ELSEVIER

Cognitive Science 26 (2002) 283–301

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Modelling asynchrony in automatic speech recognition using loosely coupled hidden Markov models

H.J. Nock*, S.J. Young

Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, UK

Accepted 22 March 2002

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 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. © 2002 Cognitive Science Society, Inc. All rights reserved.

Keywords: Automatic speech recognition; Pronunciation modelling; Loosely coupled hidden Markov models; Variational approximation

1. Introduction

Most current automatic speech recognition (ASR) systems use a statistical formulation of the ASR problem, 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 $\boldsymbol{O} = \boldsymbol{O}_1, \ldots, \boldsymbol{O}_T$, which represents the acoustic evidence upon which the recogniser makes a decision. The recogniser or *decoder* seeks the valid word sequence $\boldsymbol{W}$

* Corresponding author. Tel.: +44-914-945-3104; fax: +44-914-945-4490.

E-mail addresses: hjn11@eng.cam.ac.uk (H.J. Nock), sjy@eng.cam.ac.uk (S.J. Young).

PII: S0364-0213(02)00071-X

that maximises $p(\boldsymbol{W}|\boldsymbol{O})$ or equivalently maximises $p(\boldsymbol{O}|\boldsymbol{W})p(\boldsymbol{W})$. Probability $p(\boldsymbol{O}|\boldsymbol{W})$ is provided by an *acoustic model* and $p(\boldsymbol{W})$ 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. We assume that the reader is familiar with HMMs. 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*. 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 (e.g., Woodland, Leggetter, Odell, & Young, 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) (NIST HUB4 Results, 1999). Our experiments using a dataset comprising the *same* word-level transcript recorded in different speaking styles show that the difficulties are partly associated with changes in the acoustic realisations of words (as opposed to changes in grammar and vocabulary) (Saraclar, Nock, & Khudanpur, 2000; Weintraub, Taussig, Hunicke-Smith, & Snodgrass, 1996). Many researchers hypothesise the difficulties are specifically related to increased pronunciation change in conversational speech: the increased variability of phoneme realisations and greater phonetic and lexical deletion may not be adequately modelled by current implementations of the beads-on-a-string procedure as discussed in (e.g., Cohen, 1989; Fosler-Lussier, 1999; Greenberg, Hollenback, & Ellis, 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 realisation of sounds and mixture of Gaussian output distributions in HMMs may capture variability in segment realisations, 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 (e.g., 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 (e.g., 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 (e.g., King & Taylor, 2000; Liberman, 1998), although it has 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, e.g., CAN’T /k ae n t/ → [k $\widehat{ae}$ t], where $\widehat{ae}$ indicates nasality from the deleted segment /n/ has coloured the neighbouring vowel, and asynchronous articulation errors causing stop insertions, e.g., WARMTH /w ao m th/ → [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 (e.g., Moore, 1996; Rose, Schroeter, & Sondhi, 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 asynchrony between feature changes (e.g., Huckvale, 1994; Kirchhoff, 1996; King & 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 (e.g., Frankel, Richmond, King, & Taylor, 2000; Huckvale, 1994; Kirchhoff, 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 & Erler, 1992; Frankel et al., 2000; Kirchhoff, 1998; Richardson, Bilmes, & Diorio, 2000; Stephenson, 1998; Zweig, 1998) represent recent efforts. The contribution of this paper is to investigate and evaluate 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 observation 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. 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.

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. The opposite approach is to combine the K HMMs into a joint model: we can form a combined or *factorial* HMM in which (i) the hidden state space (or *metastate* space) is the Cartesian product of the N-state spaces of the K individual HMMs (see Fig. 1), and (ii) the observation O_t is a tuple of the individual stream observations at time t, i.e., $O_t = (o_t^1, \ldots, o_t^K)$. We assume for notational simplicity that each time series comprises D-dimensional observations.

