

A multiple-level linear/linear segmental HMM with a formant-based intermediate layer*

Martin J Russell¹ and Philip J B Jackson²

¹Electronic, Electrical and Computer Engineering,
University of Birmingham, Birmingham B15 2TT, UK

²Centre for Vision, Speech and Signal Processing,
University of Surrey, Guildford, Surrey GU2 7XH, UK

Correspondence: m.j.russell@bham.ac.uk

April 2003

*This research is part of the “Balthasar” project, which was funded by EPSRC grant GR/M87146 “An integrated multiple-level statistical model for speech pattern processing” (see <http://www.eee.bham.ac.uk/Balthasar>). The authors would like to thank the reviewers for their helpful comments, and in particular the suggestion to include the material in section 4.4.4

Abstract

A novel multi-level segmental HMM (MSHMM) is presented in which the relationship between symbolic (phonetic) and surface (acoustic) representations of speech is regulated by an intermediate ‘articulatory’ representation. Speech dynamics are characterised as linear trajectories in the articulatory space, which are transformed into the acoustic space using an articulatory-to-acoustic mapping. Recognition is then performed. The results of phonetic classification experiments are presented for monophone and triphone MSHMMs using three formant-based ‘articulatory’ parameterisations and sets of between 1 and 49 linear articulatory-to-acoustic mappings. The NIST Matched Pair Sentence Segment (Word Error) test shows that, for a sufficiently rich combination of articulatory parameterisation and mappings, differences between these results and those obtained with an optimal classifier are not statistically significant. It is also shown that, compared with a conventional HMM, superior performance can be achieved using a MSHMM with 25% fewer parameters.

1 Introduction

This paper presents a novel multi-level segmental hidden Markov model (MSHMM) in which the relationship between symbolic and acoustic representations of a speech signal is regulated by an intermediate ‘articulatory’ layer. Such an approach has many potential advantages for speech pattern processing. For example, by modelling speech dynamics directly in an articulatory domain, it may be possible to characterise the production strategies that give rise to variability in fluent, conversational speech. Furthermore, provided that the articulatory representation is sufficiently compact, there should be significant advantages for speaker adaptation, and inter-speaker differences which result from physiological factors, such as the differences between an adult’s vocal tract and that of a child, should be represented more explicitly. Thus it is hoped that such a model will improve speech recognizer performance by modelling the underlying mechanisms that cause variability, rather than relying solely on generic statistical modelling techniques.

From a mathematical perspective, our starting point is a trajectory-based segmental HMM (Digalakis 1992; Russell 1993; Gales and Young 1993; Wiewiorka and Brookes 1996). Many different types of segmental HMM have been proposed, and an overview is presented in Ostendorf et al. (1996). Unlike a conventional HMM, where states are associated with individual acoustic feature vectors, in a segmental HMM they are associated with sequences of acoustic vectors, or segments. In a trajectory-based segmental HMM, a state defines trajectories in the acoustic vector space, and a segment is treated as a ‘noisy’ realisation of such a trajectory. Several different types of trajectory have been studied, including constant (Russell 1993; Gales and Young 1993), linear (Russell 1993), linear dynamical systems (Digalakis 1992), exponential (Wiewiorka and Brookes 1996), ‘smoothed piecewise constant’ (Richards and Bridle 1999) and non-parametric (Ghitza and Sondhi 1993). In each case, the motivation for such a model is that it offers an improved model of speech dynamics, relative to a conventional HMM. However, in acoustic representations of speech (derived from short-term log-power spectra) articulator dynamics are manifested indirectly, often as movement between, rather than within, frequency bands. Intuitively, therefore, it would be much better to model dynamics directly, in an articulatory-based representation.

The particular approach used in the present paper is depicted in Figure 1a. States of the underlying Markov process are associated with trajectories in an articulatory-based feature space (the intermediate layer). These

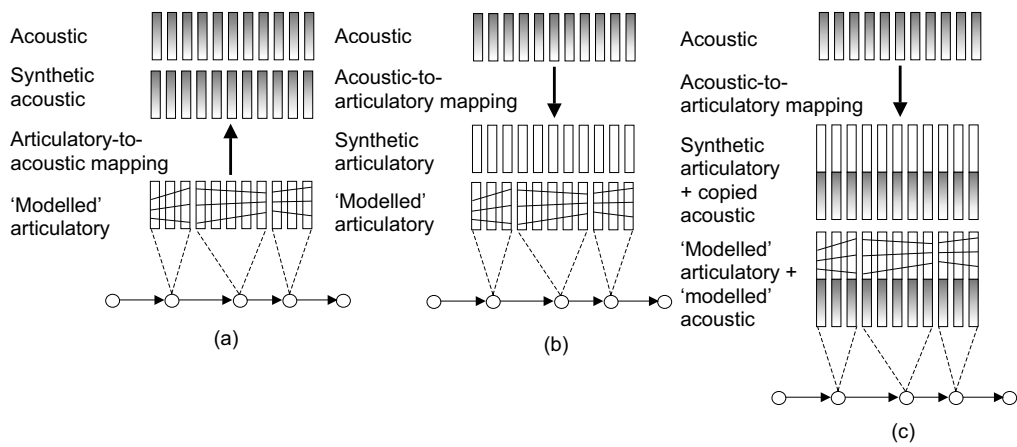


Figure 1: Illustrations of segmental models that use linear trajectories in an intermediate space and some mapping function W : (a) MSHMM, (b) classification in articulatory space, and (c) as (b) but supplemented with acoustic data.

trajectories are mapped into the acoustic feature space by an articulatory-to-acoustic mapping, where comparison is made with the acoustic feature vectors of the unknown utterance.¹ The idea of incorporating an intermediate layer into an acoustic model for speech recognition is certainly not original: see, for example (Deng and Braam 1994; Deng 1998; Richards and Bridle 1999; Gao et al. 2000; Deng and Ma 2000; Zhou et al. 2003). The novelty of the work presented here derives from the use of a particular type of segmental HMM framework, and from the validation of the model through the presentation of extensive experimental results.

An alternative approach is to use an acoustic-to-articulatory mapping to extract ‘articulatory’ features from the acoustic signal. These features are then decoded using conventional or segmental HMMs in the articulatory domain (Figure 1b). Examples include (Garner and Holmes 1998; Holmes et al. 1997; Holmes and Garner 2000) where formant data were used, and (Frankel et al. 2000) where articulatory data were extracted from the acoustic signal using an artificial neural network. However, extraction of articulatory features from an acoustic speech signal is a non-trivial pattern processing task, and irrecoverable errors are introduced into the resulting formant or articulatory representation. In addition, the chosen representation may not contain sufficient information on its own for accurate speech recognition. For these reasons the direct use of articulatory data alone as feature parameters for speech recognition typically leads to poor performance. This problem can be most simply alleviated by combining the articulatory data with the original acoustic data (Garner and Holmes 1998; Holmes et al. 1997; Holmes and Garner 2000; Frankel et al. 2000), as in Figure 1c, and by exploiting measures of confidence in the derived articulatory features (Garner and Holmes 1998; Holmes and Garner 2000; Wilkinson and Russell 2002), if such measures are available. Indeed, this use of a confidence measure can be seen as an approximation to an optimal classification strategy based on delayed decision making in which the final articulatory representation of a speech signal would be a consequence of, rather than a precursor to, the recognition process (Hunt 1987).

¹Strictly, the comparison is between the probability density function (pdf) generated by the model and the observations, but for pdfs with a single multivariate Gaussian component we can consider the comparison in terms of the Malhalanobis distance between the mean of the current state’s pdf and the observation vector, for the current frame.

The preceding paragraphs have argued that the introduction of an intermediate layer into a segmental HMM has several advantages, relative to a conventional HMM, from the perspective of speech pattern modelling. Therefore *in principle* one would expect such a model to give improved speech recognition performance. Unfortunately, the history of speech recognition contains many examples of schemes which are advantageous in principle, but fail to work in practice.

For example, a potential problem with the proposed model is that any advantage gained by the introduction of an intermediate layer may be compromised by inadequacies of the articulatory representation, limitations of the articulatory-to-acoustic mapping, or theoretical compromises made for mathematical or computational tractability. The purpose of this paper is to explore these issues. We consider a simple class of MSHMM whose trajectories in the articulatory-based representation are linear, and whose articulatory-to-acoustic mapping is realised as a set of one or more linear mappings. We refer to such a model as a *linear/linear* MSHMM. Since the resulting trajectories in the acoustic-feature space are also linear, the performance of an appropriate fixed linear-trajectory acoustic segmental HMM (FT-SHMM) (Holmes and Russell 1999) provides a theoretical upper bound on the performance of this type of MSHMM. In practice, this upper bound might only be achieved by a globally optimal acoustic FT-SHMM, and the results obtained with a particular acoustic FT-SHMM may be sub-optimal. However, this does not appear to be an issue in the experiments reported here.

