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 modeling units and long-term temporal features. We consider two alternatives for coarse scale, representing either phones, or syllable structure and lexical stress, and both fixed- and variable-rate dependencies between time scales. Experiments on conversational telephone speech (CTS) show that the proposed multi-rate approach significantly improves recognition accuracy over HMM- and other coupled HMM-based approaches (e.g. feature concatenation) for combining short- and long-term acoustic and linguistic information.

1. INTRODUCTION

The current acoustic modeling paradigm in speech recognition is largely based on representing words as a sequence of phones which are characterized by HMMs. Short-term spectral features are used, and though context-dependent phones and dynamic features implicitly incorporate information from longer time scales, current systems focus on acoustic variability over less than 100 ms. This approach has led to impressive results, 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 acoustic modeling accuracy by incorporating linguistic and acoustic information from time scales longer than phones, especially from syllables, in a multi-scale statistical modeling paradigm.

Phones are important for speech recognition, but there exists ample evidence that time scales longer than phones, especially syllables, also carry useful information. Syllables 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 discriminative information for recognizing speech extends beyond 100 ms [2].

In this paper, we incorporate long-term acoustic and linguistic information 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 acoustic model, we model speech using both the modeling units and the feature sequences corresponding to phone and syllable time scales. The fine scale in our models corresponds to the traditional phone 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 long-term features extracted from 500 ms windows. Our multi-scale approach differs from implicit approaches such as [3] and [4], where syllable features are used for acoustic model clustering or pronunciation modeling, and from HMM-based approaches such as segment models and autoregressive HMMs [5], 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. It also differs from other multi-stream approaches that incorporate long-term features [6] in that the multi-rate model reduces the redundancy of highly-correlated long-term features by downsampling. 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 specifically designed to utilize long-term features and is complementary to the research in new acoustic front-ends looking beyond the short-term spectrum, e.g. [7, 8]. 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. Statistical models and features interact and simple HMM-based combination schemes might not fully utilize complementary information in long-term features. We find that both the redundancy reduction and the selection of appropriate modeling units are important for utilizing long-term features. We also find that variable-rate sampling approaches which focus more on temporally varying regions, are particularly helpful for extracting multi-scale feature sequences, improving performance over fixed-rate approaches.

The paper proceeds with an introduction to multi-rate HMMs and their variable-rate sampling extension, followed by discussion of their applications to acoustic modeling. Then, we present experimental results on a CTS task and 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, $\{s_t\}$,

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

*Currently at International Computer Science Institute, Berkeley, CA

where the state sequence is assumed to be first-order Markov, s_{-1} is a null start state, and observations are conditionally independent of everything else given their respective states [9]. The HMM independence assumptions lead to computationally efficient probabilistic inference and parameter estimation algorithms [9], but they also limit what can efficiently be represented by HMMs. HMMs have a number of intrinsic limitations for representing multi-scale stochastic processes and long-term context. First, representation of composite state structures in an HMM requires assigning a unique state to each 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 in an HMM requires oversampling of coarser-scale sequences to make them synchronous with the finer-scale ones, resulting in skewed class posterior estimates and overconfident 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 with the time lag, due to the underlying Markov chain structure.

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 couplings. 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 $k > 1$, T_k denoting the sequence length at 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 at the $(k-1)$-th scale covering the t_k-th observation at the k-th scale. In the multi-rate HMM, statistical dependencies across time characterize the temporal dynamics of the scale-based components, whereas those across scale characterize the interaction between the components. Dependencies across time are first-order Markov; those across scale are tree structured. A K-rate HMM essentially involves K multi-length HMMs, which are coupled via their states. A graphical model illustration of the multi-rate HMM appears in Figure 1.

The various probabilistic inference tasks in multi-rate HMMs, such as evaluating marginal probability of observations and the state *a posteriori* probabilities, are solved by a multi-rate generalization of the HMM forward-backward algorithm. The overall computational cost of this 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 states into a single state and invoking the HMM forward-backward algorithm directly would induce $O(T_K N^{2K})$ cost, which is exponentially worse. The parameter estimation of multi-rate HMMs is done via the expectation-maximization (EM) algorithm, automatically dealing with hidden states in multi-rate HMMs. For details, see [10].

