Hidden Feature Models for Speech Recognition Using Dynamic Bayesian Networks

Karen Livescu, James Glass

Spoken Language Systems Group
MIT Laboratory for Computer Science
Cambridge, MA 02139 USA
{klivescu, glass}@mit.edu

Jeff Bilmes

University of Washington – SSLI Lab
Department of Electrical Engineering
Seattle, WA 98195 USA
bilmes@ee.washington.edu

Abstract

In this paper, we investigate the use of dynamic Bayesian networks (DBNs) to explicitly represent models of hidden features, such as articulatory or other phonological features, for automatic speech recognition. In previous work using the idea of hidden features, the representation has typically been implicit, relying on a single hidden state to represent a combination of features. We present a class of DBN-based hidden feature models, and show that such a representation can be not only more expressive but also more parsimonious. We also describe a way of representing the acoustic observation model with fewer distributions using a product of models, each corresponding to a subset of the features. Finally, we describe our recent experiments using hidden feature models on the Aurora 2.0 corpus.

1. Introduction

The majority of current speech recognition research assumes a model of speech consisting of a stream of contiguous segments (phones) derived from an underlying stream of basic linguistic units (phonemes), conforming with the theory of generative phonology of the 1960s and 1970s [1]. In more recent theories such as nonlinear phonology, speech is considered to be the output of *multiple* streams, or tiers, containing various *features* (e.g., [2]). These features, which are hidden from the listener (as opposed to observed acoustic features), can evolve asynchronously and may not line up to form phonetic segments. We refer to models of speech that use multiple streams of such features as *hidden feature models*.

There is also mounting evidence that the phone-based model may not be adequate for automatic speech recognition, especially in the case of spontaneous, conversational speech [3]. It has been noted, for example, that spontaneous speech is extremely difficult to transcribe phonetically, with both phone identities and phone boundaries being difficult to pinpoint [4]. Furthermore, while it has been hypothesized that pronunciation variability accounts for a large part of the performance degradation on conversational speech [5, 6], efforts to model this variability with phone-based rules or expanded lexica have had only limited success [6, 7]. One possible explanation is that phonemes affected by pronunciation rules can take on a surface form "intermediate" to the underlying phonemes and the predicted surface phones [8]. Such intermediate forms may be better represented as resulting from changes or spreading in one or more features than as entire phone changes.

There have been several previous studies using feature-based models for speech recognition, typically using articulatory features. For example, Deng *et al.* (e.g. [9]) and Richardson *et al.* [10] represent multiple features in one hidden state variable, whose evolution follows certain allowed trajectories. Kirchhoff (e.g. [11]) estimates the values of the hidden features, then treats them as observed variables and maps them to words.

A difficulty in using hidden feature models is that the most commonly used computational structures (e.g. hidden Markov models) allow for only one state variable at a time, whereas feature-based models are more naturally represented using several state variables, one for each feature stream. One framework that addresses this issue is that of graphical models (GMs) [12]. GMs, and in particular dynamic Bayesian networks (DBNs), have been gaining popularity as a modeling tool for speech recognition [13, 14, 15]. GMs allow for arbitrary sets of variables with arbitrary dependencies, making the specification of hidden feature models straightforward. In addition, if certain independencies can be assumed to hold among the variables, GMs can provide a more parsimonious representation.

In the following sections, we describe our initial experiences with DBNs for hidden feature modeling. Section 2 briefly introduces DBNs and describes a class of hidden feature models. Section 2.3 discusses the possibility of factoring the acoustic observation model into multiple small factors, each corresponding to a subset of the features. Section 3 describes initial experiments on the Aurora 2.0 corpus. Finally, Section 4 concludes with a discussion of our work thus far and possible extensions.

2. DBN-based hidden feature models

2.1. Dynamic Bayesian networks

A Bayesian network (BN) is a way of representing the conditional independence properties of a set of random variables (RVs) via a directed acyclic graph, each of whose nodes corresponds to one of the RVs. Independencies are encoded via missing edges in the graph. Specifically, for a graph over the variables $X_1, \ldots, X_N$, the joint distribution is given by

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | x_{\pi_i}), \quad (1)$$

where X_{π_i} are the parents of X_i in the graph.

Dynamic Bayesian networks (DBNs) are BNs that have a repeating structure consisting of an indefinite number of frames,

This research was supported by a grant from the Johns Hopkins Research Workshop on Human Language Technology and the NSF under subaward 8110-48270, and by Rutgers and the NSF under subaward 1483.

