MODELING HETEROGENEOUS DATA SOURCES FOR SPEECH RECOGNITION USING SYNCHRONOUS HIDDEN MARKOV MODELS

Yong Zhao and Biing-Hwang (Fred) Juang

Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA

ABSTRACT

In this paper, we propose a novel acoustic modeling framework, synchronous HMM, which takes full advantage of the capacity of the heterogeneous data sources and achieves an optimal balance between modeling accuracy and robustness. The synchronous HMM introduces an additional layer of substates between the HMM states and the Gaussian component variables. The substates have the capability to register long-span non-phonetic attributes, which are integrally called speech scenes in this study. The hierarchical modeling scheme allows an accurate description of probability distribution of speech units in different speech scenes. To address the data sparsity problem, a decision-based clustering algorithm is presented to determine the set of speech scenes and to tie the substate parameters. Moreover, we propose the multiplex Viterbi algorithm to efficiently decode the synchronous HMMs within a search space of the same size as for the standard HMMs. The experiments on the Aurora 2 task show that the synchronous HMMs produce a significant improvement in recognition performance over the HMM baseline at the expense of a moderate increase in the memory requirement and computational complexity.

Index Terms: Speech recognition, hidden Markov model, system combination, Viterbi algorithm

1. INTRODUCTION

It is widely known that the performance of the speech recognition systems often degrades dramatically if they are operated under mismatched operating conditions. A common practice to ameliorate this mismatch problem, known as multistyle training, is to collect large amounts of speech data from a variety of acoustic conditions for training the acoustic models. However, the multistyle training may not fully realize its performance potential as the HMM-based acoustic models are excessively diffused to accommodate the extraneous variabilities introduced by the tremendous amounts of speech data.

One broad class of approaches that address the modeling of heterogeneous data sources is to use an ensemble of models, each focusing on a particular acoustic condition. The simplistic way to generate multiple models is to divide the training corpus into a number of homogeneous blocks, and then train an HMM set for each block. Recognition can be performed by running multiple recognizers of these models in parallel. The recognition hypothesis is obtained by either combining the decoding outputs of the multiple recognizers in a ROVER-like paradigm [1] or choosing the one with the highest likelihood. An alternative way of combining multiple models is to preselect one model set that best matches the operating condition for recognition.

Moreover, multiple models can be combined at the frame level to achieve a more granular form of combination. The most common forms are cluster adaptive training (CAT) [2] and eigenvoice [3], where the target model is obtained as a linear interpolation of multiple speaker/cluster-dependent models. Typically, the interpolation weights are estimated from the adaptation data using the ML criterion.

The multi-model approach is an attractive scheme to address heterogeneous data sources for speech recognition. However, a number of problems may limit their usefulness. The first problem is the data sparsity in estimating parameters of multiple models. As the number of the models increases, there will be fewer data available for providing reliable estimation for each individual model. The second problem is the heavy computational load in combining multiple models. Following the classical ensemble learning theory, it is expected that the best performance should be obtained by applying the constituent models in parallel to produce a plurality of candidate hypotheses for the majority voting. Unfortunately, this introduces multiple decoding with dramatically increased computational complexity and memory requirements. The similar situation applies to CAT, which usually requires two decoding passes to accomplish. Though alternative methods such as model pre-selection can alleviate this drawback, they are at the expense of compromising the recognition accuracy.

In this paper, we consider a novel acoustic modeling framework, synchronous HMM, which takes full advantage of the capacity of the diversified speech data and achieves an optimal balance between modeling accuracy and robustness. In contrast to the conventional HMMs, the synchronous HMM introduces an additional layer of latent variables, referred to as substates, between the HMM state and the Gaussian component variables. The substates have the capability to register long-span non-phonetic attributes, such as gender, speaker identity, and environmental condition, which are integrally called speech scenes in this study.

The acoustic models built upon the synchronous HMMs can be thought of as a collection of multiple acoustic models, each corresponding to a specific speech scene. In this regard, it is related to the multi-model approaches [4], [5], [6]. However, the synchronous HMM offers a number of advantages over the conventional multi-model approaches. First, the hierarchical modeling scheme allows an accurate description of probability distribution of speech units in different speech scenes. Second, by closely incorporating the models of speech scenes as sub-models of the synchronous HMM, we can determine the model structure and estimate the model parameters in an integral and consistent manner. Furthermore, by exploiting the synchronous relationship among the speech scene sub-models, we propose the multiplex Viterbi algorithm to efficiently decode the synchronous HMM within a search space of the same size as for the standard HMM. The multiplex Viterbi can also be generalized to decode an ensemble of isomorphic HMM sets, a problem often arising in the multi-model systems.

