MULTI-RATE AND VARIABLE-RATE MODELING OF SPEECH AT PHONE AND SYLLABLE TIME SCALES

Özgür Çetin and Mari Ostendorf*

Department of Electrical Engineering, University of Washington, Seattle, WA
{cozgur,mo}@ee.washington.edu

ABSTRACT

This paper introduces a multi-rate extension of hidden Markov models (HMMs), for joint acoustic modeling of speech at multiple time scales. The approach complements the usual short-term, phone-based representation of speech with wide-context modeling units and long-term temporal features. We consider two alternatives for coarse scale modeling units and features, representing either phones, or syllable structure and stress, and both fixed- and variable-rate dependencies between two time scales. Experiments on conversational telephone speech (CTS) show that the proposed multi-rate acoustic models significantly improve recognition accuracy over HMM- and alternative coupled HMM-based approaches (e.g. feature concatenation) for combining short- and long-term acoustic information.

1. INTRODUCTION

The current acoustic modeling paradigm in speech recognition systems is largely based on representing words as a sequence of phones which are characterized by hidden Markov models (HMMs). Short-term spectral features are used, and though the use of context-dependent phones and dynamic features implicitly incorporate information from longer scales, current systems focus on acoustic variability and linguistic information over less than 100 ms. This approach has led to impressive results on large vocabulary speech recognition, but the state-of-the-art performance on conversational speech still lags far beyond that of humans. Many factors contribute to this performance gap, but the inaccuracy of HMM-based acoustic models for characterizing variability associated with conversational speech is believed to be a large contributing factor. Our goal in this paper is to improve the accuracy of acoustic models by incorporating acoustic and linguistic information from time scales longer than phones, especially from syllables, into the acoustic model in a multi-scale statistical modeling paradigm.

Phones and phone time scales are important for speech recognition, but there exists ample evidence that time scales longer than phones, especially syllables, also carry useful information for speech recognition. Syllables and syllable time scales play a central role in human speech perception of English. Pronunciation and durational variability observed in conversational speech show a high degree of dependence on syllable structure. For example, phones occurring in a syllable onset are more likely to be preserved than those in a coda, which are more likely to substituted by another phone, or completely deleted [1]. In addition, data-driven corpus studies have consistently shown that useful information for recognition lies in long time scales and wide contexts [2].

In this paper, we incorporate acoustic and linguistic information from long-time scales into speech recognition by joint statistical modeling of speech at phone and syllable time scales via a new multi-rate coupled HMMs architecture. In a 2-rate HMM-based acoustic model, we model speech using both the recognition units and the feature sequences corresponding to phone and syllable time scales. The fine scale in our models corresponds to the traditional phonetic HMMs with cepstral features, whereas for the coarse scale, we will explore two alternatives for characterizing either phones broadly or syllable structure and stress, with features extracted from 500 ms windows of speech. Our multi-scale approach differs from implicit approaches, where wide-context syllable features are used in decision-tree-based acoustic model clustering [3], or pronunciation modeling [4]. The multi-scale approach is also different from HMM-based approaches such as segment models [5], autoregressive models [6] and other temporal models, e.g. [2], which represent long-term dependence and higher-order statistics in a single stream of short-term features but do not involve any multi-scale modeling. Our approach is also different from the multi-stream approaches such as [7], which incorporate long-term features but do not reduce redundancy by downsampling the long-term features that tend to be highly correlated. As our experiments will demonstrate, such redundancy reduction is important for both confidence estimation and classification accuracy when combining information from multiple sources.

Multi-scale modeling based on multi-rate HMMs is explicitly designed to exploit long-term features, and as such, it is complementary to the research in new acoustic front-ends looking beyond the short-term spectrum, e.g. [8, 9]. The traditional approach for utilizing new features is to concatenate them with existing cepstral features after oversampling and use them in standard HMM-based models. However, HMMs have become so tuned to short-term features that their use might obscure the gains from new features, especially those from long-time scales. As our experiments show, statistical models and features interact and simple HMM-based combination approaches might not fully utilize complementary information in long-term features. In particular, we find that both the redundancy reduction and the selection of appropriate sized modeling units are important for utilizing long-term features for speech recognition. We also find that variable-rate sampling approaches [10] are very helpful for extracting multi-scale feature sequences that focus more on temporally varying regions, improving performance over fixed-rate sampling approaches.