Presented at the 8th European Conference on Speech Communication and Technology (EUROSPEECH), Geneva, Switzerland, September, 2003.

so as to model stochastic processes over time or space. A hidden Markov model (HMM) can be represented with a DBN in which each frame contains two variables (the state and the observation) and two edges (from the state to the observation, and from the state in the previous frame to the current state).

2.2. Hidden feature models

Figure 1 shows the basic structure of one class of hidden feature model for two frames. This structure does not (explicitly) model feature asynchrony; we chose to start with this model because of its reduced complexity. In each frame, there are N features $A_1, \ldots, A_N$, each of which depends on the current phonetic state S and on its own value in the previous frame. O is the vector of observations (i.e. acoustic features), which depends on the current features. The intuition for this structure is that, at any instant, each feature would like to be at the target value for the current phone, but is also affected by its own value in other frames because of inertia and continuity constraints. S and $A_1, \ldots, A_N$ are discrete with discrete parents, so their probabilities are given by (multidimensional) conditional probability tables (CPTs). O is typically continuous; we refer to its probability, conditioned on its parents, as the *observation model*.

This model could be converted to an HMM, by combining $S, A_1, \ldots, A_N$ into a single variable whose state space is the product space of the individual variables. However, the size of the state variable's CPT would be much larger than the sum of the CPT sizes for the original variables. Specifically, if the cardinalities of the original variables are $c_S, c_{A_1}, \ldots, c_{A_N}$, then the equivalent HMM state variable would have cardinality $c_S \prod c_{A_l}$; so while the original variables require about $c_S^2 + c_S \sum c_{A_l}^2$ parameters, the new state variable would need about $(c_S \prod c_{A_l})^2$, quickly leading to data sparseness. Therefore, if we can assume certain independencies among the variables, it is more parsimonious to represent this explicitly than to collapse it into an HMM. Also, a representation that explicitly encodes independencies may result in computational savings for probabilistic inference in training and decoding.

Depending on our assumptions, each a_l's CPT may be (a) dense, i.e. any value is possible with some non-zero probability; (b) sparse, i.e. only certain trajectories are allowed; (c) dense or sparse, but also independent of a_l's previous value, in which case the inter-frame edges can be removed; or (d) deterministic, i.e. the value is completely determined by the parents. In the case of (d), the model reduces to an HMM; as we discuss below, however, we further factor the observation model in this case, so it is not identical to a phone-based HMM.

This structure has been suggested previously [13, 16] but, to the authors' knowledge, no experimental investigations using such a model have been reported on. We also consider possible extensions, namely the inclusion of certain additional edges. For example, edges can be inserted between different hidden features to represent possible additional dependencies between them. Also, depending on the feature set and the allowed dependencies between features, it may be unrealistic to assume that the observation is independent of the phone given the features. In that case, an edge may be added from the phone state to the observation.

This structure defines a large class of models. We report on experiments using a subset of these, consisting of options (c) and (d) above, with or without a phone-to-observation edge.

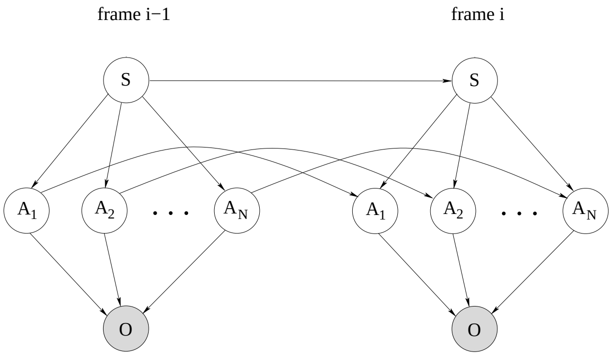

Figure 1: *A hidden feature model.*

2.3. Factoring the observation model

Another problem encountered in feature-based models is the large number of possible feature states (whether factored into multiple feature variables or represented as a single state variable). The observation model, $p(\mathbf{o}|a_1, \ldots, a_N)$, requires a separate distribution for each allowed combination of feature values $a_1, \ldots, a_N$, typically resulting in a very large number of distributions. In order to avoid the resulting data sparseness and complexity, we are exploring the possibility of factoring this probability into terms corresponding to smaller subsets, or clusters, of features. The number of distributions to be estimated during training and evaluated during decoding is then given by the sum of the cluster cardinalities, rather than by the product of the feature cardinalities.

