Dynamic Evidence Models in a DBN Phone Recognizer*

William Schuler, Tim Miller, Stephen Wu, Andrew Exley

Department of Computer Science and Engineering
University of Minnesota - Twin Cities, Minnesota, USA
{schuler,tmill,swu,exley}@cs.umn.edu

Abstract

This paper describes an implementation of a discriminative acoustical model – a Conditional Random Field (CRF) – within a Dynamic Bayes Net (DBN) formulation of a Hierarchic Hidden Markov Model (HHMM) phone recognizer. This CRF-DBN topology accounts for phone transition dynamics in conditional probability distributions over random variables associated with *observed evidence*, and therefore has less need for *hidden* variable states corresponding to transitions between phones, leaving more hypothesis space available for modeling higher-level linguistic phenomena such syntax and semantics. The model also has the interesting property that it explicitly represents likely formant trajectories and formant targets of modeled phones in its random variable distributions, making it more linguistically transparent than models based on traditional HMMs with conditionally independent evidence variables. Results on the standard TIMIT phone recognition task show this CRF evidence model, even with a relatively simple first-order feature set, is competitive with standard HMMs and DBN variants using static Gaussian mixture models on MFCC features.

Index terms: Phone recognition, Dynamic Bayes Nets, Conditional Random Fields, dynamic evidence model, phone recognition, acoustic modeling

1. Introduction

Phone recognition is hard to do well in a manner that allows lexical, syntactic, and semantic information to also be integrated. Often discriminative approaches that do well by themselves do not extend well to larger models. This paper describes a preliminary implementation of a discriminative acoustical model within a Dynamic Bayes Net (DBN) formulation of a Hierarchic Hidden Markov Model (HHMM) [1] phone recognizer. This acoustical model has the interesting property that it explicitly represents both formant trajectories and formant targets in its random variable distributions: the former in the distribution $\mathsf{P}(\boldsymbol{o}_t \mid \boldsymbol{o}_{t-1}, Q_t)$ over the observed acoustical features $\boldsymbol{o}$ at frame t given each possible phone Q; and the latter in the distribution $\mathsf{P}(F_t^{\boldsymbol{O}} \mid \boldsymbol{o}_{t-1}, Q_{t-1})$ over binary 'final state' variables used in the DBN formulation of HHMMs. This acoustical model was intended to function as a component in a larger DBN-based interface that integrates phonology, syntax, and referential semantics into a single recognition process [2]. As a component in such a large system, the acoustical model was designed to avoid the need for large sets of context-dependent (e.g. triphone) values for hidden states at the phone level, which are now common in state-of-the-art transcription systems.

1.1. DBN-based Spoken Language Interface Model

The acoustical model described in this paper is defined to function within a structural language model, which explicitly represents syntactic constituents and semantics associated with these constituents in a linear-time Dynamic Bayes Net (DBN) recognizer [3]. This network is a variant of a Hierarchical Hidden Markov Model (HHMM) topology [1], which has been modified to encode a finite-stack right-corner parser.[1] The model is factored at each time step into a finite number of stack elements, each of which has a limited number of conditional dependencies, allowing complex patterns to be learned from relatively small amounts of data. Once trained, the model can be compiled into an efficient, unfactored HMM with only one hidden random variable per time step, by multiplying out all possible combinations of individual factored random variable values, then adopting a beam-search strategy in Viterbi decoding. Thus, any reduction of phonological (or syntactic, or semantic) uncertainty in this compiled hidden variable domain makes a correct interpretation less likely to fall off the beam.

The last two (Q and $\boldsymbol{o}$) levels of this larger interface model comprise the phone recognizer described in this paper, with the evidence ($\boldsymbol{o}$) and final-state ($F^{\boldsymbol{o}}$) variables comprising the acoustical model portion of this recognizer. The experiments described in Section 3 evaluate this acoustical model and phone recognizer on the standard TIMIT phone recognition task, and therefore assume no lexical- or higher-level input from above the phone level Q. Note that the random variables in the acoustical model are discriminative in that they are conditioned on evidence (which greatly simplifies decoding since only one input value need be considered rather than a whole distribution over a high-dimensional variable domain), but they are not strictly discriminative in that they are also conditioned on hidden variables as well (albeit ones with relatively small distributions over a relatively small phone set).

*This research was supported by National Science Foundation CAREER award 0447685, and by grants from the University of Minnesota Grant-In-Aid and Digital Technology Center Initiative Programs. The views expressed are not necessarily endorsed by the sponsors.

