EXPERIMENTAL EVALUATION OF SEGMENTAL HMMS

Wendy J. Holmes and Martin J. Russell

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

ABSTRACT

The aim of the research described in this paper is to overcome important speech-modeling limitations of conventional hidden Markov models (HMMs), by developing a dynamic segmental HMM which models the changing pattern of speech over the duration of some phoneme-type unit. 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 separately from local variability that occurs within a segment. Experiments have demonstrated that the performance of segmental HMMs relative to conventional HMMs is dependent on the "quality" of the system in which they are embedded. On a connected-digit recognition task for example, static segmental HMMs outperformed conventional HMMs for triphone systems but not for a vocabulary-independent monophone system. It is concluded that static segmental HMMs improve performance, as long as the system is such that the independence assumption is a major limiting factor.

1. INTRODUCTION

In the HMM approach to speech recognition, assumptions are made which are clearly inappropriate for modeling speech patterns. The **independence assumption** states that the probability of a given acoustic vector corresponding to a given state 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, characterizing dynamic behavior 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. 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 modeling 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 behavior 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: long-term variations such as speaker identity and chosen pronunciation of a speech sound **(extra-segmental variability)**, and short-term variations which occur within a segment as a result of the continuous articulation process and other random fluctuations **(intra-segmental variability)**. 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 summarized below.

Extra-segment variation associated with state σ_i is characterized by a Gaussian pdf $N_{(\mu_i,\gamma_i)}$, termed the **state target pdf.** On arrival at state σ_i, a target c is chosen randomly according to this pdf. Any one target is described by a Gaussian pdf with fixed intra-segment variance τ_i. A duration D_i is chosen randomly according to the pdf d_i and a sequence of vectors is generated randomly and independently according to the target pdf $N_{(c,\tau_i)}$. Given a sequence of observation vectors $y = y_1,\ldots,y_T$, the probability of a particular subsequence $y_{t_{i-1}+1}^{t_i} = y_{t_{i-1}+1},\ldots,y_{t_i}$ 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 maximizes the probability of the observations. It can be shown that the value of $\hat{c}$ is given by

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

0-7803-2431-5/95 $4.00

The standard dynamic programming approach to recognition can easily be extended to segmental models, and it has been shown by Russell [3] that a Baum-Welch-type re-estimation process can be derived for the GSHMM parameters.

3. INITIAL SEGMENTAL HMM 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. 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 analyzed 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 at this stage the aim was to investigate basic segmental modeling 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 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 initialized based on a uniform segmentation of each training utterance. The means and extra-segment variances of the GSHMMs were initialized in the same way, with all intra-segment 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, and the optimized probability of the training set is greater for segmental than for conventional HMMs. A segment duration of five frames is sufficient to provide a considerable difference, and therefore all segmental recognition experiments reported in this paper used a maximum segment duration of five.

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 (previously found to be appropriate for this task), conventional HMMs gave a word accuracy of 40.4%, whereas the segmental HMMs gave only 17.5%. In view of the potential importance of model initialization strategy, a second experiment was tried in which the means and extra-segment variances of the segmental HMMs were initialized 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 by studying a very simple segmental framework.

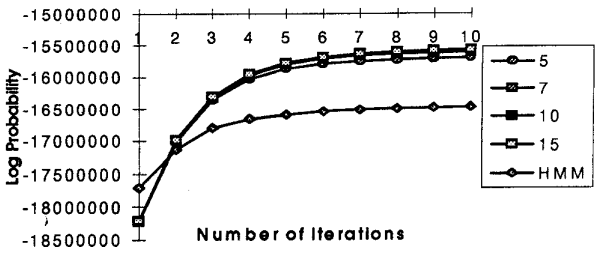


Figure 1: Log Probability of the ARM training set as a function of iteration number for conventional HMMs and for segmental HMMs with maximum segment durations of 5, 7, 10 and 15.