2. SYNCHRONOUS HMMS

In contrast to the conventional Gaussian mixture HMM, the synchronous HMM introduces an additional layer of latent variables, referred to as substates, between the HMM state and the Gaussian component variables. The substates depend on the previous substate in addition to the state that generates it. Accordingly, the model consists of a quadruple of stochastic processes $(\boldsymbol{x}_1^T, s_1^T, z_1^T, m_1^T)$, where $\boldsymbol{x}_1^T = \boldsymbol{x}_1, ..., \boldsymbol{x}_T$ is a sequence of observations of length T, and $s_1^T = s_1, ..., s_T$, $z_1^T = z_1, ..., z_T$, and $m_1^T = m_1, ..., m_T$ are sequences of latent variables of HMM states, substates, and mixture indexes, respectively. The statistical dependencies between these

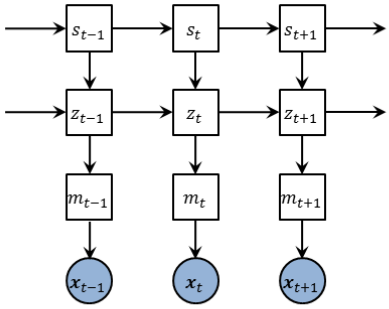

Fig. 1. Dynamic Bayesian network representation of the synchronous HMM.

variables can be represented by a DBN [7] as shown in Fig. 1.

The synchronous HMM is motivated by aiming at accurately characterizing the highly heterogeneous data sources. For speech recognition, the state layer represents the process of the canonical speech, and the substate layer represents the process of a variety of real-world speech due to speaker and environmental variations, which are referred to as speech scenes in this study.

One key property of the synchronous HMMs is that the evolution of the substate layer is synchronous with the evolution of the state layer. This effectively eliminates the possible explosion of the state space caused by introducing multiple Markov chains, as for the case of the factorial HMM [8]. Suppose that the model consists of N states and each state corresponds to K substates, which leads to NK substates in the substate layer. Naively, the state space composed of the direct product of states and substates would be of size N^2K. However, by imposing the synchronous constraint on the two Markov chains, the state space retains a size of NK. Moreover, the synchronous HMM can be interpreted as the synchronization among substates of different speech scenes. This will lead to substantial computational savings in learning and decoding the model as will be discussed later.

From the DBN in Fig. 1, the joint probability of these sequences in the synchronous model can be factored as

$$p(\boldsymbol{x}_1^T, s_1^T, z_1^T, m_1^T) = \prod_{t=1}^{T} p(s_t|s_{t-1})p(z_t|z_{t-1}, s_t) \times p(m_t|s_t, z_t)p(\boldsymbol{x}_t|s_t, z_t, m_t) \quad (1)$$

The synchronous HMM consists of the following elements: state transition probability $p(s_t = j|s_{t-1} = j') = a_{j'j}$, substate transition probability $p(z_t|z_{t-1}, s_t)$ (to be discussed shortly), prior of Gaussian component l from state $s_t = j$ and substate $z_t = k$

$$p(m_t = l|s_t = j, z_t = k) = w_{jkl}$$

and likelihood of Gaussian component l from state $s_t = j$

$$p(\boldsymbol{x}_t|s_t = j, m_t = l) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_{jl}, \boldsymbol{\Sigma}_{jl})$$

Note that, to make effective use of data, the Gaussian components for each state are shared among all substates of that state. Thus the substates from the same state differ only in the mixture weights, analogous to the method used in the semi-continuous HMMs [9].

Depending on the form of the substate transition probabilities $p(z_t|z_{t-1}, s_t)$, there are several variants of the synchronous HMM. When the observation distribution of the substates degenerates to a single Gaussian, the synchronous HMM is equivalent to the stranded HMM as described in [10]. One limitation of the stranded HMM is that the transitions between the substates (i.e., mixture components in the stranded HMM) are bounded inside an HMM, preventing it from capturing the long-span temporal dependency of speech.

