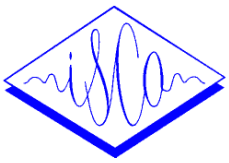

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

SPEECH RECOGNITION USING A LINEAR DYNAMIC SEGMENTAL HMM

Wendy J. Holmes and Martin J. Russell

e-mail: holmes@signal.dra.hmg.gb or russell@signal.dra.hmg.gb

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

ABSTRACT

This paper describes research into linear-trajectory dynamic segmental hidden Markov models (HMMs). The main advantage of these models over conventional HMMs is that they allow explicit modelling of speech segment dynamics. In general terms, a trajectory-based segmental HMM provides a parametric representation of the range of possible underlying trajectories for a speech sound. Acoustic feature vectors are regarded as noisy observations of a particular trajectory. This model represents an extension and generalization of the previously-developed static segmental HMM [1], which can be viewed as a constant-trajectory model. In the present paper, a linear-trajectory segmental HMM is described and Baum-Welch-type re-estimation formulae are presented. Preliminary recognition experiments on a connected-digit recognition task have demonstrated performance improvements over the previous results with static segmental HMMs, which in turn outperformed conventional HMMs.

1. INTRODUCTION

Segmental HMMs [1] are extended versions of conventional HMMs, in which states are associated with sequences of feature vectors rather than with individual vectors. The sequences of vectors are referred to as segments. The main advantage of segment-based models is that they provide a framework in which the relationship between frames comprising a segment can be modelled explicitly. It is thus possible to overcome the standard-HMM assumption that successive observations are independent, which is clearly inappropriate from a speech-modelling perspective.

The overall aim of our work is to extend the basic HMM formalism, together with its associated mathematical theory, to derive a segmental HMM which appropriately characterizes the *dynamic* behaviour of speech parameters. The first stage towards this goal was to develop a general segmental HMM framework, and to test this by implementing a simple static segmental HMM [1,2]. The model uses two processes to model variability due to long-term factors separately from local variability that occurs within a segment. The impact of the independence assumption is thus reduced by fixing the long-term factors for the duration of any one segment. Russell [1] has described the mathematical theory of static segmental HMMs and presented Baum-Welch-type re-estimation formulae for the model parameters. Holmes and Russell [2] demonstrated that these static segmental HMMs can improve speech recognition performance over that obtained with conventional HMMs. The importance of the research described in [1] and [2] is that it has demonstrated the viability of this segmental-HMM approach, so providing a foundation for the development of more sophisticated models which explicitly address speech pattern dynamics.

The present paper is concerned with the introduction of a linear-trajectory segmental model. First we give a general formalization of our segmental approach in terms of modelling variable-duration trajectories describing acoustic feature-vectors, followed by a specific formulation of a linear-trajectory model. Parameter re-estimation formulae are presented for this linear model, and some preliminary experiments are described.

2. TRAJECTORY-BASED SEGMENTAL HMMS

2.1. Modelling trajectories of speech parameters

The relationship between successive acoustic feature-vectors representing sub-phonemic speech segments can be approximated by some form of trajectory through the feature space. This idea has formed the basis for a number of segmental models, such as those proposed by Ghitza and Sondhi [3], by Goldenthal and Glass [4], by Digalakis, Rohlicek and Ostendorf [5], and by Deng, Asmanovic, Sun and Wu [6]. All these models incorporate the concept that acoustic representations of speech segments follow some underlying trajectory, combined with the notion of statistical variation in the realization of the trajectory. The main differences between alternative approaches are in the nature of the model for statistical variation and in the chosen trajectory representation itself. Ghitza and Sondhi and Goldenthal and Glass used non-parametric models, while Digalakis et al. adopted a linear model and Deng et al. have suggested including higher-order polynomials. At the Speech Research Unit, we are concentrating on developing a flexible segmental approach within an HMM framework, including the derivation of tractable training and recognition algorithms. The aim is to evaluate different parametric models for feature-vector trajectories, within a modelling approach which encompasses the total variability of natural utterances in an appropriate way.

2.2. The segmental-HMM approach

A segmental HMM for a speech sound is seen as providing a parametric representation of the range of possible underlying trajectories for that sound, where these are noisy trajectories of variable duration. Any one trajectory is considered to be a Gaussian stochastic process, with constant variance τ, whose mean changes as a function of time according to the parameters of the trajectory. The static segmental HMM described in [1] represents the simplest case, in which the underlying trajectory is assumed to be constant over time.