4. RELATIONSHIP WITH VFR ANALYSIS

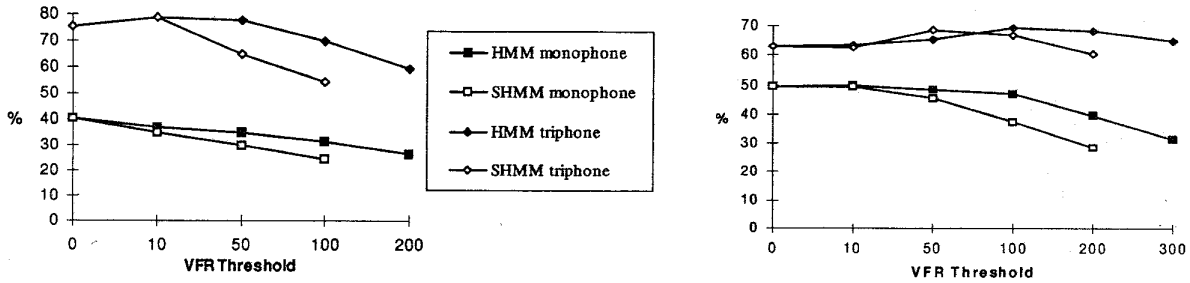
The GSHMM can be interpreted as an extension and integration of variable frame-rate (VFR) analysis and HMMs [3]. In its simplest form, the VFR algorithm removes vectors from an observation sequence, based on a distance between the current vector and the most recently retained vector. Observations are discarded if the distance is below a 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 [6]. 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.

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, performance of both systems degrades as threshold increases. The slightly faster degradation of the segmental system is to be expected, as this system measures distance from the segment mean instead of 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 important, 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



(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.

figure 2(b)): there are some improvements with triphones but not monophones. The segmental VFR HMM shows similar benefits 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.

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 modeling advantage. From studying distance scores, it became apparent that 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 modeling sophistication.

5. FURTHER SEGMENTAL HMM EXPERIMENTS

These experiments used the simpler task of connected digit recognition, to allow faster experiment turn-around time and easier analysis of recognition errors. 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 use more parameters than single-Gaussian models while retaining the conventional model format.

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). 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. The data for vocabulary-dependent training were taken from the same 225 speakers, each reading 19 four-digit strings.

5.2. Training procedure

The single-Gaussian HMM monophones were initialized from a uniform segmentation of the training data, and trained with five iterations of Baum-Welch re-estimation. The resulting models were used to initialize both two-component-mixture HMMs and GSHMMs. For the mixture models, the initialization used 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 extra-segment variances were taken from the HMM means and variances. All intra-segment variances were initialized to 0.5 (in appropriate dB-related units).

For all types of HMM, the relevant monophone models were used to initialize 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 using a range of training conditions, for segmental models compared with the different sets of standard HMMs. 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. The segmental-HMM word accuracy is similar to that obtained with two-component-mixture conventional HMMs: the mixture models perform slightly better for digit-trained triphones, but the segmental models are better for the digit-trained monophones and sentence-trained triphones.

training data	monophone				triphone			
	std	std vfr	seg	2 mix	std	std vfr	seg	2 mix
sentences	82.4	84.6	77.7	85.6	83.1	85.0	88.5	87.6
digits	82.3	84.3	87.3	86.8	86.6	88.2	89.3	89.9

Table 1: Percent word accuracy on connected-digit recognition, for segmental compared with standard HMMs with and without VFR analysis, and two-component-mixture standard HMMs.

5.4. Discussion

Effect of modeling sophistication on GSHMM 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 extra-segmental probabilities and the intra-segmental probabilities, as discussed in the following paragraph.

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 probability 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 favor long segment durations over the sometimes short durations which were required for correct recognition, due to the penalty of an additional extra-segmental probability outweighing any benefit from higher intra-segmental probabilities. It is hypothesized that this imbalance in the segmental models arose due to differences in the extent to which the two types of distributions fitted the modeling assumptions: with speaker-independent, context-independent models, the extra-segmental distributions will not be well-modeled 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 modeling the extra-segmental distribution, and the trained segmental models show a better balance between the two types of probability.

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 modeling 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-modeled by single-Gaussian distributions. The second mixture component provides a different type of modeling 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 modeling of inter-speaker variability than is possible with a single-Gaussian extra-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 modeling is sufficiently accurate for the independence assumption to be a major limitation on performance: if there are other fundamental restrictions on modeling 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 modeling tasks for which segmental models are able to operate correctly, possible refinements are being investigated: in particular, model initialization 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] V. Digalakis, J.R. Rohlicek and M. Ostendorf, "A dynamical system approach to continuous speech recognition", *Proc. IEEE ICASSP*, Toronto, pp. 289-292, 1992

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

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

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

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

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