TEMPORAL CONSTRAINTS IN VITERBI ALIGNMENT FOR SPEECH RECOGNITION IN NOISE

Nestor Becerra Yoma[1,2] *Lee Luan Ling*[1] *Sandra Dotto Stump*[2]

[1]*DECOM/FEEC/UNICAMP*
CP 6101, 13083-970 Campinas-SP, Brazil
nestor@decom.fee.unicamp.br
[2]*MACKENZIE UNIVERSITY*
Rua da Consolação 896, 01302-000 São Paulo-SP, Brazil

10.21437/Eurospeech.1999-633

ABSTRACT

This paper addresses the problem of temporal constraints in the Viterbi algorithm using conditional transition probabilities. The results here presented suggest that in a speaker dependent small vocabulary task the statistical modelling of state durations is not relevant if the max and min state duration restrictions are imposed, and that truncated probability densities give better results than a metric previously proposed [1]. Finally, context dependent and context independent temporal restrictions are compared in a connected word speech recognition task and it is shown that the former leads to better results with the same computational load.

1. INTRODUCTION

The transition probability is represented by a constant in the ordinary HMM topologies and this leads to a geometric probability density for state duration which is not accurate for most cases. Consequently, bounding or modelling state durations seems an interesting approach to reduce the error rate specially when the speech signal is corrupted by noise. Several techniques [2] [3] [4] to include state duration modelling in the HMM training procedure have been proposed but the methods require a high computation load. Including temporal constraints in the recognition process is conceptually simpler and generally requires a lower computational load. In this paper the procedure suggested by [1] was followed, where every state was associated to a gamma distribution whose parameters were estimated using the training database after the HMMs had been trained. The contributions of this paper concern: a) conditional transition probabilities to include state duration modelling; b) trucation of the conditional transition probabilities using the *max* and *min* possible state duration; c) comparison of the truncated gamma and geometric probability distributions; d) comparison of the conditional transition probabilities with the metric proposed in [1]; and e) comparison of context dependent (CD) with context independet temporal constraints.

For the task here considered this paper showes that the accurate statistical modelling of state duration (eg with gamma probability distribution) does not seem to be very relevant if max and min state duration restrictions are imposed, and that the introduction of max and min duration to states gives better results than the metric proposed in [1]. Finally, connected word recognition experiments suggest that the CD temporal constraints are much more effective than the ordinary CI temporal restrictions with the same computational load.

2. CONDITIONAL PROBABILITIES

In [1] the state duration contribution to the total Viterbi metric (given the topology shown in Fig. 1), when making a transition from state i at time t to state j at time $t+1$, was equal to

$$P_{i,j}^{(t)} = \begin{cases} \log[d_i(\tau+1)] - \log[d_i(\tau)] & \text{if } i = j \\ \log[d_i(1)] & \text{if } i \neq j \end{cases} \quad (1)$$

where τ is the number of frames in state i up to time t and $d_i(\tau)$ is the probability of state duration equal to τ. In this paper, state duration modelling is included in the Viterbi algorithm by means of the generalization of transition probabilities:

$$a_{i,j}^{(\tau)} = Prob(s_{t+1} = j | s_t = s_{t-1} = ... = = s_{t-\tau+1} = i) \quad (2)$$

where $j = i$ or $j = i+1$ given the topology shown in Fig. 1. Using these definitions for the transition probabilities, $a_{i,i}^{(\tau)}$ and $a_{i,i+1}^{(\tau)}$ can be estimated by

$$a_{i,i}^{(\tau)} = \frac{D_i(\tau) - d_i(\tau)}{D_i(\tau)} \quad (3)$$

$$a_{i,i+1}^{(\tau)} = \frac{d_i(\tau)}{D_i(\tau)} \quad (4)$$

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

where $D_i(\tau)$ is the probability of state i being active for $t \geq \tau$:

$$D_i(\tau) \quad = \quad \sum_{t=\tau}^{\infty} d_i(t) \tag{5}$$

If the original geometric distribution is used, (1) coincides with (3) and (4). If a generic probability distribution is considered, (3) and (4) lead always to $a_{i,i}^{(\tau)} + a_{i,i+1}^{(\tau)} = 1$ which is coherent with the definition of probability. On the other hand, interpreting $P_{i,i+1}^{(t)}$ in terms of transition probability leads to $a_{i,i}^{\tau} = \frac{d_i(\tau+1)}{d_i(\tau)}$ and $a_{i,i+1}^{\tau} = d_i(1)$ which do not satisfy $a_{i,i}^{(\tau)} + a_{i,i+1}^{(\tau)} = 1$ except for the geometric distribution. To include the possible *min* and *max* durations, $min_i(\tau)$ and $max_i(\tau)$ respectively, the transition probabilities were modified to:

$$a_{i,i}^{\tau} \quad = \quad \begin{cases} 1 & \text{if } \tau < t_{min_i} \\ 0 & \text{if } \tau \geq t_{max_i} \\ \frac{D_i(\tau) - d_i(\tau)}{D_i(\tau)} & \text{otherwise} \end{cases} \tag{6}$$

$$a_{i,i+1}^{\tau} \quad = \quad \begin{cases} 0 & \text{if } \tau < t_{min_i} \\ 1 & \text{if } \tau \geq t_{max_i} \\ \frac{d_i(\tau)}{D_i(\tau)} & \text{otherwise} \end{cases} \tag{7}$$

where $t_{min_i} = tol\_min \cdot min_i(\tau)$ and $t_{max_i} = tol\_max \cdot max_i(\tau)$. The constants tol_min and tol_max introduce a tolerance to the min and max duration for every state.

The discrete gamma distribution is given by:

$$d_i(\tau) \quad = \quad K \cdot e^{-\alpha \cdot \tau} \cdot \tau^{p-1} \tag{8}$$

where $\tau = 0, 1, 2, ...$ is the duration of a given state i in number of frames, $\alpha > 0$, $p > 0$ and K is a normalizing term. The mean duration ($E_i(\tau)$), the variance ($Var_i(\tau)$), and the *max* and *min* durations were computed for every state in each model by means of estimating the optimal state sequence for every training utterance using the Viterbi algorithm after training the HMMs. The parameters α and p were estimated by:

$$\alpha_i \quad = \quad \frac{E_i(\tau)}{Var_i(\tau)} \tag{9}$$

$$p_i \quad = \quad \frac{E_i^2(\tau)}{Var_i(\tau)} \tag{10}$$

The truncated transition probabilies using the geometric distrubution of the ordinary HMM topology are given by:

$$a_{i,i}^{\tau} \quad = \quad \begin{cases} 1 & \text{if } \tau < t_{min_i} \\ 0 & \text{if } \tau \geq t_{max_i} \\ a_{i,i} & \text{otherwise} \end{cases} \tag{11}$$

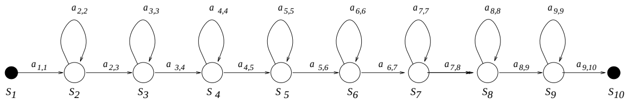

Figure 1. Eight-state left-to-right HMM without skip-state transition.

$$a_{i,i+1}^{\tau} \quad = \quad \begin{cases} 0 & \text{if } \tau < t_{min_i} \\ 1 & \text{if } \tau \geq t_{max_i} \\ a_{i,i+1} & \text{otherwise} \end{cases} \tag{12}$$

where $a_{i,i}$ and $a_{i,i+1}$ are transition probabilities estimated during the training algorithm.

3. CONTEXT-DEPENDENT STATE DURATION MODELLING

In HMM based speech recognition, word modelling (i.e. a HMM per each vocabulary word) is generally used for small vocabulary recognition systems. Consequently, in a small vocabulary connected word recognition task each HMM attempts to capture the coarticulation effect between contiguous words indepently of the context (i.e. a word is represented by only one HMM). In [1] and [5], where word modelling was used, every state duration distribution was modelled with only one probability density which is reasonable for the isolated word task but is not very accurate for connected word recognition. In the latter case, the state duration (at least for the first and last states) depends on the context and this paper proposes that the word's position in the string carries information which should be used to improve the accuracy of temporal restrictions. Table 1 showes the mean duration of the first and last states, respectively, of the word "zero" HMM in triplets (i.e. three digit strings) when the word is string-initial, when it is between two other words and when it is string-final. As can be seen in Table 1, the context-independent state duration distribution is not a precise model due to the fact that the mean state duration strongly depends on whether there is or not a "silence" interval before and/or after the word. In order to compare the context-dependent with the context-independent temporal constraints, the temporal restrictions according to (11) and (12) were used. For the task here considered, two set of state duration models were estimated: context-independent (CI) and context-dependent (CD). The CI model is composed by the overall min_i^O and max_i^O state durations independently of the word's position in the string. The CD model is composed by three set of *min* and *max* state durations: min_i^I and max_i^I, when the word is string-initial; min_i^M and max_i^M, when the word is preceded and followed by other words; and min_i^F and max_i^F, when the word is string-final.