The combined model just described is equivalent to a standard HMM in which the N^K states and KD-dimensional 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 to reduce the number of parameters and sometimes as the basis for efficient, approximate training and decoding algorithms (e.g., Ghahramani & Jordan, 1997; Saul & Jordan, 1999; Deng & Erler, 1992; Sun, Deng, & Jing, 2000; Logan & 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 choice of parameter reduction scheme; it then discusses parameter reduction schemes leading to more general models.

2.1. Special cases of the factorial model

We generalise 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

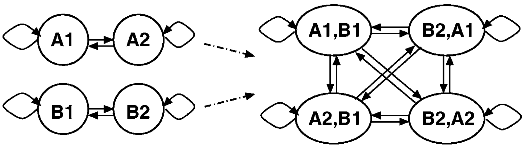

Fig. 1. Metastate space from combined ergodic HMMs A and B.

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 (pmfs) over discrete state spaces, p denotes densities over continuous observation spaces. Using this notation, the parameters to be estimated in the factorial model are $P(J|I)$ and $p(O_t|J)$. (For brevity we omit the parameters specifying the initial distribution $P(J)$.)

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

- conditional independence of metastate components given previous metastate:

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

- conditional independence of observation components given current metastate:

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

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, Jansen, Odell, Ollason, & Woodland, 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 (e.g., Mirghafori, 1999). Fig. 2(a–c) 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

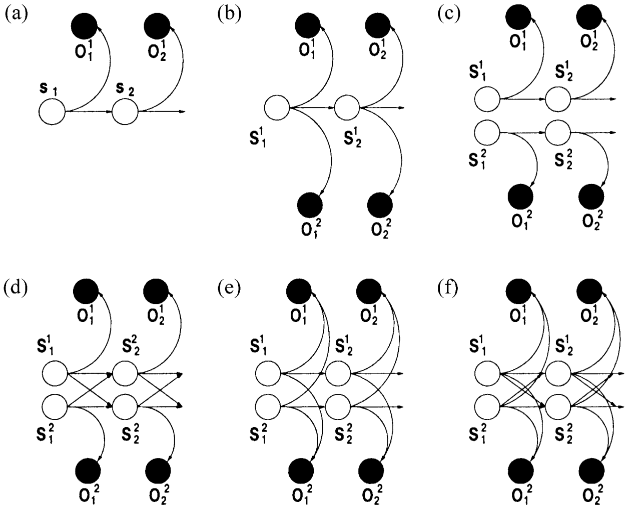

Fig. 2. (a) Hidden Markov model; (b) HTK synchronous multiple stream model; (c) independent streams model; (d) transition-only coupled MMFHMM; (e) observation-only coupled MMFHMM; (f) fully coupled MMFHMM.

reduction through maximum likelihood (ML) parameter tying (Nock, 2001), but in this paper we adopt the *Mixed-Memory Assumptions* of Saul and Jordan (1999), in which stream k's observation o_t^k or next state j^k is conditioned not on the current metastate of *all* streams (Fig. 2 b), nor necessarily just on stream k's own current state (Fig. 2 c), but on the current state of just one randomly chosen stream ℓ (Fig. 2 f):

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

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

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

$$p(o_t^k|J) = \sum_{l=1}^{K} \phi^k(l) b^{kl}(o_t^k|j^l). \tag{2}$$

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. Parameters $\psi^k(l)$, $\phi^k(l)$ are mixture weights that indicate how often stream k is conditioned on stream ℓ. 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 $\mathcal{O}(K^2(N^2+ND^2))$ parameters, versus $\mathcal{O}(N^K(N^K+K^2D^2))$ for the general factorial HMM.

Adoption of the Mixed-Memory Assumptions allows separate evaluation of the effects of making transition- 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 Fig. 2 (d–f). An *observation-only coupled MMFHMM* sets ψ to the $K \times K$ identity matrix, i.e., only observation distributions can be dependent upon metastates. A *transition-only coupled MMFHMM* sets ϕ to the $K \times K$ identity matrix, i.e., 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

ML estimation of the MMFHMM with appropriate choices of observation distribution is possible using an EM algorithm (Dempster, Laird, & Rubin, 1977) appropriate for the Mixed-Memory Assumptions. In addition to variables s_t^k encoding the metastate sequence taken through the model, the algorithm must reconstruct two new types of latent variables x_t^k, y_t^k (Saul & Jordan, 1999). The new variables encode the identity of the cross-emission distribution and cross-transition matrix (i.e., the stream ℓ that stream k depended on) used in each stream k at each t (Fig. 3).