The paper proceeds with an introduction to multi-rate HMMs and their variable-rate sampling extension, followed by discussion of their application to acoustic modeling. Then, we present ex-

*Currently at International Computer Science Institute, Berkeley, CA

perimental results on a conversational speech recognition task and finally summarize the key contributions.

2. MULTI-RATE HIDDEN MARKOV MODELS

An HMM characterizes a length T time series, $\{o_t\}$, called observations, through an underlying hidden state sequence, $\{s_t\}$,

$$p(\{o_t\},\{s_t\}) \equiv \prod_{t=0}^{T-1} p(s_t|s_{t-1})\, p(o_t|s_t)$$

in which it is assumed that the state sequence is first-order Markov, s_{-1} is a null start state, and observations are conditionally independent of everything else given their respective states [11]. The HMM independence assumptions lead to computationally efficient algorithms for probabilistic inference ($O(TN^2)$ computational cost for state cardinality N), the forward-backward algorithm, and parameter estimation, the Baum-Welch algorithm [11]. However, HMMs have a number of intrinsic limitations for representing multi-scale stochastic processes and long-term context information. First, representation of composite state structures by an HMM requires assigning a unique state to each possible state configuration, resulting in an exponential state space which increases both the computational cost of inference and the number of free parameters. Second, representation of multi-scale observation sequences by an HMM requires over-sampling of coarser-scale observations to make them synchronous with the finer scale, resulting in skewed class posterior estimates and over-confident classification decisions due to overcounting evidence from coarser scales. Lastly, the information between the past and present observations as represented by an HMM, for many state topologies, decays exponentially fast, as the time lag between them increases, due to the underlying Markov chain structure [12].

2.1. Basic Multi-rate HMM

The multi-rate HMM is a generalization of the HMM to multiple time scales. The multi-rate HMM decomposes process variability into scale-based parts, characterizing both the intra-scale dependencies and time evolution within each scale-based part as well as inter-scale interactions. In a K-rate HMM, the process is modeled at K time scales, and associated with each scale is a hidden state sequence, $\{s^k_{t_k}\}$, and an observation sequence, $\{o^k_{t_k}\}$, k denoting the scale level. Scales are organized in a hierarchical manner from the coarsest $k=1$ to the finest $k=K$, and the k-th scale is M_k times faster than the $(k-1)$-th scale, i.e. $T_k = M_k T_{k-1}$ for $k>1$ where T_k denotes the length of the observation sequence in the k-th scale. The joint distribution of state and observation sequences is modeled as

$$p(\{o^1_{t_1}\},\{s^1_{t_1}\},\ldots,\{o^K_{t_K}\},\{s^K_{t_K}\}) \equiv \prod_{k=1}^{K}\prod_{t_k=0}^{T_k-1} p(s^k_{t_k}|s^k_{t_k-1}, s^{k-1}_{\lfloor t_k/M_k \rfloor})\, p(o^k_{t_k}|s^k_{t_k}) \quad (1)$$

where s^k_{-1} is a null start state for the k-th scale, $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x, and hence $\lfloor t_k/M_k \rfloor$ is the index of the observation in the $(k-1)$-th scale covering the t_k-th observation in the k-th scale. A graphical model illustration of the multi-rate HMM is given in Figure 1.

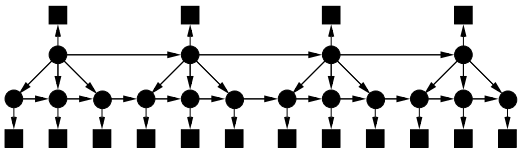

Fig. 1. Graphical model illustration of a multi-rate HMM with $K=2$ and $M_2=3$, with the coarse scale at the top. States and observations are depicted as circles and squares, respectively.

In the multi-rate HMM, statistical dependencies across time characterize the temporal dynamics of the scale-based components, whereas the hierarchical dependencies across scale characterize the interaction between the components. Dependencies across time are first-order Markov; those across scale are tree structured. The K-rate HMM essentially involves K multi-length HMMs, which are coupled via dependency of state transitions at one scale on the overlapping state variable at the next higher scale.

