ROBUST PARAMETRIC MODELING OF DURATIONS IN HIDDEN MARKOV MODELS

David Burshtein

Dept. of Electrical Engineering - Systems
Tel-Aviv University, Tel-Aviv 69978, Israel
and: DSP Telecommunications, Israel

ABSTRACT

We address the problem of explicit state and word duration modeling in hidden Markov models (HMMs). A major weakness of conventional HMMs is that they implicitly model state durations by a Geometric distribution, which is usually inappropriate. Using explicit modeling of state and word durations, it is possible to significantly enhance the performance of speech recognition systems. The main outcome of this work is a modified Viterbi algorithm that by incorporating both state and word duration modeling, reduces the string error rate of the conventional Viterbi algorithm by 29% and 43% for known and unknown string lengths respectively, for a speaker independent, connected digit string task. The uniqueness of the algorithm is that unlike alternative approaches, it adds the duration metric at each frame transition (and not at the end of a state, word or sentence), thus enhancing the performance.

1. INTRODUCTION

Conventional hidden Markov models (HMMs) implicitly model the duration probability distribution of each state $p_i(\tau)$ by a Geometric distribution, i.e. $p_i(\tau) = a_{ii}^{\tau-1}(1-a_{ii})$, where i is the state, τ is the duration ($\tau \geq 1$), and a_{ii} is the state self transition probability [8]. This exponential distribution is usually inappropriate. Instead, explicit modeling of the duration distribution was shown [7], [8] to improve the performance of speech recognition systems. In addition to that, conventional HMMs do not include modeling of word durations. Hence the decoding procedure might produce unrealistically short or unrealistically long word durations. Explicit modeling of word durations rules out such durations and thus results in performance improvements.

Several approaches to state duration modeling has been proposed. In the Ferguson model [2] each state, i, has an associated state duration probability $d_i(\tau), \tau = 1, 2, \ldots, \tau_{\max}^i$ where $\tau_{\max}^i$ is the largest duration allowed. Ferguson incorporated the estimation of $d_i(\tau)$ into the Baum Welch re-estimation algorithm. There are two disadvantages to the Ferguson algorithm. The first is the excessive computational requirements it poses. The second is the excessive amounts of training data that might be required to estimate all the duration parameters: each state i, has $\tau_{\max}^i$ duration parameters, and sufficient statistics on each duration τ needs to be collected at each state, i, so as to estimate $d_i(\tau)$ reliably. In order to accomodate the second problem, Russell and Moore [10] and Levinson [6] suggested using parametric state duration distributions. [10] applied the Poisson distribution while [6] applied the Gamma distribution. Although the second problem was eliminated, the first one was not. Rabiner et. al. [7],[8] suggested a postprocessor approach, in order to incorporate duration modeling in a computationally efficient way. Besides real time implementation difficulties, a major disadvantage of the backtracking approach is that the duration contribution to the standard Viterbi metric is only added after candidate paths have been collected. Hence the correct path might not be one of these candidates. A similar problem occurs in [4], since the duration metric is only added at the end of the state or word.

This work is focused on a practical approach to state and word duration modeling in HMMs, that avoids the above mentioned problems. In section **2** we investigate possible parametric descriptions for state and word durations. In section **3** we propose a modified Viterbi algorithm that incorporates both state and word duration modeling, and has essentially the same computational requirements of the conventional Viterbi algorithm. In section **4** we present recognition experiments that demonstrate a significant reduction in the string error rate, compared to the standard Viterbi algorithm, for a speaker independent, connected digit string task.