[1] This is simply the left-right dual of left-corner parsing used in compiler design. The advantage of the right-corner formulation is that it uses its stack to store recognized constituents rather than goal constituents, allowing semantics associated with these constituents to be remembered and used as antecedents of intra-sentential co-reference in subsequent constituents.

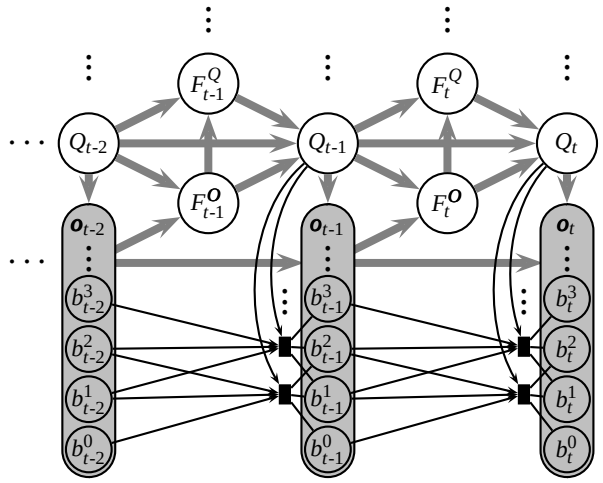

Figure 1: A graphical representation of the HHMM-based phone recognition model at three time steps (speech frames) t-2, t-1, and t, showing hidden random variables (Q) over phones or sub-phone states at each frame, instantiated evidence random variables ($\boldsymbol{o}$) over short-time FFT spectra at each frame (each subsuming a spectrum of convexity indicators $b^{1,2,3,\cdots}$), and boolean final-state variables (F^Q and $F^{\boldsymbol{O}}$) between frames, indicating whether lower-level variables can serve as final states for higher-level variables. Thick arcs represent ordinary conditional dependencies between random variables. Thin arcs and factor nodes (opaque boxes) show factor-specific dependencies in the evidence model, using an extension of factor graph notation [4] to conditional models. The set of nodes connected to a factor node by undirected edges represent maximal cliques of interdependent variables, conditioned on all of the nodes with directed arcs into the factor node. Ellipses at the top of the model denote optional interface with HHMM or other DBN language models.

2. Discriminative Models within Generative Models

The acoustical model described in this paper uses a dynamic evidence model which, unlike conventional HMMs, conditions the observed acoustical features at each time step on the acoustical features at the previous time step, in addition to the current hidden (phone) state (see Figure 1). This model maps high-dimensional inputs to high-dimensional outputs. In order to make this mapping learnable, we 1) restrict our evidence to vectors of only binary 'convexity' features on frequency domain, which highlight the formant tracks in the spectrum at various granularities; and 2) parameterize our model as a Conditional Random Field [5], which calculates probability distributions over ordered sequences (in our case, an array of spectral peak indicators at the current time step) given another ordered sequence as input (in our case, an array of spectral peak indicators at the previous time step) as a product of factors on local correlations at the same offset in the input and output sequences. Probability distributions over output sequences are then computed as a product of these factors normalized by the total probability of all possible output sequences, which can be calculated very efficiently using a dynamic programming algorithm similar to that used in HMM filtering. Conditional random fields are a good model structure for this problem because the dynamics of the speech signal are complex, and CRFs allow for heavy overlap of features without explicitly normalizing out that overlap.

One advantage of this approach is that it allows a dynamic representation of spectral information such as formant trajectories predicted by each phone, rather than a static representation of likely formant configurations at various sub-states of a phone, capturing the intuition that phones are defined not so much by a set of formants being *at* a particular configuration of frequencies, but rather by traveling *to* a particular configuration of frequencies. Similar approaches have been tried in connectionist models [6], which use Recurrent Neural Networks to simultaneously predict phones and acoustical features at each time step given the acoustical features at the previous time step. However, unlike most connectionist approaches, the DBN-based model described here also permits, in addition to formant trajectories, an explicit representation of formant *targets* in distributions associated with $F^{\boldsymbol{O}}$ random variables. These variables control whether lower-level Markov sub-models (in this case, the dynamic model of acoustical features) have reached a final state so that higher-level HMMs (in this case phones) can transition. This is simply a straightforward extension of the DBN formulation of HHMMs to cover evidence models, but intuitively this is motivated by the fact that for a sequence to be considered an instance of a particular phone, it is not enough for the formants to be on their way to a particular configuration, they also have to eventually arrive there.

