# AUTOMATIC SPEECH RECOGNITION EXPERIMENTS USING SEGMENTAL HIDDEN MARKOV MODELS

Wendy J. Holmes and Martin J. Russell

Speech Research Unit, DRA Malvern,
St Andrews Road, Malvern, Worcs WR14 3PS, UK

## ABSTRACT

This paper describes the use of a segmental hidden Markov model (HMM), to overcome some important speech-modelling limitations of the conventional HMM approach. The aim is to develop a dynamic segmental HMM to model the changing pattern of speech over the duration of some phoneme-type unit, thereby avoiding the usual assumptions that observations are independent and that speech patterns are piece-wise stationary. As a first step towards this goal, a static segmental HMM [3] has been implemented and tested. This model reduces the influence of the independence assumption by using two processes to model variability due to long-term factors (such as speaker identity) separately from local variability that occurs within a segment. A set of experiments has been carried out which demonstrates that the performance of segmental models relative to conventional HMMs is dependent on the "quality" of the system in which they are embedded. Performance of a conventional HMM with variable frame-rate (VFR) analysis has been compared with that of a segmental HMM used to simulate a VFR effect. On a speaker-independent task with a medium-sized vocabulary, both systems failed to improve the performance of a monophone-based system but did provide an advantage for a triphone system. In further experiments on a small vocabulary (connected digit) recognition system, the full segmental HMM was shown to be capable of outperforming the conventional HMM. It is concluded that there are no inherent problems with segmental models, and that static segmental HMMs provide performance improvements over conventional HMMs, as long as the system is such that the independence assumption is a major limiting factor.

## 1. INTRODUCTION

HMMs have proved to be a very successful approach to automatic speech recognition. In addition to providing a tractable mathematical framework with straightforward algorithms for training and recognition, HMMs have a general structure which is broadly appropriate for speech: the time-varying nature of spoken utterances is accommodated through an underlying Markov process, while statistical processes associated with the model states encompass short-term spectral variability. The approach does however make assumptions which are clearly inappropriate for modelling speech patterns. The **independence assumption** states that the probability of a given acoustic vector corresponding to a given state depends only on the vector and the state, and is independent of the sequence of acoustic vectors preceding and following the current vector and state. It is also assumed that a speech pattern is produced by a **piece-wise stationary** process with instantaneous transitions between stationary states. The model thus ignores the fact that a speech signal is produced by a continuously moving physical system (the vocal tract). These erroneous assumptions can be overcome by using a segment-based model, characterising dynamic behaviour over several consecutive frames. Such models include the dynamical system model of Digalakis, Rohlicek and Ostendorf [1], and the continuous-time formulation of HMMs proposed by Saerens [2].

At the Speech Research Unit we are extending the basic HMM formalism, together with its associated mathematical theory, to derive a dynamic segmental HMM which overcomes both of the limitations mentioned above while retaining the advantages of the general HMM approach. A segmental HMM framework has been developed to allow comparison between alternative models of speech dynamics. As the first stage towards this goal, a static segmental HMM [3] has been implemented to reduce the impact of the independence assumption, which should provide some modelling advantages over conventional HMMs. However, substantial improvements would not be expected until a model of the dynamics is incorporated. A similar static segmental model has been studied by Gales and Young [4], who actually reported slightly worse performance than that obtained with conventional HMMs. In the current study, experiments have therefore been carried out with the aim of thoroughly understanding the behaviour of static segmental HMMs in different situations, prior to incorporating a dynamic model.

## 2. A STATIC SEGMENTAL HMM

In a conventional HMM, the statistical process associated with a state is defined by a single probability density function (pdf), which typically has to accommodate two quite distinct types of variability:

- **Extra-segmental variability:** long-term variations such as speaker identity and chosen pronunciation of a speech sound, which are essentially fixed throughout the duration of a segment.
- **Intra-segmental variability:** short-term variations which occur within a segment as a result of the continuous articulation process and other random fluctuations.

When combined with the independence assumption, the result of using a single pdf is that the model allows extra-segmental factors such as speaker identity to change in synchrony with the frame rate of the acoustic patterns. The problem can be considerably reduced by using a segmental HMM which has an underlying semi-Markov process [5] to model speech at the segmental level and, at the state level, uses separate models for extra-segmental and intra-segmental sources of variability. This allows extra-segmental factors to be fixed throughout a state occupancy. The Gaussian segmental HMM (GSHMM) is summarised below.

