A Generative Modeling Framework for Structured Hidden Speech Dynamics

Li Deng, Dong Yu, and Alex Acero
Microsoft Research, One Microsoft Way, Redmond WA 98052
{deng,dongyu,alexac}@microsoft.com

Abstract

We outline a structured speech model, as a special and perhaps extreme form of probabilistic generative modeling. The model is equipped with long-contextual-span capabilities that are missing in the HMM approach. Compact (and physically meaningful) parameterization of the model is made possible by the continuity constraint in the hidden vocal tract resonance (VTR) domain. The target-directed VTR dynamics jointly characterize coarticulation and incomplete articulation (reduction). Preliminary evaluation results are presented on the standard TIMIT phonetic recognition task, showing the best result in this task reported in the literature without using many heterogeneous classifier combinations. The pros and cons of our structured generative modeling approach, in comparison with the structured discriminative classification approach, are discussed.

1 Introduction

Despite significant progress in automatic speech recognition (ASR) technology based on hidden Markov modeling (HMM), some fundamental and practical limitations in the technology have hindered its widespread use. Many leading researchers in the field understand the fragile nature of the current ASR system design, and have advocated new, serious research needed to overcome the basic limitations of the current, HMM-based ASR technology (e.g., [1, 8, 3, 4, 10, 12]).

The strengths of the HMM for ASR are well known, so are its weaknesses [11, 9, 10, 5]. Two major research directions are currently being pursued in the ASR research community to overcome these weaknesses. One of them is the structured classification approach where global speech feature dependency is embedded in the direct discriminative learning and its potential success will be heavily dependent on the success of "feature engineering" yet to be demonstrated. Research along this direction has been summarized in the recent paper [10]. The other direction, as is the focus in this paper (as well as in other recent publications such as [1]), overcomes the same weak feature dependency and correlation problem associated with the HMM from a different perspective. In our approach, we directly construct a detailed probabilistic generative model that embeds the underlying structured speech dynamics not captured by the HMM. Such hidden dynamics provide long-span contextual dependency, but due to the powerful constraints in speech production, compact parameterization is achieved that successfully eliminates the contextual enumeration/clustering procedure in the HMM approach.

2 Dynamic Speech Modeling — Literature Overview

As a linguistic and physical abstraction, human speech generation can be functionally represented at four distinctive but correlated levels of dynamics — phonological level, "task" level, articulatory level, and acoustic level. Many different types of computational dynamic models for speech generation in the literature that will be presented in this section will be organized in view of these functional levels of the dynamics. To make this overview most relevant to our specific HTM implementation, we will classify the models into two main categories. In the first category are the models focusing on the lowest, acoustic level of dynamics (which is also the most peripheral level for human or computer speech perception). This class of models is often called the stochastic segment models as are well known through the earlier review paper [9]. The second category consists of what is called the *hidden dynamic model* where the task dynamic and articulatory dynamic levels are functionally grouped into a single level. In contrast to the acoustic-dynamic model which represents coarticulation at the surface, observational level, the hidden dynamic model explores a deeper, unobserved (hence "hidden") level of the speech dynamic structure that regulates coarticulation and phonetic reduction.

A comprehensive review including sub-classifications of the major types of probabilistic generative models within each of these two categories will be presented.

3 Hidden Trajectory Modeling with Target Filtering and Nonlinear Cepstral Prediction

As a special type of the hidden dynamic model, the HTM presented in this section is a structured generative model, from the top level of phonetic specification to the bottom level of acoustic observations via the intermediate level of (non-recursive) FIR-based target filtering that generates hidden VTR trajectories. This section is devoted to mathematical formulation of the HTM as a probabilistic generative model. Four key aspects of mathematical formulation and model parameterization to be presented are: 1) Generating stochastic hidden VTR trajectories from the VTR target sequence; 2) Generating acoustic observation data (cepstra); 3) Linearizing cepstral prediction function; and 4) Computing acoustic likelihood by marginalizing over VTR uncertainty.

