MODELLING ASYNCHRONY IN AUTOMATIC SPEECH RECOGNITION USING LOOSELY COUPLED HIDDEN MARKOV MODELS

H.J. Nock* and S.J. Young
Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK.
{hjn11,sjy}@eng.cam.ac.uk
(* Currently visiting SSLI Lab, University of Washington, USA.)

July 6, 2001

Abstract

Hidden Markov Models (HMMs) have been successful for modelling the dynamics of carefully dictated speech, but their performance degrades severely when used to model conversational speech. Since speech is produced by a system of loosely coupled articulators, stochastic models explicitly representing this parallelism may have advantages for automatic speech recognition (ASR), particularly when trying to model the phonological effects inherent in casual spontaneous speech. This paper presents a preliminary feasibility study of one such model class: *loosely coupled HMMs*. Exact model estimation and decoding is potentially expensive, so possible approximate algorithms are also discussed. Comparison of one particular loosely coupled model with conventional speech models on an isolated word task suggests loosely coupled HMMs merit further investigation. An approximate algorithm giving performance which is almost always statistically indistinguishable from the exact algorithm is also identified, making more extensive research computationally feasible.

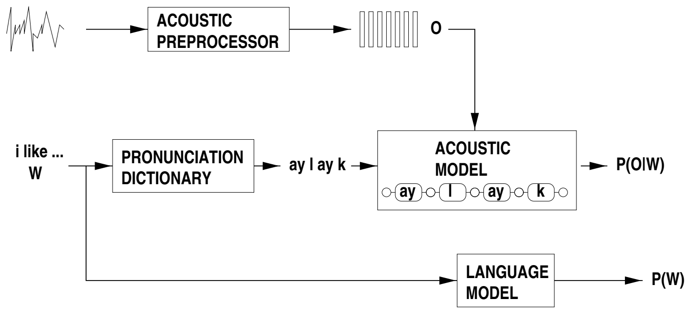

Figure 1: Statistical Framework For Speech Transcription

1 BACKGROUND AND MOTIVATION

Most current automatic speech recognition (ASR) systems use a statistical formulation of the ASR problem, due to [Bahl et al., 1983] (Figure 1), since it has consistently led to performance exceeding purely linguistically motivated approaches. An *acoustic preprocessor* converts the speech waveform into a sequence of observation vectors $\mathbf{O} = \mathbf{O}_1, \ldots, \mathbf{O}_T$, which represents the acoustic evidence upon which the recognizer makes a decision. The recognizer or *decoder* seeks the word sequence $\mathbf{W}^*$ such that

$$\begin{aligned} \mathbf{W}^* &= \arg\max_{\mathbf{W}} p(\mathbf{W}|\mathbf{O}) && (1) \\ &= \arg\max_{\mathbf{W}} p(\mathbf{O}|\mathbf{W}) p(\mathbf{W}) && (2) \end{aligned}$$

where $\mathbf{W}$ denotes a valid word sequence. Probability $p(\mathbf{O}|\mathbf{W})$ is provided by an *acoustic model* and $p(\mathbf{W})$ by a *language model*; these models are typically estimated independently.

This work focuses on the acoustic model. Most current commercial and research systems use Hidden Markov Models (HMMs), partly because of efficient algorithms for parameter estimation and decoding [Rabiner, 1989]. For isolated word tasks with sufficient training data, a single HMM is trained per word. For tasks with insufficient data or for continuous speech tasks, an overall HMM is formed for a word or word sequence using the *beads-on-a-string* procedure: HMMs of phone or phoneme-like subword units are concatenated according to the mapping from words to subword unit sequences given in the *pronunciation dictionary* (Figure 1). Subword units may be modelled directly or using *context-dependent* models such as *triphones*, in which a separate HMM is constructed for each phoneme in the context of a single preceding and following phoneme.

This approach yields very good performance when applied to dictated speech (eg. [Woodland et al., 1995]) but performance degrades severely when confronted with conversational speech. For example, the DARPA HUB4E Broadcast News Evaluation includes both spontaneous and more formal utterances in studio recording conditions: in 1998 every system performed less well on spontaneous utterances (F1) than more formal, studio speech utterances (F0) [Graff, 1997, Hub4-Results, 1999]. Our experiments using a data set comprising the *same* word-level transcript recorded in different speaking styles show the difficulties are partly associated with changes in the acoustic realizations of words (as opposed to changes in grammar and vocabulary) [Saraclar et al., 2000, Weintraub et al., 1996]. Many researchers hypothesize the difficulties are specifically related to increased pronunciation change in conversational speech: the increased variability of phoneme realizations and greater phonetic and lexical deletion may not be adequately modelled by current implementations of the beads-on-a-string procedure [Cohen, 1989, Fosler-Lussier, 1999, Greenberg et al., 1996, Keating, 1997, Saraclar, 2000]. Most pronunciation dictionaries have only one or two pronunciations per word, unlikely to cover all variants in conversational speech [Keating, 1997]; often the pronunciations included are not even frequent conversational variants, since dictionaries are often derived from dictated speech or even text-to-speech systems. Thus, there is a strong assumption that the statistical subword modelling scheme adequately captures the remaining variability. However, whilst context-dependent models do acknowledge contextual effects on the realization of sounds and mixture of Gaussian output distributions in HMMs may capture variability in segment realizations, it can be argued that neither technique is a parsimonious model of these types of change. Further, phone-level HMMs without skip transitions[1] are unlikely to adequately model phonetic and lexical deletions.

One approach to these problems extends the pronunciation lexicon with multiple, conversational pronunciations for each word, possibly weighting variants by probabilities. Unfortunately this approach often increases *confusability* by increasing the homophonous word sequences which must be distinguished solely through pronunciation and language model probabilities eg. [Riley et al., 1999, Saraclar et al., 2000]. Dynamically restricting the set of word pronunciations to those "appropriate" for each speaker and speaking style is a possibility but gains have again been limited eg. [Fosler-Lussier, 1999, Ostendorf, 2000].

A rather different approach is motivated by objections to the fundamental assumption made by the beads-on-a-string procedure: namely, the assumption that speech can be rigidly segmented into a linear sequence of (phone-like) segments. Speech scientists, linguists and engineers agree the notion of a phoneme or speech segment is not realistic eg. [King and Taylor, 2000, Liberman, 1998], although it proved an adequate assumption for dictated speech transcription systems. Speech production studies show that changes in speaking rate, manner and style can lead to variation in the amplitude of and phase relations between articulatory gestures; these changes in relative timing underlie the colouring and merging of 'segments' and the 'segment-like' insertions that extended pronunciation dictionaries attempt to capture. Examples of these effects include feature spreading eg. `CAN'T /k ae n t/` $\rightarrow$ `[ k` $\hat{ae}$ `t ]`, where

[1]Skip transitions are rarely used in state-of-the-art systems, having been found to degrade performance.