Extra-segment variability associated with a state $\sigma_i$ is characterised by a Gaussian pdf $N_{(\mu_i,\gamma_i)}$ which will be termed the **state target pdf**. On arrival at state $\sigma_i$, a target $c$ is chosen randomly according to this pdf. Any one target is described by a Gaussian pdf with fixed variance $\tau_i$. A state duration $D_i$ is chosen randomly according to the pdf $d_i$ and a sequence of vectors is then generated randomly and independently according to the target pdf $N_{(c,\tau_i)}$. Given a sequence of observation vectors $y = y_1,...,y_T$, the probability of a particular subsequence $y_{t_{i-1}+1}^{t_i} = y_{t_{i-1}+1},...,y_{t_i}$ with length $D_i$ can be defined as

$$\hat{P}_{\sigma_i}(y_{t_{i-1}+1}^{t_i}) = d_i(D_i) \cdot N_{(\mu_i,\gamma_i)}(\hat{c}) \cdot \prod_{t=t_{i-1}+1}^{t_i} N_{(\hat{c},\tau_i)}(y_t)$$

where $\hat{c}$ denotes the optimal target, which is the value of $c$ that maximises the probability of the observations. It can be shown that the value of $\hat{c}$ is given by

$$\hat{c} = \frac{\mu_i\tau_i + \sum_{t=t_{i-1}+1}^{t_i} y_t\gamma_i}{\tau_i + D_i\gamma_i}.$$

Thus, computing the probability that a particular segment of speech was produced by a certain model involves first determining the optimal target. The segment probability then comprises three components:

- the probability of the hypothesised duration
- the probability of the optimal target given the state target pdf
- the probability of each observation in the segment given the optimal target.

It has been shown by Russell [3] that a Baum-Welch-type re-estimation process can be derived for the parameters of GSHMMs. In addition, the standard dynamic programming approach to recognition can easily be extended to segmental models. In both cases, the need to explicitly consider times $t - \delta$ for all possible segment durations ($\delta = 1,2,\ldots, d_{max}$), combined with the additional computation required to determine the optimal target, leads to an increase in computational load relative to conventional HMMs.

## 3. INITIAL RECOGNITION EXPERIMENTS

### 3.1. Speech Data

The first experiments were performed on speaker-independent recognition of airborne reconnaissance mission (ARM) reports, using a 497-word vocabulary. This application has been used extensively at the Speech Research Unit for experiments with conventional sub-word HMMs [6]. Three reports from each of 61 male speakers were used for training, and three reports from a different 10 male speakers for testing. The speech was analysed using a critical-band filterbank at 100 frames/s, with output channel amplitudes in units of 0.5 dB, converted to an eight-parameter Mel cepstrum and an average amplitude parameter. Time derivatives were *not* used, as the aim was to investigate basic segmental modelling without any dynamics.

### 3.2. Model structure

Three-state context-independent monophone models and four single-state non-speech models were used (with single-Gaussian pdfs), as a baseline for comparisons between segmental and conventional HMMs. A simple left-to-right model structure was used, including self-loop transitions. The GSHMMs for these experiments were minimally different from standard HMMs: self-loop transitions were retained to allow the models freedom to represent each phone by as many 'segments' as required for the best match. In addition, all segment durations were assigned equal probability and duration distributions were not re-estimated.

### 3.3. Training procedure

The parameters of the conventional HMMs were initialised based on a uniform segmentation of each training utterance. The means and inter-variances of the GSHMMs were initialised in the same way, with all intra-variances being set to 0.5 (in dB-related units as defined by the transformed filterbank amplitudes). Figure 1 shows that the segmental training algorithm appears to operate correctly: probability increases with number of iterations in all cases, and the optimised probability of the training set is greater for segmental than for conventional HMMs. The optimised probability of the segmental models increases only slightly as maximum segment duration is increased from 5 to 15 frames. A duration of 5 frames is sufficient to provide a considerable difference from conventional HMMs, and all recognition experiments reported in this paper were therefore based on a maximum segment duration of five, in order to minimise the additional computation required over that for the standard HMM approach.

