of a segment X_n centred at frame n and of length d frames given trajectory T_k of phoneme s can be written as

$$P(X_n|T_k,d,s)=\prod_{i=0}^{Q-1}\mathcal{N}(x_{n-\frac{d}{2}+i\frac{d}{Q}};m_{k,i}^{(s)},\Sigma_{k,i}^{(s)})^{w_i} \quad (1)$$

where Q is the number of states of a trajectory, w_i's are weighting factors to compensate for the larger variance of extreme states, and $m_{k,i}^{(s)}$ and $\Sigma_{k,i}^{(s)}$ are respectively the mean vector and the covariance matrix associated with the k^{th} trajectory of the i^{th} state for the phoneme s.

The mapping from a d length segment to the Q points of a trajectory is done deterministically using a linear warping function as given in (2)

$$x_{f(n,i)}=x_{n-\frac{d}{2}+i\frac{d}{Q}} \quad (2)$$

2.2. Training

For a fixed number of trajectories the training phase consists of a K-means loop. Segment classification is done using (1) as similarity measure, and parameter estimates of a trajectory are obtained using ML estimation as given in (3) and (4). The above two steps are repeated until convergence.

$$m_{k,i}^{(s)}=\frac{1}{N_k}\sum_{n=1}^{N_k}x_{n,i} \quad (3)$$

$$\Sigma_{k,i}^{(s)}=\frac{1}{N_k}\sum_{n=1}^{N_k}(x_{n,i}-m_{k,i}^{(s)})(x_{n,i}-m_{k,i}^{(s)})^T \quad (4)$$

where N_k is the number of segments assigned to trajectory T_k.

In practice for a set of labelled utterances training is performed using the well known LBG algorithm [5] with binary splitting. We start with one trajectory for each phone, and perform splitting until a maximum prespecified number of trajectories (depending on the number of training samples of the phone) is reached.

It should be also noted that training can be accomplished using unlabelled speech by introducing an additional loop over all possible segmentations, but preliminary experimentation with our system indicated almost no performance difference.

2.3. Phoneme Recognition

We define the plausibility $\mu_{n,s}$ of phoneme s at time instant n as $\mu_{n,s} \equiv \log P(s|X_n)$. Regarding $P(X_n)$ as a constant for the purpose of phoneme recognition we can write

$$\mu_{n,s} = \log[\sum_{d} P(X_n|s,d)P(d|s)]P(s) \qquad (5)$$

for computational efficiency, we replace $\sum_d$ in (5) by $\max_d$, and we write

$$\mu_{n,s} = \log[\max_{d} P(X_n|s,d)P(d|s)]P(s) \qquad (6)$$

where we have

$$P(X_n|d,s) = \sum_{k} P(X_n|T_k,d,s)P(T_k|d,s) \qquad (7)$$

where we assume $P(T_k|d,s) = P(T_k|s) \quad \forall d$

By using (6) and (7) we can calcaulate the plausibility $\mu_{n,s}$ for each symbol s at each time instant n. Once the plausibility is calcaulated it can be used in a dynamic programming (DP) search for sentence recognition as will be discussed in the next subsection .

2.4. Sentence Recognition Formulation

Once we have calcaulated the plausibilities $\mu_{n,s}$ of the phoneme symbols at each time instant, the sentence recognition problem could be easily formulated using dynamic programming (DP).

Let

$\mathcal{F}$ be the set of all grammatical sentences

$\omega \in \mathcal{F}$ be made up of $L(\omega)$ symbols of the phonetic alphabet $\{a_j\}$ $0 \le j < L(\omega)$

and N be the number of frames of ω

In the original Vinics formulation [3], the plausibility of each sentence $\omega \in \mathcal{F}$ can be calcaulated using dynamic programming as follows

$$\Pi(l,j) = \max_{0 \le k < l} \{\Pi(k,j-1) + Pr(l-k|a_j)^{\gamma} \sum_{k \le n < l} \mu_{n,a_j}\} \qquad (8)$$

$0 \le l < N$, $0 \le j < L(\omega)$

where

γ is a duration penalty control parameter

$\Pi(l,j)$ is the cumulated plausibility for the j^{th} symbol at the l^{th} frame.