$\hat{ae}$ indicates nasality from the deleted segment `/n/` has coloured the neighbouring vowel, and asynchronous articulation errors causing stop insertions eg. `WARMTH /w ao m th/` $\rightarrow$ `[ w ao m p th ]`. When articulations become more decoupled, as in conversational speech, it becomes increasingly difficult to describe pronunciation variation at the level of segments. Motivated by *non-linear* or *autosegmental* rather than *linear* phonological models in linguistics [Goldsmith, 1999], researchers have therefore begun considering methods for modelling phonological processes that incorporate the more fundamental ideas of asynchrony between articulatory gestures or phonological features eg. [Moore, 1996, Rose et al., 1996, Russell, 1997]. In particular, several authors advocate a two-stage approach to ASR in which the acoustic signal is first mapped into an intermediate representation comprising several potentially asynchronous feature streams, such as phonologically-motivated distinctive features or articulatory parameters; this representation is then modelled using an approach incorporating the notion of asychrony between feature changes eg. [Huckvale, 1994, Kirchhoff, 1996, King and Taylor, 2000]. Thus, for example, the partial colouring of vowel `/ae/` by nasal `/n/` is modelled by timing differences in feature changes between the combinations for `/n/` and for `/ae/`. However, whilst timing changes in different feature tiers may not be fully coupled, there is still some dependence between the points at which they change.

Papers discussing articulatory or phonological speech representations are ubiquitous eg. [Ali et al., 1999, Dalsgaard, 1992, Eide et al., 1993, Frankel et al., 2000, Huckvale, 1994, King and Taylor, 2000, Kirchhoff, 1998, Niyogi et al., 1999, Saul et al., 2001, Stephenson, 1998, Stevens, 2000]; for the purposes of this work, such representations are simply thought of as multiple, discrete time series. Fewer papers consider schemes for directly modelling or otherwise incorporating these representations within a statistical ASR system, although [Deng and Erler, 1992, Kirchhoff, 1998, Stephenson, 1998, Frankel et al., 2000, Stephenson et al., 2000, Richardson et al., 2000, Zweig, 1998] represent recent efforts. The contribution of this paper is to develop an approach to modelling multiple time series that are potentially *loosely coupled* rather than assumed fully or only very weakly coupled as in previous work[2]; it then considers how these models might be made tractable for use in large vocabulary ASR.

The paper is organised as follows. Section 2 outlines the theory of loosely coupled HMMs and shows several standard speech models are special cases. It then introduces one specific loosely coupled HMM: the *Mixed-Memory Assumption Factorial HMM* (MMFHMM) and outlines an EM algorithm for estimating MMFHMMs with multivariate Gaussian distributions. This exact algorithm is potentially computationally costly so the section also considers approximate algorithms that may be necessary for applying the new models to large-vocabulary speech recognition tasks. Section 3 presents a preliminary evaluation of models and algorithms on a standard isolated word classification task. The paper ends with conclusions and outlines future research.

[2] Similar arguments might motivate loosely coupled models as a relation of the more standard multiband models used for noise robustness eg. [Mirghafori, 1999].

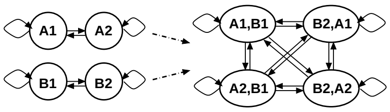

Figure 2: Metastate Space From Combined Ergodic HMMs A, B

2 THEORY OF LOOSELY COUPLED HMMS

The data to be modelled comprises K loosely coupled time series. Observations in each time series (or *stream*) k, denoted $o_1^k, o_2^k, \ldots, o_T^k$, are produced on the same time-scale and may be scalars or vectors. Each time series might correspond to an articulator trace or a phonological feature such as voicing, for example.

One scheme for modelling such data combines the K observations at each time t into a single observation vector $O_t = (o_t^1, \ldots, o_t^K)$ and then builds a standard HMM, but this may not yield a parsimonious representation of the data. Alternatively, each stream k could be modelled independently by using a single HMM per stream and the resulting likelihoods combined to give an overall score, but this fails to capture coupling between the different time series. An intermediate approach is to combine the K independent HMMs into a joint model which can capture more of the coupling between different streams. We can form a combined or *factorial* HMM in which (i) the hidden state space (or *metastate* space) is the Cartesian product of the state spaces of the individual HMMs (see Figure 2), and (ii) the observations O_t are formed by concatenating the individual stream observations at time t, ie. $O_t = (o_t^1, \ldots, o_t^K)$.

The combined model just described is equivalent to a standard HMM in which states and observations now have internal structure. However, as K and N increase, estimation of output densities and transition matrix for this *factorial* HMM will become intractable both computationally and in terms of robust parameter estimation. Recent work in the machine learning and speech literature handles these difficulties through additional conditional independence assumptions and approximations which exploit the internal, combinatorial structure of the metastates and observations firstly to reduce the number of parameters and sometimes as the basis for efficient, approximate training and decoding algorithms (eg. [Ghahramani and Jordan, 1997, Saul and Jordan, 1999, Deng and Erler, 1992, Sun et al., 2000, Logan and Moreno, 1998]). The next section will show that the general factorial or loosely coupled model contains several standard speech models as special cases under appropriate choices for K and for the parameter reduction schemes; it then discusses parameter reduction schemes leading to more general models.

2.1 Special Cases of the Factorial Model

We assume each time series comprises D-dimensional observations for notational simplicity and denote combined observations by $O_t = (o_t^1, \ldots, o_t^K)$. We generalize the presentation in the previous section to allow L underlying Markov chains, where it is not necessary that $L = K$; hidden metastates therefore comprise L hidden variables and are described by L-tuples $I = (i^1, \ldots, i^L)$ and $J = (j^1, \ldots, j^L)$. Each chain has N possible states, again for notational ease. P denotes probability mass functions (pmf), p denotes densities. Using this notation, the parameters to be estimated in the factorial model are $\pi(J)$, $P(J|I)$ and $p(O_t|J)$.

All parameter reduction schemes considered in this paper make three assumptions:

- independence of initial metastate components:

$$\pi(J) = \prod_{l=1}^{L} P(j^l) \tag{3}$$

- conditional independence of metastate components given previous metastate:

$$P(J|I) = \prod_{l=1}^{L} P(j^l|I) \tag{4}$$

- conditional independence of observation components given current metastate:

$$p(O_t|J) = \prod_{k=1}^{K} p(o_t^k|J) \tag{5}$$

Setting $K = L = 1$ in this parameter-reduced model gives the standard *HMM*. Setting $L = 1$ and K to the number of output streams gives the *HTK synchronous multiple stream model* [Young et al., 1995]. Setting $L = K$ plus additional conditional independence assumptions $P(j^k|I) = P(j^k|i^k)$ and $p(o_t^k|J) = p(o_t^k|j^k)$ gives the asynchronous *independent streams model*, which is related to the *multiband* model eg. [Mirghafori, 1999]. Figures 3, 4 and 5 illustrate these models as dynamic Bayesian networks[3]. For notational ease, henceforth $L = K$.

Our real interest is in new parameter reduction schemes giving tractable models that can still capture coupling between the K time series. Many possibilities exist, such as parameter reduction through maximum likelihood (ML) parameter

[3] *Directed Acyclic Graphical Models* (DAGM) or *Bayesian Networks* (BN) are graphical statements of conditional independence relations amongst random variables, see [Jensen, 1996, Lauritzen, 1996, Zweig, 1998].

