INITIAL EVALUATION OF HIDDEN DYNAMIC MODELS ON CONVERSATIONAL SPEECH[1]

J. Picone, S. Pike, R. Regan, T. Kamm, J. Bridle, L. Deng, Z. Ma, H. Richards, M. Schuster

1998 Workshop on Language Engineering
Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218-2686
ws98dynam@mail.clsp.jhu.edu

ABSTRACT

Conversational speech recognition is a challenging problem primarily because speakers rarely fully articulate sounds. A successful speech recognition approach must infer intended spectral targets from the speech data, or develop a method of dealing with large variances in the data. Hidden Dynamic Models (HDMs) attempt to automatically learn such targets in a hidden feature space using models that integrate linguistic information with constrained temporal trajectory models. HDMs are a radical departure from conventional hidden Markov models (HMMs), which simply account for variation in the observed data. In this paper, we present an initial evaluation of such models on a conversational speech recognition task involving a subset of the SWITCHBOARD corpus. We show that in an N-Best rescoring paradigm, HDMs are capable of delivering performance competitive with HMMs.

1. INTRODUCTION

Hidden dynamic models [1,2] (HDMs) attempt to produce acoustic likelihoods of phone-level sound units that reflect intended spectral configurations rather than likelihoods based on the actual realization of the sound in the speech data. This is a radical departure from current statistical modeling approaches that attempt to account for variation in the data by accumulating large numbers of Gaussian mixture components. It is conjectured that this approach will produce more consistent acoustic scoring for conversational speech, because sounds are rarely fully articulated in such data. Tremendous amounts of variation are observed in the speech data because of the manner in which the realization of a sound was truncated is highly context-dependent. It is the goal of this work to produce acoustic scores that reflect measurements in the hidden (or target) space, rather than directly in the feature space as is currently done in context-dependent phonetic modeling.

The work presented here was the culmination of an intense effort at the 1998 NSF Workshop on Language Engineering held at the Center for Language and Speech Processing at Johns Hopkins University. One goal of this work, which is the primary focus of this paper, was to evaluate the HDM approach on a credible conversational speech recognition task involving the SWITCHBOARD (SWB) Corpus [3].

2. HIDDEN DYNAMIC MODELS

The models presented here consist of two separate components which accommodate separate sources of speech variabilities. The first component is a smooth dynamic one, linear but nonstationary. The nonstationarity is described by a sequence of segments each corresponding to a phonological unit (phones). The second component is static and non-linear. This component handles other types of variabilities (lower-level). The two components combined form a nonstationary, non-linear dynamic system whose structure and properties are well understood in terms of the general process of human speech production. An overview of this approach is given in Figure 1.

2.1. Deterministic Hidden Dynamic Models (DHDMs)

The dynamic system in the DHMD approach [4] is basically a low-pass filter operating on components of

1. This material is based upon work supported by the National Science Foundation (NSF) under Grant No. (#IIS-9732388), and was carried out at the 1998 Workshop on Language Engineering, Center for Language and Speech Processing, Johns Hopkins University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF or The Johns Hopkins University.

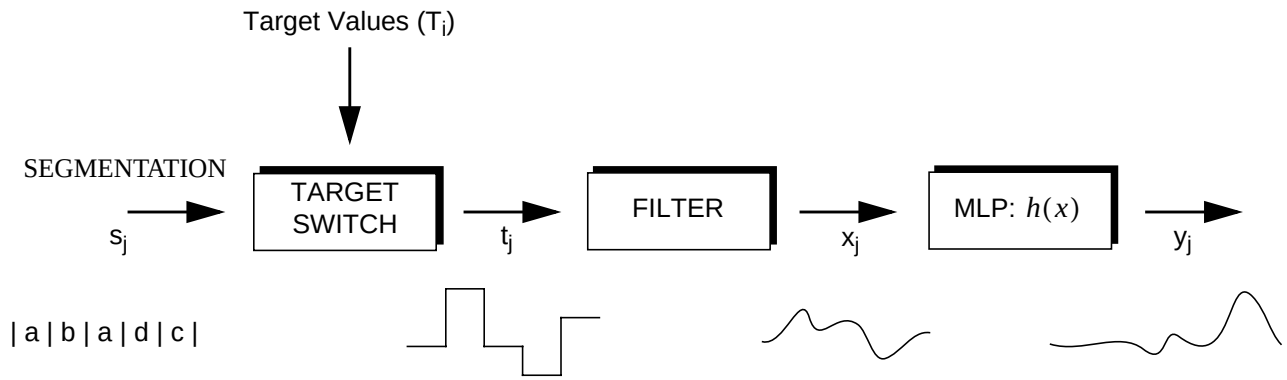

Figure 1: An overview of the hidden dynamic model (HDM) approach. Acoustic likelihood computations are performed in a hidden space that is inferred from standard recognition features (such as mel-frequency cepstra coefficients).

the target values, with time-constants as specified by the current segment. Each phone type specifies a vector of target values and a vector of time-constants. The variable low-pass filter is symmetrical so that the center of transitions occurs at phone boundaries, to agree with normal phonetic marking practice.