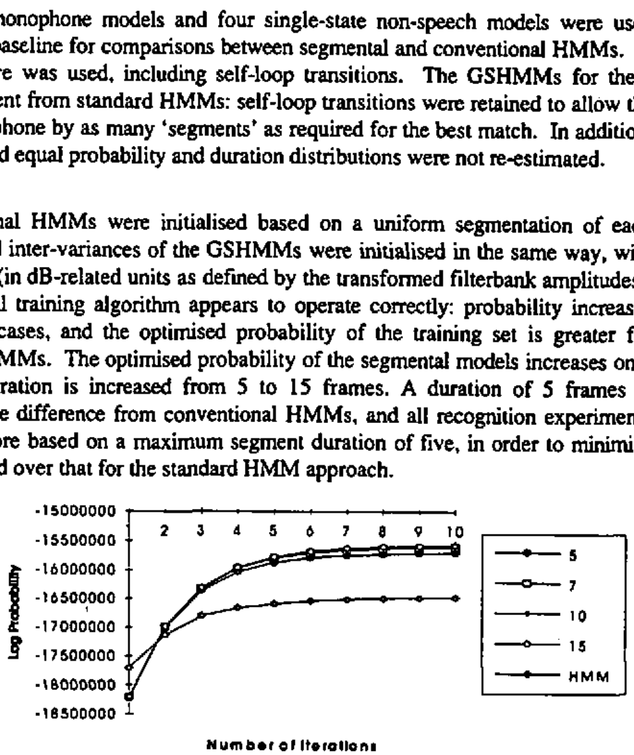


*Figure 1: Log Probability of the ARM training set as a function of iteration number for conventional HMMs and for segmental HMMs with maximum segment durations of 5, 7, 10 and 15.*

### 3.4. Recognition results

An initial evaluation was conducted on a single spoken ARM report, with the aim of verifying that the segmental HMM recognition algorithm was operating correctly. For connected word recognition with no explicit syntax and a word transition penalty of 30 (found in [6] to be appropriate for this task), conventional HMMs gave a word accuracy of 40.4% (which is consistent with the results presented in [6]), whereas segmental HMMs with a maximum state duration of five frames gave only 17.5%. In view of the potential importance of model initialisation strategy, a second experiment was tried in which the means and inter-variances of the segmental HMMs were initialised from the means and variances of trained conventional HMMs. This set of segmental models gave an improved GSHMM word accuracy of 31.6%, which is still much worse than the conventional HMM result. These very poor results were unexpected, and are much worse than the results reported by Gales and Young [4] with a similar model. Further experiments were therefore carried out to investigate the cause. Although various aspects of the detailed approach to initialisation, training and recognition with GSHMMs may not have been optimal, the difference in performance from that of standard HMMs was so large that it was decided to begin by investigating performance with a basic segmental framework.

## 4. THE RELATIONSHIP BETWEEN SEGMENTAL MODELS AND VFR ANALYSIS

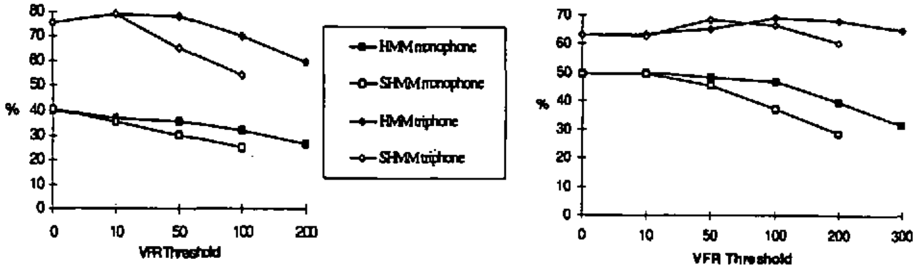
The GSHMM can be interpreted as an extension and integration of variable frame-rate (VFR) analysis and HMMs, as discussed in [3]. In its simplest form, the VFR algorithm removes vectors from an observation sequence, based on computing a distance between the current observation vector and the most recently retained vector. Observations are discarded if the distance is below a specified threshold, so compressing quasi-constant regions into one observation. It has been demonstrated that this form of VFR analysis can lead to improved recognition performance [7]. Experiments were therefore carried out to assess the effect of VFR analysis for the task and model set described in Section 3, comparing performance with that of a type of segmental HMM which effectively performs VFR analysis.