2.1. Conditional Random Fields

Conditional Random Fields [5] are probabilistic models for structured prediction which estimate probabilities of complex output states as products of exponential weights on arbitrary overlapping features of evidence and output states, and then globally normalize these products over the entire space of possible output states. This allows a CRF model to capture more long-distance dependencies (in this case, of widely separated peaks of first and second formants in certain vowel sounds) than non-parametric conditional models such as Bayes Net factorizations, which locally normalize each component feature over its overlap with other components.

This system uses a CRF to model probability distributions over the observed evidence variables (describing short-time FFT spectra) of a larger DBN model used to recognize phonemes. We calculate this conditional probability distribution as:

$$\mathsf{P}_\lambda(\boldsymbol{o}_t \mid \boldsymbol{o}_{t\text{-}1}, Q_t) = \frac{\exp\left(\sum_{r,i,j} \lambda_{r,i,j}\, f_{r,i,j}(\boldsymbol{o}_{t\text{-}1}, \boldsymbol{o}_t, Q_t)\right)}{Z(\boldsymbol{o}_{t\text{-}1}, Q_t; \lambda)} \qquad (1)$$

where $\boldsymbol{o}_t$ is the spectral evidence vector at time t (see following section), Q_t is a realization of a subphone, and $Z(\boldsymbol{o}_{t\text{-}1}, Q_t; \lambda)$ is a normalization factor.

Note that this differs from many implementations that have used CRFs to model $\mathsf{P}_\lambda(Q_{1..T} \mid \boldsymbol{o}_{1..T})$, where $Q_{1..T}$ are hidden phones in a segment and $\boldsymbol{o}_{1..T}$ are evidence in that segment. The purpose in our CRF model is not to discriminatively estimate the phone, but to generate a conditional probability table for integration into the (generative) dynamic evidence DBN model described above. In other words, most CRFs estimate the probability of *hidden states* over a segment of time, whereas we estimate the probability of generating the next observed *evidence*.

With probabilities conditioned on a hidden variable, this evidence-to-evidence model is not strictly discriminative - which usually implies a large search space when the generating state is

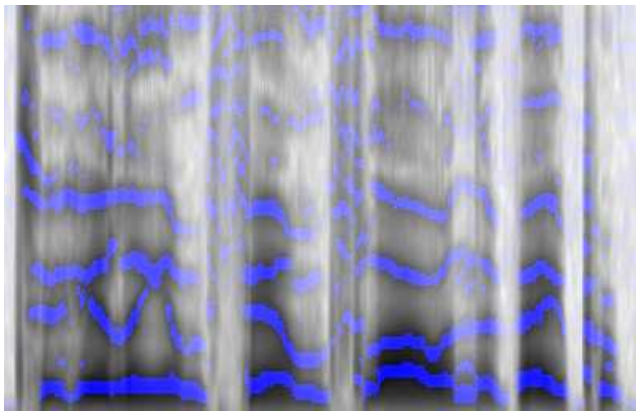

Figure 2: Convexity indicators at $r = 1$ on sample spectrogram for utterance fragment '…to helium film flow in the vapor …'

unknown (i.e. from hidden state to evidence in an HMM). However, since the portion of the generating state due to the $\boldsymbol{o}$ random variable is evidence, it is not necessary to iterate over each of $|\boldsymbol{O}||Q|$ possibilities, but only $|Q|$ (where $|\cdot|$ is the set size). Furthermore, since the generated state is also evidence, we only need to calculate one instance $\boldsymbol{o}_t$ of equation 1. Therefore, we have added a dynamic model to the evidence with little additional computational cost over a traditional HMM.

2.2. Feature Set

The features $f(\boldsymbol{o}_{t-1}, \boldsymbol{o}_t, Q_t)$ used in this work were binary values indicating the presence of $\cap$-convexities[2] (loosely, 'peaks') in the spectrum at any given time frame (see Figure 2). A $\cap$-convexity over an interval of discrete data produced by some function g is defined as

$$g[ci + (1-c)j] \geq cg[i] + (1-c)g[j] \tag{2}$$

Here, $i < j$ are both points in the domain of g; also, $c \in [0,1]$ is an arbitrary averaging factor such that $ci + (1-c)j$ is an integer. The implication is that the average value at the ends of an interval must be less than the midpoint of any two points on the interval.

