LOOSELY COUPLED HMMS FOR ASR

H.J. Nock [*] *S.J. Young*

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK.
{hjn11,sjy}@eng.cam.ac.uk

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. This paper presents a preliminary feasibility study of an alternative class of models: *loosely coupled HMMs*. 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. The paper evaluates one coupled model on a simple ASR task, using both exact and approximate estimation schemes. We conclude such models merit further investigation.

1. INTRODUCTION

Hidden Markov Models (HMMs) have been successful for modelling the dynamics of carefully dictated speech. However, their performance degrades severely when they are used to model conversational speech, and it has been widely hypothesized that more sophisticated models will be required to achieve acceptable transcription performance on this type of data. This paper describes our preliminary investigations into an alternative class of models, which we describe informally as *loosely coupled HMMs*. Since speech is produced by a system of loosely coupled articulators, stochastic models which explicitly represent this parallelism may have advantages for automatic speech recognition (ASR), particularly when trying to model the phonological effects inherent in casual spontaneous speech.

Today's large-vocabulary recognizers are constructed using the notion of phonemic segments, corresponding (roughly) to particular configurations of the articulators. We build one or more statistical models for each element of the resulting inventory of speech segments (the *phone set*) and model words as a simple concatenation of segments. However, both speech scientists and linguists agree that the notion of a phoneme or speech segment is not a realistic one. Whilst the phoneme concept may be adequate for carefully read speech, in which articulatory gestures correspond sufficiently closely to some abstract ideal, there is evidence from speech production studies showing that changes in speaking rate, manner and style can all lead to variation in the amplitude of and phase relations between articulatory gestures. These changes, whilst simple to explain in an articulatory or phonological domain, can have extreme effects on the resulting acoustic signal: there may be colouring or even merging of the underlying 'segments' due to interaction between the articulatory gestures both within and across segment boundaries. We hypothesize that these effects contribute to the poor performance of current systems on conversational speech.

One approach to modelling this variability is to extend a conventional HMM-based recognition framework with a more sophisticated state- or model-sharing scheme (eg. [5]). In contrast, we attempt to model the underlying process more directly through a two-stage approach to ASR, in which (i) the acoustic signal is mapped into an intermediate representation comprising a number of potentially asynchronous feature streams, such as cepstra derived from sub-frequency bands, phonologically-motivated distinctive features or articulatory parameters, and then (ii) the intermediate representation is modelled statistically using a technique appropriate for loosely coupled time series. The family of parallel, loosely coupled HMMs provides one source of possibilities. In principle, this type of model offers the ability to model words not as a sequence of phonemes, but as a sequence of loosely coordinated articulatory or phonological feature changes, with synchronization required only at the level of words or even utterances. This paper presents a preliminary study of coupled models for ASR. Section 2 introduces coupled models, and then describes one specific coupled model and exact and approximate estimation schemes in more detail[1]. Section 3 evaluates this model and associated algorithms on an isolated letter classification task. Section 4 draws conclusions and outlines future research.

2. THEORY

Suppose we wish to model K loosely coupled time series, where the 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. We might model such data by combining the K observations at each time t into a single observation vector $O_t = (o_t^1, \ldots, o_t^K)$ and building a standard HMM. However, the resulting model would not be a parsimonious representation of the data. Alternatively, we might model each stream k independently by using a single HMM per stream. The individual likelihood scores from the independent HMMs can be then combined in some fashion to obtain an overall score, as in the multi-

[*] H.J. Nock is supported by an EPSRC studentship. The authors thank Dr. M.J.F. Gales for helpful discussions.

[1] A different coupled model and exact estimation scheme is evaluated for ASR in [6]; another model addressing asynchrony is presented in [7].