Let F be the set of features. Define (non-intersecting) feature clusters $F_1, \ldots, F_M$, where $M \leq N$, such that $F_k \subseteq F$ and $\bigcup F_k = F$. We consider replacing $p(\mathbf{o}|a_1, \ldots, a_N)$ with

$$\frac{\prod_{k=1}^{M} p(\mathbf{o}|f_k)}{Z(\mathbf{o})}, \tag{2}$$

where f_k is a vector of values of the features in F_k, and $Z(\mathbf{o})$ is a normalizing constant. There is a number of possible modifications to the model, involving additional independence assumptions, that can be shown to lead to this factorization. We are currently working on the particular assumptions that are needed and their implications.

3. Experiments

Our experiments have been performed on the Aurora 2.0 corpus of noisy connected digits [17] using GMTK [14], a toolkit for representation of and computation with DBNs. We compare our models to a baseline 3-state, left-to-right phone-based (monophone) HMM. Our actual training and decoding structures represent S using several variables such as the current word, position within the word, and phone and word transitions, as in [16].

Our feature set consists of eight features: **voicing** (off, on), **velum** (closed, open), **manner** (closure, sonorant, fricative, burst), **place** (labial, labio-dental, dental, alveolar, post-alveolar, velar, nil), **retroflex** (off, on), **tongueBodyLowHigh** (low, mid-low, mid-high, high, nil), **tongueBodyBackFront** (back, mid, front, nil), and **rounding** (off, on). "Nil" place is used for vowels; "nil" tongue features are used for most consonants. All phones except silence are mapped to vectors of canonical feature values; silence has its own observation model.

We have experimented with model configurations (c) and (d) (see Section 2.2), with or without a phone-to-observation edge during decoding. For decoding with a phone-to-observation edge, we trained separate baseline and feature-based observation models and combined them during decoding using exponential weights on both models; the phone transition probabilities in the combined models are taken from the baseline. We also use exponential weights on all of the other local probability models to control their relative contributions. Specifically, the per-frame score used in decoding is given by

$$p(s_i|s_{i-1})^{w_s}\left(\prod_l p(a_{l,i}|s_i,a_{l,i-1})^{w_a}\right)p(\mathbf{o}|s_i)^{w_p w_o} \times p(\mathbf{o}|a_{1,i},\ldots,a_{N,i})^{(1-w_p)w_o}, \tag{3}$$

where w_o is the total weight of the observation model, which is varied to control the insertion and deletion rates, and w_p is the weight of the phone-based observation model relative to w_o.

In the case of model (d), we used a factored observation model, so that the last term in the above expression is further factored into multiple terms corresponding to feature clusters. Therefore, while this model does not implement a feature-based *phonology*, it implements a feature-based factorization. To construct the feature clusters, we used an agglomerative clustering procedure, with the average mutual information between features in the clusters as the distance measure. The mutual information estimates were obtained from forced alignments of the training data. The resulting clustering has **voicing** and **manner** in one cluster, the two tongue features in another cluster, and the four remaining features in four separate clusters.

The observation vector consists of 13 Mel-frequency cepstral coefficients plus energy, along with their derivatives and second derivatives. All observation models are implemented as Gaussian mixtures. We set the number of Gaussians, as well as the various weights, based on development set results.

3.1. Results

3.1.1. *Deterministic models*

Figure 2 shows the word error rates of the baseline phone-based HMM recognizer and the deterministic hidden feature model (model (d)) combined (via interpolation) with the phone-based observation model (referred to as the "HFM + phone" model) on three development sets with added subway noise. This configuration is very similar to the model in [18]. We compare recognizers trained on the "multi-train" set with multiple noise conditions and the "clean-train" set containing only clean speech. Error rates for the hidden feature model *alone* are: in the multi-train case, 1.9% on the clean test set, 10.0% at 10 dB, and 60.0% at -5 dB; in the clean-train case, 2.1% on the clean set, 57.3% at 10 dB, and 90.7% at -5 dB. Table 1 shows results on several independent test sets with a different noise type (babble noise) and an additional noise level, using the weights found to give the best results on the development sets.

In most cases, the HFM + phone models achieve lower error rates than the baseline for some range of w_p; the best setting of w_p, however, can be quite different from one noise level to another. The feature models by themselves are worse than the baseline, except in the -5 dB condition where they improve on the baseline appreciably.