An alternative of the synchronous HMM, which aims to take into account the long-span temporal dependency, is to deterministically specify the dependency between substates. Consider that the kth substates of all the states represent speech from a particular speech scene, and the speech scene keeps unchanged during an utterance. The substate transition probability can be written as

$$p(z_t = k|z_{t-1} = k', s_t = j) = \delta(k', k) \quad (2)$$

where $\delta(\cdot, \cdot)$ denotes the Kronecker delta function. In addition, we can maintain the speech scene dependency across the models by creating multiple dummy substates for each dummy state and drawing links between the corresponding ones. Thus, the kth substates of all the states form a separate sub-model representing the kth speech scene. The substate transition diagram of the synchronous HMM with separate speech scenes is illustrated in Fig. 2.

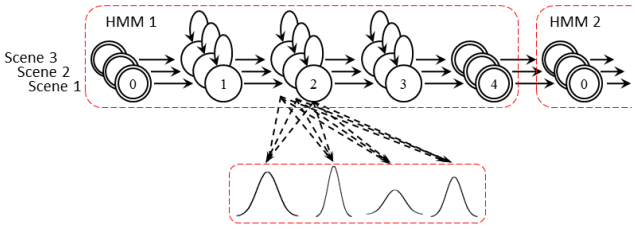

Fig. 2. Illustration of substate transitions and observation distributions for the synchronous HMM. The substates of three speech scenes are connected separately within and across the models. The processes of different speech scenes are synchronous with each other. The substates from the same state share the pool of Gaussian components and differ in the mixture weights.

Moreover, the speech scenes can be allowed to switch at boundaries of HMMs or words. The synchronous HMM with switching speech scenes can capture the speech under the influence of slowly varying factors, such as non-stationary environmental noise. However, this paper will be focused on investigating the synchronous HMM with separate speech scenes.

The synchronous HMM can be learned using an EM algorithm similar to the regular HMM, but the estimation of the scene-specific model parameters (i.e., mixture weights) needs to be addressed with care. Basically, we start with a set of well-trained Gaussian mixture HMMs. We then divide the training corpus into a number of homogeneous subsets based on some criterion, and generate multiple speech scene sub-models by updating the mixture weights of a sub-model based on each subset. Finally, several runs of full-scale re-estimation are carried out without explicitly associating the subsets to the speech scene sub-models.

3. SPEECH SCENE DECISION TREE

One issue in employing the synchronous HMM for speech recognition is the data sparsity. When the number of the speech scenes increases, there will be fewer data available for providing reliable estimation of the scene-specific model parameters. We present a decision tree-based algorithm to address this problem, analogous to the phonetic decision tree used for clustering context-dependent phone models [11]. Suppose that the utterances are tagged with the acoustic conditions, such as gender, speaker identity, and environmental condition. We first produce the synchronous HMMs with as many sub-models as the distinct acoustic conditions in the training corpus. Then the decision tree-based clustering is applied globally or at the substates of individual states.

The global decision tree is aimed to cluster the speech scene sub-models, and thus to determine the set of sub-models in the final synchronous HMMs. The tree is built in a top-down fashion with the questions relating to the acoustic condition tags. Nodes are iteratively split at each iteration by finding a node and an associated question that jointly produce the maximum increase in the log-likelihood on the training data. If we assume that during clustering, the Gaussian components remain unchanged, then only the mixture weights of the clustered sub-models need to be calculated. It turns out that the criterion of maximum increase in the log-likelihood is equivalent to the maximum reduction of the weighted entropy of the mixture weights. On completion of the clustering, the speech scene sub-models in the same cluster can be merged. This leads to a compact set of speech scenes in the synchronous HMM.

Alternatively, we can cluster and tie the substates of individual states for improved efficiency. We may apply the decision tree for each state following the same top-down clustering procedure. However, building separate trees for different states hinders the possible merging of the speech scene sub-models, because the sub-models can be merged only if they share common clusters for all their substates. To address this problem, we cluster the substates of individual states by trimming the global decision tree in a bottom-up fashion. We initialize a decision tree of the substates for each state by cloning the topology of the global decision tree. The trimming process starts with pairs of sibling leaf nodes, which are merged if the log-likelihood reduction is less than some threshold, or some leaf nodes lack sufficient data to support themselves. After iteratively merging all of such sibling pairs, we note that some rarely seen leaf nodes have not yet been trimmed, because their siblings are non-leaf nodes. We then merge these dangling leaf nodes with other leaf nodes that result in the minimum reduction in the log-likelihood.