The various inference problems in multi-rate HMMs such as evaluating marginal likelihood of observations and calculating the state *a posteriori* probabilities are performed with a multi-rate extension of the HMM forward-backward algorithm. The overall cost of the resulting algorithm is $O(T_K N^{K+1})$ for a K-rate HMM (assuming that the state cardinality at each scale is equal to N), whereas collapsing multiple state components into a single state variable and invoking the HMM forward-backward algorithm directly would lead to a $O(T_K N^{2K})$ algorithm, which is exponentially worse. The parameter estimation of multi-rate HMMs is done via the expectation-maximization (EM) algorithm [13], automatically dealing with hidden states in multi-rate HMMs. For details, see [14].

State and observation factoring is also used in variations of HMMs for single-rate processes, including factorial HMMs [15], mixed memory models [16] and coupled HMMs [17], where single-rate multiple state and observation sequences are involved, in part to reduce the number of free parameters (for more robust estimation) and in part to allow asynchrony between state processes associated with different observation streams. The multi-rate state factoring benefits from these advantages, but it is also better suited for modeling long-term context (captured in the more slowly changing coarse scale state) and reducing feature redundancy. Two-dimensional multi-resolution HMMs [18] and hierarchical HMMs [19] are examples of multi-scale models involving scale-based state and observation sequences. Like the multi-rate HMM, these models use tree-structured coarse-to-fine dependencies to characterize inter-scale dependencies, but they are more restrictive in terms of the assumptions they make: state and observation variables at a given scale are conditionally independent of their distant relatives given their parent or ancestor state variables, and the state sequence at a given scale is disconnected and not a Markov chain.

2.2. Variable-rate Extension

In the basic multi-rate HMM, we assume that each observation and state at a scale k covers a fixed number, M_k, of observations and states at the next finest scale, the $(k-1)$-th scale. A fixed-rate downsampling ratio in the multi-rate HMM implies that features in each scale uniformly cover the original physical process. However, phone durations in English range from $10-30$ ms for

stop consonants to $50-150$ ms for vowels, and a time-invariant downsampling might not be of sufficient resolution to recognize phones with very short duration. In addition, a variable-rate feature extraction method can tailor the signal analysis to focus more on information-bearing regions such as phone transitions.

Hence, we extend the basic multi-rate HMM paradigm to allow for time-varying sampling rates between scales, so the number of observations at one scale corresponding to an observation at the next coarsest scale can vary. For example, the original 2-rate factorization assuming a fixed sampling ratio, M, is modified to

$$p(\{o^1_{t_1}\},\{s^1_{t_1}\},\{o^2_{t_2}\},\{s^2_{t_2}\}|\{M_{t_1}\}) \equiv \prod_{t_1=0}^{T_1-1} p(s^1_{t_1}|s^1_{t_1-1}) \times p(o^1_{t_1}|s^1_{t_1}) \prod_{t_2=l(t_1)}^{l(t_1)+M_{t_1}-1} p(s^2_{t_2}|s^1_{t_1},s^2_{t_2-1})\, p(o^2_{t_2}|s^2_{t_2}) \quad (2)$$

where M_{t_1} and $l(t_1)$ denote the number and starting index, respectively, of observations at the fine scale corresponding to the t_1-th observation at the coarse scale. Notice that we necessarily have $l(t_1) = \sum_{\tau_1=0}^{t_1-1} M_{\tau_1}$ and $T_2 = \sum_{t_1=0}^{T_1-1} M_{\tau_1}$. A graphical model illustration of the variable-rate extension is given in Figure 2.

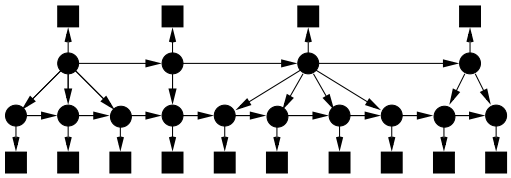

Fig. 2. Graphical model illustration for a variable-rate sampling 2-rate HMM with $M_0 = 3$, $M_1 = 1$, $M_2 = 4$, and $M_3 = 2$.

In the variable-rate factorization of Equation 2, we have assumed that sampling rates $\{M_{t_1}\}$ are given (deterministic). The variable-rate sampling framework is partially motivated to focus on interesting regions over the signal space, where such regions are determined during the signal processing stage. We note that it is possible to make M_t random, as in stochastic segment models [5], but the cost is higher and it is not implemented in this work.