10.21437/Eurospeech.1995-390

For any segment of observations $y = y_0,\ldots,y_T$ and a given state of a model, the joint probability of y and a trajectory f_a (ignoring any duration-probability contribution) is specified by the equation

$$P(y, f_a) = P(f_a)\ .\ P(y \mid f_a) = P(f_a)\ .\ \prod_{t=0}^{T} P(y_t \mid f)\ .$$

$P(f_a)$ is the probability that f_a is a valid trajectory for the segment corresponding to the given state, while $P(y|f_a)$ is the probability that the time series y is a valid instantiation of that trajectory. Intuitively, $P(f_a)$ accounts for the **extra-segmental** variations (such as differences between speakers) which would lead to different trajectories for the same sub-phonemic unit, while $P(y|f_a)$ accounts for the detailed **intra-segmental** variations in the realization of a particular trajectory. The extra-segmental variance will generally be much larger than the intra-segmental variance. Thus, although the probability calculation treats the individual elements of y independently, the dependence on the trajectory f_a provides a fairly tight constraint due to its relatively small variance. The separation of the two types of variability is an important characteristic of the model: it reduces the impact of the independence assumption to a much greater extent than is possible if all the variance is represented (time-independently) around a single trajectory (as in the model described by Deng et al. [6]). At the same time, the constraints provided by the segmental HMM avoid the need for modelling error-correlations between successive observations (as used by Goldenthal and Glass [4] for example).

The probability of the segment y given the segmental HMM state can be obtained by integrating (or summing) over the set of all trajectory parameters A, thus:

$$P(y) = \int_{a \in A} P(y, f_a)\ .$$

An alternative is to only consider the *optimal trajectory*, whose parameters are those which maximize the joint probability of the observations and trajectory, given the model parameters, thus:

$$\hat{P}(y) = \max_{a \in A} P(y, f_a)\ .$$

The form of the optimal trajectory depends on the parameterization which is adopted. The constant-trajectory case has been described by Russell in [1], and the linear model is discussed in the following sections.

3. A LINEAR DYNAMIC SEGMENTAL HMM

In the linear segmental HMM, a segment is modelled as a noisy, linear trajectory of variable duration. This model is closely related to the dynamical systems model proposed by Digalakis et al. [5]. A trajectory $f_{(m,c)}$ is defined by its slope m and the segment mid-point value c, such that $f_{(m,c)}(t) = c + m(t - \frac{T}{2})$. In the current formulation, m and c are vectors from the acoustic feature space. It is well known that the slope $m'(y)$ and mid-point value $c'(y)$ of the linear trajectory which provides the best fit to the data y (in a least-squared error sense) are given by:

$$m'(y) = \frac{\sum_{t=0}^{T} (t - \frac{T}{2}) y_t}{\sum_{t=0}^{T} (t - \frac{T}{2})^2} \text{ and } c'(y) = \bar{y} = \frac{\sum_{t=0}^{T} y_t}{T+1}\ .$$

Now suppose that the distribution of trajectory parameters for a given state is defined by Gaussian distributions $N_{(\mu,\gamma)}$ and $N_{(\nu,\eta)}$ (with diagonal covariance matrices) for the slope and mid-point respectively. The intra-segmental distributions are assumed to be Gaussian with diagonal covariance τ. In addition to any duration-probability component, the probability of the sequence y given a particular Gaussian segmental HMM (GSHMM) can be defined as

$$\hat{P}(y) = N_{(\mu,\gamma)}(\hat{m})\ .\ N_{(\nu,\eta)}(\hat{c})\ .\ \prod_{t=0}^{T} N_{(f_{(\hat{m},\hat{c})},\tau)}(y_t)$$

where $\hat{m}$ and $\hat{c}$ are the values of the slope and mid-point which together maximise the joint probability of the observations and the trajectory. As in the case of the constant-trajectory model [1], it can be shown that these quantities are a weighted sum of the parameters $m'(y)$ and $c'(y)$ (which are optimal with respect to the data) and their expected values as defined by the model, thus:

$$\hat{m} = \frac{\left(\sum_{t=0}^{T} (t - \frac{T}{2}) y_t\right)\gamma + \mu\tau}{\left(\sum_{t=0}^{T} (t - \frac{T}{2})^2\right)\gamma + \tau} \text{ and } \hat{c}(y) = \frac{\left(\sum_{t=0}^{T} y_t\right)\eta + \nu\tau}{(T+1)\eta + \tau}\ .$$