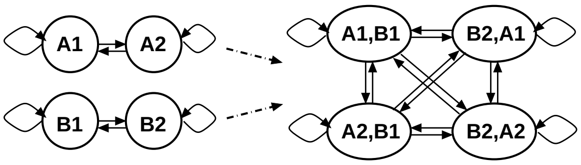

Figure 1: Metastate Space From Combined Ergodic HMMs A, B

band framework (eg. [8]). However, this scheme fails to represent any coupling between the different time series[2]. An intermediate approach is to combine the K independent HMMs into a joint model which can capture something of the correlations between different streams. We can form a combined HMM in which (i) the hidden state space is the Cartesian product of the state spaces of the individual HMMs (see Figure 1), and (ii) the observations O_t are formed by concatenating the individual stream observations at time t, ie. $O_t = (o_t^1, \dots, o_t^K)$. We refer to the Cartesian product hidden state space as the *metastate space*, to distinguish it from the state spaces of the original independent HMMs for each stream.

If we assume that each independent HMM has N states, then for moderate K and N, estimation of output densities and a transition matrix for the combined HMM will be intractable both computationally and in terms of robust parameter estimation. Several recent schemes handle these difficulties through additional conditional independence assumptions and approximations which exploit the internal, combinatorial structure of the metastates and observations both to reduce the number of parameters and as the basis for efficient, approximate training and decoding algorithms (eg. [10], [3]). The form of model simplification most appropriate for speech is an open research question. In this paper, we simply adopt the scheme proposed in [10] in order to gain insight into the issues involved. This *Mixed-Memory Approximation* for reducing the number of parameters is described in Section 2.1; some (more generally applicable) decoding and estimation schemes of differing computational complexity are discussed in Section 2.4.

2.1. The Mixed-Memory Approximation

For notational convenience, we assume that each of the K time series comprises D-dimensional observations and is initially modelled by a single N-state HMM. These HMMs are combined as described above to produce a model in which hidden metastates comprise K hidden variables and each observation vector has K sub-vectors. K-tuples $I = (i^1, \dots, i^K)$ and $J = (j^1, \dots, j^K)$ denote metastates. Combined observations are denoted $O_t = (o_t^1, \dots, o_t^K)$. P denotes probability mass functions, p denotes densities.

Saul and Jordan [10] propose the following simplifications:

- assume conditional independence of state components given previous state:

$$P(J|I) = \prod_{k=1}^{K} P(j^k|I) \tag{1}$$

- approximate these conditional probabilities with a convex combination of *cross-transition* matrices:

$$P(j^k|I) = \sum_{l=1}^{K} \psi^k(l) a^{kl}(j^k|i^l) \tag{2}$$

- conditional independence of observation components given current state:

$$p(O_t|J) = \prod_{k=1}^{K} p(o_t^k|J) \tag{3}$$

- approximate these conditional probabilities with a convex combination of *cross-emission* distributions:

$$p(o_t^k|J) = \sum_{l=1}^{K} \phi^k(l) b^{kl}(o_t^k|j^l) \tag{4}$$

The 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. They are fixed for a single model, and give a measure of the dependency between different streams, using a total of $2K^2$ parameters. This model thus has $\mathcal{O}(K^2(N^2+ND^2))$ free parameters vs. $\mathcal{O}(N^K(N^K+K^2D^2))$ for the combined, full metastate space model.

Approximation (2) has limitations for speech modelling, since speech HMMs are typically constrained *a priori* to have a left-to-right transition structure. This is achievable under (2) only when ψ is the identity matrix I (ie. transitions in stream k depend only on the previous state in stream k). This study also investigates the case $\psi \neq I$, introducing coupling through transition probabilities. When we do so, we use a^{kl} matrices which are individually left-to-right. This limits backwards transitions in the metastate space, but is not as strong as the standard left-to-right constraint.

2.2. Special Cases

The experimental section will discuss various special cases of the loosely coupled model presented above. The standard HMM is obtained by setting $K = 1$, $\phi^1(1) = \psi^1(1) = 1$. If we generalize the loosely coupled model to the case where observations have K subvectors but metastates comprise L hidden variables, and allow $L \neq K$, we can obtain the HTK [12] synchronous multiple stream model by setting $L = 1$, K=number of output streams, $\psi^1(1) = 1$ and $\phi^k(1) = 1$ for all k. The asynchronous multiband model (eg. [8]) is obtained by setting $\psi^k(k) = 1$, $\phi^k(k) = 1$, and $\psi^k(l) = 0$, $\phi^k(l) = 0$ for $l \neq k$. We further distinguish three forms of loosely coupled model: an observation-only coupled model sets ψ to the $K \times K$ identity matrix; a transition-only coupled model sets ϕ to the $K \times K$ identity matrix; a fully-coupled model refers to the general case of unrestricted ϕ, ψ.