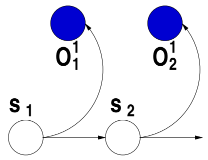

Figure 3: Hidden Markov Model

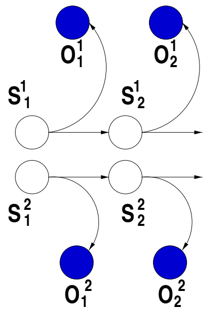

Figure 4: Independent Streams Model

tying [Nock, 2001], but in this paper we adopt the *Mixed-Memory Assumptions* of [Saul and Jordan, 1999]:

- parameterize prior-related probabilities by a convex combination of *cross-prior* matrices:

$$P(j^k) = \sum_{l=1}^{K} \psi^k(l)\pi^{kl}(j^k) \tag{6}$$

- parameterize transition-related conditional probabilities by a convex combination of *cross-transition* matrices:

$$P(j^k|I) = \sum_{l=1}^{K} \psi^k(l)a^{kl}(j^k|i^l) \tag{7}$$

- parameterize observation-related conditional probabilities by a convex combination of *cross-emission* distributions:

$$p(o_t^k|J) = \sum_{l=1}^{K} \phi^k(l)b^{kl}(o_t^k|j^l) \tag{8}$$

Parameters $\pi^{kl}(j^k)$ are K^2 elementary $1 \times N$ cross-stream prior matrices, a total of K^2N prior parameters. Parameters $a^{kl}(j^k|i^l)$ are K^2 elementary $N \times N$ cross-transition matrices, a total of K^2N^2 transition parameters. The $b^{kl}(o_t^k|j^l)$ are K^2N cross-emission output densities; for D-dimensional observations and full covariance gaussians, a total of $K^2ND(1+D)$ observation-related parameters.

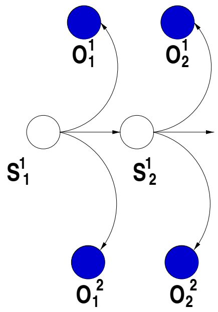

Figure 5: HTK Multiple Streams Model

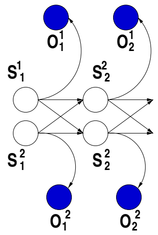

Figure 6: Transition-Only Coupled MMFHMM

Parameters $\psi^k(l)$, $\phi^k(l)$ are mixture weights. They are fixed for a single model, and give a measure of the dependency between different streams, using a total of $2K^2$ parameters. The MMFHMM thus has $\mathcal{O}(K^2(N^2 + ND^2))$ free parameters vs. $\mathcal{O}(N^K(N^K + K^2D^2))$ for the general factorial HMM.

Adoption of the Mixed-Memory Assumptions allows separate evaluation of the effects of making transition-related or observation-related probabilities dependent upon full metastate identity, as well as the case where both are metastate dependent. We use the following terminology for these three cases, illustrated as dynamic Bayesian networks in Figures 6, 7 and 8. An *observation-only coupled MMFHMM* sets ψ to the $K \times K$ identity matrix, ie. only observation distributions can be dependent upon metastates. A *transition-only coupled MMFHMM* sets ϕ to the $K \times K$ identity matrix, ie. only transition distributions can be dependent upon metastates. Finally, a *fully coupled MMFHMM* is the general case of unrestricted ϕ, ψ, where both observation and transition distributions may depend upon metastates.

2.2 Maximum Likelihood MMFHMM Estimation

Maximum likelihood (ML) estimation of the MMFHMM with appropriate choices of observation distribution is possible using an EM algorithm [Dempster et al., 1977]. Equations (7) and (8) are viewed as mixture models, introducing two new types of latent variables x_t^k, y_t^k in addition to variables s_t^k denoting the metastate sequence taken through the model [Saul and Jordan, 1999]. The new variables

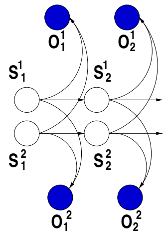

Figure 7: Observation-Only Coupled MMFHMM

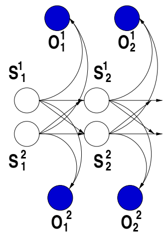

Figure 8: Fully Coupled MMFHMM

encode the identity of the cross-emission distribution and cross-transition matrix (ie. the distributions within each mixture model) used in each stream k at each t (Figure 9).

The EM parameter update equations for the case where $b^{kl}(o_t^k|j^l)$ are modelled by full-covariance Gaussian densities $\mathcal{N}(\mu_j^{kl}, \Sigma_j^{kl})$ are:

$$\hat{\psi}^k(l) = \frac{\sum_t P(x_t^k = l|O)}{\sum_{\nu,t} P(x_t^k = \nu|O)} \tag{9}$$

$$\hat{\phi}^k(l) = \frac{\sum_t P(y_t^k = l|O)}{\sum_{\nu,t} P(y_t^k = \nu|O)} \tag{10}$$

$$\hat{\pi}^{kl}(j^k) = \frac{\sum_t P(x_t^k = l, s_1^k = j^k|O)}{p(x_t^k = l|O)} \tag{11}$$

$$\hat{a}^{kl}(j^k|i^l) = \frac{\sum_t p(x_t^k = l, s_t^k = j^k, s_{t-1}^l = i^l|O)}{\sum_t p(x_t^k = l, s_{t-1}^l = i^l|O)} \tag{12}$$

$$\hat{\mu}_j^{kl} = \frac{\sum_t p(y_t^k = l, s_t^l = j^l|O) o_t^k}{\sum_t p(y_t^k = l, s_t^l = j^l|O)} \tag{13}$$

$$\hat{\Sigma}_j^{kl} = \frac{\sum_t p(y_t^k = l, s_t^l = j^l|O)(o_t^k - \hat{\mu}_j^{kl})(o_t^k - \hat{\mu}_j^{kl})'}{\sum_t p(y_t^k = l, s_t^l = j^l|O)} \tag{14}$$

where $O = O_1, \ldots, O_T$ denotes the current utterance and $S_t = (s_t^1, \ldots, s_t^K)$ denotes the metastate at time t. Summations over t run from 1 to T, except in equation 12 which runs from 2 to T; summations over ν run from 1 to K. $\hat{\theta}$ denotes an updated parameter θ. See [Nock and Young, 2000] for the derivation

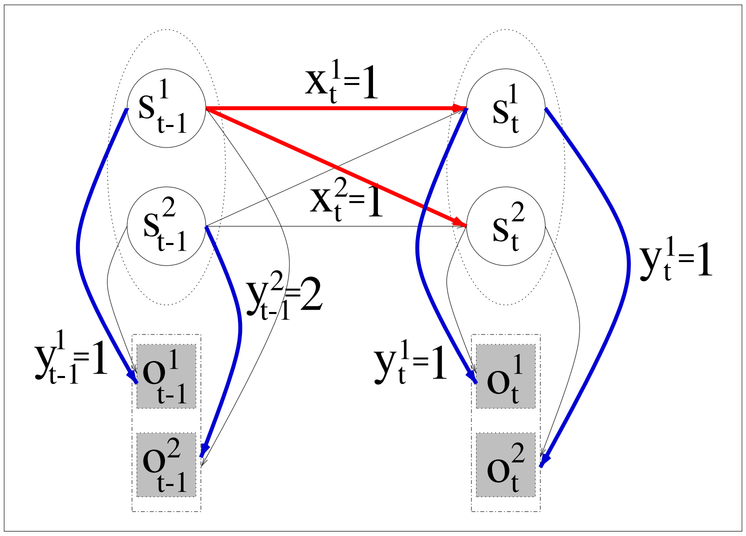

Figure 9: Bold lines illustrate information specified by hidden variables y_t^k and x_t^k.

and procedure for calculating the necessary posterior probabilities.

2.3 Approximations for Model Estimation and Decoding

Likelihood calculations and EM estimation require forward and backward probabilities in the metastate-space of size N^K, which could become intractable as K or N increase. Phonological or articulatory feature sets typically involve $K > 5$; allowing asynchrony within word or larger modelling units increases the required N. Thus more efficient, perhaps approximate, decoding and estimation schemes may be required. Two alternative approaches are suggested[4].