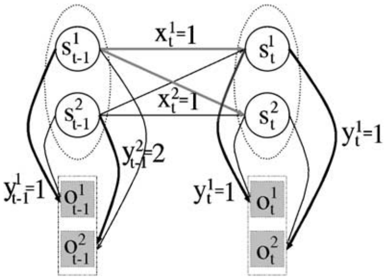

Fig. 3. Bold lines illustrate information specified by hidden variables y_t^k and x_t^k.

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 as follows:

$$\hat{\psi}^k(l) = \frac{\sum_t P(x_t^k = l|O)}{\sum_{\nu,t} P(x_t^k = \nu|O)} \tag{3}$$

$$\hat{\phi}^k(l) = \frac{\sum_t P(y_t^k = l|O)}{\sum_{\nu,t} P(y_t^k = \nu|O)} \tag{4}$$

$$\hat{a}^{kl}(j^k|i^l) = \frac{\sum_t p(x_t^k = l, s_t^k = j^k, s_{t-1}^l = i^l|O)}{\sum_t p(x_t^k = l, s_{t-1}^l = i^l|O)} \tag{5}$$

$$\hat{\mu}_j^{kl} = \frac{\sum_t p(y_t^k = l, s_t^l = j^l|O)o_t^k}{\sum_t p(y_t^k = l, s_t^l = j^l|O)} \tag{6}$$

$$\hat{\Sigma}_j^{kl} = \frac{\sum_t p(y_t^k = l, s_t^l = j^l|O)(o_t^k - \hat{\mu}_j^{kl})(o_t^k - \hat{\mu}_j^{kl})'}{\sum_t p(y_t^k = l, s_t^l = j^l|O)} \tag{7}$$

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 Eq. (5) which runs from 2 to T; summations over ν run from 1 to K. $\hat{\theta}$ denotes an updated parameter θ. See Nock (2001) for the derivation 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 words 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 state (or in this case, metastate) sequence $\boldsymbol{S}^*$ given the data, which is obtained using the Viterbi algorithm (Viterbi, 1967). However, the Viterbi algorithm also operates in the metastate space of size N^K. Saul and Jordan (1999) propose a more efficient *Chain Viterbi* scheme for approximating the metastate sequence $\boldsymbol{S}^*$ 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 states of the other chains.[5] The state space is thus reduced to size N when doing the optimisations for chain k. Iteration through all K chains continues until convergence, which is not necessarily to $\boldsymbol{S}^*$ (see Nock, 2001 for a counter-example). For a more formal presentation, see Nock (2001). *Assuming* the resulting sequence is similar to the Viterbi sequence $\boldsymbol{S}^*$ leads to an approximate, Viterbi-like estimation scheme: the associated parameter update equations are obtained by conditioning posterior probabilities in Eqs. (3)–(7) on $\boldsymbol{S}^*$ as well as observations $\boldsymbol{O}$.

2.3.2. Mean-Field variational approximation

Variational methods exploit a lower bound on the data likelihood for approximating model likelihoods and for 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, Ghahramani, Jaakkola, & Saul, 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 $\boldsymbol{Y} = \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T$ and $\boldsymbol{S} = \boldsymbol{S}_1, \ldots, \boldsymbol{S}_T$, where $\boldsymbol{Y}_t = (y_t^1, \ldots, y_t^K)$ and $\boldsymbol{S}_t = (s_t^1, \ldots, s_t^K)$, the variational lower bound is:

$$\mathcal{L}(\lambda) = \ln p(\boldsymbol{O}|\lambda) = \ln \left\{ \sum_{\boldsymbol{S},\boldsymbol{Y}} p(\boldsymbol{O}, \boldsymbol{S}, \boldsymbol{Y}|\lambda) \right\} = \ln \left\{ \sum_{\boldsymbol{S},\boldsymbol{Y}} Q(\boldsymbol{S}, \boldsymbol{Y}|\Psi) \frac{p(\boldsymbol{O}, \boldsymbol{S}, \boldsymbol{Y}|\lambda)}{Q(\boldsymbol{S}, \boldsymbol{Y}|\Psi)} \right\}$$

$$\geq \sum_{\boldsymbol{S},\boldsymbol{Y}} Q(\boldsymbol{S}, \boldsymbol{Y}|\Psi) \ln \frac{p(\boldsymbol{O}, \boldsymbol{S}, \boldsymbol{Y}|\lambda)}{Q(\boldsymbol{S}, \boldsymbol{Y}|\Psi)} = \mathcal{L}_Q(\Psi, \lambda) \tag{8}$$

where $\mathcal{L}(\lambda)$ denotes the likelihood function, $Q(\boldsymbol{S}, \boldsymbol{Y}|\Psi)$ 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. Note that $\mathcal{L}(\lambda)$ exceeds $\mathcal{L}_Q(\Psi, \lambda)$ by exactly the Kullback–Leibler (KL) divergence $\mathrm{KL}[Q(\boldsymbol{S}, \boldsymbol{Y}|\Psi)||p(\boldsymbol{S}, \boldsymbol{Y}|\boldsymbol{O}, \lambda)]$ between the distributions, which is non-negative. This lower bound on likelihood may be tightened for each observation sequence $\boldsymbol{O}$ by adjusting the *variational parameters* Ψ of the *variational distribution* Q to minimise the KL divergence. The lower bound can also be used in estimation: iterative coordinate ascent in the lower bound, first maximising 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 $\mathcal{L}_Q(\Psi, \lambda)$ 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 $Q(\boldsymbol{S}, \boldsymbol{Y}|\Psi)$ encompasses all distributions over the hidden variables, this learning procedure is equivalent to the standard EM algorithm (Dempster et al., 1977; Neal & Hinton, 1998).

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

$$Q(\boldsymbol{S}, \boldsymbol{Y}|\Psi) = \prod_{t=1}^{T} \left\{ \prod_{k=1}^{K} Q_t^{Sk}(s_t^k|\Psi_t^{Sk}) Q_t^{Yk}(y_t^k|\Psi_t^{Yk}) \right\} \tag{9}$$

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 outcome 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 outcome the l-th mixture component.

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

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

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

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

Lower bound maximisation with respect to parameters Ψ can be implemented using basic gradient descent, although solution via fixed point iteration may give faster convergence (Attias, 2000); maximisation with respect to parameters λ is similar to standard ML estimation for multivariate Gaussian distributions. See Nock (2001) for details.

3. Preliminary evaluation using ISOLET

These preliminary experiments use the OGI ISOLET database (Cole, Muthusamy, & Fanty, 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 (e.g., Mirghafori, 1999; Tomlinson, Russell, Moore, Buckland, & Fawley, 1997), 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 2–8 kHz, with cepstral truncation (7,6), yielding a 39-d combined observation vector $\boldsymbol{O}_t$. Observations for three-stream experiments comprise cepstra from three subbands 0–0.9, 0.8–2.7, 2.7–8 kHz, with cepstral truncation (5,4,4), again yielding a 39-d combined observation vector $\boldsymbol{O}_t$.

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, i.e., utterance $\boldsymbol{O}$ is allocated to the class W that maximises $p(\boldsymbol{O}|W)$ (equivalent to the Bayes minimum error decision rule for this task since class 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 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 is 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 utilise any form of stream-weighting.