4. PARAMETER ESTIMATION FOR LINEAR SEGMENTAL HMMS

It has been shown [1] that Baum-Welch parameter re-estimation can be extended from conventional to constant-trajectory GSHMMs. A similar extension applies to linear-trajectory GSHMMs. Given a sequence of observation vectors $y = y_1,\ldots,y_T$ and a linear-trajectory GSHMM M, a new linear-trajectory GSHMM $\overline{M}$ can be derived such that $P(y|\overline{M}) \geq P(y|M)$. The parameters of the GSHMM $\overline{M}$ are defined as follows:

$$\bar{\mu}_i = \frac{\sum_{x \in S_i} P(y, x \mid M)\ \kappa_{x,i} \sum_{t=t_i}^{t_{i+1}-1} (t - t_i^{\text{mid}}) y_t}{\sum_{x \in S_i} P(y, x \mid M)\ \kappa_{x,i} \sum_{t=t_i}^{t_{i+1}-1} (t - t_i^{\text{mid}}) y_t}$$

$$\bar{\gamma}_i = \frac{\sum_{x \in S_i} P(y, x \mid M)(\bar{\mu}_i - \hat{m}_{x,i}(y))^2}{\sum_{x \in S_i} P(y, x \mid M)}$$

$$\bar{\nu}_i = \frac{\sum_{x \in S_i} P(y,x|M)\lambda_{x,i} \sum_{t=t_i}^{t_{i+1}-1} y_t}{\sum_{x \in S_i} P(y,x|M)\lambda_{x,i} d_{x,i}}$$

$$\bar{\eta}_i = \frac{\sum_{x \in S_i} P(y,x|M)(\bar{\nu}_i - \hat{c}_{x,i}(y))^2}{\sum_{x \in S_i} P(y,x|M)}$$

$$\bar{\tau}_i = \frac{\sum_{x \in S_i} P(y,x|M) \sum_{t=t_i}^{t_{i+1}-1} (f_{(\hat{m}_{x,i}(y),\hat{c}_{x,i}(y))}(t) - y_t)^2}{\sum_{x \in S_i} P(y,x|M) d_{x,i}}$$

where S_i is the set of state sequences of length T which include state i. For any one state sequence x, the state duration $d_{x,i}$ is the time spent in state i, t_i defines the time of entry to state i and $t_i^{\text{mid}} = (t_i + t_{i+1})/2$ is the mid-point of the time interval over which state i is occupied. The factors $\kappa_{x,i}$ and $\lambda_{x,i}$ are defined by

$$\kappa_{x,i} = \frac{1}{\sum_{t=t_i}^{t_{i+1}-1} (t - t_i^{\text{mid}})^2 \bar{\gamma}_i + \bar{\tau}_i}$$

$$\lambda_{x,i} = \frac{1}{d_{x,i}\bar{\eta}_i + \bar{\tau}_i} .$$

The procedure for deriving the re-estimation formulae is basically the same as the one presented in [1] for constant-trajectory GSHMMs, which in turn was based on the conventional-HMM method described by Liporace [7]. In the case of linear-trajectory GSHMMs, as with the constant-trajectory models, the right-hand sides of the resulting re-estimation formulae include the re-estimated parameter values. In practice, workable re-estimation formulae are obtained by replacing $\bar{\mu}_i, \bar{\gamma}_i, \bar{\nu}_i, \bar{\eta}_i$ and $\bar{\tau}_i$ with $\mu_i, \gamma_i, \nu_i, \eta_i$ and τ_i on the right-hand sides of the equations. Both the experiments described in [2] and those discussed in the following section suggest that the iterative application of these formulae leads to a monotonic increase in $P(y|M)$.

5. PRELIMINARY EXPERIMENTS

The aims of these initial experiments with linear-trajectory GSHMMs were to establish that the training algorithm behaves in an appropriate way and to demonstrate the potential recognition-performance advantages over the more-limited constant-trajectory model. The experiments were performed on digit recognition using vocabulary-dependent monophone models, which was one of the experimental conditions used in the previous experiments with static segmental models [2]. This condition was chosen in order to minimize the computational requirements for segmental training while providing a task for which constant-trajectory GSHMMs are known to outperform standard HMMs [2]. The experimental conditions used for the linear-trajectory case were basically identical to those established in the previous experiments, with the only difference being in the addition of the slope parameters.

5.1. Speech data

The test data were three lists of 50 digit triples spoken by each of 10 male speakers. The training data were taken from 225 different male speakers, each reading 19 four-digit strings. 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 the aim was to investigate the effect of incorporating the dynamic representation into the model itself.

5.2. Model structure

Three-state context-independent monophone models and four single-state non-speech models were used, all with single-Gaussian pdfs and diagonal covariance matrices. A simple left-to-right model structure was used, including self-loop transitions. The maximum segment duration was set to five frames, all segment durations were assigned equal probability and duration distributions were not re-estimated. This is the model structure which was used in the baseline static-GSHMM experiments and may well not be optimal, particularly for dynamic models.

5.3. Training procedure

As before, the static GSHMMs were initialized from trained conventional HMMs: 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). A corresponding set of linear-trajectory GSHMMs was initialized by using the HMM means and variances as initial estimates for the trajectory mid-point parameters, and setting the slope means to zero and the slope variances to a small constant value (0.05). Hence the starting points for the constant-trajectory and linear-trajectory GSHMMs can be considered to be the same, although the linear model has the option of diverging from the zero-slope condition within the first iteration. Ten training iterations were applied for both types of segmental model.