2.3.1 Chain Viterbi Algorithm

In decoding it is typically assumed that the total likelihood of data is well approximated by the likelihood calculated along one particular state sequence: the most likely metastate sequence $\mathbf{S}^*$ given the data, which is obtained using the Viterbi algorithm [Viterbi, 1967] but the Viterbi algorithm also operates in the metastate space of size N^K. Saul and Jordan [Saul and Jordan, 1999] propose a more efficient *Chain Viterbi* scheme for approximating $\mathbf{S}^*$ when the K time series are *assumed* weakly coupled. Starting from some initial metastate sequence, the algorithm iterates through each chain k in turn, finding the optimal sequence of hidden states through chain k given fixed values for the hidden

[4] Note also exact likelihood calculations can made more efficient for observation-only coupled MMFHMMs eg. [Ghahramani and Jordan, 1997].

states of the other chains[5]. The state space is thus reduced to size N when doing the optimizations for chain k. Iteration through all K chains continues until convergence, which is not necessarily to $\mathbf{S}^*$ (see [Nock and Young, 2000] for a counter-example). For a more formal presentation, see [Nock, 2001]. *Assuming* the resulting sequence is similar to the Viterbi sequence $\mathbf{S}^*$ leads to an approximate, Viterbi-like estimation scheme: the associated parameter update equations are obtained by conditioning posterior probabilities in equations (9)-(14) on $\mathbf{S}^*$ as well as observations $\mathbf{O}$.

2.3.2 Mean-Field Variational Approximation

Variational methods exploit a lower bound on the data likelihood for approximating model likelihoods and in model estimation. Such methods are currently popular in the graphical models community, where ML estimation using the EM algorithm is often intractable[6]. This section outlines the basic arguments behind variational approximations specifically for the observation-only coupled MMFHMM; [Jordan et al., 1998] is a more general presentation.

For the observation-only coupled MMFHMM with parameters λ, which for an utterance of length T has hidden variables $\mathbf{Y} = \mathbf{Y}_1, \ldots, \mathbf{Y}_T$ and $\mathbf{S} = \mathbf{S}_1, \ldots, \mathbf{S}_T$, where $\mathbf{Y}_t = (y_t^1, \ldots, y_t^K)$ and $\mathbf{S}_t = (s_t^1, \ldots, s_t^K)$, the variational lower bound is:

$$\begin{aligned}
\mathcal{L}(\lambda) = \ln p(\mathbf{O}|\lambda) &= \ln\{\sum_{\mathbf{S},\mathbf{Y}} p(\mathbf{O}, \mathbf{S}, \mathbf{Y}|\lambda)\} \\
&= \ln\{\sum_{\mathbf{S},\mathbf{Y}} Q(\mathbf{S}, \mathbf{Y}|\Psi) \frac{p(\mathbf{O}, \mathbf{S}, \mathbf{Y}|\lambda)}{Q(\mathbf{S}, \mathbf{Y}|\Psi)}\} \\
&\geq \sum_{\mathbf{S},\mathbf{Y}} Q(\mathbf{S}, \mathbf{Y}|\Psi) \ln \frac{p(\mathbf{O}, \mathbf{S}, \mathbf{Y}|\lambda)}{Q(\mathbf{S}, \mathbf{Y}|\Psi)} = \mathcal{L}_Q(\Psi, \lambda)
\end{aligned}$$

where $\mathcal{L}(\lambda)$ denotes the likelihood function, $Q(\mathbf{S}, \mathbf{Y}|\Psi)$ is a distribution over the hidden variables with parameters Ψ, and $\mathcal{L}_Q(\Psi, \lambda)$ denotes the lower bound of interest. The inequality in the third line follows by Jensen's inequality eg. [Ross, 1997]. Note that $\mathcal{L}(\lambda)$ exceeds $\mathcal{L}_Q(\Psi, \lambda)$ by exactly the Kullback-Leibler (KL) divergence $KL[Q(\mathbf{S}, \mathbf{Y}|\Psi)||p(\mathbf{S}, \mathbf{Y}|\mathbf{O}, \lambda)]$ between the distributions, which is non-negative [Cover and Thomas, 1991]. This lower bound on likelihood may be tightened for each observation sequence $\mathbf{O}$ by adjusting the *variational parameters* Ψ of the *variational distribution* Q to minimize the KL divergence. The lower bound can also be used in estimation: iterative coordinate ascent in the lower bound, first maximizing with respect to the parameters λ of model p and then with respect to parameters Ψ of variational distribution Q, is guaranteed to increase the lower bound $\mathcal{L}_Q(\Psi, \lambda)$

[5] Softer versions of this iterative scheme are possible, fixing subsets of chains and optimizing over the remainder.

[6] Inference in the general case is NP-hard [Cooper, 1990].

on the likelihood at each step, although not necessarily the likelihood $\mathcal{L}(\lambda)$. Convergence of this procedure can be assessed by monitoring changes in the lower bound. Where $Q(\mathbf{S},\mathbf{Y}|\Psi)$ encompasses all distributions over the hidden variables, this learning procedure is equivalent to the standard EM algorithm [Dempster et al., 1977, Neal and Hinton, 1998].

The variational lower bound is useful when likelihood calculations or EM estimation are intractable. Family $Q(\mathbf{S},\mathbf{Y}|\Psi)$ is chosen to allow more tractable inference than p. For example, when working with graphical models Q often makes independence assumptions additional to those made by the family p. The variational approximation Q used here is the simplest, completely factorized approximation in which all hidden variables are assumed independent given the observations[7]. This *Mean-Field* approximation can be written:

$$Q(\mathbf{S},\mathbf{Y}|\Psi) \quad = \quad \prod_{t=1}^{T}\{\prod_{k=1}^{K} Q_t^{Sk}(s_t^k|\Psi_t^{Sk})Q_t^{Yk}(y_t^k|\Psi_t^{Yk})\}$$

where

- $Q_t^{Sk}(s_t^k|\Psi_t^{Sk})$ denotes a pmf with parameters $\Psi_t^{Sk} = \{\Psi_{tj^k}^{Sk}|j^k \in \Theta_k\}$; $\Psi_{tj^k}^{Sk}$ denotes the probability of state j^k;
- $Q_t^{Yk}(y_t^k|\Psi_t^{Yk})$ denotes a pmf with parameters $\Psi_t^{Yk} = \{\Psi_{tl}^{Yk}|1 \leq l \leq K\}$; Ψ_{tl}^{Yk} denotes the probability of l-th mixture component.

To simplify maintenance of positivity, ensure appropriate normalization and guarantee that no hidden event has probability zero, a softmax form is assumed for variational pmf's:

$$Q_t^{Sk}(s_t^k = j^k|\Psi_t^{Sk}) \stackrel{\text{def}}{=} \frac{\exp \Psi_{tj^k}^{Sk}}{\sum_{i^k \in \Theta_k} \exp \Psi_{ti^k}^{Sk}}$$

and for each $1 \leq l \leq K$

$$Q_t^{Yk}(y_t^k = l|\Psi_t^{Yk}) \stackrel{\text{def}}{=} \frac{\exp \Psi_{tl}^{Yk}}{\sum_{\nu=1}^{K} \exp \Psi_{t\nu}^{Yk}}$$

Lower bound maximization with respect to parameters Ψ can be implemented using basic gradient-descent, although solution via fixed point iteration often gives faster convergence [Attias, 2000]; maximization with respect to parameters λ is similar to standard ML estimation for multivariate Gaussian distributions. See [Nock, 2001] for details.