At this point, it is worth noting that a linear/linear system is inadequate for speech pattern modelling (Richards and Bridle 1999). Consider a case where speech is represented in the acoustic domain as the output of a set of A uniformly-spaced band-pass filters spanning frequencies up to 4 kHz, and a single, hypothetical, ‘formant’ trajectory f , with unit amplitude, whose frequency increases continuously from 100 Hz to 4 kHz. The corresponding trajectory in acoustic space is a complex path over the surface of the A dimensional unit sphere, which passes through each of the axes in turn, and cannot be realised as the image of f under a linear mapping. At the opposite extreme, if the intermediate representation is equal to the acoustic representation and speech dynamics are represented directly and accurately (and therefore non-linearly) in this domain, then a linear mapping (the identity mapping) is sufficient. The linear trajectory model used in the current work is simpler and less realistic than the intermediate-layer models of dynamics described in (Deng and Braam 1994; Deng 1998; Richards and Bridle 1999;

Gao et al. 2000; Deng and Ma 2000; Zhou et al. 2003). However, a key motive in the current work is to introduce an articulatory-related intermediate space in which dynamics can be modelled simply. From this perspective it can be argued that the limitation exposed here should be seen as a consequence of the linearity of the mapping rather than the linearity of the trajectory.

All of the intermediate representations we consider are based on formant frequencies. The simplest intermediate representation consists of the first three formant frequencies, while the most complex comprises the twelve synthesizer control parameters from the Holmes-Mattingly-Shearman (HMS) formant synthesizer (Holmes et al. 1964). Although these representations are implicitly (rather than explicitly) articulatory, they will be referred to as ‘articulatory’ throughout this paper.

This paper demonstrates unambiguously that, in the case of TIMIT phone classification, speech recognition performance is not compromised by the introduction of an intermediate layer into a segmental HMM. Even with the simple linear/linear system considered here, the upper bound on performance can be achieved by appropriate choice of articulatory representation and articulatory-to-acoustic mappings. It has been shown elsewhere that a fixed linear-trajectory SHMM can outperform a conventional HMM (Holmes and Russell 1999). Hence the results presented in this paper give confidence that substantial improvements in performance relative to a conventional HMM can be achieved through the use of appropriate non-linear trajectories or non-linear articulatory-to-acoustic mappings. More immediately, the results also show that a conventional HMM system can be out-performed by a MSHMM system with 25% fewer parameters.

2 Theory

2.1 Linear-trajectory segment models

In the terminology of Holmes and Russell (1999), the model that concerns us is a fixed, linear trajectory segmental HMM (FT-SHMM). In such a model, a state treats an acoustic speech segment as a variable-duration, noisy function of a linear trajectory. Each state s_i is identified with a midpoint vector \mathbf{c}_i and slope vector \mathbf{m}_i , whose dimension N is that of the acoustic-feature space. A

segment trajectory of duration τ is defined by $\mathbf{f}_i(t) = (t - \bar{t}) \mathbf{m}_i + \mathbf{c}_i$, where $\bar{t} = (\tau + 1)/2$, and, given the state s_i , the probability of the sequence of acoustic vectors $Y_1^\tau = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_\tau]$ is

$$b_i(Y_1^\tau) = c_i(\tau) \prod_{t=1}^{\tau} \mathcal{N}(\mathbf{y}_t; \mathbf{f}_i(t), V_i), \quad (1)$$

where $\mathcal{N}(\mathbf{y}_t; \mathbf{f}_i(t), V_i)$ denotes the multivariate Gaussian probability density function (pdf), with mean $\mathbf{f}_i(t)$ and diagonal $N \times N$ covariance V_i , evaluated at \mathbf{y}_t , and c_i is the duration pdf. The case $\mathbf{m}_i = \mathbf{0}$ corresponds to a constant trajectory SHMM. If, in addition, c_i is a geometric pdf then this is functionally identical to a conventional HMM except for an upper bound τ_{\max} on state duration.

Now consider the case where a trajectory \mathbf{f}_i is realised in the M -dimensional intermediate (articulatory) space, and projected onto the acoustic layer by an ‘articulatory’-to-acoustic mapping function \mathcal{W} . In general, it could be arbitrary, but in this paper \mathcal{W} will be assumed to be piecewise linear. The probability of the acoustic segment Y_1^τ becomes

$$b_i(Y_1^\tau) = c_i(\tau) \prod_{t=1}^{\tau} \mathcal{N}(\mathbf{y}_t; \mathcal{W}(\mathbf{f}_i(t)), V_i). \quad (2)$$

Let \mathcal{M} be an S -state MSHMM (for simplicity, it is assumed that the probability of a transition from s_i to state s_j is zero unless $j \geq i$). Suppose that the acoustic sequence $Y = [\mathbf{y}_1, \dots, \mathbf{y}_T]$ comprises several segments. Then \mathcal{M} can only explain Y via a state sequence $\mathbf{x} = [x_1, \dots, x_T]$, which can be written in the form $\mathbf{x} = [d_1 \otimes z(1), \dots, d_L \otimes z(L)]$, For each $l \in \{1, \dots, L\}$, $z(l) = s_i$ for some $i \in \{1, \dots, S\}$, and $d_l \otimes z(l)$ represents a duration d_l spent in state $z(l)$. Thus, the joint density has the form,

$$p(Y, \mathbf{x} | \mathcal{M}) = \pi_{z(1)} b_{z(1)} \left(Y_{t_1}^{t_2-1} \right) \prod_{l=2}^L a_{z(l-1), z(l)} b_{z(l)} \left(Y_{t_l}^{t_{l+1}-1} \right), \quad (3)$$

where $\pi_{z(1)}$ is the probability that the state sequence begins in state $z(1)$; $b_{z(l)}$ denotes the acoustic segment pdf associated with state $z(l)$; $a_{z(l-1), z(l)}$ denotes the transition probability from $z(l-1)$ to $z(l)$; t_l is the time at which the state sequence \mathbf{x} enters state $z(l)$, and $t_{L+1} = T + 1$.²

²The introduction of symbol $z(l)$ to denote a state simplifies subsequent notation, in

2.2 The segmental Viterbi decoder

A simple extension of the segmental Viterbi decoder (see, for example, Holmes and Russell 1999) can be used to compute the optimal state sequence $\hat{\mathbf{x}}$ for the given model \mathcal{M} , such that

$$\hat{p}(Y|\mathcal{M}) = p(Y, \hat{\mathbf{x}}|\mathcal{M}) = \max_{\mathbf{x}} p(Y, \mathbf{x}|\mathcal{M}). \quad (4)$$

For completeness, a brief description of the segmental Viterbi decoder is included. By analogy with the notation for the forward probability used in the case of a conventional HMM (see, for example, Holmes and Holmes 2001), let

$$\hat{\alpha}_j(t) = p(\mathbf{y}_1, \dots, \mathbf{y}_t; x_t = s_j, x_{t+1} \neq s_j). \quad (5)$$

The final condition, $x_{t+1} \neq s_j$ is included to ensure that only segments which are complete at time t are considered. Then, it can be shown that,

$$\hat{\alpha}_j(t) = \max_i \max_{\tau} \begin{cases} \pi_i b_i(Y_1^\tau) & \text{for } t = \tau \\ \hat{\alpha}_i(t - \tau) a_{i,j} b_j(Y_{t-\tau+1}^t) & \text{for } t > \tau \end{cases} \quad (6)$$

where $1 \leq \tau \leq \tau_{\max}$ and π_i is the probability that the state sequence begins in the i^{th} state. The requirement in Equation 6 to optimise over all possible segment durations, τ , and to evaluate segmental state output probabilities, $b_j(Y_{t-\tau+1}^t)$, leads to a substantial increase in computational load relative to the normal Viterbi decoder. As in conventional Viterbi decoding for continuous speech recognition, this algorithm is applied to a single, integrated MSHMM in which the individual word- or phone-level MSHMMs are connected according to a grammar (Bridle et al. 1983). Thus, in the case of phone recognition and a bigram language model, the result of decoding is the sequence of phones $[\rho_1, \dots, \rho_\Phi]$ and phone boundaries $[t_1, \dots, t_\Phi]$ such that the joint probability,

$$\hat{p}(Y; t_1, \dots, t_\Phi; \rho_1, \dots, \rho_\Phi) = \hat{p}(Y_1^{t_1} | \rho_1) \prod_{\phi=2}^{\Phi} P(\rho_\phi | \rho_{\phi-1})^\lambda \hat{p}(Y_{t_{\phi-1}+1}^{t_\phi} | \rho_\phi), \quad (7)$$

is maximised. Here, λ is the language model scale factor.

particular Eq. 3. In the symbol x_t , the time index t is in synchrony with the observation sequence \mathbf{y}_t ; whereas for $z(l)$, the index l is in synchrony with the state transitions. Unlike in a conventional HMM, the two are not generally the same in a SHMM.