Table 1. Mean duration of the first and last states of the word "zero" HMM according to the word's position in the string.

State	*string-initial*	*middle*	*string-final*
First state	15.0	5.3	6.2
Last state	2.0	2.4	10.1

4. EXPERIMENTS

The proposed methods were tested with speaker-dependent isolated and connected word (English digits from 0 to 9) recognition experiments. The tests were carried out employing the two speakers (one female and one male), and the car noise from the Noisex database [6]. Where convolutional noise experiments were performed a +6dB/oct spectral tilt was applied. The signals were downsampled to 8000 samples/sec. The signal was divided in 25ms frames with 12.5ms overlapping. Each frame was processed with a Hamming window before the spectral estimation. The band from 300 to 3400 Hz was covered with 14 Mel DFT filters. At the output of each channel the energy was computed, SS according to [7] was applied and the log of the energy was estimated. In every frame 10 cepstral coefficients were computed.

Each word was modelled using an 8-state left-to-right topology (see Fig. 1) without skip-state transition, with a single multivariate Gaussian density per state and a diagonal covariance matrix. The HMM's were estimated by means of the clean signal utterances. $E_i(\tau)$, $Var_i(\tau)$, $max_i(\tau)$ and $min_i(\tau)$ were estimated using the training database after the HMMs had been trained by means of Viterbi alignment. Due to the fact that the experiments were speaker dependent, in some cases it was observed that the variation in state duration was very low, which resulted in a low $Var(\tau)$ which in turn caused a low recognition accuracy. To counteract this, a threshold was introduced to set a floor for $Var(\tau)$ in the parameter estimation for the gamma distribution (9) (10).

Four experiments were done with isolated word recognition: the ordinary Viterbi algorithm Vit; the Viterbi algorithm with max and min state duration plus state duration distribution with gamma pdf (6) (7) $Vit - Mm - Gamma$; the Viterbi algorithm with max and min state duration plus the ordinary geometric distribution (11) (12) $Vit - Mm - Geom$; and finally, the Viterbi algorithm with the metric (1) proposed in [1], $Vit - metric$. Results are shown in Fig. 2 and Table 2.

In order to test the validity of context-dependent (CD) state duration modelling from the connected word speech recognition point of view, three experiments were performed: the ordinary Viterbi algorithm Vit; and the Viterbi algorithm with max and min state duration plus the ordinary geometric distribution (11) (12) using context-independent $(Vit - CI - Mm - Geom)$ and context-dependent $(Vit - CD - Mm - Geom)$ temporal restrictions. Results are shown in Tables 3 and 4.

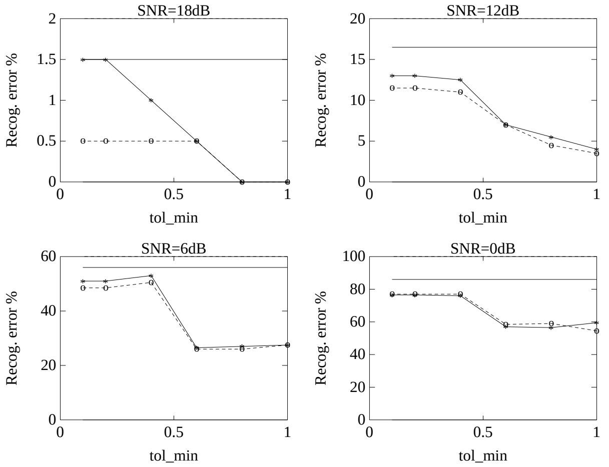

Figure 2. Recognition error rate (%) vs *tol_min* ($tol\_max = 1.5$) **in isolated word recognition experiments for speech signal corrupted by additive noise (car): (–),** *Vit* **; (-o-),** *Vit-Mm-Gamma* **; and (-*-),** *Vit-Mm-Geom* **.**

5. DISCUSSION AND CONCLUSION

As can be seen in Fig. 2 and Table 2, the introduction of temporal constraints $Vit - Mm - Gamma$ and $Vit - Mm - Geom$ substantially reduced the error rate when compared to the ordinary Viterbi algorithm with both noises and at all the SNR's. According to Fig.2, the introduction of t_{min_i} abruptly increased the recognition accuracy. The gamma and geometric distributions were not very sensitive to t_{max_i} and this should be due to the fact that both distributions are monotonically decreasing for $\tau \to \infty$. However, it is worth mentioning that a low t_{max_i} can reduce the number of computations for the gamma distribution without affecting the error rate. Another result is that the state duration modeling using the gamma