[7] As observed by Hagai Attias and an anonymous reviewer, the Chain Viterbi procedure for likelihood approximation is also a variational approximation in which Q puts all its probability mass on a single metastate sequence.

3 Preliminary Evaluation Using ISOLET

These preliminary experiments use the OGI ISOLET database [Cole et al., 1990], which comprises wideband recordings of isolated utterances of single letters of the alphabet. Whilst far from the conversational speech motivating the research, ISOLET is adequate for an initial feasibility study of novel models and algorithms without the additional complications introduced by continuous speech tasks. We use *Isolet1-4* (6240 utterances) to train and the speaker-disjoint *Isolet5* (1560 utterances) to test. Our baseline HMM performance using a 39-dimensional observation vector of *full-band* cepstra (including 0th) with delta and acceleration coefficients is between 96.2% (3 state HMM) and 96.6% (10 state HMM). The experimental task investigated is that of modelling cepstra derived from frequency subbands (eg. [Mirghafori, 1999, Tomlinson et al., 1997, McMahon et al., 1998]), rather than a more speculative articulatory or phonological representation. Some evidence of asynchrony between different frequency bands exists [Mirghafori, 1999]; however, there is likely to be more asynchrony in articulatory or phonological representations, where the advantages of loosely coupled models may be more evident.

3.1 Procedure for Subband Cepstra Extraction

25ms windows of speech are Fourier-transformed and filtered through a bank of 20 overlapping, equally mel-spaced, filters giving a vector of log spectral energies $E = [e_1, \ldots, e_{20}]$. A choice of V frequency subbands subdivides E into V subvectors E_v. A DCT D_v is applied to each E_v to yield a vector of cepstra $C_v = D_v E_v$ for subband v. Decreasing D_v row dimensionality effects cepstral truncation, reducing the dimensionality of C_v from that of E_v: a V-tuple $(\#_1, \ldots, \#_V)$ denotes the truncation scheme, where $\#_v$ indicates retention of cepstra $0, \ldots, \#_v - 1$ in subband v. Finally, observations for the v-th subband stream (o_t^v in our earlier notation) are formed by appending the appropriate delta and acceleration coefficients to C_v.

Our experiments use cepstra from two and from three frequency subbands. Observations for the two-stream experiments comprise cepstra from two subbands 0-2 and 2-8kHz, with cepstral truncation (7,6), yielding a 39-d combined observation vector $\mathbf{O}_t$. Observations for three-stream experiments comprise cepstra from three subbands 0-0.9, 0.8-2.7, 2.7-8kHz, with cepstral truncation (5,4,4), again yielding a 39-d combined observation vector $\mathbf{O}_t$.

3.2 Comparison: Model Structures

This subsection compares classification performance of loosely coupled models with more conventional speech models. Classification uses an ML decision rule ie. utterance $\mathbf{O}$ is allocated to class W^* where $W^* = \arg\max_W p(\mathbf{O}|W)$ (equivalent to the Bayes' minimum error decision rule for this task since class

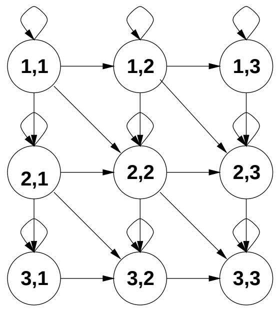

Figure 10: Left-to-right Metastate Space Topology

priors are equal in the ISOLET test set). Performance is compared against two baselines. The first is a standard HMM-based system trained on the combined observations. However, since HMM and MMFHMM-based classifiers are quite different in their use of free parameters, additional comparisons are made against *HTK multiple stream* and *independent streams* models. These "conventional" models are configured to be comparable with the loosely coupled models not only in terms of the total number of parameters (as for the HMM baseline) but also in their usage of parameters. These three types of model differ in the degree of asynchrony allowed between streams. To reflect this, results are ordered in terms of increasing potential asynchrony: the synchronous *HTK multiple stream* model are followed by the loosely coupled models and then the completely asynchronous *independent streams* model. Note that none of the HTK multiple stream, loosely coupled or independent streams results utilize any form of stream-weighting.

3.2.1 Experimental Setup

Experiments will investigate the effect of introducing coupling between streams by making transition probabilities dependent upon metastates. This introduces a difficulty, since approximation (7) has a limitation when applied to speech modelling. Speech HMMs are typically constrained *a priori* to have a left-to-right transition structure, but this is achievable under (7) only when ψ is the identity matrix I (ie. transitions in chain k depend only on the previous state in chain k). In this study, when we investigate coupling through transition probabilities in mixed-memory models (achieved by setting $\psi \neq I$), we use a^{kl} matrices which are individually left-to-right to limit backwards transitions in the metastate space. Figure 10 shows a left-to-right metastate space topology; Figure 11 provides a contrast by showing the connections *removed* from an ergodic metastate space topology when the underlying cross-transition matrices are left-to-right in a two-stream, three-state per stream, model. This illustrates that whilst the left-right constraint is not fully enforced, many backwards transitions are prevented.

MMFHMMs with metastate dependent observation and/or transition proba-

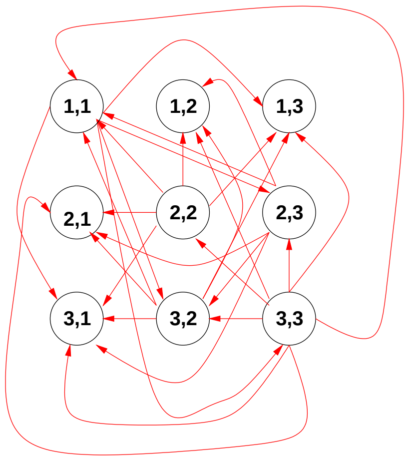

Figure 11: Restricted Metastatespace Transition Topology

Model (# states)	# Parameters	%C
HMM (3)	4686	96.3
HMM (6)	9327	96.1
HMM (8)	12496	96.4
HMM (10)	15620	96.7

Table 1: HMM Baseline (2 subbands)

bilities that model K observation streams use K underlying chains. Gaussians are full covariance, initialized using the global mean and covariance of the training set. Models using cross-emission or cross-transition dependencies are constructed incrementally: first, an HMM is trained for each stream independently; then, cross-stream dependencies are introduced gradually, with two training iterations between the addition of one cross-dependency per stream and/or per chain. Training termination uses an absolute threshold on the gain in likelihood.

3.2.2 Experimental Results

Table 1 gives baseline percentage correct ($\%C$) performance of standard HMMs for modelling combined observation vectors formed by concatenating the two subband cepstra streams.