Our current software uses a single implementation of the segmental Viterbi decoder for embedded and non-embedded training, phone classification and phone recognition. This is achieved by introducing a time-indexed array of breakpoints that specify, at each time t , whether a phone boundary is obligatory, possible or illegal.

2.3 Estimation of the articulatory-to-acoustic mapping

In general, the phone set is partitioned into K phone categories, and a separate mapping W_k is estimated for each one, $k \in \{1, \dots, K\}$. In the work described here it is assumed that corresponding articulatory and acoustic data are available for learning each W_k . This is not necessary, as the W_k s could be optimised in Baum-Welch style re-estimation, along with the other model parameters. However, using matched data preserves the strict articulatory interpretation of the models in the intermediate layer. Suppose that $Y = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T]$ and $R = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_T]$ denote matched sequences of N -dimensional acoustic and M -dimensional articulatory feature vectors, respectively, corresponding to the same phone category k . We would like to find the $N \times M$ matrix W_k that minimises

$$E = (Y - W_k R)' V^{-1} (Y - W_k R), \quad (8)$$

where $'$ denotes the transpose.

In the case where the covariance matrix V is assumed to be diagonal, with diagonal elements denoted by v_n , this reduces to

$$E = \sum_{n=1}^N \frac{1}{v_n} \sum_{t=1}^T (y_t^n - W_k r_t^n)^2, \quad (9)$$

where y_t^n and $W_k r_t^n$ are the n^{th} elements of \mathbf{y}_t and $W_k \mathbf{r}_t$, respectively. Hence minimising E is equivalent to minimising

$$\tilde{E} = (Y - W_k R)' (Y - W_k R). \quad (10)$$

This is a standard least-squares problem, with solution

$$W_k = Y R^\dagger, \quad (11)$$

where Y is the $N \times T$ matrix whose t^{th} column is \mathbf{y}_t , R is the $M \times T$ matrix whose t^{th} column is \mathbf{r}_t , and R^\dagger is the pseudo-inverse of R (see, for example, Bishop 1995). This can be solved using standard techniques from linear algebra, e.g., by singular value decomposition.

2.4 Model parameter estimation

The MSHMM model parameters are optimised using an estimation-maximisation (EM) scheme based on segmental Viterbi decoding. Suppose that we have a set of phone-level MSHMMs which includes an S -state MSHMM \mathcal{M}_ϕ for a phone ϕ in class k . Let $\hat{\mathbf{x}}$ be an optimal state sequence between the sequence of models corresponding to the label for $t \in \{1, \dots, T\}$, and the observation sequence $Y = [\mathbf{y}_1, \dots, \mathbf{y}_T]$. Following the derivations in the Appendix, it can be shown that maximum-likelihood estimates of the midpoint $\hat{\mathbf{c}}_i$ and slope $\hat{\mathbf{m}}_i$ for the i^{th} state, s_i , are given by:

$$\hat{\mathbf{c}}_i = \frac{1}{T_i} \sum_{j=1}^{J_i} \sum_{t=t_j^s}^{t_j^e} (D_i W_k)^\dagger D_i \mathbf{y}_t \quad (12)$$

$$\hat{\mathbf{m}}_i = \frac{\sum_{j=1}^{J_i} \sum_{t=t_j^s}^{t_j^e} (t - \bar{t}_j) (D_i W_k)^\dagger D_i \mathbf{y}_t}{\sum_{j=1}^{J_i} \sum_{t=t_j^s}^{t_j^e} (t - \bar{t}_j)^2}, \quad (13)$$

respectively, where $D_i = V_i^{-\frac{1}{2}}$, and T_i is the total number of frames spent in state s_i , J_i is the number of occurrences of the state in $\hat{\mathbf{x}}$, t_j^s and t_j^e are the start and end times for the j^{th} occurrence of state s_i in $\hat{\mathbf{x}}$, respectively, and $\bar{t}_j = (t_j^s + t_j^e)/2$. Note that if $N = M$ and the rank of W_k is N , then $(D_i W_k)^\dagger = W_k^\dagger D_i^\dagger$ and the D_i terms disappear from both equations.

Intuitively, Equations 12 and 13 state that the maximum likelihood estimates of the trajectory parameters in the articulatory domain are those which give the best linear fit to the (pseudo)inverse-transformed acoustic observation vectors, taking into account the covariance information in the matrix D_i . They are the basis for an iterative EM optimisation algorithm. Given an initial model set \mathcal{M}^0 and training data Y , the segmental Viterbi algorithm is used to compute an optimal state sequence $\hat{\mathbf{x}}^0$. The model parameters are then re-estimated to give a new model set \mathcal{M}^1 such that

$$\hat{p}(Y|\mathcal{M}^1) \geq \hat{p}(Y|\mathcal{M}^0). \quad (14)$$

The process is repeated until convergence is achieved, according to some stopping criterion.

3 Method

3.1 The TIMIT speech corpus

All of the experiments used the TIMIT speech corpus (Garofolo et al. 1993). Speech from all male subjects in the TIMIT training and test sets was down-sampled to 8 kHz for compatibility with the formant analyser (Holmes 2001). We focused on male speakers to reduce the variability that our experimental system would need to accommodate, and because we were concerned that the formant frequency estimates might have been less accurate for female speakers. The data were partitioned into three sets: a *training* set, comprising speech from all male speakers in the TIMIT training set except for the first speaker in each dialect region (318 subjects, 3,180 sentences, 121,400 phones); an *evaluation* set, comprising all of the speech from the first male speaker in each of the eight dialect regions (8 subjects, 80 sentences, 3003 phones); and a *test* set comprising speech from all male speakers in the TIMIT test set (112 subjects, 1,120 sentences, 42,421 phones).

Acoustic features (13 MFCCs, including zeroth) were obtained using HTK (25 ms window, 10 ms fixed frame rate; Young et al. 1997), while formant-based parameters for the intermediate layer were extracted using the Holmes formant analyser (Holmes 2001). Note that only static MFCCs were used in the observation vectors. Three formant-based parameterisations were considered, each with a 10 ms fixed frame rate to ensure synchrony with the acoustic data: (a) 3FF, the first 3 formant frequencies (25 Hz resolution); (b) 3FF+5BE, first 3 formant frequencies plus 5 frequency-band energies; (c) 12PFS, the 12 control parameters from Holmes-Mattingly-Shearme parallel formant synthesizer, which include the first three formant frequencies and amplitudes (Holmes et al. 1964). Thus the dimensions of the intermediate spaces for schemes (a), (b) and (c) were $M = 3, 8$ and 12 , respectively.

Linear articulatory-to-acoustic mappings were estimated with matched sequences of formant-based and acoustic data, as described in section 2.3. In the case of parameterisations (a) and (b), the analyser also returned a confidence measure for each formant estimate, based on the curvature and relative amplitude of a candidate formant peak (Holmes 2001). It has been shown previously that the performance of a formant-based phone recognition system can be improved by weighting the influence of the formant features according to this confidence measure (Garner and Holmes 1998; Wilkinson

and Russell 2002). The comparative performance of linear and nonlinear formant-to-acoustic mappings is studied in (Jackson et al. 2002).

3.2 Phone categories

With one mapping per category, a set of articulatory-to-acoustic mappings $\{W_1, \dots, W_K\}$ was obtained for each of the following categorisations of the phones (the number of mappings K is given in parentheses): A. (1) all data; B. (2) speech, silence/non-speech; C. (6) linguistic categories; D. (10) as in Deng and Ma (2000); E. (10) discrete articulatory regions (Jackson et al. 2002); F. (49) individual phones.