2.3. Maximum Likelihood Estimation

Maximum-likelihood estimation of the model above may be achieved using the EM algorithm [2]. Latent variables x_t^k, y_t^k are introduced to encode the missing data (namely, the cross-transition and cross-emission mixture components used in stream

[2] At least between points at which streams are forced to synchronize.

k at each time t, respectively). The following update equations, for the case where each $b^{kl}(o_t^k|j^l)$ is modelled by a full-covariance Gaussian density $\mathcal{N}(\mu_j^{kl}, \Sigma_j^{kl})$, are derived in [9]:

$$\hat{\psi}^k(l) = \frac{\sum_t P(x_t^k = l|O)}{\sum_{\nu,t} P(x_t^k = \nu|O)} \tag{5}$$

$$\hat{\phi}^k(l) = \frac{\sum_t P(y_t^k = l|O)}{\sum_{\nu,t} P(y_t^k = \nu|O)} \tag{6}$$

$$\hat{a}^{kl}(j^k|i^l) = \frac{\sum_t p(x_t^k = l, s_t^k = j^k, s_{t-1}^l = i^l|O)}{\sum_t p(x_t^k = l, s_{t-1}^l = i^l|O)} \tag{7}$$

$$\hat{\mu}_j^{kl} = \frac{\sum_t p(y_t^k = l, s_t^l = j^l|O)o_t^k}{\sum_t p(y_t^k = l, s_t^l = j^l|O)} \tag{8}$$

$$\hat{\Sigma}_j^{kl} = \frac{\sum_t p(y_t^k = l, s_t^l = j^l|O)(o_t^k - \hat{\mu}_j^{kl})(o_t^k - \hat{\mu}_j^{kl})'}{\sum_t p(y_t^k = l, s_t^l = j^l|O)} \tag{9}$$

where $O = O_1, \dots, O_T$ denotes the current utterance and $S_t = (s_t^1, \dots, s_t^K)$ denotes the metastate at time t. Summations over t run from 1 to T; summations over ν run from 1 to K. $\hat{\theta}$ denotes an updated parameter θ. The posterior probabilities necessary for these updates can be calculated using a (notationally complicated) generalization of the Forward-Backward Algorithm, also in [9].

2.4. Approximations for Decoding/Training

EM estimation requires forward and backward probabilities in the metastate-space of size N^K, which becomes intractable as K or N increase. Phonological or articulatory feature sets typically involve $K > 5$; allowing stream asynchrony within words or larger modelling units increases the required N. Thus more efficient, perhaps approximate, decoding and estimation schemes are essential. We will evaluate two alternative schemes[3].

Viterbi State Sequences The most-likely metastate sequence given the data, S^*, can be obtained through standard Viterbi decoding in the metastate space. S^* may also be used in estimation, akin to Viterbi training of HMMs. The associated parameter update equations are obtained by conditioning posterior probabilities in equations (5)-(9) on S^* as well as observations O. However, obtaining Viterbi sequences in the N^K metastate space might again be intractable.

Chain Viterbi State Sequences Saul and Jordan [10] also propose a more efficient scheme for approximating S^* when the K time series are *assumed* weakly coupled. The algorithm iterates through each stream k in turn, finding the optimal sequence of hidden states through stream k given fixed values for the hidden states of the other streams[4]. The state space is thus reduced to size N when doing the optimizations for stream k. The algorithm can be initialized by (for example) computing a Viterbi state sequence for each chain individually or by assuming a uniform segmentation of the observations for each stream. Iteration through all K streams continues until convergence, which is not necessarily to S^* (see [9] for a counter-example). *Assuming* the resulting sequence is similar to the Viterbi sequence S^* leads to an approximate, Viterbi-like estimation scheme.

