# DURATIONAL MODELLING FOR IMPROVED CONNECTED DIGIT RECOGNITION

*Kevin Power*

Speech Technology Unit,
BT Laboratories,
Martlesham Heath,
Ipswich, IP5 7RE, UK

## ABSTRACT

A durational modelling technique is proposed for CDHMM-based connected digit recognition. This reduces the insertion error rate, which is typically the most frequent recognition error observed when no grammar constraint is applied. Insertion errors can be attributed in part to the acknowledged weakness of the acoustic models for accurate temporal modeling of speech signals. Two forms of durational model are investigated: an expanded-state model and an explicit model. Both forms of model significantly reduce the number of insertion errors and hence the digit string error rate. A modification to the explicit model which also accounts for speaking rate is described.

## 1. INTRODUCTION

HMMs provide a powerful framework for modeling time-varying signals such as speech. Both the acoustic and durational properties of speech sounds are modelled as separate stochastic processes thus ensuring effective modelling of both sources of variability.

A well known weakness of conventional HMMs is that the inherent modelling of the durations of different acoustic regions is unrealistic. Within the conventional HMM the probability $P_i(d)$ of staying in state *i* for *d* frames is given by the geometrically decreasing distribution $P_i(d) = a_{ii}^{d-1}(1-a_{ii})$ where the free parameter $a_{ii}$ is the self-loop transition probability for state *i*. This has disastrous consequences when HMMs are applied to connected word tasks: during the matching process, word strings where the associated models have short state durations, tend to be favoured over competing strings with fewer words but longer state durations. This effect can be observed in a connected digit recognition task with no grammar constraints where the number of insertion errors greatly exceeds that of deletions and substitutions. The disproportionately high insertion error rate can be partly attributed to inaccurate durational modelling by the digit HMMs.

It has been shown that state durations are more accurately described by a gamma distribution peaking at a duration of several frames [1] and a well known technique to extend the HMM formalism to more accurately model such distributions is the hidden semi-Markov model [2][3]. This incorporates temporal properties into the HMM framework with parameters that can be trained using a modified Baum-Welch algorithm.

In this paper two alternative durational models are compared: an expanded-state HMM which models state durations implicitly and an explicit state and word duration model. Results are presented in both cases for isolated and connected digit tasks.

## 2. DURATIONAL MODELLING TECHNIQUES

### 2.1. Expanded-state duration model

In the expanded-state HMM each individual state is replaced by multiple states, each sharing the original state observation pdf [4][5]. This can implicitly model more complex state duration distributions than the geometrically decreasing distribution of standard HMMs. An example is shown in Figure 1 where a state is replaced by a 3-state ensemble. The transition probabilities $a_1$ to $a_6$ of the multi-state system effectively become parameters of a duration distribution and can be estimated using standard techniques. A probability distribution $P_i(d)$, of duration d for state i, is modelled by the system in Figure 1 in the following way:

$P_i(1) = 1-a_1-a_2-a_3$

$P_i(2) = a_1(1-a_1-a_2-a_3) + a_2(1-a_4-a_5) + a_3(1-a_6)$

$P_i(3) = a_1^2(1-a_1-a_2-a_3) + a_1a_2(1-a_4-a_5) + a_2a_6(1-a_6) + a_1a_3(1-a_6) + ...$

*etc.*

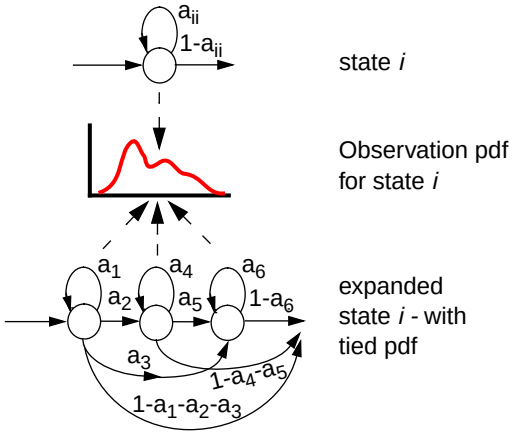


**Figure 1:** Expanded-state duration model.

## 2.2. Explicit duration model

In the explicit form of durational model state duration penalties can be directly applied during the acoustic matching process. In addition, word duration penalties can also be applied. Word duration partly accounts for state duration correlation, which might be expected to result from effects such as variation in speaking rate and stressed/unstressed syllable pronunciation.