### 4.1. Performing VFR analysis with a segmental HMM

In segmental HMM terms, the single observation vector can be regarded as the target for the quasi-stationary segment which it replaces, while the threshold and the distance metric together play the role of the intra-segmental pdf. Thus an integrated form of VFR HMM recognition can be performed with segmental models, by modifying the definition of the optimal target to be the segment mean, and replacing the Gaussian intra-segmental pdf by a uniform pdf with radius specified by a threshold parameter. This segmental VFR scheme differs from conventional VFR approaches only in that the retained information is the mean rather than the first observation, and in that the segmentation is integrated into the dynamic programming process, rather than being performed as a pre-processing stage. Both these aspects should be improvements over the conventional approach.

### 4.2. Recognition results

Based on the single ARM report used in the initial experiments, figure 2(a) illustrates performance with both conventional VFR and the segmental VFR HMM as a function of VFR threshold. For monophone models, both systems show a degradation in performance as threshold increases. The slightly faster degradation of the segmental system is to be expected, as this system measures distance from the segment mean whereas conventional VFR measures distance from the initial segment vector. Thus, for a given threshold, segmental VFR permits more compression than conventional VFR. The poor performance of both VFR systems on this report is an important result, as it suggests that any form of segment-based approach will perform poorly on this data and model set. When the experiment was repeated using triphone models, the results show the expected performance improvements at low VFR thresholds, followed by a fall in performance for larger values which permit too much compression. The pattern of results is the same, although less extreme, when taken over the complete ARM evaluation set (see figure 2(b)): there are some improvements with triphones but not monophones. The segmental VFR HMM shows similar improvements to the conventional VFR system, so demonstrating that there is no intrinsic problem with a segment-based model whereby the segmentation is integrated into the dynamic programming.



*(a) ARM report used for initial experiments* *(b) Complete ARM evaluation set*

*Figure 2: Word accuracy as a function of VFR threshold for conventional VFR analysis and VFR segmental HMMs.*

### 4.3. Discussion

The VFR results suggest that, for reasonably good models (i.e. task-dependent triphones), the temporal independence assumption does indeed limit the performance of conventional HMMs. For a simple monophone system however, the disadvantage of discarding data (which happens explicitly in VFR and implicitly in segmental HMMs) outweighs any modelling advantage. From studying the distance scores for alignments of models with data using correct and incorrect transcriptions, it became apparent that the distributions for the monophone models were so broad that there were only small differences between the scores for alternative recognition possibilities and so the discrimination ability of the models was very poor. All the data frames were therefore required to contribute individually to the distance calculation in order to obtain maximum cumulative discrimination. It seems probable that this is the reason for the poor performance of both conventional and segmental VFR schemes with monophone models. The possibility that a similar pattern might be seen for the full GSHMM was therefore investigated by performing comparative experiments between segmental and conventional HMMs for systems with varying degrees of modelling sophistication.

## 5. FURTHER SEGMENTAL HMM RECOGNITION EXPERIMENTS

These experiments with the full GSHMM used the simpler task of connected digit recognition, to allow faster experiment turn-around time and make analysis of recognition errors more straightforward. Experiments were carried out with vocabulary-dependent versus vocabulary-independent training and context-dependent versus context-independent models. Using single-Gaussian models, the performance of segmental HMMs was compared with that of conventional HMMs with and without VFR analysis. As the segmental models require an increased number of parameters over conventional HMMs, comparisons were also carried out with two-component-mixture HMMs, which also use more parameters than single-Gaussian models while retaining the conventional model format.

It should be noted that, for many conditions, the results were improved by the use of a word transition penalty. Although the precise value of penalty was not critical, it was found that the performance was noticeably worse if the penalty was a long way from the optimum value. The best value was dependent on the type of training data used and, to a lesser extent, on the type of models (single-Gaussian HMM, two-component-mixture HMM or GSHMM). For the purposes of performing experimental comparisons, results are quoted with the best word transition penalty for each condition.