3. MULTI-RATE HMM ACOUSTIC MODELS

We use 2-rate HMMs for joint acoustic modeling of speech at two time scales in two alternative ways. In both applications, the goal is to complement the usual cepstrum-based sub-phone models at the fine scale with long-term temporal features (in particular, a variation on TempoRAl PatternS – TRAPS [8], the so-called hidden activation TRAPS – HAT [9]) and wide modeling units at the coarse scale. (TRAPS are a relatively new method for data-driven, posterior-based feature extraction from very long time windows using neural network classifiers trained to predict, for example, phones, see [8, 9] for details.) The two applications differ in the phenomenon they characterize in the coarse scale. In the first case, the coarse scale characterizes phones broadly using HATs trained on phone targets, whereas in the second case the coarse scale characterizes a larger-scale phenomena, lexical stress and syllable structure, using HATs trained on such targets.

3.1. 2-rate HMM Phone Models

The 2-rate HMM phone models characterize phones at two time scales. The fine scale corresponds to the traditional HMM-based phone models, where we use a three-state left-to-right state transition topology and cepstral features. The coarse scale broadly characterizes phones using long-term temporal features (HATs trained on phone targets). Each state in the coarse scale is associated with a whole phone (not just part of it as in fine-scale states). Similar to context-dependent modeling in phonetic HMMs, we use cross-word, left and right context-dependent modeling units in both fine and coarse chains of the 2-rate HMM phone models.

3.2. 2-rate HMM Joint Syllable/Stress and Phone Models

In the 2-rate HMM joint syllable/stress and phone models, the fine scale again represents the usual phones using short-term cepstral features, but the coarse scale represents lexical stress and syllable structure (namely onset, nucleus, and coda). Specifically, the coarse scale involves: 1) acoustic units corresponding to the syllable structure (onset, coda, and ambisyllabic consonants; stressed and unstressed vowels) with two additional classes for silence and non-speech sounds such as noise; and 2) acoustic features (HATs) which are trained to predict to these seven classes of sounds. The models for words are composed by gluing together 2-rate HMMs corresponding to syllable constituents in that word. For example, the model sequence corresponding to the word `seven` is constructed by decomposing it using a stress- and syllable-marked dictionary:

```
seven [ s + eh [ v ] . ax n ]
```

where the phone sequence between an open bracket and a closed one corresponds to a syllable, the "[v]" is ambisyllabic, and '+' and '.' are stress and no-stress markers, respectively, for vowels. Using this pronunciation of `seven`, the coarse and fine state sequences in the composite 2-rate HMM model are obtained as:

```
  CO    |   V1    |  CA    |   V0    |  CC
s(1-2-3)|eh(1-2-3)|v(1-2-3)|ax(1-2-3)|n(1-2-3)
```

where | denotes a 2-rate HMM boundary; `CO, CC`, and `CA` denote onset, coda, and ambisyllabic consonants, respectively; `V0` and `V1` denote unstressed and stressed vowels, respectively; and `x(1-2-3)` denotes the 3-state sequence corresponding to the phone `x`. In our implementation, we again use context-dependent modeling units at both rates.

4. EXPERIMENTS

4.1. Task

We perform recognition experiments in a medium vocabulary CTS task, that of recognizing speech using a vocabulary of the $2,500$ most frequent words in Switchboard. It is a scaled-down version of the 2001 NIST Hub-5 recognition task used for evaluating large vocabulary speech recognizers, in terms of both amount of training data and vocabulary coverage during testing. The training data consists of 69 hours of speech from Switchboard and Callhome corpora, and the testing data is an $.9$ hour subset of the 2001 NIST Hub-5 evaluation data. There is a 1% out-of-vocabulary (OOV) rate on the test set with respect to the $2,500$ word lexicon, which is similar to the OOV rates in large-vocabulary speech recognition tasks. It has been shown that new techniques proven to be useful on this task are likely to be useful on full-vocabulary CTS tasks [20]. The language model (LM) is fixed to a bigram LM

which has been trained on Switchboard, Callhome, and Broadcast News transcriptions.

We use 12 perceptual linear prediction (PLP) coefficients and the logarithm of energy as well as their first- and second-order derivatives as our short-term features. For the coarse-rate features, we use HATs which are trained on either phone targets (resulting in a 23-dimensional feature after principle component analysis), or seven broad categories related to the syllable structure and stress (resulting in a 21-dimensional feature after concatenating two derivatives). Per-side mean subtraction and variance normalization have been applied to both PLP and HAT feature sequences.