In the explicit model, state duration penalties can be applied at each state transition [1] but informal experiments suggest that applying the state and word penalties at each model transition is just as effective. In the experiments described in Section 5 penalties are applied to the output probability $P(O|\lambda)$ of observing feature vector sequence O given model $\lambda$:

$$P(O|\lambda) \rightarrow P(O|\lambda) \cdot P_w(d_w) \cdot \prod_{i=1}^{N} P_i(d_i)$$

and $$d_w = \sum_{i=1}^{N} d_i$$

where $d_w$ is word duration, $d_i$ is duration within state i, N is the number of model states and $P_i(d_i)$, $P_w(d_w)$ are state and word durations penalties respectively.

Directly applying duration penalties during recognition is potentially more computationally efficient than for the expanded-state equivalent with the added advantage that extremely long or short durations can be more effectively penalised.

## 3. EXPERIMENTAL SETUP

Recognition experiments were performed for isolated and connected digit tasks. In both cases the recognition vocabulary consisted of words *one* to *nine*, *zero*, *nought*, *oh* and *double*.

For isolated digit experiments vocabulary examples were taken from a 1000 talker telephony database which consisted of several repetitions of each word spoken in isolation collected over the UK public telephone network. The data was partitioned into 6199 training and 1206 test utterances.

Connected digit experiments were performed using the serial numbers collected with the BT Subscriber database [6]. This was collected from different speakers across the UK over the telephone network — each reciting a unique serial number with an average length of seven digits. The data was partitioned into 455 training and 479 testing examples. An unconstrained recognition grammar is assumed throughout all connected digit experiments.

A simple baseline model topology of 6-state, 7-mode, left-to-right with no skips was chosen for the connected and isolated vocabulary models. The same single-state line noise model was used for both tasks as durational modelling was applied to vocabulary models only. The acoustic features were generated by a standard cepstral-based front end at a 16ms frame interval.

The baseline isolated recogniser has an accuracy of 96.02% at a 95% confidence interval of ±1.10%. Performance figures for the baseline connected recogniser are shown in Table 1 (abbreviations "sub", "del" and "ins" referring to substitution, deletion and insertion errors respectively).

| %word accuracy | 95% confidence interval | #sub | #del | #ins |
|---|---|---|---|---|
| 77.0% | ±1.42% | 332 | 28 | 411 |

**Table 1:** Baseline UK-English connected-digit recognition figures.

The relatively low word accuracy for the connected digit task is due both to the paucity of representative training examples combined with the wide accent variation for UK-English. This work, however, focuses on comparing different durational models rather than on optimising the acoustic models.

## 4. EXPANDED-STATE DURATION MODEL

Three expanded-state topologies, shown in Figure 2, are examined. In (a) each state is replaced by a 2-state system 3 durational parameters per state. The system in (b) is a second-order HMM [5] and the 3-state expansion (c) has 6 free parameters. The results are shown in Table 2.

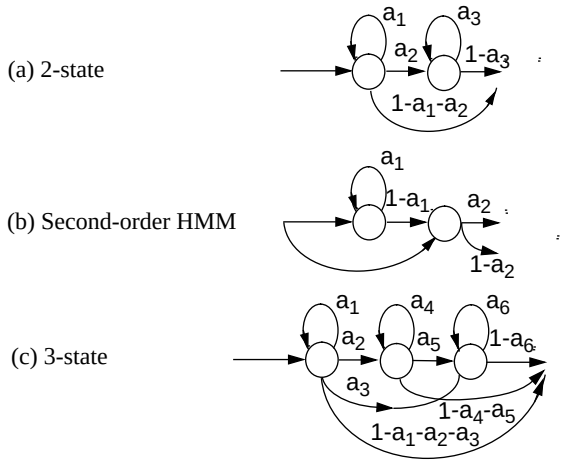


**Figure 2:** Expanded-state model topologies.

Each form of duration model greatly improves the unconstrained connected-digit word accuracy. The 2-state system gives best results - although it may be that the 3-state system is undertrained on this training set.

| Duration model | Isolated | Connected | | | |
|---|---|---|---|---|---|
| | %acc | %word acc | #sub | #del | #ins |
| 2-state | 96.10 | 80.58 | 312 | 30 | 309 |
| 2nd-order | 96.35 | 80.20 | 307 | 31 | 326 |
| 3-state | 96.27 | 80.14 | 306 | 27 | 333 |

**Table 2:** Results for expanded-state duration models.

In Figure 3 baseline and expanded-state duration distributions are plotted for a typical state. This illustrates how the improved durational model of the expanded-state systems peak at durations greater than one frame. It is interesting to compare these with the explicitly estimated duration probabilities shown in Figure 5.