Table 2 gives performance of MMFHMMs with coupling through transition probabilities only, ie. transition probabilities depend upon metastates, again for two observation streams. The MMFHMM and more conventional models in each block of the table are ordered using allowable asynchrony between streams: the synchronous *HTK multiple stream* model precedes the MMFHMM with metastate dependent transition probabilities, which precedes the asynchronous

Model (# states per chain)	# Parameters	%C
HTK multiple stream (3)	2418	94.2
MMFHMM, transition probability metastate dependence (3)	2440	94.1
independent streams (3)	2424	93.9
HTK multiple stream (6)	4836	94.9
MMFHMM, transition probability metastate dependence (6)	4876	95.0
independent streams (6)	4848	94.8
HTK multiple stream (8)	6448	95.4
MMFHMM, transition probability metastate dependence (8)	6500	95.3
independent streams (8)	6464	95.8

Table 2: Results: Transition-Only Coupled Models (2 streams)

# states per chain	HTK Multiple Stream (# states)	Independent Stream (# states per chain)	HMM (3-state)	HMM (6-state)
3	NO (3)	NO (3)	$p = 1.2 \times 10^{-4}$	$p = 1.5 \times 10^{-3}$
6	NO (6)	NO (6)	NO	NO
8	NO (8)	NO (8)	NO	NO

Table 3: Significance Test Outcomes for 2-Stream, Transition-Only Coupled Model Results ($\alpha = 0.01$ level); p-values specified where results differ significantly

Model (# states per chain)	# Parameters	%C
HTK multiple stream (3)	4842	94.6
MMFHMM, output + transition probability metastate dependence (3)	4856	94.7
MMFHMM, output probability metastate dependence (3)	4840	94.9
independent streams (3)	4848	94.0
HTK multiple stream (6)	9684	96.2
MMFHMM, output + transition probability metastate dependence (6)	9704	95.8
MMFHMM, output probability metastate dependence (6)	9676	96.7
independent streams (6)	9696	95.3
HTK multiple stream (8)	12912	96.2
MMFHMM, output + transition probability metastate dependence (8)	12936	96.2
MMFHMM, output probability metastate dependence (8)	12900	96.0
independent streams (8)	12928	96.3

Table 4: Results: Observation-Only and Fully Coupled Models (2 streams)

Model # states per chain	HTK Multiple Stream (# states)	Independent Stream (# states per chain)	observation-only coupled MMFHMM (# states per chain)	HMM (# states)
3	NO (3)	NO (3)	NO (3)	$p = 0.03$ (3)
6	NO (6)	NO (6)	NO (6)	NO (6)
8	NO (8)	NO (8)	NO (8)	NO (8)

Table 5: Significance Test Outcomes for 2-Stream, Fully Coupled Model Results ($\alpha = 0.01$ level); p-values specified where results differ significantly

# states per chain	2-Stream Observation-Only Coupled Models	3-Stream Observation-Only Coupled Models
3	19%	34%
6	35%	51%
8	38%	58%

Table 6: % "Asynchronous" Metastates in Training Set Viterbi Metastate Sequences

independent streams model. Table 3 analyses significance of performance differences for models with comparable numbers of parameters using the McNemar test [Gillick and Cox, 1989].

Table 4 gives performance of MMFHMMs with coupling through observation probabilities only, where observation probabilities depend upon metastates, and then for systems coupled through both observation and transition probabilities, again for two observation streams. Results are again ordered by allowable asynchrony. Each state in *HTK multiple stream* and *independent streams* models uses a two Gaussian mixture to model the data from a single stream. The number of observation-related parameters in these systems is thus comparable with the MMFHMMs with metastate dependent observation probabilities, which use a single Gaussian for each $b^{kl}(o_t^k|i^l)$ distribution. Table 5 analyses significance of performance differences for models with comparable numbers of parameters using the McNemar test [Gillick and Cox, 1989].

The overall results suggest that in most cases the performance of the various models does not differ significantly on the task of frequency subband modelling. Similar trends were seen when repeating experiments with three cepstral subbands.

Further analysis examined whether the potential asynchrony between state chains is utilized. A Viterbi decoding of each training utterance under the *correct* observation-only coupled model gives the optimal metastate sequence for that utterance; the resulting metastate sequences were examined to determine the percentage of "asynchronous" metastates used (ie. for a two-stream system, metastates (i, j) where $i \neq j$; for a three-stream system, metastates (i, j, k) where it is not the case that $i = j = k$). Table 6 shows asynchronous metastates are indeed used.

3.3 Comparison: Exact and Approximate Decoding Algorithms

This subsection considers quality of likelihood approximations given by exact and approximate decoding algorithms.

3.3.1 Experimental Setup

Exact and approximate algorithms are used to decode a fixed set of observation-only coupled models originally trained using the EM algorithm.

Chain Viterbi Initialization: two procedures were investigated. The first used a uniform segmentation of stream k observations against states in chain k; the second used the segmentation obtained by doing a Viterbi decoding of stream k observations using the chain k parameters only (for each k). Preliminary experiments found the algorithm insensitive to initialization; results below use per-chain Viterbi initialization.

Mean-Field Initialization and Step-Sizes: two initializations of $Q_t^{Sk}(j)$ distributions were investigated. Initial per-stream state sequences were obtained using the uniform or the per-chain Viterbi decoding schemes as for Chain Viterbi initialization; each $Q_t^{Sk}(j^k)$ distribution was then initialized with a soft version of this distribution, assigning mass 0.8 to the state occupied in the Viterbi or uniform segmentation, and distributing mass equally amongst the remaining states. $Q_t^{Yk}(l)$ distributions were initialized uniformly. Preliminary experiments found per-chain Viterbi initialization gave considerably better results and it is used below. The naive gradient-descent implementation also requires a step-size: a brute force search over a range of values was used and the results below correspond to the stepsize yielding the highest value for the lower bound on test set likelihood (*not* the stepsize giving the best *classification performance*, since this would constitute cheating).

3.3.2 Experimental Results

Figure 12 shows total test set likelihood and the Viterbi approximations for each class *A*-*Z* as calculated for models of three subband cepstral streams which used eight states per chain; these should be compared with the values obtained from the Chain Viterbi approximation and from the Mean-Field Variational lower bound. Similar trends were seen when algorithms were compared using models with different numbers of states and models of two subband cepstral streams. Table 7 shows classification performance when using the approximations with a ML decision rule.

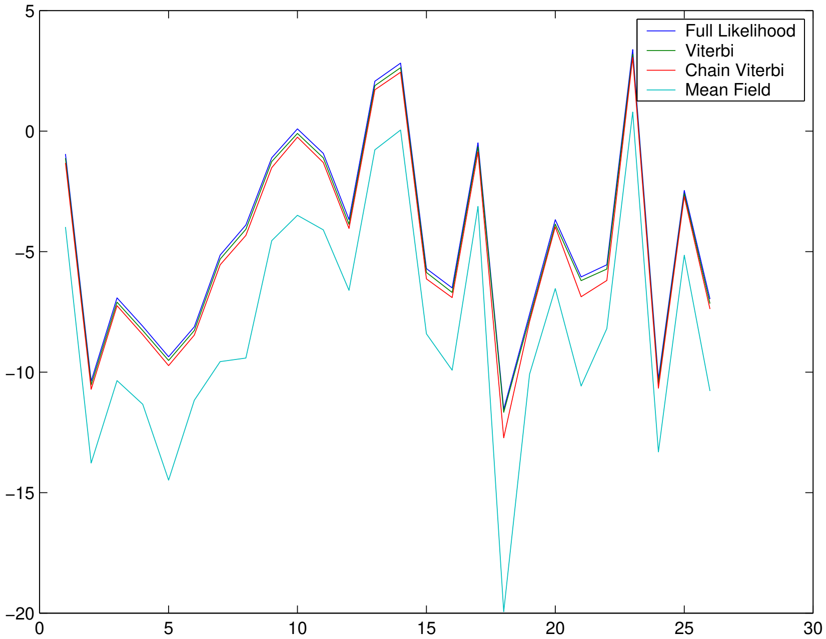

Figure 12: Approximations To Test-Set Likelihoods, 3 streams, 8-state models (X-axis corresponds to classes A-Z)

states per stream	Full Likelihood %C	Viterbi %C	Chain Viterbi %C	Mean Field %C
3	94.9	95.0	96.2	91.7*(p=0)
6	96.4	96.3	95.0	95.2*($p = 9.4 \times 10^{-3}$)
8	96.4	96.3	96.2	95.3*($p = 7.6 \times 10^{-3}$)

Table 7: 3-Stream Results: Decoding Schemes (*=significantly different at $\alpha = 0.01$)

The approaches marked '*' are significantly different from the exact algorithm at level $\alpha = 0.01$ using the McNemar test [Gillick and Cox, 1989]. The graph explains this: the Chain Viterbi likelihood approximation is much closer to the exact likelihoods than the Mean-Field variational lower bound. The Chain Viterbi procedure typically converges within 3-4 iterations and has proven more efficient than the gradient-descent-based implementation of the Mean-Field approach. It is our algorithm of choice for future work.