4.2. Training and Testing Procedures

We used the following procedure to train and test the 2-rate HMM acoustic models in Graphical Models Toolkit (GMTK) [21]. The 2-rate HMM parameters are booted from the HMM systems corresponding to fine and coarse scale chains in the 2-rate HMM. We first train separate HMM systems using features and sub-word modeling units corresponding to each chain and determine tying of context-dependent of states in Hidden Markov Toolkit (HTK) [22]. We then transfer these HMM systems into GMTK and continue independent EM training with mixture splitting until 8 mixture components per state are achieved. The state-conditional output distributions in these fine- and coarse-scale HMMs are used to initialize the state-conditional output distributions in the fine and coarse chains in our 2-rate HMM systems. Then, we jointly train all parameters until the desired number of mixture components per state are achieved. All recognition experiments are performed by using GMTK in rescoring 500-best lists generated by HTK (which have an oracle word error rate (WER) of 19.3% and 1-best WER of 42.5%). The reported HMM systems in GMTK are the full-trained versions of HMM systems that were used to initialize the fine and coarse chains in our 2-rate HMM systems. The baseline HMM system with PLPs uses 32 mixture components per state, and the total number of parameters in all systems (except HMM system with C/V HATs, see Table 2 caption) are roughly made equal by adjusting the number of mixture components per state in each system.

In experiments with the basic multi-rate HMMs, we used a fixed-rate downsampling ratio of 3, whereas in experiments with the variable-rate extension of multi-rate HMMs, we dynamically sampled the coarse features (phone or C/V HATs) only when they significantly differ from the ones occurring before so that on average one in three coarse feature frames is kept.

4.3. Results

The WER performance of the 2-rate HMM system modeling phones and related HMM systems are presented in Table 1. In this table, we also report results with the 2-stream coupled HMMs [7], which are similar to 2-rate coupled HMMs, but do not reduce the redundancy of coarse features by downsampling. The 2-rate HMM system (MHMM) gives equivalent performance improvements to the HMM-based feature concatenation and score combination, or the multi-stream modeling approaches, but with smaller models, whereas the variable-rate system (VHMM) gives a further improvement, clearly outperforming all other systems.

The WER results of the 2-rate HMM system modeling syllable structure and stress in its coarse scale and the related HMM and multi-stream systems are reported in Table 2. We find that neither the HMM combination nor the multi-stream approaches are successful in utilizing complementary information in C/V HATs, which is unlike the multi- and variable-rate approaches that significantly improve over the baseline HMM system using PLPs. The 2-rate HMM systems achieve significant gains from representing syllable structure and stress, with a very small increase in parameters, they are not as large as those from representing phones broadly. The low temporal complexity of syllable phenomena (as represented in our models) seemed to be a limiting factor.

system	*WER %*	*# of states*
HMM-PLP	42.3	4986
HMM-HAT	44.7	4788
HMM-PLP/HAT	40.1	7906
HMM-PLP+HAT	41.1	4986/3163
MSTREAM (1-state)	40.1	4986/3163
MSTREAM (3-state)	39.7	4986/4788
MHMM	39.9	4986/1438
VHMM	39.1	4986/1438

Table 1. WER and number of states for different models using PLP and/or phone-based HATs, where the number of states for the multi-stream (MSTREAM) models is per phone at the coarse scale. The PLP/HAT and PLP+HAT indicate state-level feature concatenation vs. utterance-level score combination. The pairs in the number of states column denote those in the HAT and PLP streams of the corresponding system.

system	*WER %*	*# of states*
HMM-PLP	42.3	4986
HMM-HAT	50.4	95
HMM-PLP/HAT	42.8	7091
HMM-PLP+HAT	42.1	4986/95
MSTREAM	42.0	4986/95
MHMM	40.9	4986/84
VHMM	40.6	4988/84

Table 2. WER and number of states for different models using PLP and/or C/V HATs, designed to predict C/V classes. See Table 1 caption for further details.