## 4.1. De-emphasising acoustic likelihoods

In practice the temporal modelling effect within CDHMM-based recognition tends to be over-shadowed by the spectral model. A convenient mechanism to emphasise the temporal model is to apply an appropriate exponent weight, often referred to as the stream weight, to the state output likelihoods [7].

In Figure 4 the system stream weight is varied for connected recognition with the expanded-state models - where a stream weight of 1.0 corresponds to the figures in Table 2.

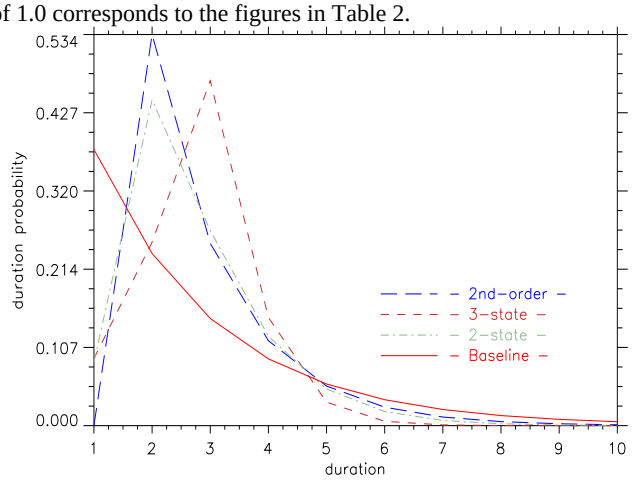


**Figure 3:** Baseline and expanded-state distributions of duration (in frames) for state 3 of connected model *double*.

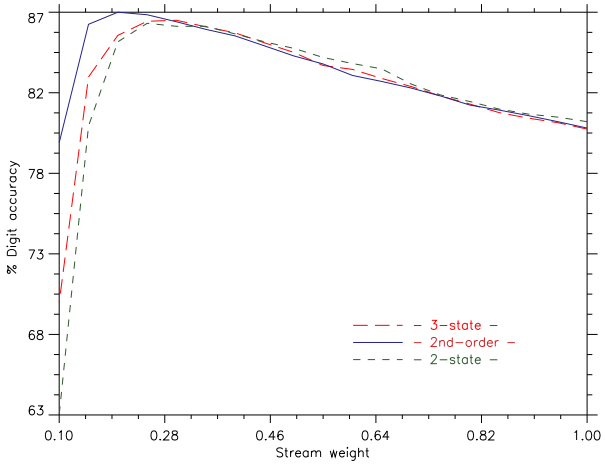


**Figure 4:** Varying the connected-digit system stream weight.

# 5. EXPLICIT DURATION MODEL

The simplest form of explicit model is the bounded duration model in which duration distributions are approximated by the penalty function

$$P(d) = \begin{cases} 1 & d^{min} \le d \le d^{max} \\ 0 & otherwise \end{cases}$$

where d is a state or word duration, $d^{min}$ and $d^{max}$ are the minimum and maximum corresponding durations observed on the training data. In a system, proposed by Gu *et al* [8], the state duration $d_i$ is replaced by state duration normalised by state sequence length. This form of model concentrates on detecting extremes of duration which might result when the wrong model matches to data.

In Table 3 results are shown for applying bounded duration penalties. The model topology in all explicit duration experiments is the same as for the baseline system. As can be seen this form of penalty degraded connected recognition performance - suggesting that durational extremes are not necessarily due to incorrect model matches.

| **penalty** | **Isolated** | **Connected** | | | |
|---|---|---|---|---|---|
| | %acc | %word acc | #sub | #del | #ins |
| bounded state | 94.86 | 74.32 | 334 | 24 | 503 |
| bounded word | 96.35 | 75.90 | 330 | 26 | 452 |
| bounded state & word | 94.86 | 75.54 | 328 | 22 | 470 |
| bounded normalised-state | 96.77 | 74.19 | 328 | 22 | 482 |

**Table 3:** Results for explicit bounded duration penalties.

A more accurate duration penalty, using the actual duration distributions directly is next applied. State and word distributions are estimated from the training data and implemented as smoothed histograms. Results for this form of penalty are given in Table 4 and duration penalties plotted for a typical state in Figure 5. With explicit modelling of durations the temporal model is conveniently emphasised by raising penalties to an exponent weight - the last row of Table 4 gives results for the optimal weight value.

| **duration penalty** | **Isolated** | **Connected** | | | |
|---|---|---|---|---|---|
| | %acc | %word acc | #sub | #del | #ins |
| state | 96.27 | 82.91 | 304 | 43 | 226 |
| word | 96.10 | 78.41 | 332 | 29 | 363 |
| state & word | 96.27 | 84.28 | 295 | 48 | 184 |
| state&word (weighted) | 96.52 | 86.91 | 277 | 89 | 73 |

**Table 4:** Results for smoothed histogram duration penalties.

The most significant (connected) performance improvement results from combining the state and word histograms. This is better than using state duration histograms alone which may indicate that inter-state correlations are important and are weakly accounted for by combining the word histogram into the duration penalty.