Adapting this definition for each frequency point in our spectral data, we consider only the two nearest neighbors and introduce a threshold $\gamma \geq 0$ to reject convexities produced by noise. To obtain features that encode characteristics of the data at a variety of scales, we performed this convexity detection on different decimated versions of the spectrum. We lowpass-filtered the spectrum using 2^{r+1}-tap triangle filters, then decimated by $2^r, r = 0,..,5$, producing the spectra g_r. We then define binary convexity indicators at each frequency bin i as follows:

$$b^{r,i}(g) = \begin{cases} 1 & \text{if } (g_r[i] - g_r[i-1]) - (g_r[i+1] - g_r[i]) > \gamma \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The (also binary) features used in the CRF model are then defined on paired triples of adjacent binary convexity indicators at the current and previous frame:

$$f_{r,i,j}(\boldsymbol{o}_{t\text{-}1}, \boldsymbol{o}_t, Q_t) = \begin{cases} 1 & \text{if } j = \langle b^{r,i}(\boldsymbol{o}_{t\text{-}1}), \ldots, b^{r,i+2}(\boldsymbol{o}_{t\text{-}1}) \\ & \qquad b^{r,i}(\boldsymbol{o}_t), \quad \ldots, b^{r,i+2}(\boldsymbol{o}_t) \rangle \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The result is a feature set that is sensitive to upward- and downward- tending formant tracks at overlapping frequency bands at various granularities.

Decimation on time signals typically reduces the necessary bandwidth; in our case, decimation to a spectral signal rejects the noise at higher cepstral quefrencies. The resulting spectra have lower resolutions, which are useful for detecting characteristics like formants or frication. Combining the data from different levels of spectra, we have $\sum_{r=0}^{5} 2^{8-r} - 2(r+1) = 378$ features in our feature vector, where two endpoint convexities are undefined and unused in each r level.

This choice of features departs from typical Mel-frequency cepstral coefficient (MFCC) feature vectors for several important reasons. MFCCs aim to minimize the size of the feature vector and are known to produce good results with only about 12 cepstral coefficients (about 60 features overall). To implement this, the mel-frequency spectrum is organized into up to 40 nonlinear frequency bins.

Our approach relies on the observation that the formants in vowels increase or decrease monotonically over time towards some target configuration - movements which should be observed with high resolution. These criteria do not hold for the mel-frequency spectrum, so we maintain a linear scale for frequency bins. The linear scale also simplifies the hierarchy of the r level spectra, preserving more of the relevant data.

Another interesting point is that using convexity detection encodes a more general notion of 'peaks' than local maxima, e.g. distinguishing formants which are close together in the spectrum, one of which slightly dominates the other.

A final, most tangible benefit to using convexity detection is that binary-valued functions are compatible with the CRF model. Although the number of features in MFCCs is small, the features themselves are continuous; this complicates the formulation of the CRF evidence model, which can tractably perform normalization using dynamic programming only if there is a finite set of possible values to store and share.

2.3. Final-State Variables

The Murphy-Paskin formulation of Hierarchic HMMs defines boolean final-state variables at each hidden level d in the HMM hierarchy, which indicate whether the HMM at depth d can serve as a final state for the HMM above it (at depth $d-1$). This is done in order to ensure that the higher-level HMMs in the hierarchy transition only when the lower-level HMMs have concluded. The model described in this paper extends this formulation by introducing final-state variables at the evidence level as well, indicating whether the observed evidence can serve as a final state (i.e. formant target) for the lowest-level hidden state above it (over phones or sub-phones). In order to model formants that are sustained at a target configuration, all sub-phone states are allowed to self-transition with non-zero probability.

These evidence-level final-state variables $F^{\boldsymbol{O}}$ are implemented as single neuron models (equivalent to a degenerate CRF with a one-bit output sequence), which can be trained relatively quickly using gradient descent. Since distributions over these final-state

[2] The notation $\cap$-convex and $\cup$-convex help disambiguate confusing mathematical definitions. In this paper, 'convex' and 'convexity' exclusively refer to $\cap$-convex functions, so that formant 'peaks' are *convex*. Note that this is considered concave, not convex, in e.g. optimization theory.

random variables $\mathsf{P}(F_t^{\boldsymbol{O}} \mid \boldsymbol{o}_{t\text{-}1}, Q_{t\text{-}1})$ are conditioned on (rather than generating) the observed evidence $\boldsymbol{o}$, they may use whatever features of this evidence provide the most help, covering as many preceding frames as desired, as in any discriminatively trained model. However, for simplicity (and because of limited target-annotated data), the final-state models used in this implementation were defined only on the convexity indicator spectra generated by the evidence model at the immediately previous speech frame.

3. Evaluation

The test system was trained on the TIMIT corpus of phonetically transcribed continuous speech. Because it models phones as *culminating in* particular formant targets, the dynamic evidence model defined above dictates an approach to annotation that differs from that used in the TIMIT corpus, in which sonorant phone labels are placed *around* the formant target, with the formant target in the center.