The parametric form of stochastic target filtering is

$$\boldsymbol{z}(k) = h_{s(k)} * \boldsymbol{t}(k) = \sum_{\tau=k-D}^{k+D} c_{\boldsymbol{\gamma}} \boldsymbol{\gamma}_{s(\tau)}^{|k-\tau|} \boldsymbol{t}_{s(\tau)}, \tag{1}$$

where $h_{s(k)}$ is the impulse response function of a non-causal FIR filter, and $c_{\boldsymbol{\gamma}}$ is the normalization constant that ensures target undershooting for fast speech and avoids target overshooting. The segmental input target (random) vector values $\boldsymbol{t}_{s(\tau)}$ typically takes not only those associated with the current home segment, but also those associated with other segments in the speech utterance, with exponentially decaying weights characterized by segment-dependent vector $\boldsymbol{\gamma}_{s(\tau)}$. The latter case happens when the time τ in (1) goes beyond the home segment's boundaries; i.e., when the segment $s(\tau)$ occupied at time τ switches from the home segment to adjacent segments.

The parametric form for the observation cepstral data generation is

$$\boldsymbol{o}(k) = \mathcal{F}[\boldsymbol{z}(k)] + \boldsymbol{\mu}_{r_{s(k)}} + \boldsymbol{v}_s(k),$$

where the observation "noise" $\boldsymbol{v}_s(k) \sim \mathcal{N}(\boldsymbol{v}_s; \mathbf{0}, \boldsymbol{\Sigma}_{r_{s(k)}})$. The each component of the

vector-valued nonlinear mapping function above is

$$\mathcal{F}_j(k) = \frac{2}{j} \sum_{p=1}^{P} e^{-\pi j \frac{b_p(k)}{f_s}} \cos(2\pi j \frac{f_p(k)}{f_s}), \tag{2}$$

where f_s is the sampling frequency, P is the highest VTR order, and j is the cepstral order.

4 Phonetic Recognition Experiments

We have carried out phonetic recognition experiments to evaluate the HTM. The experiments are based on the standard phonetic recognition task on the core test set of TIMIT as described in [6]. Overall, the lattice-based HTM system (75.07% accuracy) gives 13% fewer errors than our HMM system implemented by HTK. (71.43% accuracy). This performance is also better than any other HMM systems in other labs worldwide on the same task as summarized in [6].

Table 1: *TIMIT phonetic recognition performance comparisons between an HMM system and three versions of the HTM system. HTM-1: N-best rescoring with HTM scores only; HTM-2: N-best rescoring with weighted HTM, HMM, and LM (Language Model) scores; HTM-3: Lattice-constrained A* search with weighted HTM, HMM, and LM scores. Identical acoustic features (frequency-warped LPC cepstra) are used.*

	Acc %	Corr %	Sub %	Del %	Ins %
HMM	**71.43**	73.64	17.14	9.22	2.21
HTM-1	**74.31**	77.76	16.23	6.01	3.45
HTM-2	**74.59**	77.73	15.61	6.65	3.14
HTM-3	**75.07**	78.28	15.94	5.78	3.20

5 Summary and Discussion

Modeling dynamic structure of speech is a novel paradigm in speech recognition research within the generative modeling framework, and it offers a potential to overcome limitations of the current hidden Markov modeling approach. Analogous to structured language model where syntactic structure is exploited to represent long-distance relationships among words [2], the structured speech model described in this paper makes use of the dynamic structure in the hidden VTR space to characterize long-span contextual influence among phonetic units. An general overview is provided on hierarchically classified types of dynamic speech models in the literature. A detailed account is then given for a specific model type, HTM, and we outline model construction and the parameter estimation algorithms in this extended abstract. In the full paper, we show how the use of resonance target parameters and their temporal filtering enables joint modeling of long-span coarticulation and phonetic reduction effects that the conventional HMM is unable to achieve. Experiments on phonetic recognition evaluation demonstrate superior recognizer performance over a modern HMM-based system. Error analysis shows that the greatest performance gain occurs within the sonorant speech class.