3. APPLICATION TO MODELLING OF FREQUENCY SUB-BAND CEPSTRA

Several authors advocate the use of cepstra derived from frequency subbands for ASR (eg. [8], [11]). We use this representation for our experiments, rather than a more speculative articulatory or phonological representation.

3.1. Experimental Setup

The OGI ISOLET database consists of wideband recordings of single letters of the alphabet. We use *Isolet1-4* (6240 utterances) to train and speaker-disjoint *Isolet5* (1560 utterances) to test. The performance of our baseline HMMs using a 39-d 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) for this task, comparable to results reported previously (eg. [1]). Subband cepstra extraction proceeds as follows. 25ms windows of speech are Fourier-transformed and filtered through a bank of 20 overlapping, equally mel-spaced, filters. Filtering produces a vector of log spectral energies $E = [e_1, \dots, e_{20}]$. 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 $(\#_1, \dots, \#_V)$ denotes the truncation scheme, where $\#_v$ indicates retention of cepstra $0, \dots, \#_v - 1$ 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. All Gaussians are full covariance, initialized using the global mean and covariance of the training set. Model sets using cross-emission or cross-transition dependencies are initialized in stages: first, independent HMMs are trained for each stream as in a multiband system; then, cross-stream dependencies are introduced gradually, with two training iterations between the addition of one cross-dependency per stream. Training continues until likelihood gains fall below a pre-specified threshold.

3.2. Experimental Results

The first set of experiments compares model structures. Observations comprise cepstra from two subbands 0-2 and 2-8kHz, with cepstral truncation (7,6), yielding a 39-d combined observation vector O_t. Table 1 gives baseline percentage correct (*%C*) performance of standard *HMM*s. Table 2 examines coupling through the transition matrices, using single Gaussian output distributions. The models in each block of the table are ordered in terms of the allowable asynchrony between streams: the synchronous *HTK stream* model is followed by the transition-only (*trans-*) coupled model and then the completely asynchronous *multiband* model. Table 3 examines the observation-only (*obs-*) and *fully-*coupled model structures, again ordering models via increasing asynchrony. Each state in *HTK stream* and *multiband* 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 *fully-* and *obs*-coupled models, which use a single Gaussian to model each $b^{kl}(o_t^k|j^l)$ distribution. The tables show that the performance of loosely coupled models is comparable to that of other conventional models on this task; further experiments comparing three-stream model structures using three cepstral subbands show similar behaviour (see [9]).

[3] We are also investigating mean-field and more structured variational methods from the graphical models community as the basis of efficient algorithms.

[4] We note that softer versions of this algorithm in which subsets of streams are fixed whilst others are decoded are also possible.

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

Model (states per stream)	# Parameters	%C
HTK stream (3)	2418	94.2
trans-coupled (3)	2440	94.1
multiband (3)	2424	93.9
HTK stream (6)	4836	94.9
trans-coupled (6)	4876	95.0
multiband (6)	4848	94.8
HTK stream (8)	6448	95.4
trans-coupled (8)	6500	95.3
multiband (8)	6464	95.8
HTK stream (10)	8060	96.4
trans-coupled (10)	8124	95.6
multiband (10)	8080	95.9

Table 2: Results: Transition-Coupled Models

The second set of experiments compares decoding and estimation schemes using two sub-band data as above with two-stream models. The first three columns of Table 4 compare three decoding schemes using *EM*-trained, fully-coupled models: full likelihood (*FL*), Viterbi (*Vit*) and Chain Viterbi (*CVit*) algorithms. The McNemar test [4] finds no significant differences between the *FL* and *Vit* schemes; only the '*'-ed *CVit* result differs significantly from *FL* (at significance level 0.05, but not at 0.01). The first and final two columns of the table compare three matched training and decoding schemes: *EM*-training and *FL* classification, Viterbi (*Vit*) training and decoding, and Chain Viterbi (*CVit*) training and decoding. The McNemar test finds no significant differences between the *Vit* or *CVit* schemes and *EM/FL Scheme* (at significance level 0.05). Similar results are obtained when the decoding (or training and decoding) schemes are applied to transition-only or observation-only coupled models.