We note that we found similar (development set) results when using just a small subset of the features. For example, a model using the tongue feature cluster *alone* still outperforms the baseline at -5 dB; when combined with the phone-based observation model, it achieves improvements in both the clean and the -5 dB conditions.

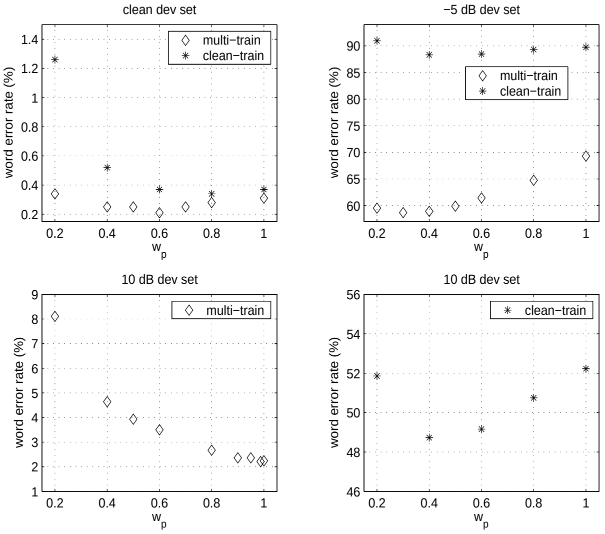

Figure 2: *Results of HFM + phone models on Aurora set A with added subway noise. $w_p = 1$ corresponds to the baseline ($w_p = 0$ would not correspond to the feature model alone since it uses the baseline phone transition probabilities). The 10 dB results have been split into two plots for clearer viewing.*

Train set	**Test set**	**Baseline WER (%)**	**HFM+phone WER (%)**	**% reduction**
multi-train	clean	0.4	0.3	25
	10 dB	1.8	1.8	0
	0 dB	46.0	43.7	5
	-5 dB	115.7	101.1	13
clean-train	clean	0.4	0.5	-25
	10 dB	57.6	49.8	14
	0 dB	94.4	85.4	10
	-5 dB	94.8	89.5	6

Table 1: *Test set word error rates (WERs) on Aurora set A with added babble noise. The last column shows the relative WER reduction of the HFM + phone model over the phone baseline.*

3.1.2. *Non-deterministic models*

We have experimented with model (c) using an unfactored observation model. Since this model is more computation-intensive than the deterministic one, we have trained it on a random 1000-utterance subset of the multi-train set, and used additional random 1000-utterance training subsets as development data. We compare this model to the baseline recognizer trained on the same subset. In order to help constrain the features to remain associated with their intended "meanings", we used a two-step training procedure: first, we fixed the Gaussians to be identical to a set of Gaussians trained with the deterministic model, and trained only the CPTs of the feature variables; second, we kept the CPTs fixed while training the Gaussians.

Development set results with varying w_p show a similar pattern to that obtained with the deterministic models. Table 2 shows test set results obtained using weight values tuned on the development set.

An encouraging note is that many of the learned feature CPTs behave as expected, suggesting that they have retained

Test set	Baseline WER (%)	HFM+phone WER (%)	% reduction
clean	1.3	1.2	8
10 dB	3.7	3.6	3
0 dB	47.4	47.2	0
-5 dB	116.1	108.6	6

Table 2: *Test set word error rates (WERs) on Aurora set A with added babble noise, using model (c) trained on a 1000-utterance subset of the multi-train set.*

their intended meanings. Specifically, on a task with such a small vocabulary and phone set, one can make some predictions, based solely on phone identities, for how the feature probabilities should behave. For example, we expect the phone [ɔ], and especially its final state, to be retroflexed more often than other phones are, as it only occurs before [r] (in the word "four"). We have, in fact, found such patterns in the trained feature CPTs, including: retroflexion of the final state of [ɔ]; nasalization of the final states of [ɪ] and [ʌ], both of which occur just before [n] (in "seven" and "one"); rounding of the first state of [ʌ], which occurs after [w] in "one"; and the realization of the middle states of [t] and [k] as both closures and fricatives (the canonical realization being a burst).

4. Discussion

We have described a class of hidden feature models that is quite rich, with special cases ranging from completely deterministic features to features that can range over all possible values and have dependencies on the previous frame and on other features. We have found that certain models in this class can achieve (slightly) improved performance on a simple recognition task when combined with a phone-based observation model, indicating that the hidden features contain information not present in the phone-based representation. Anecdotal evidence suggests that the trained models retain the intended meanings of at least some of the features even when they are given some freedom to stray from their canonical realizations.