4. MULTIPLEX VITERBI DECODING

The decoding process using the synchronous HMM is to find the best path that matches the given observation sequence, through the search space spanned by states and substates. A straightforward decoding method is to perform the Viterbi algorithm through the search space comprising the substates of the model. However, this method leads to a dramatic increase in memory requirements and computational complexity, which roughly correspond to K times of decoding the standard HMM.

We propose a novel multiplex Viterbi algorithm that performs an effective decoding on the synchronous HMM by keeping the search space of the same size as for the standard HMM. The search space is constructed based on the model states, except that each state node is compounded by all the substates of that state. At each time step, the substates of a state share the same path, but keep individual records of the sub-path scores, which are cumulated over the substate sequence following the shared path. The path score of the state takes the highest sub-path score from the constituent substates, and is used to represent the fitness of that state in the Viterbi decoding. Fig. 3 shows the trellis diagram for the multiplex Viterbi algorithm.

More formally, let $\tilde{\delta}_t(j,k)$ be the sub-path likelihood for substate k following the best partial path ending in state j at time t. By induction, we have the recursion formula

$$\tilde{\delta}_t(j,k) = \tilde{\delta}_{t-1}(i^*,k)a_{i^*j}b_{jk}(\boldsymbol{x}_t) \tag{3}$$

$$i^* = \arg\max_i \left\{ \max_k \tilde{\delta}_{t-1}(i,k)a_{ij}b_{jk}(\boldsymbol{x}_t) \right\} \tag{4}$$

where i^* is the preceding state that leads to the best path ending in state j at time t. Note that in a strict sense, i^* should be written as

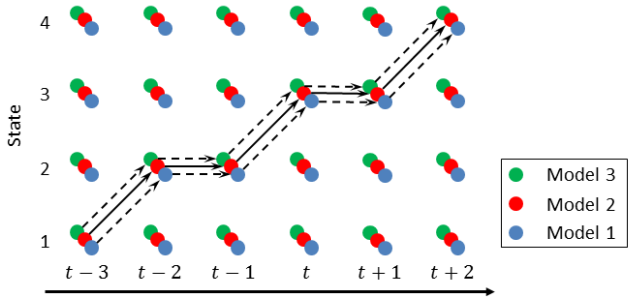

Fig. 3. Illustration of the multiplex Viterbi algorithm. The substates of a state share the same path, but keep individual sub-path scores.

$i_t^*(j)$ to indicate their dependence on time t and state j. Nevertheless, we apply the shorter notation for ease of understanding (3).

It should be noted that the multiplex Viterbi finds an approximate solution to the best state sequence, because the state path is jointly determined by the constituent sub-paths. We here assume that if a substate sequence can successfully match the observation sequence, the resulting state/observation alignment should also be appropriate for aligning the observations with other substate sequences following the same state sequence.

Remarkably, the multiplex Viterbi can be generalized to decode an ensemble of isomorphic standard HMM sets, a problem often arising in the multi-model systems. These HMM sets share the same search space, and equal in the state transition probabilities, but differ in the observation probability distributions. Typically, we are interested in the best HMM set and the best state sequence that jointly achieve the highest likelihood given the observation sequence. The multiplex Viterbi can efficiently address this decoding problem.

Moreover, we can apply the beam search strategy [12] to prune less likely speech scenes during the multiplex Viterbi to accelerate the decoding speed. We first determine, at each time step, the highest sub-path score for each speech scene. Then the speech scenes whose highest scores fall short of the score of the best speech scene by more than a fixed factor are pruned from further consideration.

The major advantage of the multiplex Viterbi is that it significantly reduces the memory requirements and computational complexities in comparison with the standard Viterbi algorithm for decoding K HMM sets. First, by constructing a search space of the same size as for the standard HMM, the multiplex Viterbi eliminates the memory and computational overhead in constructing and maintaining a K-times increased search space. Moreover, the use of the speech scene pruning strategy further saves a considerable computational load.