The recognized sentence is chosen as

$$\omega^* = \arg\max_{\omega \in \mathcal{F}} \frac{\Pi(N-1, L(\omega)-1)}{N} \qquad (9)$$

3. NONLINEAR TIME ALIGNMENT

3.1. Formulation

In this section we want to replace the linear transformation in eqn. (2) by a nonlinear warping function with a generic form of $f_w(i)$, and we write

$$x_{f(n,i)} = x_{n-\frac{d}{2}+f_w(i)} \qquad (10)$$

As we previously noted our motivations to use the nonlinear warping function are:

- We want to further optimize the model parameters with respect to the warping function.
- We provide a new degree of freedom to the models which could result in better representation of data.
- We are encouraged by the powerful nonlinear alignment technique of HMM.

3.2. Training

Once trajectories are established using the training algorithm discussed in section (2.2), training for nonlinear alignment is performed in two steps using the same set of labelled utterances as shown below

- We estimate $m_{k,i}^{(s)}$, and $\Sigma_{k,i}^{(s)}$ given a warping function $f_w(i)$ as discussed in section (2.2).
- Given $m_{k,i}^{(s)}$, and $\Sigma_{k,i}^{(s)}$ we find $f_w(i)$ to maximize $P(X_n|T_k,d,s)$ using dynamic programming (DP).

and we repeat the abovetwo steps until convergence.

By taking logarithms in eqn. (1) and after some trivial simplifications, we can deduce that finding the optimal warping function in step 2 above reduces to the minimization of $l_k^*(X_n|T_k,d,s)$ given by

$$l_k^*(X_n|T_k,d,s) = \sum_{i=0}^{Q-1} w_i (x_{n,i} - m_{k,i}^{(s)})^T \Sigma_{k,i}^{(s)-1} (x_{n,i} - m_{k,i}^{(s)}) \qquad (11)$$

Noting that $(x_{n,i} - m_{k,i}^{(s)})^T \Sigma_{k,i}^{(s)-1} (x_{n,i} - m_{k,i}^{(s)})$ is the Mahalanobis distance d_M between frames $x_{f_w(i)}$ of the segment and m_i of the mean template of the trajectory, we can write (11) as

$$l_k^*(X_n,T_k,d,s) = \sum_{i=0}^{Q-1} w_i d_M(x_{f_w(i)}, m_i) \qquad (12)$$

where dependence on k and s has been dropped from the notation for simplicity and is to be understood from the sequel.

The minimization in (12) can be viewed as a classical dynamic time warping problem for the segment and the mean template using the Mahalanobis distance measure. The $w_i's$ can be considered as a weighting function on the warping path reflecting the relative importance of different points on the path. Recently the optimization of the values of these weights has been considered using generalized probabilistic descent (GPD), and connectionist approaches [6].

In this work we use a heuristic function reflecting the importance of the centre part of the segment. In fact we have tried using a connectionist framework for the optimization of the $w_i's$ but we abondened it due to degraded recognition results.

In order to avoid singular warpings that may result in recognition errors, we must choose a suitable set of constraints on the warping path. We use the well known parallelogram constraints [7], but we parametrize the slopes of the lines forming the parallelogram to be α and $\frac{1}{\alpha}$ (usually they are 2 and $\frac{1}{2}$). We empirically determine the value of the warping parameter α (where α is a positive integer) giving best recognition accuracy.

In the case of a segment length requiring a warping function outside the range of our constraints we linearly downsample the corresponding segemnt to the maximum permissible length given the set of constraints. This motivated us to modify the parallelogram constraints (as

discussed above) to avoid (as much as possible) the downsampling process, hoping to obtain the best data representation in the states of the trajectory.

We modify the usual DP equations to

$$D(i,j) = d_M(i,j) + \min_{0\le\beta\le\alpha} D(i-1, j-\beta)W(i,j,\beta) \quad (13)$$

$i = 0,..,Q-1$ & $j = 0,..,d-1$
and

$$W(i,j,\beta) = \begin{cases} \infty & \text{if } P(i-1,j) = j \ \& \ \beta = 0 \\ 1 & \text{otherwise} \end{cases} \quad (14)$$

where
$D(i,j)$ is the cumulative distance at frame i of mean template and frame j of segment,
$d_M(i,j)$ is the Mahalanobis distance between frame i of the template and frame j of the segment.
and

$$P(i,j) = j - \arg\min_{0\le\beta\le\alpha} D(i-1, j-\beta)W(i,j,\beta) \quad (15)$$

We impose that $D(0,0) = d_M(0,0)$ and we get $l_k^*(X_n|T_k,d,s) = D(Q-1,d-1)$. After an alignment, the best path $f_w(i)$ could be easily obtained by backtracking the predecessor array P.