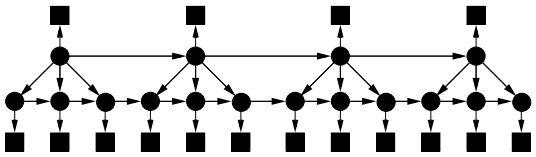

Fig. 1. Graphical model illustration of a multi-rate HMM with $K = 2$ and $M_2 = 3$, with the coarse scale at the top. States and observations are depicted as circles and squares, respectively.

State and observation factoring are also used in variations of HMMs for single-rate processes, e.g. factorial HMMs [11] and coupled HMMs [12], where multiple same-rate 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 across sequences. The multi-rate HMMs also benefit from these advantages but are better suited for modeling long-term context (captured in the coarse scales) and reducing feature redundancy. Two-dimensional multi-resolution HMMs [13] and hierarchical HMMs [14] are examples of multi-scale models similar to multi-rate HMMs. Like the multi-rate HMMs, these models employ tree-structured coarse-to-fine dependencies to characterize inter-scale dependencies, but they are more restrictive in terms of the assumptions they make: states at a given scale are conditionally independent of their distant relatives given their parent or ancestor states, and conditionally, the state sequence at a given scale is disconnected and not a Markov chain, unlike in multi-rate HMMs.

2.2. Variable-rate Extension

In the basic multi-rate HMM, we assume that each observation at the scale k covers a fixed number, M_k, of observations at the next finest scale, the $(k-1)$-th scale. A fixed-rate downsampling ratio 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 sampling might not be of sufficient resolution to recognize phones with very short duration. In addition, a variable-rate sampling method can tailor the signal analysis to focus more on information-bearing regions such as transitions. Thus, 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 temporally varies. For example, the original 2-rate HMM 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. We require $l(t_1) = \sum_{\tau_1=0}^{t_1-1} M_{\tau_1}$.

In the variable-rate factorization of Equation 2, we 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 signal processing. It is straightforward to make M_t stochastic, as in segment models [5] and hierarchical HMMs [14], 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 subphone models at the fine scale with long-term temporal features and wide-context modeling units at the coarse scale. We use a variation of on TempoRAl PatternS (TRAPS) [7], the so-called hidden activation TRAPS (HAT) [8]) as long-term features. (TRAPS are a method for data-driven, posterior-based feature extraction from very long time windows using neural network classifiers trained to predict, for example, phones. See [7, 8] 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 separate 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 phones using short-term cepstral features, but the coarse scale represents syllable structure and lexical stress. Specifically, the coarse scale involves: (1) acoustic units corresponding to the syllable structure and stress (onset, coda, and ambisyllabic consonants; and, stressed and unstressed vowels) with two additional classes for silence and non-speech sounds such as noise (C/V classes in short); and (2) acoustic features (HATs) trained to predict to these seven classes of sounds. The models for words are composed by gluing together the 2-rate HMMs corresponding to syllable constituents in words. 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. Using this pronunciation of `seven`, the coarse and fine state sequences in its composite 2-rate HMM are obtained as:

$$\overset{\texttt{CO}}{\texttt{s(1-2-3)}}\,\big|\,\overset{\texttt{V1}}{\texttt{eh(1-2-3)}}\,\big|\,\overset{\texttt{CA}}{\texttt{v(1-2-3)}}\,\big|\,\overset{\texttt{V0}}{\texttt{ax(1-2-3)}}\,\big|\,\overset{\texttt{CC}}{\texttt{n(1-2-3)}}$$

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 three-state sequence corresponding to the phone `x`. In our implementation, we again use context-dependent modeling units at both scales.