4. CONCLUSIONS AND FUTURE WORK

Coupled models are theoretically appropriate for learning asynchronous behaviour from data. This paper has shown empirically that coupled models can perform as well as more conventional models on a simple task, and has identified approximate estimation schemes which make more extensive experimental evaluation tractable. Further analysis suggests that the models do capture information about asynchrony between streams. Although ISOLET is clearly a limited task, we conclude that loosely coupled models merit further investigation. Future work will develop alternative coupled models. These will circumvent the difficulties that equations (2) and arguably (4) present when modelling speech, whilst retaining the attractive properties of the model: the tractable numbers of free parameters, efficient decoding/estimation algorithms and the possibility of incorporating exponent stream weights[5].

Model (states per stream)	# Parameters	%C
HTK stream (3)	4842	94.6
fully-coupled (3)	4856	94.7
obs-coupled (3)	4840	94.9
multiband (3)	4848	94.0
HTK stream (6)	9684	96.2
fully-coupled (6)	9704	95.8
obs-coupled (6)	9676	95.7
multiband (6)	9696	95.3
HTK stream (8)	12912	96.2
fully-coupled (8)	12936	96.0
obs-coupled (8)	12900	95.8
multiband (8)	12928	96.3

Table 3: Results: Observation- and Fully-Coupled Models

states	Tr=EM	Tr=EM	Tr=EM	Tr=Vit	Tr=CVit
per	D=FL	D=Vit	D=CVit	D=Vit	D=CVit
stream	%C	%C	%C	%C	%C
3	94.7	94.7	*94.0	94.6	94.1
6	95.8	95.8	95.6	95.9	95.8
8	96.0	95.9	95.9	95.8	96.4

Table 4: Results: Training (Tr) and Decoding (D) Schemes

5. REFERENCES

1. R Cole, Y Muthusamy, and M Fanty. The ISOLET spoken letter database. Technical Report CSE 90-004, OGI, 1990.
2. AP Dempster, NM Laird, and DB Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.
3. Z Ghahramani and MI Jordan. Factorial Hidden Markov Models. *Machine Learning*, 29:245–273, 1997.
4. L Gillick and SJ Cox. Some Statistical Issues In The Comparison Of Speech Recognition Algorithms. In *Proc ICASSP*, pages 532–535, 1989.
5. T Hain and PC Woodland. Dynamic HMM Selection for Continuous Speech Recognition. In *Proc Eurospeech*, pages 532–535, 1999.
6. BT Logan and PJ Moreno. Factorial HMMs for Acoustic Modeling. In *Proc ICASSP*, pages 813–816, 1998.
7. S Matsuda, M Nakai, H Shimodaira, and S Sagayama. Asynchronous-Transition HMM. In *Proc ICASSP*, pages 1001–1004, 2000.
8. N Mirghafori. *A Multi-Band Approach to Automatic Speech Recognition*. PhD thesis, ICSI, UC Berkeley, 1999.
9. HJ Nock and SJ Young. Loosely Coupled HMMs For ASR: A Preliminary Study. Technical Report CUED/F-INFENG/TR386, CUED, 2000.
10. LK Saul and MI Jordan. Mixed Memory Markov Models. *Machine Learning*, 37:75–87, 1999.
11. MJ Tomlinson, MJ Russell, RK Moore, AP Buckland, and MA Fawley. Modelling Asynchrony in Speech Using Elementary Single-Signal Decomposition. In *Proc ICASSP*, pages 1247–1250, 1997.
12. S Young, J Jansen, J Odell, D Ollason, and P Woodland. *The HTK Book (Version 2.0)*. ECRL, 1995.

[5] Such weights appeal for modelling phonological or articulatory features, where only subsets of *critical* articulators may be necessary to distinguish certain sounds.