5. CONCLUSIONS

In this paper, we proposed multi-rate and variable-rate acoustic modeling schemes for speech recognition, based on a multi-scale extension of HMMs, multi-rate HMMs. In the proposed models, the usual sub-phone recognition units and short-term spectral features are complemented with coarser recognition units and long-term temporal features. We used the coarse scale in multi-rate acoustic models in two alternative ways to represent phones, and syllable structure and stress. Experimental results in a challenging CTS task showed that the variable-multi-rate HMM significantly improves recognition accuracy over baseline HMM and multi-stream systems. We have found that significant WER improvements are possible via representing both syllables and phones in the multi-rate modeling framework with a very small increase in parameters, though the gains so far are not as large as those from representing phones at multiple scales.

Acknowledgments

The authors thank J.Bilmes of U. of Washington for GMTK and B.Y. Chen of ICSI for the HAT features. This work was supported by the DARPA Grant MDA972-02-1-0024. The opinions and conclusions are those of the authors and not necessarily endorsed by the US Government.

6. REFERENCES

[1] S. Greenberg, "Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, pp. 159–176, 1999.

[2] J. Bilmes, *Natural Statistical Models for Automatic Speech Recognition*, Ph.D. thesis, U. of California Berkeley, 1999.

[3] I. Shafran and M. Ostendorf, "Acoustic model clustering based on syllable structure," *Computer Speech and Language*, vol. 17, pp. 311–328, 2003.

[4] E. Fosler-Lussier *et al.*, "Incorporating contextual phonetics into automatic speech recognition," *Proc. of ESCA Workshop on Modeling Pronun. Variation for ASR*, pp. 611–614, 1999.

[5] M. Ostendorf *et al.*, "From HMMs to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. on ASSP*, vol. 4, pp. 369–378, 1996.

[6] B.H. Juang and L.R. Rabiner. Mixture autoregressive hidden Markov models for speech signals. *IEEE Trans. on ASSP*, vol. 6, pp. 1404–1413, 1985.

[7] S. Dupont and H. Bourlard, "Using multiple time scales in a multi-stream speech recognition system," in *Proc. of Eurospeech*, 1997, pp. 3–6.

[8] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in noisy speech," in *Proc. of ICASSP*, 1999, pp. 289–292.

[9] B.Y. Chen *et al.*, "Learning long term temporal feature in LVCSR using neural networks," in *Proc. of ICSLP*, 2004, pp. 612–615.

[10] A. Alwan *et al.*, Human and machine recognition of speech sounds in noise. In *Proc. of World Multi-conf. on Systems, Cybernetics, and Information*, vol.8, pp. 218–223, 2001.

[11] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, vol. 77, pp. 257–286, 1989.

[12] Y. Bengio and P. Frasconi. Diffusion of context and credit information in Markovian models. *J. of Artificial Intelligence Research*, vol. 3, pp. 249–270, 1995.

[13] A.P. Dempster *et al.*, Maximum likelihood from incomplete data via the EM algorithm. *J. of Royal Statistical Society Series B*, vol.39, pp. 185–197, 1977.

[14] Ö. Çetin and M. Ostendorf, "Multi-rate hidden Markov models for monitoring machining tool wear," Tech. Rep. UWEETR-2004-0011, U. of Washington Dept. of EE, 2003.

[15] Z. Ghahramani and M.I. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, pp. 245–273, 1997.

[16] L.K. Saul and M.I. Jordan, "Mixed memory Markov models: Decomposing complex stochastic processes as mixtures of simple ones," *Machine Learning*, vol. 37, pp. 1–11, 1998.

[17] M. Brand and N. Oliver, "Coupled hidden Markov models for complex action recognition," in *Proc. of IEEE Conf. Comp. Vision and Pattern Recog.*, 1997, pp. 994–999.

[18] J. Li *et al.*, "Multiresolution image classification by hierarchical modeling with two dimensional hidden Markov models," *IEEE Trans. on IT*, vol. 46, pp. 1826–1841, 2000.

[19] S. Fine *et al.*, "The hierarchical hidden Markov model: Analysis and applications," *Machine Learning*, vol. 32, pp. 41–62, 1998.

[20] B.Y. Chen *et al.*, "A CTS Task for Meaningful Fast-Turnaround Experiments," *Proc. of DARPA RT04 WS*, 2004.

[21] J. Bilmes and C. Bartels, "On triangulating dynamic graphical models," in *Proc. of UAI*, 2003, pp. 47–56.

[22] S. Young *et al.*, *The HTK Book (for HTK Version 3.2)*, Cambridge University, 2002.