	Discrete	Continuous
Source	phonation [voiced/voiceless]; plosion.	aspiration; frication.
Filter	velum [open/closed]; obstruction in vocal tract, e.g., at the glottis, tongue, teeth, or lips [open/part/closed].	tongue height; tongue position; jaw angle; lip rounding.

Table 1: *Dissection of speech production mechanisms into discrete and continuous features, based on source-filter theory.*

Categorisation E is motivated by a classification of speech production mechanisms into discrete features based on source-filter theory (see Table 1). Many combinations of the acoustic source and vocal-tract configurations outlined in Table 1 are prohibited by physical constraints during speech production, but the standard repertory of English phonemes contains seven or eight discrete groups. Yet although aspiration and frication are continuous as acoustic sources (in that they can be produced at any relative amplitude), linguistically they act to make a phonetic distinction. This paradox highlights the problem of representation, related to ‘overshoot’ of the frication noise

and devoicing in voiced fricatives. Ultimately, we would like an articulatory model that was continuous over turbulence-noise source amplitude. However, for the purposes of this study, we have used the TIMIT labels to deduce articulatory features, and hence frication was treated as a separate discrete source. Using the discrete vocal-tract filter attributes, in conjunction with the various source combinations (and neglecting the fricative component of affricates), data were classified as belonging to one of the ten classes defined in Table 2 (categorisation E).

	Features	Description
1	+voice, VT open	vowel, glide
2	+voice, VT part	liquid, approximant
3	+voice, VT closed, +velum	nasal
4	+voice, VT closed	voiced stop closure
5	−voice, VT closed	voiceless stop closure
6	+voice, VT open, +plosion	voiced stop release
7	−voice, VT open, +plosion	voiceless stop release
8	+voice, VT part, +fric/asp	voiced fricative
9	−voice, VT part, +fric/asp	voiceless fricative
10	−voice	silence, non-speech

Table 2: *Classification of phones according to discrete, acoustic source and filter features, where the states of voicing and the vocal tract (VT) are indicated, followed by additional features: +velum denotes the velum is open, and +plosion and +fric/asp indicate plosive and fricative/aspirative sources, respectively.*

3.3 MSHMM initialisation and re-estimation

The parameters of a conventional, three-state (single component) Gaussian monophone acoustic HMM were estimated for each symbol in the TIMIT 49-phone set using the tools in HTK (Young et al. 1997) and the training set specified above. For each conventional HMM \mathcal{M}_ϕ , representing a phone ϕ in class k , a MSHMM \mathcal{A}_ϕ was created as follows:

- The mid-point vector \mathbf{c}_i for the i^{th} state of \mathcal{A}_ϕ in the articulatory domain was defined as $\mathbf{c}_i = W_k^\dagger \boldsymbol{\mu}_i$, where $\boldsymbol{\mu}_i$ is the mean vector for the i^{th} state of \mathcal{M}_ϕ .
- The slope vector \mathbf{m}_i was set to zero.
- The variance vector \mathbf{v}_i for the i^{th} state of \mathcal{A}_ϕ in the acoustic domain was set equal to the variance vector for the i^{th} state of \mathcal{M}_ϕ .
- The transform W_k and its pseudo-inverse W_k^\dagger were appended to the model.

Given these initial models, Viterbi alignment and Eqs. 12 and 13 were used to re-estimate the MSHMM state parameters. The maximum state duration was set to 15 frames ($\tau_{\max} = 15$), as this was sufficient to accommodate all TIMIT phone labels.

3.4 Triphone MSHMM parameter estimation

The monophone MSHMMs from section 3.3 were used to provide initial parameters for a set of triphone MSHMMs, which were then re-estimated using Equations 12 and 13. The triphone set was defined by a simple ‘backoff’ scheme driven by a parameter n_{\min} , in which a triphone MSHMM was created if at least n_{\min} examples of the relevant context-dependent triphone occurred in the training set, otherwise the corresponding left-context biphone MSHMM was considered. Similarly, if the number of examples of this biphone context in the training set was less than n_{\min} , then the monophone MSHMM was used instead. Small values of n_{\min} will lead to large numbers of models and more accurate modelling of contextual effects. However, if n_{\min} is too small then there will be insufficient training data to ensure robust training.

3.5 Language model

A phone-level probabilistic bigram language model was estimated using all of the TIMIT label files in the training set. Since acoustic and language model probabilities are not necessarily compatible, it is common practice to apply a language model scale factor, denoted by λ in Equation 7.

4 Phone classification results

All experiments reported in this paper are phone classification experiments, in which the TIMIT annotations were used to give the start and end times of each phone.

4.1 Implementation Issues

As explained in section 2.2, our implementation of the segmental Viterbi decoder includes a time-indexed “breakpoint” array which specifies, for each time t , whether a phone boundary is *obligatory*, *possible* or *illegal*. For a classification experiment, *breakpoint*[t] is set to *obligatory* if and only if t corresponds to a phone start time according to the TIMIT annotation, and *illegal* otherwise (for a phone recognition experiment, *breakpoint*[t] is set to *possible* for all values of t). In all cases, the Viterbi decoder is applied to complete sentences. However, the maximisation over τ in Equation 6 stops whenever *breakpoint*[$t - \tau + 1$] is *obligatory* (so that a segment never goes over an obligatory breakpoint), and for this value of τ the maximisation over i is restricted to phone exit states.

For phone classification or recognition experiments, all of the phone models are combined into a single model. In the case of monophones, a transition is allowed from the final null state of a model corresponding to phone ρ_i to the initial null state of a model corresponding to any phone ρ_j , and the scaled language model probability $P(\rho_j|\rho_i)^\lambda$ (see Equation 7) is associated with this transition. However, for context sensitive phone models, such a transition is only permitted if any right-context of the former model \mathcal{M}_{ρ_i} matches ρ_j and if the left-context of the latter model \mathcal{M}_{ρ_j} matches ρ_i .

Thus, the results of Viterbi decoding are the sequence of phone models $[\rho_1, \dots, \rho_\Phi]$ and phone boundaries $[t_1, \dots, t_\Phi]$ such that Equation 7 is max-

imised, subject to the constraints that any right-context of \mathcal{M}_{ρ_i} matches any left-context of $\mathcal{M}_{\rho_{i+1}}$ $\forall i \in \{1, \dots, \Phi - 1\}$ and, in the case of a classification experiment, the phone boundaries $[t_1, \dots, t_\Phi]$ correspond to the phone start times from the TIMIT annotation. The usual reduced TIMIT set of 39 phones, plus silence, was used for scoring.

4.2 Monophone MSHMM performance

4.2.1 Monophone FT-SHMM baseline, no language model

Baseline phone classification experiments were conducted using FT-SHMMs, with no intermediate layer (Holmes and Russell 1999; Jackson and Russell 2002). The results are presented in the first column of Table 3 for constant, ID0(m), and linear, ID1(m), trajectory FT-SHMMs, and are consistent with (Holmes and Russell 1999). Thus the performance upper bound for the monophone MSHMM experiments with no language model was 54.3% phones correct.

4.2.2 Monophone MSHMM results, no language model

The remaining entries in Table 3 are phone classification results for monophone MSHMMs with no language model (Jackson and Russell 2002). The results correspond to different combinations of the number of mappings and the type of formant-based intermediate layer. In these experiments, the D_i term in Eqs. 12 and 13 was ignored, so that the mappings W_k were optimised according to the minimum mean squared error criterion. In general, improved results were obtained by either increasing the dimension of the intermediate layer (i.e., from column (a) towards column (c)), or increasing the number of mappings (downwards to the bottom row). In particular, by using the 12PFS representation or a large set of mappings (e.g., 49 for categorisation F), near optimal performance was achieved.

4.2.3 Effect of formant-frequency confidence

Since optimisation of the mappings W_k will be compromised by errors in the formant parameter estimates, experiments were conducted in which formant

Map.	Base	(a)	(b)	(c)	(a)*
		3FF	3FF+5BE	12PFS	3FF
ID0(m)	52.9	-	-	-	-
ID1(m)	54.3	-	-	-	-
A (1)	-	47.7	53.1	54.0	47.7
B (2)	-	47.5	53.2	53.9	-
C (6)	-	48.9	53.3	53.7	48.6
D (10)	-	48.9	52.9	53.5	48.7
E (10)	-	49.1	53.2	52.7	49.3
F (49)	-	52.9	53.9	54.1	53.1

Table 3: *Classification accuracy (%) for TIMIT tests using monophone models with no language model: ID0(m), identity mapping with constant trajectory; ID1(m), identity mapping with linear trajectory; A, linear mapping with linear trajectory; B, two lin. map. & lin. traj.; C, six lin. map. & lin. traj.; D, ten lin. map. & lin. traj.; E, ten lin. map. & lin. traj.; F, lin. map. per phone & lin. traj. The parameterisations were: MFCCs for the Baseline; (a) 3FF, formant frequencies; (b) 3FF+5BE, formants and band energies; (c) 12PFS, synthesis control parameters. *Maximum-likelihood estimation of the trajectory parameters.*

and MFCC vectors were only used in estimation of the mappings W_k if the formant confidence was greater than a threshold (Table 4). The results show that as the threshold was increased, classification accuracy decreased. It appears that any benefit gained from only training on robust formant estimates was negated by the reduction in training data (Jackson and Russell 2002).