0-7803-2431-5/95 $4.00

2. PARAMETRIC MODELING OF THE DURATIONS

As we indicated in the previous section, parametric modeling of state and word duration distributions reduce the amount of training data that is required for proper training of the distributions. Hence parametric duration distributions possess improved robustness features. Since several parametric distributions have been proposed in the past for duration modeling [6], [10], [7], [8], we found it important to investigate which is the most appropriate. For that purpose we used an HMM speech recognition system, that models each word by an 8 state left to right HMM (the system will be described in section 4). A supervised Viterbi segmentation of the training set was carried out, and an histogram of the duration was collected for each state and word, from which we obtained the empirical state and word probability distributions. The Gamma distribution

$$p(x) = \frac{\alpha^p}{\Gamma(p)} exp\{-\alpha x\} x^{p-1} \qquad 0 < x < \infty$$

($\alpha > 1$ and $p > 0$) was found to produce a high quality fit to the empirical distributions, for describing state and word durations. The Gauss distribution

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{x^2}{2\sigma^2}\}$$

produces high quality approximations for word durations, but is however inferior in its ability to describe certain state durations. The Geometric distribution is inferior to the Gamma distribution in its ability to describe both state and word durations.

Figure 1 displays the empirical duration distribution, and its Gauss, Gamma and Geometric fit for the 7'th state of the word 'seven'.

As can be seen, here the Gamma and Gauss fit are both superior to the Geometric fit. Figure 2 displays the same data for the 3'rd state of the the word 'oh'.

Here both the Gamma and Geometric fit are superior to the Gauss fit. Figure 3 demonstrates that both the Gamma and Gauss fit are superior to the Geometric fit for describing the duration of the word 'seven',

Note that the Gamma distribution is capable of describing both the monotonic character of the Geometric distribution, and the unimodal character of the Gauss distribution. Hence, the Gamma distribution is capable of describing both monotonic and unimodal distributions. This capability is very desirable for proper modeling of state and word durations. In addition, the Gamma distribution assigns zero probability to negative x's, which is appropriate for duration distribution-

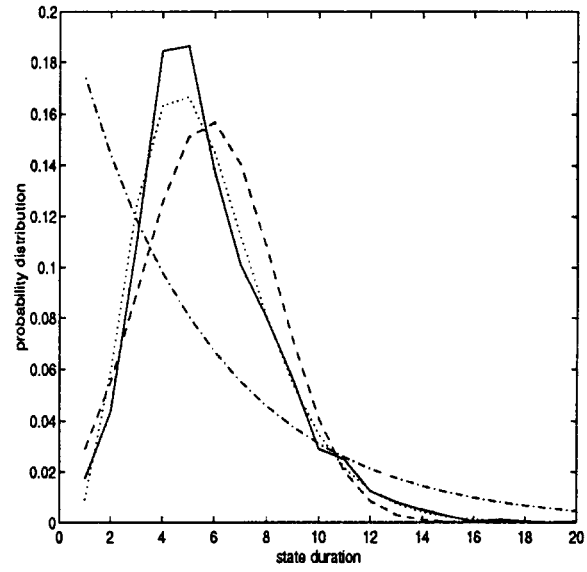

Figure 1: Duration distributions for the 7'th state of the word 'seven': a) empirical distribution (solid line) b) Gauss fit (dashed line) c) Gamma fit (dotted line) d) Geometrical fit (dash-dot line)

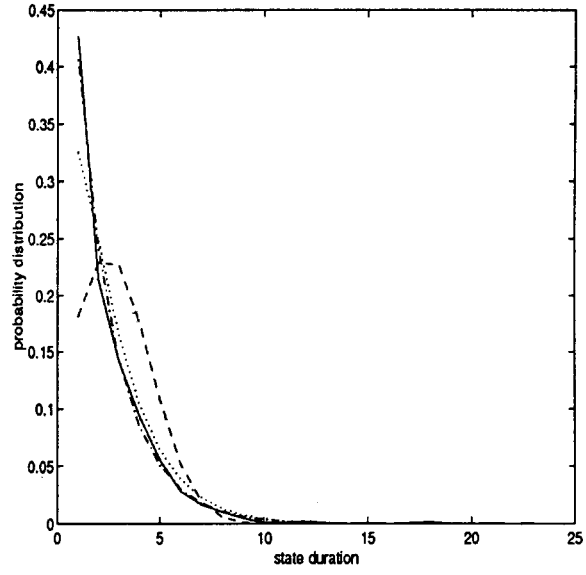

Figure 2: Same as Figure 1 for the 3'rd state of the word 'oh'