### 5.1. Speech data

The test data were three lists of 50 digit triples spoken by each of 10 male speakers. Vocabulary-independent training was based on recordings of 225 different male speakers each reading 10 phonetically-rich sentences selected from a set of 460 such sentences [8]. The data for vocabulary-dependent training were taken from the same 225 speakers, each reading 19 four-digit strings.

### 5.2. Training procedure

The parameters of the single-Gaussian standard HMM monophones were initialised based on a uniform segmentation of the training data. They were then trained with five iterations of Baum-Welch re-estimation. The resulting models were used to initialise both the two-component-mixture HMMs and the GSHMMs. For the mixture models, the initialisation was achieved using the conventional approach of splitting the single component into two and perturbing the means slightly. In the case of the GSHMMs, the initial values for the means and inter-variances were taken from the HMM means and variances respectively. All intra-variances were initialised to 0.5 (in the appropriate dB-related units).

For all types of HMM, the relevant monophone models were used to initialise triphone models which were then trained with three iterations. When performing recognition with the vocabulary-independent triphones, any triphones which had not occurred in the training data were replaced by the relevant monophone.

### 5.3. Recognition results

Table 1 shows percentage word accuracy with a range of training conditions, for segmental models compared with conventional HMMs and with VFR analysis using the optimum value of VFR threshold. In all cases, using VFR analysis improved the performance of the conventional HMM. For all conditions except the vocabulary-independent monophone training, the segmental HMMs perform better than the conventional HMMs even with optimum VFR analysis.

| training | monophone | | | triphone | | |
|---|---|---|---|---|---|---|
| data | std | std vfr | seg | std | std vfr | seg |
| sentences | 82.4 | 84.6 | 77.7 | 83.1 | 85.0 | 88.5 |
| digits | 82.3 | 84.3 | 87.3 | 86.6 | 88.2 | 89.3 |

*Table 1: Percentage word accuracy on a connected-digit recognition task, for segmental HMMs compared with standard HMMs with and without VFR analysis.*

Table 2 compares the performance of the segmental HMMs with that of the two-component-mixture conventional HMMs. For all conditions except the vocabulary-independent monophones, the word accuracy is similar for both types of model: the mixture models perform slightly better for digit-trained triphones, but the segmental models are better for the digit-trained monophones and the sentence-trained triphones.

*Table 2: Percentage word accuracy on a connected-digit recognition task, for segmental HMMs compared with two-component-mixture standard HMMs.*

| training | monophone | | triphone | |
|---|---|---|---|---|
| data | 2-mix std | seg | 2-mix std | seg |
| sentences | 85.6 | 77.7 | 87.6 | 88.5 |
| digits | 86.8 | 87.3 | 89.9 | 89.3 |

## 5.3. Discussion

### Effect of modelling sophistication on GSHMM recognition performance

The GSHMMs have performed better than the single-Gaussian conventional HMMs for both sets of triphones and for the digit-trained monophones. With digit training, even the "monophone" models will have been trained in only the appropriate contexts and the "triphone" models will in fact be word-dependent. It therefore appears that, as postulated in Section 4.3, the full static segmental HMM offers advantages when the acoustic representations in the models are reasonably accurate and so the independence assumption is a major limiting factor. The likely explanation is related to the balance between the inter-segmental probabilities and the intra-segmental probabilities. In conventional HMM-based classification, the probability of any model having produced a particular utterance of length $T$ is obtained as the product of exactly $T$ frame-state probabilities. In a segmental model however, any one segment probability consists of the product of two different types of probabilites and different explanations of the data may use different numbers of the two types (depending on the preferred number of segments). Recognition performance is therefore dependent on the correct balance between the two types of probability contribution. In the case of both the segmental ARM monophone models and the segmental vocabulary-independent monophone models, this correct balance had apparently not been achieved: there was a strong tendency to favour long segment durations over the sometimes short durations which were required for correct recognition, due to the penalty of an additional inter-segmental probability outweighing any benefit from higher intra-segmental probabilities. It is hypothesised that this imbalance in the segmental models arose due to differences in the extent to which the two types of distributions fitted the modelling assumptions: with speaker-independent, context-independent models, the inter-segmental distributions will not be well-modelled by a single Gaussian, whereas the intra-segmental distributions should fit quite well to the Gaussian assumption. When context-dependent models are used, a single Gaussian is not so inappropriate for modelling the inter-segmental distribution, and the trained segmental models show a better balance between the two types of probabilities.