Threshold	0.0	0.2	0.4	0.6	0.8
A (1)	53.1	53.1	53.0	52.8	52.3
F (49)	53.9	53.5	53.4	52.7	52.7

Table 4: *Classification accuracy (%) for TIMIT tests varying the threshold for the minimum acceptable confidence, using formants and band energies (3FF+5BE), with 1 mapping and 49. Key as for Table 3.*

4.2.4 Including variance in parameter re-estimation

The final column of Table 3 shows the results of including an estimate of the matrix D_i in the re-estimation formulae (Eqs. 12 and 13), based on the expected value of E over matched pairs of acoustic and articulatory vectors from category k , \mathbf{y}_t and \mathbf{r}_t respectively (Eq. 8). The results show no significant advantage over ignoring D_i , and suggest that any benefit from including D_i is offset by inaccuracy in its estimation (Jackson and Russell 2002).

4.3 Monophone MSHMM performance with bigrams

4.3.1 Effect of language model scale factor

Phone classification experiments were conducted on the evaluation set to determine an appropriate value for the language model scale factor λ . The experiments included all combinations of the three intermediate representations: (a) 3FF, (b) 3FF+5BE, (c) 12PFS, and the six phone categorisations, with constant (zero slope) and linear (non-zero slope) trajectories. Integer

values of λ from 1 to 40 were considered, giving a total of 1440 experiments. The results of these experiments are summarised in Figure 2 (Russell et al. 2003). For each value of λ , the figure shows the phone classification accuracy for each intermediate representation, and for zero and non-zero trajectory slope, averaged over the six different phone categorisations. The relationships between the different graphs reflect the performance differences from section 4.2.2. It is clear from the results that inclusion of the language model had a substantial effect on results. For example, in the case of the 3FF intermediate representation and phone categorisation C (six linguistic categories), phone classification accuracy increased from 48.9% with no language model to a maximum of 66.7% for $\lambda = 13$. Overall, the results show that setting $\lambda = 10$ was close to optimal in all cases, and so this value was chosen for all subsequent monophone experiments. Phone classification results for the test set with the language model are given in Table 5.

4.3.2 Monophone FT-SHMM baseline, with language model

The baseline phone classification experiments using FT-SHMMs, with no intermediate articulatory layer, were repeated with a language model. The results are presented in the first column of Table 5 for constant, ID0(m+lm), and linear, ID1(m+lm), trajectories. The results show that use of the language model leads to reductions in error rates of 32% (ID0(m+lm)) and 34% (ID1(m+lm)).

4.3.3 Monophone MSHMM results, with language model

The remaining entries in Table 5 are the phone classification results for different combinations of the number of mappings and the type of formant-based intermediate layer (Russell et al. 2003). Again the mappings W_k were optimised according to the minimum mean squared error criterion. As in Table 3, improved results were obtained by either increasing the richness of the intermediate layer, or increasing the number of mappings.

The results in Table 5 were analysed using the NIST implementation of the Matched Pair Sentence Segment test on word errors (MPSSWE; NIST 2000), in order to assess the significance of differences between the performance of the FT-SHMM (ID1(m+lm)) system (at 69.7%) and that of each of the MSHMMs. For column (a), with 3 formant frequencies in the inter-

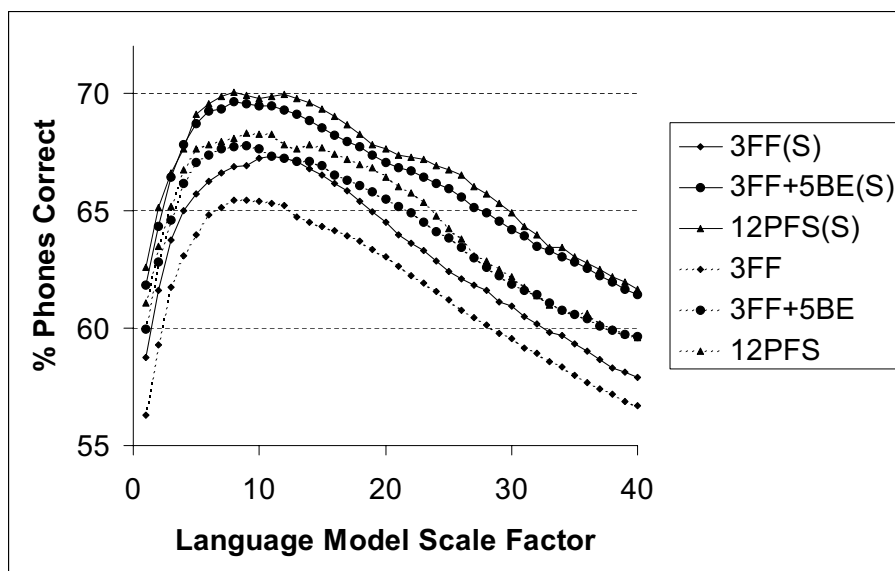


Figure 2: *Phone classification results for different values of the language model scale factor λ . Each curve shows average number of phones correct (%) for the intermediate representations indicated in the key: (a) 3FF, (b) 3FF + 5BE, (c) 12PFS, averaged over the six different phone categorisations. The symbol (S) in the legend indicates linear trajectories in the articulatory domain (i.e., with slope); otherwise the trajectories were constant.*

mediate representation, the performance of all MSHMMs was significantly worse than the FT-SHMM. However, for column (b), where the intermediate representation also included 5 band energies, the MPSSWE test indicated that there was no significant difference between performance for 49 phone classes (categorisation F) and that of the FT-SHMM. Finally, for an intermediate representation comprising 12 synthesizer control parameters (c), the MPSSWE test indicated that the performances for 1, 10(E) and 49 phone classes, and the FT-SHMM were statistically the same.

Map	Base	(a)	(b)	(c)
		3FF	3FF+5BE	12PFS
ID0(m+lm)	68.1	-	-	-
ID1(m+lm)	69.7	-	-	-
A(1)	-	64.6	68.6	69.7
C(6)	-	65.9	69.1	69.5
D(10)	-	65.4	69.3	69.4
E(10)	-	66.1	69.2	69.7
F(49)	-	68.7	69.5	69.7

Table 5: *Phone classification results using monophone MSHMMs with a language model. ID0(m+lm) and ID1(m+lm) used acoustic (i.e., no intermediate layer) fixed trajectory monophone SHMMs with zero (constant) and non-zero (linear) slopes, respectively.*

4.4 Triphone MSHMM performance

4.4.1 Determination of the number-of-examples threshold

Before classification experiments on the test set using triphone MSHMMs could be conducted, it was necessary to determine an appropriate value for the ‘back-off’ parameter n_{\min} , which provides a threshold for the minimum number of instances required to create a context-sensitive model. Figure 3 shows phone classification accuracy on the evaluation set as a function of n_{\min} for triphone FT-SHMMs (i.e., no intermediate articulatory layer) with linear trajectories. Surprisingly, classification accuracy was maintained, or improved, for progressively smaller values of n_{\min} , even for as few as five examples. However, these small values resulted in a large number of models and, hence, in excessive computational load. The results were analysed using the MPSSWE test, which indicated that there were no significant differences between the performance obtained with $n_{\min} = 30$ and that obtained with smaller values of the threshold. Hence, n_{\min} was set to 30, giving a set of exactly 1400 triphones. Although this value of n_{\min} was chosen empirically using sets of triphone FT-SHMMs, it was assumed that the same value would be appropriate for a range of MSHMMs.