5. EXPERIMENTAL RESULTS

The proposed algorithm is evaluated on the Aurora 2 database [13] of connected digits. The multistyle training set is used to train the acoustic models. It consists of noisy data involving four types of noise (subway, babble, car, and exhibition hall) at four SNRs (20, 15, 10, and 5 dB), along with clean data, totaling 17 noise conditions. We further split the training set by gender and obtain 34 subsets as the basic speech scenes for the synchronous HMMs.

The baseline HMMs are obtained following the standard Aurora 2 recipe for the complex back end. Each digit is modeled by the whole word left-to-right HMM, consisting of 16 states and 20 Gaussian components per state. Besides, a 3-state silence model and a 1-state short pause model with 36 Gaussian components per state are used. Each feature vector consists of 13 mel-cepstral coefficients (including zeroth order for the energy term), and their delta

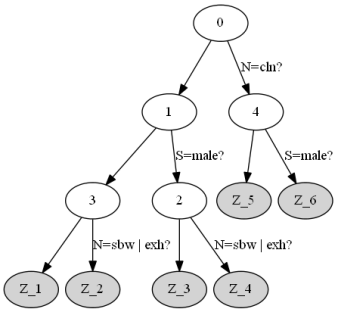

Fig. 4. Example of a speech scene decision tree. The questions used to split the nodes relate to speaker gender (S), noise type (N), and noise levels (SNR), separately. The non-leaf nodes are indexed in the order of splitting precedence.

and delta-delta coefficients. Features are normalized by CMN in the sentence level. The HMM baseline yields word error rate (WER) of 7.53% by averaging over SNRs between 20 and 0 dB of three test sets.

First, we generate the synchronous HMMs using the global speech scene decision tree. Fig. 4 shows an example of the decision tree which results in six speech scenes. As can be seen, the most useful questions that split the trees are concerned with noise types and speaker genders. Also, the decision tree grows in a symmetric fashion. In particular, the left subtree of the root node, which involves all the noise-corrupted speech, is expanded layer by layer (shown partially in Fig. 4).

By modifying the stop criterion for the speech scene clustering, the synchronous HMMs with different numbers of speech scenes are prepared for the evaluation, as shown in Table 1. Due to the symmetric structure of the decision tree, we can encode the clustering results for those selective numbers of speech scenes with a shorthand notation as in the second column of the table. It is observed that as the number of the speech scenes increases, the performance of the synchronous models is gradually improved. Specifically, the synchronous HMMs using 18 speech scenes achieve the lowest WER of 6.25%, 17% relative reduction over the baseline HMMs. We do not see that the speech scene clustering would improve the recognition performance over the unclustered 34-scene system. This implies that the training data in the Aurora 2 corpus are sufficient for reliably estimating the mixture weights of 34 speech scene sub-models. We also observe that the multiplex Viterbi algorithm greatly improves the decoding speed in comparison with the standard Viterbi for decoding K HMM sets.

Table 1. WER (%) and decoding time (times the HMM baseline) of the synchronous HMMs with different numbers of speech scenes on the Aurora 2 task.

# of scenes	Set of scenes	WER	Avg. time
1	—	7.53	1.0
2	{cln, no cln}	7.07	1.6
6	{Female, Male} × {cln, sbw|exh, bbl|car}	6.61	2.5
10	{Female, Male} × {cln, {sbw|exh, bbl|car} × {15-20 dB, 5-10 dB}}	6.38	3.3
18	{Female, Male} × {cln, {sbw, bbl, car, exh} × {15-20 dB, 5-10 dB}}	6.25	5.0
34	All combinations	6.27	7.9

Since the speech scene sub-models in the synchronous HMMs differ only in the mixture weights, it is worth investigating the distribution of the mixture weights. Fig. 5 shows the mixture weights of the substates for a particular state in the 34-scene system. We see that the mixture weights are sparse and each substate only relates to a small number of Gaussian components. This is in contrast to the conventional Gaussian mixture HMMs, where the mixture weights for each state are approximately equal and noninformative. Moreover, the substates for female and male speech largely involve different Gaussian components.