3.2.1. Experimental setup

Speech HMMs are typically constrained *a priori* to have a left-to-right transition structure. Fig. 4(b) shows a metastate space topology in which the left-to-right property is enforced for each stream separately: metastate (i, j) can transition only to metastates in $\{i, i+1\} \times \{j, j+1\}$. Unfortunately this intuitive arrangement is not possible in MMFHMMs with coupled transitions ($\psi \neq I$ in approximation (1); see Fig. 2 d and f), since then with some probability a stream’s next state is independent of its current state (depending instead on a *different* stream’s

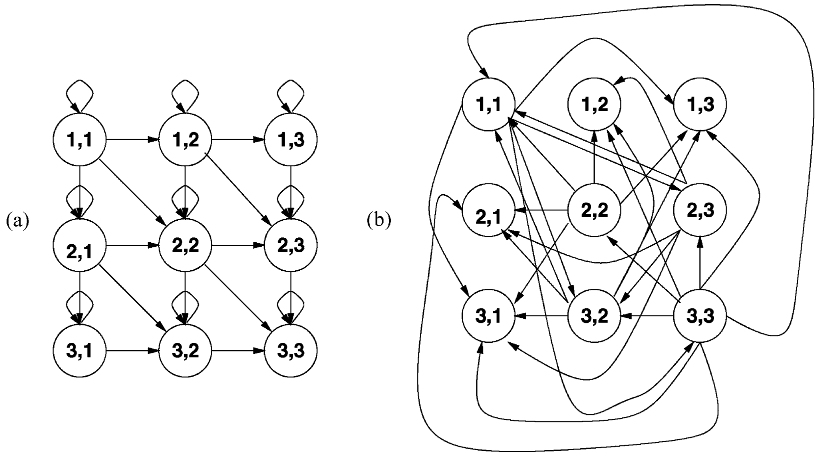

Fig. 4. (a) Left-to-right metastate space topology with no transition coupling; (b) connections *disallowed* when transition coupling is constrained by "left-to-right" (upper bidiagonal) cross-transition matrices.

current state). In this case we do the next best thing, requiring the $N \times N$ cross-transition matrices a^{kl} to have the same upper-bidiagonal form that characterises left-to-rightness in the within-stream transition matrices a^{kk}. This allows metastate (i, j) to transition to metastates in $\{i, i+1, j, j+1\} \times \{i, i+1, j, j+1\}$, e.g., metastate (3,1) can transition to (2,3). Fig. 4 (b) illustrates that, although the left-to-right constraint is then not fully enforced within each stream, many backwards transitions have nonetheless been prevented.

Gaussian emission densities are full covariance, initialised 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 *HMM*s for modelling combined observation vectors formed by pairing the two subband cepstra streams at each time step.

Table 1
HMM baseline (two subbands)

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 2
Results: transition-only coupled models (two streams)

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 gives performance of MMFHMMs with coupling through transition probabilities only, i.e., 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 *independent streams* model. Table 3 analyses significance of performance differences for models with comparable numbers of parameters using the McNemar test (Gillick & 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 is analogous to Table 3.

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.

Table 3
Significance of differences among two-stream, transition-only coupled models ($\alpha = 0.01$)[a]

# States per chain	HTK multiple stream (# states)	Independent stream (# states per chain)	HMM (three-state)	HMM (six-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

[a] Each row compares an MMFHMM model from Table 2 with four other models from Tables 1 and 2. The p-values are specified where results differ significantly.

Table 4
Results: observation-only and fully coupled models (two streams)

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 5
Significance of differences among two-stream, fully coupled models ($\alpha = 0.01$)

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)

Further analysis examined whether the potential asynchrony between state chains is utilised. 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 (i.e., 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 that asynchronous metastates are indeed used.

3.3. *Comparison: exact and approximate decoding algorithms*

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

Table 6
Percentage of "asynchronous" metastates in training set Viterbi metastate sequences

# States per chain	Two-stream observation-only coupled models (%)	Three-stream observation-only coupled models (%)
3	19	34
6	35	51
8	38	58

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.

3.3.1.1. Chain Viterbi initialisation. 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 initialisation; results below use the per-chain Viterbi initialisation.

3.3.1.2. Mean-Field initialisation and step-sizes. Two initialisations 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 in the previous paragraph; each $Q_t^{Sk}(j^k)$ distribution was then initialised with a soft version of this segmentation, 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 initialised uniformly. Preliminary experiments found per-chain Viterbi initialisation gave considerably better results and it is used below. The naive gradient descent implementation also requires a stepsize: 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

Fig. 5 shows total test set likelihood and the Viterbi approximation for each class A through Z as calculated for models of three subband cepstral streams with 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.