4.4.2 Triphone MSHMM results, with language model

Table 6 shows the phone classification accuracy achieved using triphone MSHMM systems on the test set. Our baseline result for a triphone FT-SHMM with linear trajectories, ID1(t+lm), was 75.3% (see Table 6). As in the monophone case, putting aside considerations of local optima, this represents a theoretical upper bound for the performance of all the linear/linear triphone MSHMM systems. The third, fourth and fifth columns of Table 6 correspond to intermediate representations comprising: (a) three formant frequencies, 3FF; (b) three formant frequencies plus five band energies, 3FF+5BE; (c) twelve parallel formant synthesizer control parameters, 12PFS. As with the monophone systems, with the exception of phone categorisation D, classification accuracy for the three-formant-frequency representation (a) increased with the number of categories. Again, this demonstrated that shortcomings in the intermediate representation could be offset by increasing the number of phone categories. According to the MPSSWE

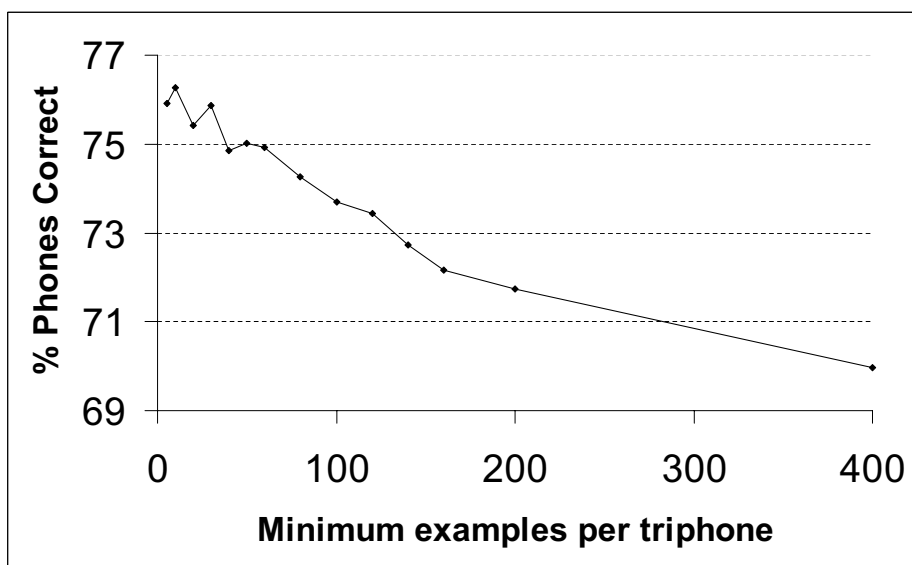


Figure 3: *Phone classification results on the TIMIT male evaluation set for different values of the threshold n_{\min} .*

test, the final result in column (a), corresponding to 3FF and 49 phone categories (F), was significantly better than the ID0(m+lm) baseline, which approximated that of a conventional HMM. A similar pattern of results occurred in column (b), corresponding to an intermediate representation with 3FF+5BE.

The final column (c) of Table 6 shows phone classification results for triphone MSHMMs in which the intermediate representation consisted of the twelve control parameters from the HMS parallel formant synthesizer (12PFS). In this case, the NIST MPSSWE test indicated that, with the exception of the phone categorisation D, the differences between the theoretically-optimal baseline FT-SHMM performance, ID1(t+lm), and the performance of any of the MSHMM systems were not statistically significant. The MPSSWE test also revealed that there were no significant differences between the performances achieved with phone categorisations A, E and F.

Map	Base	(a)	(b)	(c)
		3FF	3FF+5BE	12PFS
ID0(t+lm)	72.7	-	-	-
ID1(t+lm)	75.3	-	-	-
A(1)	-	70.3	74.3	75.4
C(6)	-	70.6	74.3	75.0
D(10)	-	70.2	74.5	74.9
E(10)	-	71.1	74.4	75.5
F(49)	-	73.3	75.0	75.4

Table 6: *Phone classification results on the TIMIT male test set using triphone MSHMMs with language model. ID0(t+lm) and ID1(t+lm) used acoustic (i.e., no intermediate layer) fixed-trajectory triphone SHMMs with zero (constant) and non-zero (linear) slopes, respectively.*

4.4.3 Previous results on TIMIT

For comparison, the best phone *recognition* performance achieved on the full TIMIT core test set (male and female speakers) using optimised conventional HMMs is 72.9% phone accuracy (Lamel and Gauvain 1993). More recently, Halberstadt and Glass have reported a phone recognition accuracy of 75.6% on the same test set (Halberstadt and Glass 1998). A summary of reported phone recognition error rates on the TIMIT core test set is presented in (Glass 2003). These results are clearly superior to the best result presented here. However, our emphasis in section 4 has been to study an evolving sequence of systems with increasing complexity, and in each case to demonstrate the effect of introducing multiple-level, segmental structure. We believe that this is appropriate for initial development of an experimental model. Competing with the state-of-the-art is a challenge for future work.

4.4.4 Comparison of model sizes

This section compares the numbers of parameters in the various systems. Let M denote the number of acoustic models, I and J the dimensions of the intermediate and acoustic spaces, respectively, T the number of “free” transition probabilities in a single conventional HMM, N the number of emitting states, and K the number of different articulatory-to-acoustic mappings. Then the number of parameters in the acoustic component of a conventional HMM system is given by

$$P_{\text{HMM}} = M \times [N \times (2 \times J) + T] \quad (15)$$

and the corresponding figure for a MSHMM system is

$$P_{\text{MSHMM}} = M \times [N \times (2 \times I + J) + (T - N)] + K \times (I + 1) \times J \quad (16)$$

Assuming that the number of models M is the same, the MSHMM system will have significantly fewer parameters than the conventional HMM system if the dimension of the intermediate representation (I) is significantly less than that of the acoustic representation (J), and the number of models (M) is significantly greater than the number of mappings (K). For monophone systems, the second condition fails. In fact, only the monophone MSHMM systems with the 3FF intermediate representation and 1, 6 or 10 phone categories

have fewer parameters than the conventional monophone HMM system, and the performances of these three systems are poor.

In the case of triphone systems, the number of models is consistently much larger than the number of mappings. Consequently, all of the triphone systems with the 3FF intermediate representation have significantly fewer parameters than the conventional HMM system. In particular, compared with the latter, the triphone MSHMM system with the 3FF intermediate representation and phone categorisation F (49 mappings), has 25% fewer parameters and gives better performance. A full comparison of the numbers of parameters in the different triphone systems is presented in Figure 4.

5 Discussion and further work

In view of the limitations of the linear/linear system which are described in the Introduction, it is natural to consider ‘non-linear/linear’ MSHMMs, in which the trajectories in the articulatory space are non-linear, or ‘linear/non-linear’ MSHMMs, in which the articulatory-to-acoustic mapping is non-linear. Of these, the case for non-linear articulatory-to-acoustic mappings is more compelling, and recent research at Birmingham (Jackson et al. 2002; Lo and Russell 2003) has focussed on the use of radial basis function (RBF) networks for this purpose (see, for example Bishop 1995). Estimation of an RBF-based articulatory-to-acoustic mapping using matched data is straightforward, but the optimisation of the model trajectory parameters is not (cf. Equations 12 and 13 for the linear case). Applying the pseudo-inverse of the linear component of the RBF network to acoustic data results in a set of ‘target’ values for the RBF kernel functions. The trajectory parameters must then be adjusted to maximise their fit to these targets using gradient ascent (Lo and Russell 2003). This research is ongoing.