In the recently proposed structured classification approach [10, 7], the deficiency of the HMM is overcome by direct discriminative learning, replacing the need for a probabilistic generative model by the assumption that the conditional class distribution is an exponential one on flexibly selected "features". Since the conditioning is made on the feature sequence and the "features" can be designed with long-contextual-span properties, the conditional independence assumption made in HMM can be conceptually removed — provided that right

"features" can be designed. In contrast, the structured speech modeling approach presented in this paper generalizes the HMM by embedding long-contextual-span properties directly into the generative model's structure. Compared with structured discriminative classification approach, our generative modeling technique enables easier incorporation of scientific principles governing speech coarticulation, speaker/environment variation, phonetic reduction, and pronunciation variation. Systematic knowledge accumulated spectrogram reading can also be applied to generative models in a conceptually clean manner, as we have done in our HTM where target-guided formant/VTR transition forms the basis of the model structure. Generative models also have added advantages for conceptually straightforward model/feature/results analysis, for verification/evaluation of assumptions, and for diagnosis of model implementation. On the other hand, generative models may suffer from the possibility of focusing on non-essential aspects of the speech process when the sole purpose is discrimination. Such "redundant" work may be avoided in the discriminative classification framework. Further, in addition to the difficulty of using non-homogeneous features, the implementation constraints of complex generative models may force model simplification that leads to performance degradation. How to combine the advantages of the structured discriminative classification framework and the structured probabilistic generative modeling framework for advancing the current ASR technology is an important future research direction.

References

[1] J. Bilmes and C. Bartels. "Graphical model architectures for speech recognition," *IEEE Signal Processing Magazine*, Vol. 22, No. 5, Sept. 2005, pp. 89-100.

[2] C. Chelba and F. Jelinek. "Structured language modeling," *Computer Speech and Language*, October 2000, pp. 283-332.

[3] L. Deng, K. Wang, and W. Chou. "Speech Technology and Systems in Human-Machine Communication —Guest editors' editorial," *IEEE Signal Processing Magazine*, Vol. 22, No. 5, Sept. 2005, pp. 12-14.

[4] L. Deng and X.D. Huang. "Challenges in adopting speech recognition," *Communications of the ACM*, Vol. 47, No. 1, January 2004, pp. 69-75.

[5] L. Deng and Doug O'Shaughnessy. SPEECH PROCESSING —A Dynamic and Optimization-Oriented Approach, Marcel Dekker Inc., New York, 2003.

[6] J. Glass. "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language*, Vol. 17, No. 2-3, pp. 137-152.

[7] A. Gunawardana, M. Mahajan, A. Acero, and J. Platt. "Hidden Conditional Random Fields for Phone Classifi cation", *Proc. Interspeech*, Lisbon, Sept 2005.

[8] C.-H. Lee. "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next-generation automatic speech recognition, *Proc. ICSLP*, Jeju Island, October, 2004, pp. 109-111.

[9] M. Ostendorf, V. Digalakis, and J. Rohlicek. "From HMMs to segment models: A unifi ed view of stochastic modeling for speech recognition" *IEEE Trans. Speech Audio Proc.*, Vol. 4, 1996, pp. 360-378.

[10] F. Pereira. "Linear models for structure prediction", *Proc. Interspeech*, Lisbon, Sept 2005, pp. 717-720.

[11] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[12] E. Shriberg. "Spontaneous speech: How people really talk and why engineers should care", *Proc. Interspeech*, Lisbon, Sept 2005, pp. 1781-1784.

[13] G. Zweig. "Bayesian network structures and inference techniques for automatic speech recognition", *Computer Speech and Language*, Vol. 17, No. 2-3, 2003, pp. 173-193.