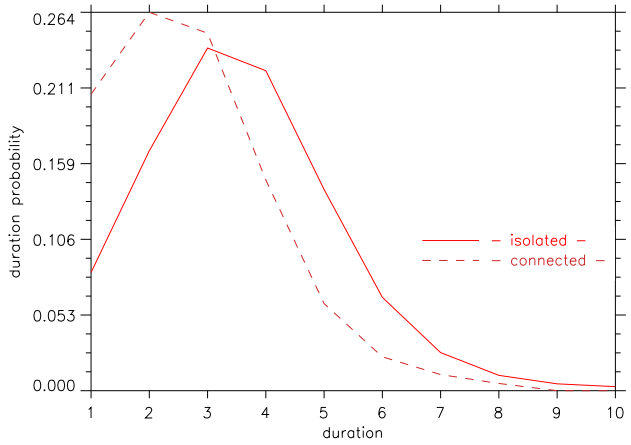


**Figure 5:** Explicit penalties for duration (in frames) for state 3 of model *double*.

A further refinement can be obtained by normalising the state durations by the word duration, thus allowing for variation in speaking rate in the durational model. This should model durations better because the normalised duration does not treat each acoustic duration as an independent event. In Table 5 results for normalised-state and combined normalised-state and word penalties are summarised.

.

| **Duration penalty** | **Isolated** | **Connected** | | | |
|---|---|---|---|---|---|
| | %acc | %word acc | #sub | #del | #ins |
| normalised-state | 96.35 | 85.51 | 294 | 68 | 124 |
| normalised-state (weighted) | 96.60 | 86.58 | 268 | 111 | 71 |
| normalised-state & word | 96.52 | 86.22 | 284 | 81 | 97 |
| normalised-state & word (weighted) | 96.77 | 86.40 | 262 | 133 | 61 |

**Table 5:** Results for explicit normalised distribution penalties.

Comparing results from Table 4 and Table 5 the normalised-state duration penalty significantly outperforms the corresponding absolute-duration system for both isolated and connected tasks when no weight is applied. However, applying a single weight did not increase performance as markedly in the normalised-state system. This may be due to the effects of combining pdfs with probability distributions for normalised-state and word penalties respectively.

## 6. CONCLUSION

Two duration models, the expanded-state and explicit duration model, were investigated. Both models greatly improve connected digit recognition performance.

The best expanded-state model (following optimisation) is the second-order HMM. This increased unconstrained digit accuracy from 77.0% to 80.6% with further improvement to 87.0% by adjusting the balance of temporal and acoustic models. The most effective explicit duration models used combined state and word duration penalties. The normalised-state and word form of explicit model increases connected digit accuracy from 77.0% to 86.2%. Applying a weight improves this to 86.4% - further gains may result from an alternative weighting strategy.

## 7. REFERENCES


1. D. Burshtein, “Robust Parametric Modelling of Durations in Hidden Markov Models”, proc. ICASSP-95.
2. M. Russell, R. Moore, “Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition”, Proc. ICASSP-85.
3. S. Levinson, “Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition”, Computer Speech and Language, vol. 1, no. 1, 1986.
4. T. H. Crystal, A.S.House, “Segmental durations in connected-speech signals: current results”, J. Acoust. Soc. Am., 83(4), April 1988.
5. J-F Mari, J-P Haton, “Automatic Word Recognition Based on Second Order Hidden Markov Models”, Proc. ICSLP-94.
6. A.D.Simons, K.Edwards, “Subscriber - A Phonetically Annotated Telephony Database”, Proc. Institute of Acoustics Speech and Hearing, 1992.
7. Y. Normandin, R. Cardin, R. De Mori, “High-Performance Connected Digit Recognition using Maximum Mutual Information Estimation”, IEEE Trans. on Speech and Audio Processing, vol. 2, no. 2, 1994.
8. H. Gu, C. Tseng, L. Lee, “Isolated-Utterance Speech Recognition using Hidden Markov Models with Bounded State Durations”, IEEE Trans. on signal processing, vol. 39, no. 8, August 1991.