3.3. Phoneme Recognition

We keep the definition of symbol plausibility as discussed in Section (2.2), but now we must do the nonlinear warping at each frame n for all possible lengths d and all trajectories k to calcaulte (6). The previous process is computationally expensive to implement in practice, but by storing the frame probabilities for each state i and trajectory k we can perform the nonlinear warping using table lookup and some additions and comparisons. Thus the nonlinear warping can be done in practice using a small extra computational overhead. The extension to sentence recognition is the same as that for the linear case, and was discussed in Section (2.4).

4. EXPERIMENTAL RESULTS

Experiments deal with a 850 word continuous speech recognition task. The grammar used has a word perplexity of 22.5. We present word recognition rates for 10 speakers. For each speaker 140 short sentences of automatically labelled speech are used for training, and 160 sentences are used for testing.

Speech recordings are collected by reading text in an office environment, using a SUN desktop omni-directional microphone, stucked on the terminal. The average distance between a speaker and the microphone is about 40 cm and it increases to 60 cm for some speakers (jel,crm). The average SNR of the recordings is about 15 db. Three of the speakers (yig,ols,syc) are experienced with speech recognition projects, while others are not, and they often make pauses between words, which are not modelled by the grammar. Speech is sampled at 16 KHz, and 32 ms frames separated by 10 ms intervals are used for the analysis.

We use 13^{th} order mel cepstrum coefficients [8] as recognition features. The basic units are 32 context independent phone trajectory models. Each phone is represented by up to 8 ($K \le 8$) trajectories, and each trajectory consists of 5 states (Q=5) and has diagonal covariance matrices.

The results shown in table 1 are the word recognition rate for each speaker, for values of the warping parameter α ranging from 1 to 4 (where $\alpha = 1$ corresponds to linear warping). We didnot use larger warping parameters because a value of 4 is sufficient to allow almost all permissible segment lengths to be included within the parallelogram.

		ALPHA			
		1	2	3	4
SPK	brs	88.92	89.73	90.37	90.21
	crm	80.21	82.02	81.12	79.91
	jel	60.07	61.74	62.75	62.89
	lar	98.54	98.39	98.39	98.39
	ols	96.74	97.39	97.55	97.55
	sat	80.36	81.18	80.69	80.52
	std	86.46	86.30	86.00	86.00
	syc	76.48	79.20	79.36	77.44
	vil	90.40	91.55	91.69	90.97
	yig	89.31	89.62	89.62	90.23
	AVG	84.71	85.71	85.75	85.41

Table 1 Word recognition rate for various warping parameters

The results presented in table 1 show a slight improvement in the word recognition rate (about 1 %) which is almost constant for various degrees of nonlinearity.

5. CONCLUSION

We have presented a nonlinear alignment technique which can be used in the framework of segment modelling for continuous speech recognition. We showed how to obtain the model estimates and how to use the technique in recognition with slight computational overhead. We also presented multispeaker word recognition results that indicated a slight improvement in the recognition accuracy when using the proposed approach. We conclude that for our experimental conditions the linear warping assumption is accurate and can be used in further investigations of trajectory models.

REFERENCES

[1] X.Huang et al., "The SPHINX-II speech recognition system:an overview", Computer Speech and Language, No. 2, pp. 137-148, Apr. 1993.

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

[3] Y.Gong,
and J.P.Haton, "Stochastic trajectory modelling for speech recognition", in Proc. ICASSP'94, Adelaide, Australia, Apr. 1994.

[4] O.Ghitza, and M.M.Sondhi, "Hidden Markov models with templates as non-stationary states:an application to speech recognition", Computer Speech and Language, No.2, pp.101-119, Apr. 1993.

[5] Y.Linde, A.Buzo, and R.M.Gray, "An algorithm for the vector quantizer design", IEEE Trans. Commun., Vol.28, No.1, pp.84-95, Jan. 1980.

[6] P.C.Chang, S.H.Chen, and B.H.Juang, "Discriminative analysis of distortion sequences in speech recognition", IEEE Trans. Speech and Audio Processing, Vol.1, No.3, pp.326-333, July 1993.

[7] F.Itakura, "Minimum prediction residual principle applied to speech recognition", IEEE Trans. ASSP, Vol.23, No.1, pp.67-72, Feb. 1975.

[8] S.B.Davis, and P.Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. ASSP, Vol.28, No.4, pp.357-365, Apr.1980.