Table 2. Comparison of temporal constraints. Recognition error rate(%) in isolated word recognition experiments for speech corrupted by car noise ($tol\_min = 0.8$ **and** $tol\_max = 1.5$)**.**

SNR	*18dB*	*12dB*	*6dB*	*0dB*
Vit	1.5	16.5	66	86
Vit-Mm-Gamma	0	4.5	26	59
Vit-Mm-Geom	0	5.5	27	56.5
Vit-metric	0.5	14.5	53	86

Table 3. Recognition error rate (%) in connected word recognition experiments for speech signal corrupted by additive noise (car).

SNR	*18dB*	*12dB*	*6dB*	*0dB*
Vit	21.0	44.7	84.0	95.0
Vit-CI-Mm-Geom	11.0	30.7	55.4	79.3
Vit-CD-Mm-Geom	2.0	7.7	26.7	62.7

Table 4. Recognition error rate (%) in connected word recognition experiments for speech signal corrupted by convolutional noise (6dB/oct spectral tilt).

SNR	*6dB/oct spectral tilt*
Vit	4.4
Vit-CI-Mm-Geom	2.7
Vit-CD-Mm-Geom	0.4

distribution almost did not improve the recognition accuracy when compared with the ordinary geometric one using the same max and min state durations. This result suggests that in a speaker dependent task the statistical modelling of state durations is not relevant and the restrictions imposed by the max and min durations are the main factor responsible for the reduction in the error rate. It is also shown (Table 2) that the temporal constraints according to (6) (7) and (11) (12), $Vit-Mm-Gamma$ and $Vit-Mm-Geom$ respectively, gave better results than the metric in (1), $Vit-metric$, which confirms the validity of imposing max and min for state duration in a speaker dependent task.

Connected word experiments with speech signals corrupted by only additive or convolutional noise (Tables 3 and 4) show that CD temporal restrictions gave a higher recognition accuracy than CI temporal constraints. As can be seen in Table 3 (additive noise), when compared with the ordinary Viterbi algorithm CI temporal constraints led to reductions as high as 91, 83 and 68% in the error rate at SNR=18, 12 and 6dB, respectively. According to Table 4 (convolutional noise) almost a 100% reduction in the error rate was achieved when CI temporal constraints were employed.

To conclude, the results presented in this paper suggest that in a speaker dependent task temporal constraints can lead to reductions in the error rate as high as 50 or 80% at SNR's equal to 18, 12 and 6dB, and the accurate statistical modelling of state duration (eg with gamma probability distribution) does not seem to be very relevant if max and min state duration restrictions are imposed. The introduction of max and min duration to states gives better results than the metric proposed in [1], at least for the task here considered. Finally, connected word recognition experiments showed that CD temporal constraints are much more effective than the ordinary CI temporal restrictions with the same computational load. The results presented in this paper are very encouraging and further work is currently in progress to evaluate applicability of the approach here proposed in more complex tasks such as speaker independent speech recognition. Due to the fact that state duration modelling does not need any information about the testing environment, it is an interesting technique from the practical application point of view.

6. ACKNOWLEDGMENT

Nestor Becerra Yoma was supported by a grant from FAPESP, Sao Paulo, Brasil.

REFERENCES

[1] D.Burshtein. Robust parametric modeling of durations in hidden Markov models. *IEEE Trans. Speech and Audio Processing*, 4(3), 1996.

[2] J.D.Ferguson. Variable duration models for speech. In J.D. Ferguson, editor, *Proc. Symp.Applic. Hidden Markov Models Text Speech*, pages 143–179, Princeton,NJ, 1980.

[3] M.J.Russell and R.K.Moore. Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition. In *Proceedings ICASSP'85*, pages 5–8, 1985.

[4] S.E.Levinson. Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, 1:29–45, 1986.

[5] N.B. Yoma, F. McInnes, and M. Jack. Weighted viterbi algorithm and state duration modelling for speech recogniton in noise. In *Proceedings ICASSP'98*, pages 709–712, 1998.

[6] A. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones. The Noisex-92 study on the effect of additive noise in automatic speech recognition. Technical report, DRA, UK, 1992.

[7] D. Van Compernolle. Noise adaptation in a hidden Markov model speech recognition system. *Computer Speech and Language*, 3:151–167, 1989.