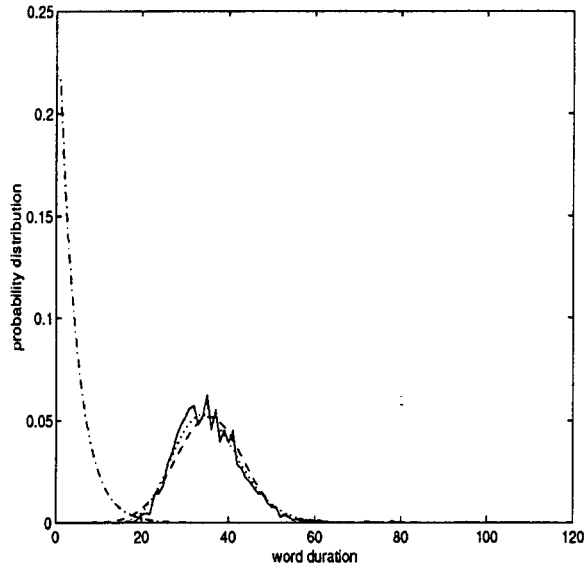

Figure 3: Same as Figure 1 for the word duration of the word 'seven'

s. This does not apply to the Gauss distribution. Finally, the Gamma distribution possesses slower decay rate, $exp\{-\alpha x\}$, which is more appropriate for duration modeling than the fast $exp\{-\frac{x^2}{2\sigma^2}\}$ decay of the Gauss distribution.

Careful examination of the Kullback Leibler (KL) distance measure defined for the densities $p(x)$ and $q(x)$ by,

$$\mathrm{KL}(p||q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx \ ,$$

showed that the Gamma fit is almost always closer to the empirical distribution than the other parametric approximations examined, although the difference from the Gauss distribution is small for word durations (but not for certain state durations).

To estimate the free parameters of the Gamma distribution α and p, note that the Gamma distribution has mean and variance values given by, (e.g., [9][p. 164]),

$$\mathrm{E}\{X\} = \frac{p}{\alpha} \qquad \mathrm{VAR}\{X\} = \frac{p}{\alpha^2}$$

Hence, α and p are estimated using the empirical expectation $\widehat{E}(x)$, and empirical variance $\widehat{\mathrm{VAR}}(x)$ of the duration, by

$$\hat{\alpha} = \frac{\widehat{E}\{X\}}{\widehat{\mathrm{VAR}}\{X\}} \qquad \hat{p} = \frac{\widehat{E}^2\{X\}}{\widehat{\mathrm{VAR}}\{X\}}$$

3. A SYNCHRONOUS FRAME BY FRAME IMPLEMENTATION

Our modified Viterbi algorithm keeps track of the duration $D_s(t)$ of each state s at time t. Let M_s denote the maximal value of the Gamma distribution, at state s, and let $l(u) = \log p(u)$, where $p()$ here is the Gamma density. The duration penalty, P, of making a transition from state s at time t to state $\hat{s}$ at time $t+1$ is given by,

$$P = \begin{cases} 0 & \text{if } D_s(t) < M_s \text{ and } \hat{s} = s \\ l(D_s(t+1)) - l(D_s(t)) & \text{if } D_s(t) \geq M_s \text{ and } \hat{s} = s \\ l(D_s(t)) & \text{if } D_s(t) < M_s \text{ and } \hat{s} \neq s \\ l(M_s) & \text{if } D_s(t) \geq M_s \text{ and } \hat{s} \neq s \end{cases}$$

The motivation is the following: before the duration is M_s, we do not penalize for remaining in the same state (unlike the conventional Viterbi algorithm). After the duration is M_s, we penalize gradually, unlike the backtracking approach, thus avoiding the problems of the backtracking approach, mentioned above. Note also that at the end of the state, at t_0, the overall duration metric penalty is $l(D_s(t_0))$, as it should be. Word durations are updated in essentially the same manner.