To make the TIMIT annotation compatible with our model, a modified training corpus was constructed in which sonorant segments were shifted backward by half the length of the corresponding segment in the original TIMIT transcript. These automatically aligned phone targets were then manually checked and adjusted in the DR1 subset of the TIMIT training set.

This need to model formants as monotonically increasing or decreasing toward a target during each annotated phone segment also motivated 1) a decomposition of diphthongs into start and end phones (which were approximated to the existing set of monophthong sonorants), and 2) the introduction of explicit stop onsets, in which sonorant formants would converge in a predictable manner before a plosive or other closure began.

After training on the formant-target-aligned DR1 subset and testing on the entire TIMIT corpus, the CRF-DBN model achieved phone recognition accuracy of 59% on the standard TIMIT test,[3] with a 54% phone error rate (computed as the sum of substitutions, insertions, and deletions). This compares to previously published phone error rates for similar approaches of 47% [7] and 46% [8] for context-independent phone recognition:

Method	Corr	Subs	Del	Ins	PER
HMM(L&M)	na	na	na	na	42%
SFHMM(L&M)	na	na	na	na	46%
Sphinx(L&H)	64%	26%	10%	11%	47%
CRF-DBN	59%	28%	13%	13%	54%
LFHMM(L&M)	na	na	na	na	71%

4. Conclusion

This paper has presented a novel acoustical model in which probability distributions over acoustical evidence, abstracted as discrete spectra of boolean convexity indicators, can be efficiently dynamically estimated using CRFs given a hypothesized phone target and any number of preceding observed spectra. Tractably estimating probability distributions over high-dimensional evidence variable domains, using dynamic programming in a CRF, requires that these domains be made discrete, weakening the sensitivity of the model. In particular, the convexity indicator spectra described in this paper make the model almost completely insensitive to relative magnitudes of spectral peaks (limited to the convexity threshold γ). Nevertheless, the model described in this paper performs competitively with conventional, static MFCC-based HMM approaches under similar conditions, suggesting that it is mostly the location of spectral peaks, and not their relative magnitudes, which is phonologically salient. This model also has a number of potential advantages over MFCC-based HMM or RNN approaches:

- it is a well-formed probability model that can be extended naturally to subsume more complex Hierarchic HMM or other DBN language models without thresholding or ignoring dependency assumptions;
- it allows both the evidence and final-state distributions ($\mathsf{P}(\boldsymbol{o}_t \mid \boldsymbol{o}_{t\text{-}1}, Q_t)$ and $\mathsf{P}(F_t^{\boldsymbol{o}} \mid \boldsymbol{o}_{t\text{-}1}, Q_{t\text{-}1})$) to be summarized using a probability vector of linear size on $|Q|$, allowing a clean separation of computation-intensive CRF inference in a networked implementation;
- it accounts for phone transition dynamics in the evidence model, and therefore may eliminate the need for sub-phone states in the hidden variable model, leaving more hypothesis space for higher-level linguistic phenomena such as syntax and semantics;
- it is relatively transparent (and thus relatively easy to extend), in that parameter weights in the evidence ($\boldsymbol{o}$) and final-state ($F^{\boldsymbol{O}}$) models correspond to linguistic intuitions about where formants should be, whereas parameters of neural nets or Gaussian mixture models in static HMMs are often relatively opaque to linguistic interpretation;
- and finally, the fact that this model achieves competitive recognition results using a very different feature set from MFCCs suggests that exploring a hybrid approach might be an attractive avenue of research.

[3] Except for the addition of stop onsets and the decomposition of diphthongs, as noted above.

5. References

[1] Kevin P. Murphy and Mark A. Paskin, "Linear time inference in hierarchical HMMs," in *Proceedings of Neural Information Processing Systems*, 2001, pp. 833–840.

[2] William Schuler and Tim Miller, "Integrating denotational meaning into a dbn language model," in *Proceedings of Eurospeech/Interspeech*, Lisbon, Portugal, 2005.

[3] Tom Dean and Keiji Kanazawa, "A model for reasoning about persitence and causation," *Computational Intelligence*, vol. 5, no. 3, pp. 142–150, 1989.

[4] Frank R. Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[5] John Lafferty, Andrew McCallum, and Fernando Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001, pp. 282–289.

[6] Tony Robinson, "An application of recurrent nets to phone probability estimation," in *IEEE Transactions on Neural Networks*, 1994.

[7] Kai-Fu Lee and Hsiao-Wuen Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, 1989.

[8] Beth Logan and Pedro Moreno, "Factorial hmms for acoustic modeling," in *Proceedings ICASSP*, 1998, pp. 813–816.