Only the Mean-Field approximations were shown significantly worse than the exact algorithm (again using the McNemar test, $\alpha = 0.01$). The graph illustrates why: 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.

Table 7
Three-stream results: decoding schemes

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

[a] Significantly different from full likelihood at $\alpha = 0.01$.

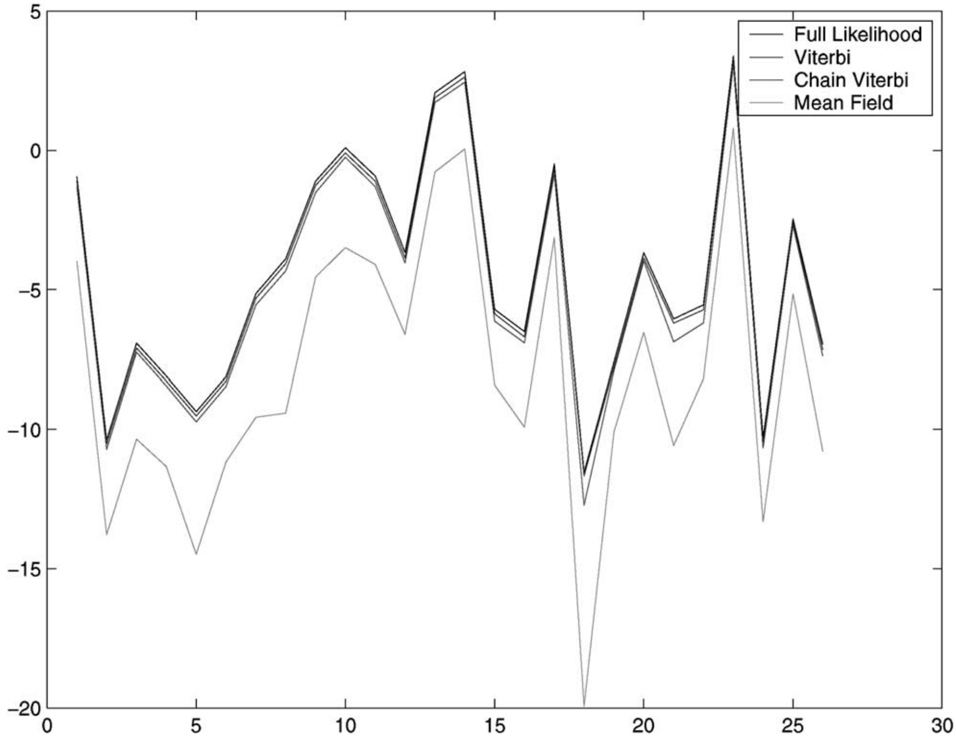

Fig. 5. Approximations to test-set likelihoods, three streams, eight-state models (X-axis corresponds to classes A–Z).

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, i.e., EM training with full likelihood (FL) classification, Chain Viterbi training with a Chain Viterbi approximation in classification and so on. Viterbi Training results are also presented for completeness, although the algorithmic cost 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 $b^{kl}(o_t^k|i^l)$, but not the transition parameters; it is viewed by the authors as a computationally cheap means of integrating K independent per-stream HMMs.

3.4.1. Experimental setup

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

3.4.1.1. Chain Viterbi initialisation. 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 initialisation in the earlier decoding-only experiments.

Table 8
Two-stream results: matched training/decoding schemes

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

[a] Significantly different from full likelihood at $\alpha = 0.01$.

3.4.1.2. Mean-Field initialisation and step-sizes. Initial per-stream state sequences were obtained using the *per-chain Viterbi decoding* schemes as in Chain Viterbi initialisation; each $Q_t^{Sk}(j^k)$ distribution was then initialised with a soft version of this segmentation, 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 initialised to the 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 that classification performance using approximate algorithms is similar to the exact scheme for the two-stream case. No significant differences between the 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 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.