4. RECOGNITION EXPERIMENTS

For Recognition Experiments we used our continuous word HMM based speech recognition system. The acoustic front end comprises an FFT based filterbank, that calculates the energy within each of 18 overlapping Mel scale filters spanning the frequency range from 200Hz to 3200Hz. These energies are then converted to the first 9 cepstral values $c_0, c_1, \ldots, c_8$ and their time derivatives $\Delta c_0 \Delta c_1, \ldots, \Delta c_8$, using a cosine transform applied to the logarithm of the energies. The resulting 18 dimensional feature vector is modeled by a tied mixture of diagonal covariance Gaussians [1], [3], using a codebook of 200 Gaussians. The mixture size is 8.

Long term adaptive spectral equalization of the filter energies was shown to improve performance over different speakers and acoustic conditions, and was hence incorporated to the system. Each word in the vocabulary was modeled by one, 8 state, left to right HMM.

The training was performed using a variant of the segmental K means algorithm [8], whose computational complexity is significantly reduced compared to the Baum Welch algorithm, and the performance is essentially the same.

The decoding algorithm was a conventional Viterbi algorithm for the baseline experiments and a modified Viterbi algorithm incorporating state and/or word duration modeling in the other experiments.

Experiment	*Training Set*		*Testing Set*	
	UL	KL	UL	KL
No duration modeling	3.94	1.84	4.77	2.20
State Duration modeling (in test)	2.32	1.24	2.86	1.60
State and word duration modeling (in test)	2.23	1.15	2.78	1.59
State and word duration modeling (train and test)	2.09	1.16	2.73	1.56

Table 1: String error rate results for the Training and Testing sets, both for Known Length (KL) Strings and for Unknown Length (UL) Strings

The speech database was the speaker independent, high quality connected digits recorded at TI [5]. The database is divided into training and testing digit strings uttered by 225 adult talkers (we did not use the sentences uttered by children talkers). The first experiment was a baseline experiment carried out using the conventional Viterbi algorithm with no duration modeling (i.e. implicitly using a Geometric distribution for the duration). In the second experiment state duration modeling was added, after estimating the free parameters of the Gamma distribution from the mean and variance of the duration at each state, from a supervised Viterbi segmentation of the training set. All other parameters were taken from the estimation performed at the baseline experiment. In the third experiment word duration modeling was added to the second experiment. In the last experiment, both training and testing were performed with duration modeling. Table 1 summarizes the string error rate results obtained. As can be seen, from the results of the testing set, our modified Viterbi algorithm reduces the string error rate by 43% for the unknown string length case, and by 29% for the known string length case. State duration is the dominant source of improvement. Further addition of word duration modeling or further training, incorporating our enhanced Viterbi algorithm into the segmental K means algorithm, does not yield any significant additional improvements.

5. REFERENCES

[1] J. R. Bellegarda and D. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 2033-2045, 1990.

[2] J. D. Ferguson, "Variable duration models for speech", *Proc. Symp. on the Application of Hidden Markov Models to Text and Speech*, J. D. Ferguson ed., pp. 143-179, Princeton, New-Jersey, 1980.

[3] X. D. Huang, "Phoneme classification using semicontinuous hidden Markov models", *IEEE Trans. Signal Processing*, vol. 40, pp. 1062-1067, 1992.

[4] C. H. Lee and L. R. Rabiner, "A frame synchronous network search algorithm for connected word recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 1649-1658, 1989.

[5] R. G. Leonard, "A database for speaker independent digit recognition", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 42.11.1-4, Mar. 1984.

[6] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition", *Computer Speech and Language*, vol. 1, pp. 29-45, 1986.

[7] L. R. Rabiner, J. G. Wilpon and F. K. Soong, "High performance connected digit recognition using hidden markov models", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 1214-1225, Aug. 1989.

[8] L. R. Rabiner, "A Tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, vol. 77, pp. 257-286, Feb. 1989.

[9] C. R. Rao, *Linear Statistical Inference and Its Applications*, Wiley, 1973.

[10] M. J. Russell and R. K. Moore, "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 5-8, 1985.