The output non-linearity is a single MLP, and the criterion for model training (and for recognition) is weighted mean-square error in the smoothed log power spectrum. (This is proportional to a log-likelihood, assuming a generative stochastic model with a simple Gaussian distribution at the acoustic level.) The weights of the MLP are optimized at the same time as the targets and time-constants, using a form of gradient descent. The derivatives with respect to the acoustic error are back-propagated through the MLP and through the filter to the target values and the time-constants.

2.2. Statistical Hidden Dynamic Models (VTRs)

In this approach, we use a statistical nonlinear dynamic system model [2] to describe vocal tract resonances (VTRs) dynamics. The VTRs are pole locations of the VT configured to produce speech sounds, and have acoustic correlates of formants which are directly measurable for vowel/glide sounds, but often are hidden or perturbed for consonantal sounds due to the concurrent spectral zeros and turbulence noises. A noisy, causal and linear dynamic system is used to describe the VTR dynamics. The output nonlinearity is multiple, switching MLPs, with each MLP associated with distinct manner of articulation of a phone. The criterion for model training (and for recognition) is maximum likelihood on MFCCs only (not on VTRs).

2.3. Significant Contrasts

The VTR and DHMD approaches, though similar in many aspects, differ significantly in a few key areas. First, the VTR system uses formant-like resonances as its internal model, while the HDM uses an unconstrained hidden state. Second, error computations for the VTR model are based on internal model "strain," while the DHDM uses a deterministic model with a single Gaussian output model. Third, the VTR system uses a goal-based speech production theory and a causal second-order low-pass filter to constrain its dynamics, while the DHDM system uses a zero-phase second-order low-pass filter. Finally, the VTR system uses a generalized EM with an extended Kalman filter while the DHDM uses a single conjugate gradient descent algorithm. Though these systems differ in their implementations, the overall philosophies are quite similar, and hence, they can be evaluated in a common framework as presented below.

3. N-BEST RESCORING EXPERIMENTS

An overview of our evaluation paradigm is shown in Figure 2. HDM systems ultimately can be viewed as performing an enhanced likelihood computation in the acoustic space. By replacing the standard likelihood computation by a new one from the HDM system, we can easily insert this system into a conventional N-best rescoring paradigm. The essential inputs to the HDM system are an N-best list with phone-level time alignments, and the corresponding speech data. The HDM systems rescore the sequence of phones in this time alignment, and produce an overall sentence

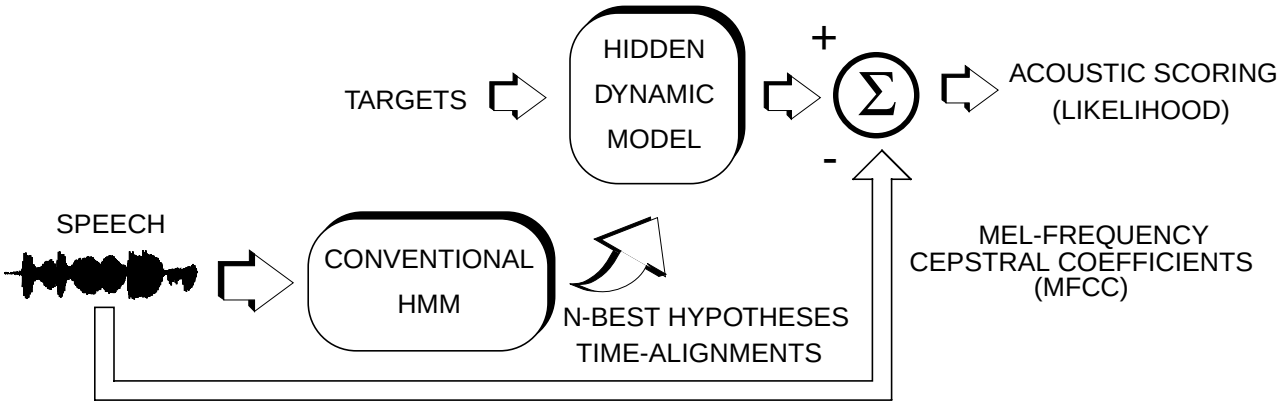

Figure 2: An overview of the N-best rescoring paradigm used to evaluate the HDMs on conversational speech. A conventional context-dependent HMM system was used to generate N-best lists, and the HDM systems rescored these lists.

likelihood. Note that in this study the HDM systems do not realign the transcription.