The results show that loosely coupled models merit further investigation. However, ISOLET is an isolated word classification task involving limited variability. Applying these techniques to conversational ASR poses further research questions. First, the acoustic preprocessor should extract observation streams corresponding not to different frequency subbands (used here for convenience while investigating practical issues), but rather to articulator traces or other phonologically motivated feature streams (Frankel et al., 2000; Kirchhoff, 1999). Second, the inability to use a left-to-right metastate space topology (Section 3.2.1) is a potential problem for larger vocabulary tasks; Nock (2001) proposes one possible solution. Finally, the approach must be extended to large-vocabulary continuous speech recognition without training a separate model of each utterance. In one possible scheme, a pronunciation dictionary is used to map a word sequence into K strings of beads (each bead being an HMM model of some feature value), one string for each of the K phonological tiers of interest. A full acoustic model is formed by loosely coupling these K featural streams, allowing asynchrony between them as in WARMTH in the introduction. However, whilst the transition coupling parameters should act to constrain the amount of asynchrony between streams, allowing this much feature asynchrony is still only likely to work over short units such as syllables. A successful solution to this particular problem will require both theoretical and engineering ingenuity.

Notes

1. Skip transitions are rarely used in state-of-the-art systems, having been found to degrade performance.
2. Similar arguments might motivate loosely coupled models as a relation of the more standard multiband models used for noise robustness (e.g., Mirghafori, 1999).
3. *Directed Acyclic Graphical Models* (DAGM) or *Bayesian Networks* (BN) are graphical statements of conditional independence relations amongst random variables (see Jensen, 1996; Zweig, 1998).
4. Note also exact likelihood calculations can made more efficient for observation-only coupled MMFHMMs (e.g., Ghahramani & Jordan, 1997).
5. Softer versions of this iterative scheme are possible, fixing subsets of chains and optimising over the remainder.
6. Inference in the general case is NP-hard (Cooper, 1990).
7. Although, as observed by Hagai Attias and an anonymous reviewer, the Chain Viterbi procedure for likelihood approximation can also be viewed as a variational approximation in which Q puts all probability mass on a single metastate sequence.

Acknowledgments

H.J. Nock was supported by an EPSRC studentship and DARPA Grant No. N660019928924. She is now at Audio–Visual Speech Technologies Group, IBM TJ Watson Research Center, Route 134, Yorktown Heights, NY 10528, U.S.A. The authors thank Dr. M.J.F. Gales (Cambridge University Engineering Dept., UK) and the SSLI Lab (University of Washington, USA) for helpful discussions, and Jason Eisner for several editorial suggestions.

References

Attias, H. (2000). Personal communication.

Cohen, M. H. (1989). *Phonological structures for speech recognition*. Unpublished doctoral dissertation, Computer Science Division, Department of Electrical Engineering and Computer Science, University of California, CA, USA.

Cole, R., Muthusamy, Y., & Fanty, M. (1990). *The ISOLET spoken letter database* (Tech. Rep. No. CSE 90-004). OGI.

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

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39*(1), 1–38.

Deng, L., & 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.

Fosler-Lussier, J. (1999). *Dynamic pronunciation models for automatic speech recognition*. Unpublished doctoral dissertation, ICSI, UC Berkeley, CA, USA.

Frankel, J., Richmond, K., King, S., & 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* (pp. 254–257). Beijing, China.

Ghahramani, Z., & Jordan, M. (1997). Factorial hidden Markov models. *Machine Learning*, *29*, 245–273.

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

Goldsmith, J. (Ed.). (1999). *Phonological theory: The essential readings*. New York: Blackwell.

Greenberg, S., Hollenback, J., & Ellis, D. (1996). Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In *Proceedings of ICSLP* (pp. S24–S27).

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

Jensen, F. (1996). *An introduction to Bayesian networks*. Berlin: Springer.

Jordan, M., Ghahramani, Z., Jaakkola, T., & Saul, L. (1998). An introduction to variational methods for graphical models. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 105–161). Dordrecht: Kluwer Academic Press.