We have also seen that the feature model configurations we have tested alone (without combining them with the phone-based edge observation model) usually perform worse than the baseline, suggesting that the strong assumptions in these models may be unreasonable. Additional model configurations using the same basic structure, but including inter-frame and inter-feature edges, may ameliorate this. It is clear, from linguistic and physiological considerations, that the features are not independent given the phone state. Some additional edges can be added manually based on such considerations, but a more attractive approach may be to learn the dependencies from data using structure-learning algorithms [16].

While the model class we have presented is very flexible, it has the drawback that the features still depend on the phone, making it impossible to explicitly model feature asynchrony. We are currently exploring the possibility of an alternate class of model with less or no dependence between the phone and the features, which would more closely adhere to the ideas of nonlinear phonology.

We are also currently further studying issues in the factorization of the observation model. In particular, we are interested in gaining a better understanding of the minimal assumptions needed to obtain the factorization and their implications, as well as developing clustering techniques that match the assumptions.

Finally, we plan to apply our techniques in domains with greater phonological variation, which may stand to gain more from the freedom of non-deterministic hidden feature models.

Acknowledgments. This work began as part of the graphical models project at the 2001 Johns Hopkins Summer Workshop. The feature set and phone-to-feature mappings were constructed with help from Katrin Kirchhoff and Eva Holtz. We also thank Geoff Zweig and Nati Srebro for helpful discussions.

5. References

[1] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York: Harper and Row, 1968.

[2] J. A. Goldsmith, *Autosegmental and Metrical Phonology*. Cambridge, MA: B. Blackwell, 1990.

[3] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," *Proc. IEEE ASRU Workshop*, Keystone, CO, 1999.

[4] E. Fosler-Lussier, S. Greenberg, and N. Morgan, "Incorporating contextual phonetics into automatic speech recognition," *Proc. Int. Congress of Phonetic Sciences*, San Francisco, CA, 1999.

[5] S. Greenberg, S. Chang, and J. Hollenback, "An introduction to the diagnostic evaluation of the Switchboard-corpus automatic speech recognition systems," *Proc. NIST Speech Transcription Workshop*, College Park, MD, 2000.

[6] D. McAllaster, L. Gillick, F. Scattone, and M. Newman, "Studies with fabricated Switchboard data: Exploring sources of model-data mismatch," *Proc. DARPA Workshop on Conversational Speech Recognition*, Lansdowne, VA, 1998.

[7] M. Riley and A. Ljolje, "Automatic generation of detailed pronunciation lexicons," in C.-H. Lee, F. K. Soong, and K. K. Paliwal, editors, *Automatic Speech and Speaker Recognition*. Boston: Kluwer Academic Publishers, 1996.

[8] M. Saraclar and S. Khudanpur, "Properties of pronunciation change in conversational speech recognition," *Proc. NIST Speech Transcription Workshop*, College Park, MD, 2000.

[9] L. Deng, G. Ramsay, and D. Sun, "Production models as a structural basis for automatic speech recognition,", *Speech Communication*, vol. 33, no. 2–3, pp. 93–111, Aug. 1997.

[10] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-articulator Markov models for speech recognition," *Proc. ISCA Workshop on ASR*, Paris, 2000.

[11] K. Kirchhoff, "Syllable-level desynchronisation of phonetic features for speech recognition," *Proc. ICSLP*, Philadelphia, 1996.

[12] S. L. Lauritzen, *Graphical Models*. Clarendon Press, Oxford, 1996.

[13] G. Zweig, *Speech Recognition Using Dynamic Bayesian Networks*. Ph.D. thesis, University of California, Berkeley, 1998.

[14] J. Bilmes and G. Zweig, "The Graphical Models Toolkit: An open source software system for speech and time-series processing," *Proc. ICASSP*, Orlando, 2002.

[15] T. Stephenson, M. Mathew, and H. Bourlard, "Modeling auxiliary information in Bayesian network based ASR," *Proc. Eurospeech*, Aalborg, Denmark, 2001.

[16] J. Bilmes, *et al.*, "Structurally discriminative graphical models for speech recognition," JHU CLSP Summer Workshop Final Report, 2001.

[17] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proc. ISCA Workshop on ASR*, Paris, 2000.

[18] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," *Proc. ICSLP*, Denver, CO, 2002.