A number of other issues also warrant further study. First, the use of articulatory-to-acoustic mappings that are ‘pre-trained’ on matched data is unlikely to be optimal. Therefore, unsupervised iterative training of the mappings should be advantageous (using the Viterbi-aligned sequences of ‘synthetic’ articulatory data generated by the MSHMMs, and the observed acoustic data). However, the explicit ‘articulatory’ interpretation of the intermediate layer would be lost. Second, the basic segment probability function (Equation 1) includes a duration probability term $c_i(\tau)$ which has not

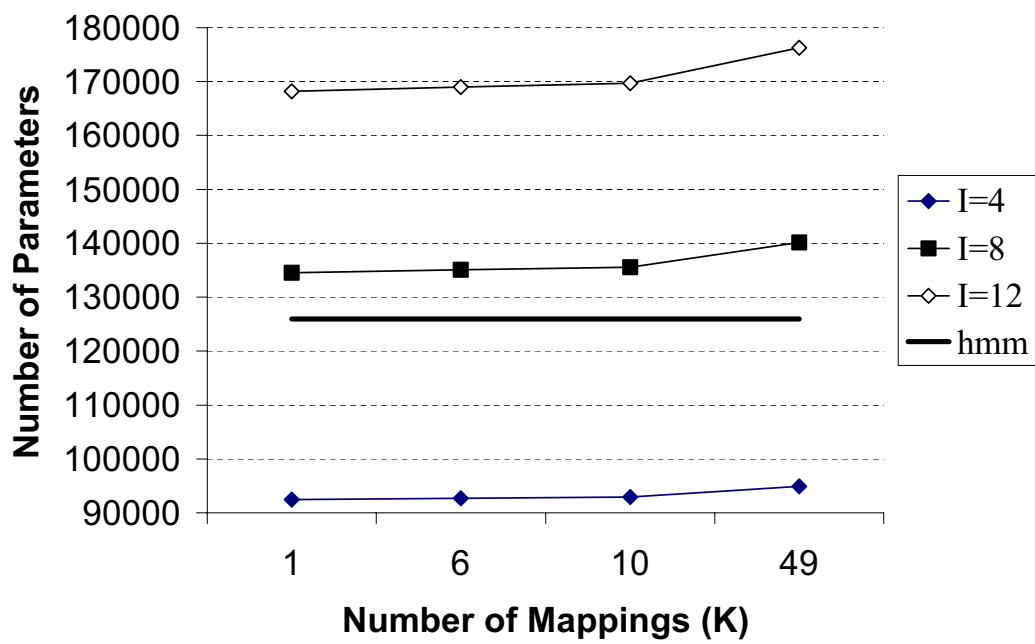


Figure 4: Comparison of the numbers of free parameters in the various triphone-based systems. I indicates the dimension of the intermediate representation. The horizontal line shows the number of parameters in the conventional HMM system.

been investigated in this paper. Experiments reported elsewhere (Jackson 2003) have demonstrated that a more accurate model of segment duration can result in improvements in phone classification accuracy.

A limitation of the experiments presently reported is that they are restricted to static cepstral parameters, whereas conventional state-of-the-art HMM systems also employ the first and second time-derivatives of the cepstrum (the so-called delta and acceleration parameters). Of course, deltas and accelerations of the articulatory parameters could also be computed and incorporated into the intermediate articulatory representation of a MSHMM. However, for consistency, linear trajectories in the articulatory delta space should correspond to quadratic and constant trajectories in its static and acceleration spaces, respectively. There is much scope here for further research. Another apparent limitation of our MSHMMs is that variability in the acoustic space is modelled using a (single-component) multivariate Gaussian pdf, whereas conventional HMMs employ multiple component Gaussian mixture state-output pdfs to considerable advantage. Two remarks are pertinent here. First, it is not immediately clear how the current use of a Gaussian pdf in a MSHMM can be extended to mixture pdfs. Second, Gaussian mixture pdfs represent a generic approach to accommodating unknown variability, whereas the goal of the current research is to understand and model the mechanisms which give rise to that variability. Thus it could be argued that the use of mixture pdfs would be an admission of defeat!

Finally, the introduction of triphone linear/linear MSHMMs significantly increases in computational load, and the use of non-linear articulatory-to-acoustic mappings would compound this problem. In order to continue this research, techniques for reducing the computational requirements of the segmental Viterbi decoder (Equation 6), such as a ‘segmental’ version of beam-pruning, will need to be investigated.

6 Conclusion

A theoretical and experimental study of a simple class of ‘linear/linear’ multi-level segmental hidden Markov models has been presented, in which speech dynamics are modelled as linear trajectories in an intermediate, articulatory-based representation, and mapped into acoustic space using a set of one or more linear transformations. It has been demonstrated that, provided the

intermediate layer is sufficiently rich, or the articulatory-to-acoustic mapping is sufficiently flexible, the TIMIT phone classification rate for such a model is not significantly different from the best that can be achieved with linear trajectories. The significance of this result is that it provides a solid theoretical foundation for the development of richer classes of multi-level models, which include non-linear models of dynamics, alternative articulatory representations, sets of non-linear articulatory-to-acoustic mappings, and integrated optimisation schemes that support unsupervised learning of the trajectory, intermediate representation and mapping parameters.

The incorporation of a low dimensional, articulatory-based intermediate representation in a segmental HMM has many attractions. It provides a compact and interpretable framework for speaker adaptation and for meaningful characterisation of the mechanisms that give rise to variability, such as in conversational speech, and for incorporation of articulatory constraints on the recognition process. It also has implications for model-based speech synthesis, and advances the goal of developing unified, trainable models which can support both recognition and synthesis.

References

- Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, UK: OUP.
- Bridle, J. S., M. D. Brown, and R. M. Chamberlain (1983). Continuous connected word recognition using whole-word templates. *Radio Engineer* 53, 167–177.
- Deng, L. (1998). A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. *Speech Communication* 24(4), 299–323.
- Deng, L. and D. Braam (1994, October). Context-dependent markov model structured by locus equations: Applications to phonetic classification. *J. Acoust. Soc. Am.* 96(4), 2008–2025.
- Deng, L. and J. Ma (2000). Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics. *J. Acoust. Soc. Am.* 108(6), 3036–3048.

- Digalakis, V. (1992). *Segment-based stochastic models of spectral dynamics for continuous speech recognition*. Ph. D. thesis, Boston University, MA.
- Frankel, J., K. Richmond, S. King, and P. Taylor (2000). An automatic speech recognition system using neural networks and linear dynamic models to recover the model articulatory traces. *Proc. Int. Conf. on Spoken Lang. Proc.*, Beijing.
- Gales, M. J. F. and S. J. Young (1993). Segmental hidden markov models. In *Proc. Eurospeech '93*, Berlin, Germany, pp. 1579–1582.
- Gao, Y., R. Bakis, J. Huang, and B. Zhang (2000). Multistage coarticulation model combining articulatory, formant and cepstral features. In *Proc. Int. Conf. on Spoken Lang. Proc.*, Beijing, Volume 1, pp. 25–28.
- Garner, P. N. and W. J. Holmes (1998). On the robust incorporation of formant features into hidden Markov models for automatic speech recognition. In *Proc. IEEE-ICASSP*, Seattle, WA.
- Garofolo et al., J. S. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Univ. Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.
- Ghitza, O. and M. Sondhi, M (1993). Hidden markov models with templates as non-stationary states: an application to speech recognition. *Comp. Speech & Lang.* 2, 101–119.
- Glass, J. (2003, April-July). A probabilistic framework for segment-based speech recognition. *Comp. Speech & Lang.* 17(2-3), 137–152.
- Halberstadt, A. and J. Glass (1998). Heterogeneous measurements and multiple classifiers for speech recognition. *Proc. Int. Conf. on Spoken Lang. Proc.*, Sydney, Australia, 995–998.
- Holmes, J. and W. Holmes (2001). *Speech synthesis and recognition* (2nd ed.). London and New York: Taylor and Francis.
- Holmes, J. N. (2001). *Speech processing system using formant analysis*. US Patent. US6292775.
- Holmes, J. N. and P. N. Garner (2000). Using formant frequencies in speech recognition. In *Proc. IEEE-ICASSP*, Istanbul, Turkey, Volume 3, pp. 1347–1350.

- Holmes, J. N., W. J. Holmes, and P. N. Garner (1997). Using formant frequencies in speech recognition. In *Proc. Eurospeech '97*, Rhodes, Greece, pp. 2083–2086.
- Holmes, J. N., I. G. Mattingly, and J. N. Shearme (1964). Speech synthesis by rule. *Language & Speech* 7, 127–143.
- Holmes, W. J. and M. J. Russell (1999). Probabilistic-trajectory segmental HMMs. *Comp. Speech & Lang.* 13(1), 3–37.
- Hunt, M. J. (1987). Delayed decision making in speech recognition — the case of formants. *Pattern Recognition Letters* 6, 121–137.
- Jackson, P. J. B. (2003). Improvements in classification accuracy through modelling duration. In *Proc. ICPHS*, Barcelona.
- Jackson, P. J. B., B.-H. Lo, and M. J. Russell (2002). Data-driven, non-linear, formant-to-acoustic mapping for ASR. *El. Lett.* 38(13), 667–669.
- Jackson, P. J. B. and M. J. Russell (2002). Models of speech dynamics in a segmental-HMM recogniser using intermediate linear representations. In *Proc. Int. Conf. on Spoken Lang. Proc.*, Denver, CO, pp. 1253–1256.
- Lamel, L. F. and J. L. Gauvain (1993). High performance speaker-independent phone recognition using cdhmm. In *Proc. Eurospeech '93*, Berlin, Germany, pp. 121–124.
- Lo, B. H. and M. J. Russell (2003). Speech recognition using an intermediate articulatory layer and non-linear articulatory-to-acoustic mapping. In *One day meeting for young speech researchers, University College London*.
- NIST (2000). *Benchmark tests: Significance tests for ASR*. Gaithersburg, MD: National Institute of Standards and Technology, (Speech Group). [<http://www.nist.gov/speech/tests/sigttests/sigttests.htm>].
- Ostendorf, M., V. V. Digalakis, and O. A. Kimball (1996). From HMM's to segmental models: a unified view of stochastic modeling for speech recognition. *IEEE Trans. on Spch. & Aud. Proc.* 4(5), 360–378.
- Richards, H. B. and J. S. Bridle (1999). The HDM: a segmental Hidden Dynamic Model of coarticulation. In *Proc. IEEE-ICASSP*, Phoenix, AZ, pp. 357–360.
- Russell, M. J. (1993). A segmental HMM for speech pattern modelling. In *Proc. IEEE-ICASSP*, Minneapolis, MN, pp. 499–502.