Keating, P. (1997). Word-level phonetic variation in large speech corpora. In B. Pompino-Marschal (Ed.), *ZAS working papers in linguistics*. Available: http://www.humnet.ucla.edu/humnet/linguistics/people/keating/berlin1.pdf

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

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

Kirchhoff, K. (1998). *Robust speech recognition using articulatory information* (Tech. Rep. Nos. 98-036). ICSI, UC Berkeley, CA, USA.

Kirchhoff, K. (1999). *Robust speech recognition using articulatory information*. Unpublished doctoral dissertation, University of Bielefeld, Germany.

Liberman, M. (1998). *Proposal for research topic at the Johns Hopkins University (Center for Language and Speech Processing) Summer Research Workshop: Acoustic–phonetic feature detectors*. Unpublished manuscript.

Logan, B., & Moreno, P. (1998). Factorial HMMs for acoustic modelling. In *Proceedings of ICASSP* (pp. 813–816).

Mirghafori, N. (1999). *A multi-band approach to automatic speech recognition*. Unpublished doctoral dissertation, ICSI, UC Berkeley, CA, USA.

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, R., & Hinton, G. (1998). A view of the EM algorithm that justifies incremental and other variants. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 355–370). Dordrecht: Kluwer Academic Press.

NIST English Broadcast News Transcription (HUB4) Benchmark Test Results. (1999, January). ftp://jaguar.ncsl.nist.gov/csr99/bn98en_official_scores_990112/readme.htm

Nock, H. (2001). *Techniques for modelling phonological processes in automatic speech recognition*. Unpublished doctoral dissertation, Cambridge University Engineering Dept., Cambridge, UK.

Ostendorf, M. (2000). Incorporating linguistic theories of pronunciation variation into speech recognition models. *Philosophical Transactions of the Royal Society*, *358*, 1325–1338.

Richardson, M., Bilmes, J., & Diorio, C. (2000). Hidden-articulator Markov models: Performance improvements and robustness to noise. In *Proceedings of ICSLP* (pp. 131–134). Beijing, China.

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

Rose, R., Schroeter, J., & 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.

Russell, M. (1997). Progress towards speech models that model speech. In *Proceedings of IEEE Workshop ASRU* (pp. 115–123). CA, USA.

Saraclar, M. (2000). *Pronunciation modelling for conversational speech recognition*. Unpublished doctoral dissertation, The Johns Hopkins University, MD, USA.

Saraclar, M., Nock, H., & Khudanpur, S. (2000). Pronunciation modelling by sharing Gaussian densities across phonetic models. *Computer Speech and Language*, *14*(2), 137–160.

Saul, L., & Jordan, M. (1999). Mixed memory Markov models. *Machine Learning*, *37*(1), 75–87.

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

Stevens, K. (2000, October). From acoustic cues to segments, features and words. *Plenary Lecture, Proceedings of ICSLP*. Beijing, China

Sun, J., Deng, L., & Jing, X. (2000). Data-driven model construction for continuous speech recognition using overlapping articulatory features. In *Proceedings of ICSLP* (Vol. 1, pp. 437–440).

Tomlinson, M., Russell, M., Moore, R., Buckland, A., & Fawley, M. (1997). Modelling asynchrony in speech using elementary single-signal decomposition. In *Proceedings of ICASSP* (pp. 1247–1250).

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

Weintraub, M., Taussig, K., Hunicke-Smith, K., & Snodgrass, A. (1996). Effect of speaking style on LVCSR performance. In *Proceedings of ICSLP* (pp. S16–S19 (addendum)). Philadephia, USA.

Woodland, P., Leggetter, C., Odell, J., & Young, S. (1995). The development of the 1994 HTK large vocabulary speech recognition system. In *Proceedings of ARPA spoken language systems technology workshop* (pp. 110–115).

Young, S., Jansen, J., Odell, J., Ollason, D., & Woodland, P. (1995). *The HTK book (version 2.0)*. ECRL.

Zweig, G. (1998). *Speech recognition with dynamic Bayesian networks.* Unpublished doctoral dissertation, UC Berkeley, CA, USA.