3.4 Comparison: Exact and Approximate Estimation Algorithms

This subsection compares classification performance of observation-only coupled MMFHMMs trained and tested using *matched* exact or approximate algorithms ie. EM training with full likelihood (FL) classification, Chain Viterbi training with a Chain Viterbi approximation in classification etc. Viterbi Training results are also presented although the computational cost of a Viterbi decoding is of the same order as the forward-backward algorithm. Note that the variational approximation is used to estimate only observation-related parameters $\phi^k(l)$ and $b^{kl}(o_t^k|i^l)$, not transition parameters and is viewed here as a computationally cheap means of integrating K per-stream independent HMMs.

3.4.1 Experimental Setup

Training algorithms stop one iteration after the relative gain in likelihood or variational lower bound falls below 1%.

Chain Viterbi Initialization: an initial metastate sequence was obtained by doing a Viterbi decoding of stream k observations using the chain k parameters only, since this proved a useful initialization in the earlier decoding-only experiments.

Mean-Field Initialization and Step-Sizes: initial per-stream state sequences were obtained using the *per-chain Viterbi decoding* schemes as in Chain Viterbi initialization; each $Q_t^{Sk}(j^k)$ distributions was then initialized with a soft version of this distribution, assigning mass 0.8 to the state occupied in the Viterbi or uniform segmentation, and distributing mass equally amongst the remaining states. $Q_t^{Yk}(l)$ distributions were initialized as a uniform distribution. The gradient-descent stepsize used in training and decoding was fixed to the value that was most effective in the decoding-only experiments.

3.4.2 Experimental Results

Table 8 shows classification performance using approximate algorithms is similar to the exact scheme for the two stream case; cases marked '*' are significantly different from the exact case using the McNemar test at level $\alpha = 0.01$ [Gillick and Cox, 1989]. Similar trends and no significant differences between exact and approximate algorithms were found in the three-stream case. On average, the EM, Viterbi and Chain Viterbi algorithms all take a similar number of iterations to fall below the relative change training termination threshold; the Mean-Field scheme takes fewer. Despite this, our current implementation of the Chain Viterbi scheme has proven more efficient than the gradient-descent-based mean-field approximation and is again our algorithm of choice for future work.

4 CONCLUSIONS AND FUTURE WORK

Speech is produced by a system of loosely coupled articulators. Stochastic models explicitly representing this parallelism may have advantages for ASR, particularly for modelling phonological effects in conversational speech. This

states per stream	Full Likelihood %C	Viterbi %C	Chain Viterbi %C	Mean Field %C
3	94.9	94.9	94.2*($p = 1.9 \times 10^{-2}$)	93.9
6	95.3	95.4	95.2	95.0
8	96.0	95.8	95.9	96.0

Table 8: 2-Stream Results: Matched Training/Decoding Schemes (*=significantly different at $\alpha = 0.01$ level, with corresponding p value in brackets)

paper has considered one possible model family, *loosely coupled HMMs*; it has shown empirically that loosely coupled models can perform as well as similar conventional models on a frequency subband speech modelling task and has identified an approximate estimation scheme making more extensive experimentation tractable. Although ISOLET is clearly a limited task, we conclude loosely coupled models merit further investigation. Our ongoing work examines (1) alternatives to the a-priori mixed-memory assumptions for factorial HMM parameter reduction (2) articulatory or phonological representations of speech, perhaps used in combination with a frequency subband representation of speech to improve robustness to both phonological variability and band-limited noise and (3) conversational speech corpora.

5 Acknowledgements

H.J. Nock is supported by an EPSRC studentship and DARPA Grant No. N660019928924. The authors thank Dr. M.J.F. Gales (Cambridge University Engineering Dept, UK) and the SSLI Lab (University of Washington, USA) for helpful discussions.

References

[Ali et al., 1999] Ali, A., Van der Spiegel, J., Mueller, P., Haentjens, G., and Berman, J. (1999). An acoustic-phonetic feature-based system for automatic phoneme recognition in continuous speech. In *Proceedings of ISCAS*, pages 118–121.

[Attias, 2000] Attias, H. (2000). Personal Communication.

[Bahl et al., 1983] Bahl, L., Jelinek, F., and Mercer, R. (1983). A maximum-likelihood approach to continuous speech recognition. *IEEE Transactions PAMI*, 5(2):179–90.

[Cohen, 1989] Cohen, M. H. (1989). *Phonological Structures for Speech Recognition.* PhD thesis, Computer Science Division, Department of Electrical Engineering and Computer Science, University of California, CA, USA.

[Cole et al., 1990] Cole, R., Muthusamy, Y., and Fanty, M. (1990). The isolet spoken letter database. Technical Report CSE 90-004, OGI.

[Cooper, 1990] Cooper, G. (1990). The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42(2-3):393–405.

[Cover and Thomas, 1991] Cover, T. and Thomas, J. (1991). *Elements Of Information Theory.* John Wiley and Sons.

[Dalsgaard, 1992] Dalsgaard, P. (1992). Phoneme label alignment using acoustic-phonetic features and gaussian probability density functions. *Computer Speech And Language*, 6:303–329.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

[Deng and Erler, 1992] Deng, L. and Erler, K. (1992). Structural design of a hidden markov model based speech recognizer using multi-valued phonetic features: Comparison with segmental speech units. *Journal of the Acoustical Society of America*, 92(92):3058–3067.

[Eide et al., 1993] Eide, E., Robin Rohlicek, J., Gish, H., and Mitter, S. (1993). A linguistic feature representation of the speech waveform. In *Proceedings of ICASSP*, pages 483–486.

[Fosler-Lussier, 1999] Fosler-Lussier, J. (1999). *Dynamic Pronunciation Models For Automatic Speech Recognition.* PhD thesis, ICSI, UC Berkeley, CA, USA.

[Frankel et al., 2000] Frankel, J., Richmond, K., King, S., and Taylor, P. (2000). An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces. In *Proceedings of ICSLP*, pages IV:254–257.

[Ghahramani and Jordan, 1997] Ghahramani, Z. and Jordan, M. (1997). Factorial hidden markov models. *Machine Learning*, 29:245–273.

[Gillick and Cox, 1989] Gillick, L. and Cox, S. (1989). Some statistical issues in the comparison of speech recognition algorithms. In *Proceedings of ICASSP*, pages 532–535.

[Goldsmith, 1999] Goldsmith, J., editor (1999). *Phonological Theory: The Essential Readings.* Blackwell.