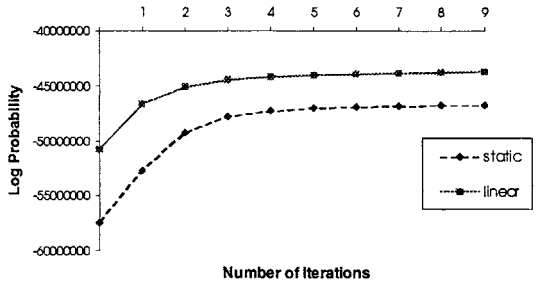

Figure 1: Log Probability of the digit training set as a function of iteration number for constant-trajectory and linear-trajectory GSHMMs with equivalent initial conditions.

It can be seen from Figure 1 that both training algorithms show the expected pattern over time: the probability of the training set increases monotonically as a function of iteration

number and converges after between five and ten iterations. This suggests that replacing the re-estimated parameters with the corresponding initial parameters on the right-hand side of the re-estimation formulae does not cause any problems. It is encouraging that the probability of the training set is consistently higher for the linear model than for the static model.

5.3. Recognition results and discussion

A recognition experiment was performed to compare the performance of the two sets of trained segmental HMMs on the connected-digit test set. Table 1 shows the percentage word accuracy after five and after ten training iterations. It can be seen that the linear models performed better than the static models, which in turn outperformed standard HMMs. It is interesting that increasing the number of training iterations from five to ten was beneficial for the linear models but did not affect performance for the static models. This would suggest that, at least with the initial estimates used in these experiments, more training may be required to optimise the parameters of the more flexible model.

A further experiment was performed in which a set of linear GSHMMs was initialized from the five-iteration trained static GSHMMs. The initialization strategy was the same as that described in Section 5.3, except that in this case the *trained* GSHMM parameters provided the initial estimates for the trajectory mid-point parameters and for the intra-segment variances. The recognition performance for these models is also included in Table 1. The performance of the models has improved by allowing the trajectory slope to vary. However, for the same total number of segmental-training iterations (i.e. ten), better performance is achieved when this flexibility is incorporated at the same time as introducing the segmental structure.

It was also considered important to compare the performance obtained by modelling the dynamics with that achieved if some form of dynamics is incorporated in the front-end representation. When the acoustic representation for the standard HMMs was augmented with time derivatives (computed for each frame as the difference between the parameters for the preceding and following frames), their recognition performance improved from 82.3% to 88.6% word accuracy. This is however not as good as the performance of the linear GSHMMs, for which the best figure obtained so far is 90.8% word accuracy.

Standard HMM	82.3
Static GSHMM (5 training iterations)	87.3
Static GSHMM (10 training iterations)	87.3
Linear GSHMM (5 training iterations)	88.9
Linear GSHMM (10 training iterations)	90.8
Linear GSHMM (5 training iterations, initialized from trained static GSHMMs)	88.8

Table 1: Percent word accuracy for recognition of the connected-digit test set, for standard HMMs and different types of segmental HMM.

6. CONCLUSIONS

A segmental HMM based on a linear-trajectory segment model has been successfully introduced. This model is an extension of the static model described in [1] and [2]. Baum-Welch-type parameter re-estimation formulae have been implemented and found to demonstrate the expected training pattern over time.

The incorporation of linear dynamics into a segmental HMM has been shown to improve recognition performance over that obtained with a static version of the model. Furthermore, performance is better than that obtained when simple time differences are incorporated into the front-end representation used with a conventional HMM. The potential of the linear-GSHMM approach has thus been demonstrated within a modelling framework based on the one used for the static models. The next stage is to determine the most appropriate model structure and training strategy for these dynamic models. It is also intended to investigate the use of alternative front-end representations such as those based on formants, which capture the dynamic nature of speech more explicitly than a cepstral representation.

7. REFERENCES

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

[2] W.J. Holmes and M.J. Russell, "Experimental evaluation of segmental HMMs, *Proc. IEEE ICASSP*, Detroit, pp. 536-539, 1995.

[3] O. Ghitza and M.M. Sondhi "Hidden Markov models with templates as non-stationary states: an application to speech recognition", *Computer Speech and Language*, 2, pp. 101-119, 1993.

[4] W.D. Goldenthal and J.R. Glass, "Statistical trajectory models for phonetic recognition", *Proc. ICSLP*, Yokohama, pp. 1871-1873, 1994.

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

[6] L. Deng, M. Asmanovic, D. Sun and J. Wu, "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states", *IEEE Trans. SAP*, 2, no. 4, pp. 507-520, 1994.

[7] L.A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources", *IEEE Trans. Information Theory*, IT-28, pp. 729-734, 1982.