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 recognizers, in terms of both amount of training data and vocabulary coverage during testing. The training data consists of 69 hours of speech from the Switchboard and Callhome corpora, and the testing data is a 0.9 hour subset of the 2001 NIST Hub-5 evaluation data. The resulting out-of-vocabulary (OOV) rate on the test set is 1%, which is similar to the OOV rates in full-vocabulary tasks. It has been shown that new techniques proved to be useful on this task are likely to transfer to the full-vocabulary tasks [15]. The language model (LM) is fixed to a bigram LM 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 sound classes 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 features.

4.2. Training and Testing Procedures

We used the following procedure to train and test the 2-rate HMM acoustic models. We first train separate HMM systems using features and subword modeling units corresponding to each chain and determine context-dependent state tying in the Hidden Markov Toolkit (HTK) [17]. We then transfer these HMM systems into the Graphical Models Toolkit (GMTK) [16] and continue independent EM training with mixture splitting until obtaining 8 mixture components per state. The state-conditional output distributions in these fine- and coarse-scale HMMs are used to initialize the output distributions in the fine and coarse chains in the 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 GMTK rescoring of 500-best lists generated by HTK (with a 19.3% oracle word error rate (WER) and 42.5% 1-best WER). 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 the 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 the HMM system with C/V HATs) are roughly made equal by adjusting the number of mixture components per state.

In experiments with the basic multi-rate HMMs, we used a fixed-rate downsampling ratio of three, whereas in experiments with the variable-rate extension of multi-rate HMMs, we dynamically sampled the coarse features (phone or C/V HATs) 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 WERs of the 2-rate HMM system modeling phones and related HMM systems are presented in Table 1. This table also re-

System	*WER %*	*# of tied 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. The performance of 2-rate HMM system modeling phones, and related systems, using PLPs and/or phone HATs. The number of states for the MSTREAM systems is per phone at the HAT stream. In the system names, $\cdot/\cdot$ and $\cdot+\cdot$ indicate state-level feature concatenation vs. utterance-level score combination. The pairs in the last column denote states in PLP vs. HAT streams.

System	*WER %*	*# of tied 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 (1-state)	42.0	4986/95
MHMM	40.9	4986/84
VHMM	40.6	4988/84

Table 2. The performance 2-rate HMM system modeling syllables, and related systems, using PLPs and/or C/V HATs, designed to predict C/V classes. See Table 1 caption for further details.

ports results with the 2-stream coupled HMMs (MSTREAM) [6], 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 and the multi-stream approaches, but with smaller models, whereas the variable-rate system (VHMM) gives a clear improvement.

The WERs of the 2-rate HMM system modeling syllable structure and stress and the related HMM and multi-stream systems are reported in Table 2. Neither the HMM combination nor the multi-stream modeling approaches are successful in utilizing C/V HATs, which is unlike the multi- and variable-rate modeling approaches that significantly improve over the baseline HMM system. The 2-rate HMM systems achieve significant gains from representing syllable structure and stress, with a very small increase in parameters, though the improvements are not as large as those from representing phones broadly. The low temporal complexity of syllable phenomena, as represented in our models, seems to be limiting.

5. CONCLUSIONS

This paper proposed multi- and variable-rate acoustic models for speech recognition, based on a multi-scale extension of HMMs, multi-rate HMMs. The usual subphone modeling units and short-term spectral features are complemented with wide-context modeling units and long-term temporal features. The novel coarse scale in these models is used to represent either phones, or syllable structure and stress. Experimental results on a challenging CTS task showed that the multi- and variable-rate HMMs significantly improve recognition accuracy over the HMM and alternative multi-stream systems. Significant improvements with a very small number of additional parameters were found from representing syllable structure, though the gains so far are not as large as those from representing phones. Further gains could be possible from increasing number of parameters in the proposed models.

Acknowledgments The authors thank J. Bilmes of UW for GMTK and B.Y. Chen of ICSI for HATs. 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] S. Dupont and H. Bourlard, "Using multiple time scales in a multi-stream speech recognition system," in *Proc. of Eurospeech*, 1997, pp. 3–6.

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

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

[9] 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.

[10] Ö. Çetin, *Multi-rate Modeling, Model Inference, and Estimation for Statistical Classifiers*, Ph.D. thesis, UW, 2004.

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

[12] 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.

[13] 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.

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

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

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

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