- Russell, M. J., P. J. B. Jackson, and L. P. Wong (2003). Development of articulatory-based multi-level segmental HMMs for phonetic classification in ASR. In *Proc. EC-VIP-MC '03*, Zagreb, Croatia.
- Wiewiorka, A. and D. M. Brookes (1996). Exponential interpolation of states in a hidden Markov model. *Proc. Institute of Acoustics* 18(9), 201–208.
- Wilkinson, N. and M. J. Russell (2002). Improved phone recognition on TIMIT using formant frequency data and confidence measures. In *Proc. Int. Conf. on Spoken Lang. Proc.*, Denver, CO, pp. 2121–2124.
- Young, S. J., J. Odell, D. Ollason, V. Valtchev, and P. Woodland (1997). *The HTK Book* (v2.1 ed.). Cambridge, UK: Entropic Camb. Res. Lab.
- Zhou, J., F. Seide, and L. Deng (2003). Coarticulation modeling by embedding a target-directed hidden trajectory model into hmm - modeling and training. In *Proc. IEEE-ICASSP*, Hong Kong, Volume 1, pp. 744–747.

A Optimal parameter estimation

A.1 Estimation of trajectory parameters

This section of the appendix describes the derivation of the maximum likelihood (ML) estimates for the midpoint and slope, $\hat{\mathbf{c}}$ and $\hat{\mathbf{m}}$ respectively.

A.1.1 Definitions

The linear trajectory for state s_i is defined as

$$\mathbf{f}_i(t) = \mathbf{m}_i(t - \bar{t}) + \mathbf{c}_i \quad (17)$$

where $\bar{t} = (\tau + 1)/2$. Thus, the probability of the observation sequence for that state, with linear mapping W_k and covariance matrix V_i (in the acoustic space), is

$$\mathbf{b}_i(\mathbf{y}_1^\tau) = \sqrt{\frac{\det |V_i^{-1}|}{(2\pi)^N}} \prod_{t=1}^{\tau} \exp -\frac{1}{2} (\mathbf{y}_t - W_k \mathbf{f}_i(t))' V_i^{-1} (\mathbf{y}_t - W_k \mathbf{f}_i(t)) \quad (18)$$

where $'$ denotes the transpose. For the multivariate Gaussian pdf, maximising the likelihood is equivalent to minimising the Malhalanobis distance:

$$E = \sum_{t=1}^{\tau} \left((\mathbf{y}_t - W_k \mathbf{f}_i(t))' V_i^{-1} (\mathbf{y}_t - W_k \mathbf{f}_i(t)) \right). \quad (19)$$

A.1.2 ML estimate of the midpoint

Hence, the maximum of the likelihood function can be found from the point at which E 's gradient is zero, for which we calculate the derivative,

$$\begin{aligned} \frac{\partial E}{\partial \mathbf{c}_m} &= \sum_{t=1}^{\tau} \frac{\partial}{\partial \mathbf{c}_m} \left((\mathbf{y}_t - W_k \mathbf{f}_i(t))' V_i^{-1} (\mathbf{y}_t - W_k \mathbf{f}_i(t)) \right) \\ &= 2 \sum_{t=1}^{\tau} \mathbf{w}_m V_i^{-1} (\mathbf{y}_t - W_k \mathbf{f}_i(t)) \end{aligned} \quad (20)$$

where \mathbf{w}_m is the m^{th} row of the mapping matrix W_k . Therefore at the optimum, we have

$$\mathbf{w}_m V_i^{-1} \sum_{t=1}^{\tau} (\mathbf{y}_t - W_k (\mathbf{m}_i (t - \bar{t}) + \hat{\mathbf{c}}_i)) = 0. \quad (21)$$

The maximum likelihood solution can also be expressed as

$$\begin{aligned} 0 &= \mathbf{w}_m V_i^{-1} \sum_{t=1}^{\tau} \mathbf{y}_t - \mathbf{w}_m V_i^{-1} \sum_{t=1}^{\tau} W_k \mathbf{m}_i (t - \bar{t}) - \mathbf{w}_m V_i^{-1} \sum_{t=1}^{\tau} W_k \hat{\mathbf{c}}_i \\ \mathbf{w}_m V_i^{-1} W_k \hat{\mathbf{c}}_i &= \frac{1}{\tau} \mathbf{w}_m V_i^{-1} \sum_{t=1}^{\tau} \mathbf{y}_t, \end{aligned} \quad (22)$$

since $\sum_{t=1}^{\tau} (t - \bar{t})$ sums to zero. The result can then be vectorised and rearranged to give

$$\begin{aligned} \hat{\mathbf{c}}_i &= \frac{1}{\tau} \left[W_k' V_i^{-1} W_k \right]^{-1} W_k' V_i^{-1} \sum_{t=1}^{\tau} \mathbf{y}_t \\ &= \frac{1}{\tau} [D_i W_k]^\dagger D_i \sum_{t=1}^{\tau} \mathbf{y}_t, \end{aligned} \quad (23)$$

where $D_i = V_i^{-\frac{1}{2}}$ and † denotes the pseudoinverse, leading to the expression used in Eq. 12.

A.1.3 ML estimate of the slope

Similarly, differentiating Eq. 19 with respect to the gradient parameter yields

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{m}_m} &= \sum_{t=1}^{\tau} \frac{\partial}{\partial \mathbf{m}_m} \left((\mathbf{y}_t - W_k \mathbf{f}_i(t))' V_i^{-1} (\mathbf{y}_t - W_k \mathbf{f}_i(t)) \right) \\ &= 2 \sum_{t=1}^{\tau} (t - \bar{t}) \mathbf{w}_m V_i^{-1} (\mathbf{y}_t - W_k \mathbf{f}_i(t)).\end{aligned}\quad (24)$$

At the optimum, we have

$$0 = \mathbf{w}_m V_i^{-1} \left[\sum_{t=1}^{\tau} (t - \bar{t}) \mathbf{y}_t - W_k \sum_{t=1}^{\tau} \hat{\mathbf{m}}_i (t - \bar{t})^2 - W_k \sum_{t=1}^{\tau} \mathbf{c}_i (t - \bar{t}) \right], \quad (25)$$

and, as before, this reduces to just two terms:

$$\mathbf{w}_m V_i^{-1} \sum_{t=1}^{\tau} (t - \bar{t}) \mathbf{y}_t = \mathbf{w}_m V_i^{-1} W_k \sum_{t=1}^{\tau} \hat{\mathbf{m}}_i (t - \bar{t})^2. \quad (26)$$

Finally, vectorising and rearranging gives

$$\begin{aligned}\hat{\mathbf{m}}_i &= \frac{1}{\sum_{t=1}^{\tau} (t - \bar{t})^2} \left[W_k' V_i^{-1} W_k \right]^{-1} W_k' V_i^{-1} \sum_{t=1}^{\tau} (t - \bar{t}) \mathbf{y}_t \\ &= [\mathbf{t} D_i W_k]^\dagger D_i \mathbf{y}_t,\end{aligned}\quad (27)$$

where $\mathbf{t} = [1 - \bar{t}, 2 - \bar{t}, \dots, \tau - \bar{t}]'$, which provides the result used in Eq.13:

$$\hat{\mathbf{m}}_i = \frac{1}{\sum_{t=1}^{\tau} (t - \bar{t})^2} [D_i W_k]^\dagger D_i \sum_{t=1}^{\tau} (t - \bar{t}) \mathbf{y}_t. \quad (28)$$