[Graff, 1997] Graff, D. (1997). The 1996 broadcast news speech and language-model corpus. In *Proceedings of DARPA Speech Recognition Workshop.*

[Greenberg et al., 1996] Greenberg, S., Hollenback, J., and Ellis, D. (1996). Insights into spoken language gleaned from phonetic transcription of the switchboard corpus. In *Proceedings of ICSLP*, pages S24–27.

[Hub4-Results, 1999] Hub4-Results (1999). Nist english broadcast news transcription (hub4) benchmark test results. ftp://jaguar.ncsl.nist.gov/csr99/bn98en_official_scores_990112/readme.htm.

[Huckvale, 1994] Huckvale, M. (1994). Word recognition from tiered phonological models. In *Proceedings of Institute of Acoustics Conference on Speech and Hearing*, volume 16(5), pages 163–170.

[Jensen, 1996] Jensen, F. (1996). *An Introduction To Bayesian Networks.* Springer.

[Jordan et al., 1998] Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1998). An introduction to variational methods for graphical models. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 105–161. Kluwer Academic Press Press.

[Keating, 1997] Keating, P. (1997). Word-level phonetic variation in large speech corpora. In Pompino-Marschal, B., editor, *ZAS Working Papers In Linguistics.* Available at `http://www.hummet.ucla.edu/humnet/linguistics/people/keating/berlin1.pdf`.

[King and Taylor, 2000] King, S. and Taylor, P. (2000). Detection of phonological features in continuous speech using neural networks. *Computer Speech And Language*, 14(4):333–353.

[Kirchhoff, 1996] Kirchhoff, K. (1996). Syllable-level desynchronization of phonetic features for speech recognition. In *Proceedings of ICSLP*, pages 2274–2276.

[Kirchhoff, 1998] Kirchhoff, K. (1998). Robust speech recognition using articulatory information. Technical Report 98-036, ICSI, UC Berkeley.

[Lauritzen, 1996] Lauritzen, S. (1996). *Graphical Models.* Oxford: Clarendon.

[Liberman, 1998] Liberman, M. (1998). Proposal for research topic atthe johns hopkins university (center for language and speech processing) summer research workshop: Acoustic-phonetic feature detectors. Unpublished.

[Logan and Moreno, 1998] Logan, B. and Moreno, P. (1998). Factorial hmms for acoustic modeling. In *Proceedings of ICASSP*, pages 813–816.

[McMahon et al., 1998] McMahon, P., McCourt, P., and Vaseghi, S. (1998). Discriminative weighting of multi-resolution sub-band cepstral features for speech recognition. In *Proceedings of ICSLP*, pages 1055–1058.

[Mirghafori, 1999] Mirghafori, N. (1999). *A Multi-Band Approach to Automatic Speech Recognition.* PhD thesis, ICSI, UC Berkeley, CA, USA.

[Moore, 1996] Moore, R. K. (1996). Critique: The potential role of speech production models in automatic speech recognition. *Journal of the Acoustical Society of America*, 99(3):1710–1712.

[Neal and Hinton, 1998] Neal, R. and Hinton, G. (1998). A view of the em algorithm that justifies incremental and other variants. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 355–370. Kluwer Academic Press.

[Niyogi et al., 1999] Niyogi, P., Burges, C., and Ramesh, P. (1999). Distinctive feature detection using support vector machines. In *Proceedings of ICASSP*, pages 425–428.

[Nock, 2001] Nock, H. (Submitted, to be examined August 2001.). *Techniques for Modelling Phonological Processes in Automatic Speech Recognition*. PhD thesis, Cambridge University Engineering Dept, Cambridge, UK.

[Nock and Young, 2000] Nock, H. and Young, S. (2000). Loosely coupled hmms for asr: A preliminary study. Technical Report CUED/F-INFENG/TR386, CUED.

[Ostendorf, 2000] Ostendorf, M. (2000). Incorporating linguistic theories of pronunciation variation into speech recognition models. In *Philosophical Transactions of Royal Society*, volume 358, pages 1325–1338. London, UK.

[Rabiner, 1989] Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of IEEE*, 77(2):257–285.

[Richardson et al., 2000] Richardson, M., Bilmes, J., and Diorio, C. (2000). Hidden-articulator markov models: Performance improvements and robustness to noise. In *Proceedings of ICSLP*, pages III:131–134.

[Riley et al., 1999] Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., and Zavaliagkos, G. (1999). Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication*, 29:209–224.

[Rose et al., 1996] Rose, R., Schroeter, J., and Sondhi, M. (1996). The potential role of speech production models in automatic speech recognition. *Journal of the Acoustical Society of America*, 99(3):1699–1709.

[Ross, 1997] Ross, S. (1997). *A First Course In Probability*. Prentice Hall.

[Russell, 1997] Russell, M. (1997). Progress towards speech models that model speech. In *Proceedings of IEEE Workshop ASRU*.

[Saraclar, 2000] Saraclar, M. (2000). *Pronunciation Modeling For Conversational Speech Recognition*. PhD thesis, The Johns Hopkins University, MD, USA.

[Saraclar et al., 2000] Saraclar, M., Nock, H., and Khudanpur, S. (2000). Pronunciation modeling by sharing gaussian densities across phonetic models. *Computer Speech And Language*, 14(2):137–160.

[Saul and Jordan, 1999] Saul, L. and Jordan, M. (1999). Mixed memory markov models. *Machine Learning*, 37:75–87.

[Saul et al., 2001] Saul, L., Rahim, M., and Allen, J. (2001). A statistical model for robust integration of narrowband cues in speech. *Computer Speech And Language*, 15(2):175–194.

[Stephenson, 1998] Stephenson, T. (1998). Speech recognition using phonetically-featured syllables. Master's thesis, Centre For Cognitive Science, University of Edinburgh.

[Stephenson et al., 2000] Stephenson, T., Bourlard, H., Bengio, S., and Morris, A. (2000). Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory variables. In *Proceedings of ICSLP*, pages II:951–954.

[Stevens, 2000] Stevens, K. (2000). From acoustic cues to segments, features and words. Plenary Lecture, Proceedings of ICSLP.

[Sun et al., 2000] Sun, J., Deng, L., and Jing, X. (2000). Data-driven model construction for continuous speech recognition using overlapping articulatory features. In *Proceedings of ICSLP*, volume 1, pages 437–440.

[Tomlinson et al., 1997] Tomlinson, M., Russell, M., Moore, R., Buckland, A., and Fawley, M. (1997). Modelling asynchrony in speech using elementary single-signal decomposition. In *Proceedings of ICASSP*, pages 1247–1250.

[Viterbi, 1967] Viterbi, A. (1967). Error bounds for convolutional codes and an asymmetrically optimum decoding algorithm. *IEEE Transactions Information Theory*, IT-13:260–267.

[Weintraub et al., 1996] Weintraub, M., Taussig, K., Hunicke-Smith, K., and Snodgrass, A. (1996). Effect of speaking style on lvcsr performance. In *Proceedings of ICSLP*, pages S16–S19 (addendum).

[Woodland et al., 1995] Woodland, P., Leggetter, C., Odell, J., and Young, S. (1995). The development of the 1994 htk large vocabulary speech recognition system. In *Proceedings of ARPA Spoken Language Systems Technology Workshop*, pages 110–115.

[Young et al., 1995] Young, S., Jansen, J., Odell, J., Ollason, D., and Woodland, P. (1995). *The HTK Book (Version 2.0)*. ECRL.

[Zweig, 1998] Zweig, G. (1998). *Speech Recognition With Dynamic Bayesian Networks*. PhD thesis, UC Berkeley, CA, USA.