### Comparing GSHMMs with conventional VFR

For instances where the segmental HMMs offer better performance than the conventional HMMs, this advantage is greater than that obtained from applying VFR analysis to the HMMs. This finding implies that, provided a useful model can be obtained, it is better to actually model the relationship between observations within a segment than to simply condense them into one observation. It is interesting that the vocabulary-independent monophones showed some performance improvements with VFR analysis but not with GSHMMs. It therefore appears that the simple VFR approach of discarding observations can be beneficial with a lower-quality system than is required for the segmental modelling to be successful.

### Comparing GSHMMs with two-component-mixture HMMs

It is not surprising that the use of an additional mixture component has improved the performance of the conventional HMMs, as this provides more parameters to describe the extensive variability which will not be well-modelled by single-Gaussian distributions. The second mixture component provides a different type of modelling improvement to that offered by segmental models: additional parameters are used to improve the approximation of each state distribution rather than to constrain the underlying model for the nature of speech variability. In some respects the mixture HMM therefore allows better modelling of inter-speaker variability than is possible with a single-Gaussian inter-segmental distribution. It should also be noted that the two-component-mixture models use more parameters per state (two sets of means and two sets of variances) than the GSHMMs (one set of means and two sets of variances). In view of these aspects of the mixture approach, it is encouraging that the GSHMMs provide a similar level of performance (except in the case of vocabulary-independent monophones). Interestingly, the GSHMM system actually performs better than the mixture system in the case of the digit-trained monophones, where there should be no danger of insufficient examples to train the required numbers of parameters. The improvements from using GSHMMs rather than conventional HMMs are therefore not simply due to increasing the number of model parameters, but result from the more appropriate nature of the underlying model.

## 6. CONCLUSIONS

A static segmental HMM has been shown to improve recognition performance over that obtained with conventional HMMs, provided that modelling is sufficiently accurate for the independence assumption to be a major limitation on performance: if there are other fundamental restrictions on modelling capabilities, these have an overriding influence and it is not possible to derive a useful static segmental model. This is probably the cause of the poor results reported by Gales and Young [4], who used segmental monophones to model TIMIT data. Having gained an understanding of the modelling tasks for which segmental models are able to operate correctly, possible refinements are being investigated: in particular, model initialisation strategy and the effect of training duration distributions. The next stage is to incorporate a model of speech dynamics, which should enable the full advantages of the segmental framework to be achieved.

## 7. REFERENCES

[1] Digalakis, V., Rohlicek, J.R. and Ostendorf, M. "A dynamical system approach to continuous speech recognition", *Proc. IEEE ICASSP*, Toronto, pp. 289-292, 1992

[2] Saerens, M. "A continuous-time dynamic formulation of Viterbi algorithm for one-Gaussian-per-state hidden Markov models", *Speech Communication*, 12, pp. 321-333, 1993.

[3] Russell, M.J. "A segmental HMM for speech pattern modelling", *Proc. IEEE ICASSP*, Minneapolis, pp. 499-502, 1993.

[4] Gales, M.J.F. and Young, S.J. "Segmental hidden Markov models", *Proc. Eurospeech-93*, Berlin, pp. 1579-1582, 1993.

[5] Russell, M.J. and Moore, R.K. "Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition", *Proc. IEEE ICASSP*, Tampa, pp. 5-8, 1985.

[6] Russell, M.J. "The development of the speaker independent ARM speech recognition system", *Proc. IOA*, Vol. 14: part 6, pp. 17-24, 1992.

[7] Peeling, S.M. and Ponting, K.M. "Variable frame rate analysis in the ARM continuous speech recognition system", *Speech Communication*, 10, pp. 155-162, 1991.

[8] Browning, S.R., McQuillan, J., Russell, M.J. and Tomlinson, M.J. "Texts of material recorded in the SI89 speech corpus", SP4 Research Note No. 142, RSRE, 1991.