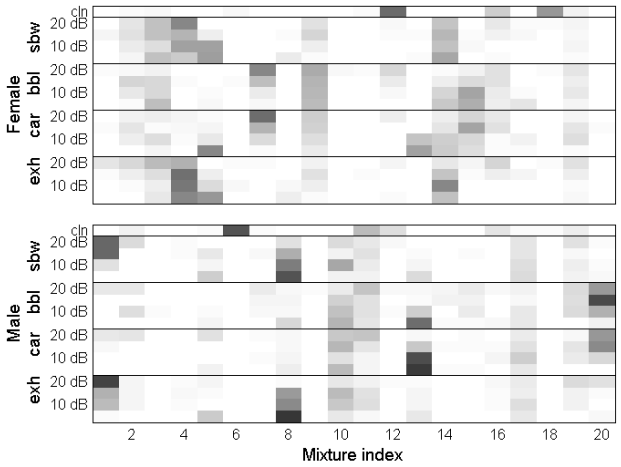

Fig. 5. Magnitudes of the mixture weights of the substates for a particular state in the 34-scene system. The rows correspond to the speech scenes, which are sorted according to gender, noise type, and noise level. The darker the color, the more prominent the weight is.

The multiplex Viterbi search can be accelerated by pruning unlikely speech scenes. Table 2 presents the WER and the decoding time of the synchronous system with varying scene pruning thresholds. The system is constructed by first generating 18 speech scenes using the global decision tree and then clustering the substates of individual states, which produces 8.8 tied substates per state in average. It is shown that few search errors occur when the scene pruning threshold is 50 or larger. In particular, at the pruning threshold of 100, the decoding search takes 2.0 times the HMM baseline decoding time, saving the computational cost with a factor of 9 compared with the simple multi-model approach.

Table 2. WER (%) and decoding time for the multiplex Viterbi with different scene pruning thresholds on the Aurora 2 task. The synchronous HMMs consist of 18 speech scenes, 8.8 tied substates per state in average.

Scene pruning threshold	WER	Avg. time
—	6.25	3.0
100	6.25	2.0
50	6.35	1.7
20	6.81	1.4

6. CONCLUSION

In this paper, we have proposed the synchronous HMM by introducing an additional substate layer into the standard HMM. The speech scene decision tree is proposed to determine the optimal set of speech scenes and tie the substate parameters. Moreover, we propose a novel multiplex Viterbi algorithm that performs an effective decoding on the synchronous HMM. Our experiments on the Aurora 2 database have showed the synchronous HMMs achieve the lowest WER of 6.27%, 17% relative reduction over the baseline HMMs. By jointly applying the speech scenes decision tree, the multiplex Viterbi, and the speech scene pruning, the decoding time of the 18-scene synchronous models is reduced to 2.0 times the HMM baseline decoding time, saving the computational cost with a factor of 9 compared with the simple multi-model approach.

7. REFERENCES

[1] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. ASRU*, 1997, pp. 347–354.

[2] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 417–428, July 2000.

[3] R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 695–707, Nov. 2000.

[4] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large vocabulary continuous speech recognition using HTK," in *Proc. ICASSP*, 1994, pp. 125–128.

[5] M. Akbacak and J. H. L. Hansen, "Environmental sniffing: Noise knowledge estimation for robust speech systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 465–477, Feb. 2007.

[6] H. Xu, P. Dalsgaard, Z. H. Tan, and B. Lindberg, "Noise condition-dependent training based on noise classification and SNR estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2431–2443, Nov. 2007.

[7] K. P. Murphy, *Dynamic Bayesian networks: representation, inference and learning*, Ph.D. thesis, Univ. Calif. Berkeley, 2002.

[8] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, no. 2, pp. 245–273, 1997.

[9] X. D. Huang and M. A. Jack, "Semi-continuous hidden Markov models for speech signals," *Comput. Speech Lang.*, vol. 3, no. 3, pp. 239–251, 1989.

[10] Y. Zhao and B. H. Juang, "Stranded Gaussian mixture hidden Markov models for robust speech recognition," in *Proc. ICASSP*, 2012, pp. 4301–4304.

[11] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. HLT*, 1994, pp. 307–312.

[12] H. Ney and S. Ortmanns, "Dynamic programming search for continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 16, no. 5, pp. 64–83, Sept. 1999.

[13] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR*, 2000, pp. 181–188.