There are some practical reasons why we chose this limited rescoring methodology. First, integration of the HDM rescoring module into a lattice rescoring paradigm is difficult (and this is the only thing that is practical for SWB evaluations). Computational requirements prohibited large-scale rescoring experiments. Second, we ignored language model (LM) scores in order to focus on acoustic modeling issues. Integration of LM scores is a research topic in itself. Rather than deal with this complex issues in this study, we chose to focus solely on improvements in acoustic scoring. Any improvements due to LM effects, should however be equally applicable to HDM systems. Finally, no conventional speaker adaptation or normalization algorithms were used. Instead, for some experiments, a simple frequency warping method of speaker normalization was used that was anticipated to be sufficient for speakers of the same sex. Effective speaker normalization is important to the HDM model, but not something extensively explored in this study.

A plausible solution to these constraints was to select the male speaker subset [6] from the WS'97 DevTest, and to reserve 10 utterances from each test speaker for adaptation. This resulted in 1241 utterances consisting of 23 speakers, 24 conversation sides, and approximately 50 minutes of speech. We used a baseline context-dependent phone HMM system [6] to generate N-best lists and time alignments for the reference transcription and the 100-best hypotheses. The HDM systems rescored these hypotheses, and the resulting sentence hypotheses were scored using standard NIST scoring software and presented in terms of word error rate (WER). In Figure 3, we demonstrate that the 100-best lists used in this study are sufficiently rich, in that the overall WER can be reduced from 52% to 32% if the best sentence hypothesis is always selected.

We evaluated the HDM systems under three conditions: ref+5 — selecting from the reference transcription and the top 5 most likely hypotheses; 5-best — selecting from only the top 5 hypotheses (performance is expected to be worse), and 100-best. The latter condition is very close to a realistic rescoring experiment. Note that the HMM systems are handicapped in these evaluations since the hypotheses being rescored are highly confusable for the HMM systems, but not necessarily for the HDM systems.

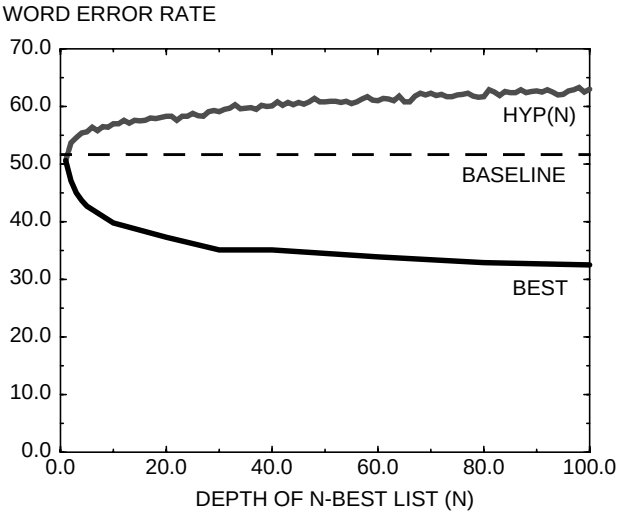

Figure 3: A demonstration of the richness of the N-best lists used in our rescoring experiments.

1241 Male Speaker SWB Subset			
System	**Ref+5**	**5-Best**	**100-Best**
Bounds:			
Oracle	0.0	42.7	32.5
Chance:	45.0	54.0	60.2
HMM:			
Baseline	48.2	52.0	56.9
Syllable	40.1	50.9	54.6
Small	44.8	52.6	58.9
DHDM:			
6 dims., 40 hidden units	32.7	52.6	59.4
standard, no warping	34.7	53.0	n/a
VTR:			
warping:			
low variance	32.4	54.4	60.7
mid variance	32.2	54.5	59.7
high variance	33.1	54.8	61.2
no warping:			
low variance	37.3	54.1	60.3
mid variance	32.3	54.3	60.9
high variance	32.2	54.6	61.0

Table 1: A analysis of word error rate (WER) for an N-best rescoring task involving 1241 male speakers from DevTest'97. The HDMs, when exposed to the reference transcription, outperform comparable HMMs and chance.

The results of these evaluations are presented in Table 1.

The first two lines in Table 1 represent bounds on performance: Oracle refers to always choosing the transcription with the lowest WER; Chance refers to randomly choosing from amongst the alternatives. We see that the HDM systems perform comparable to chance on the 100-best evaluation, and perform significantly better than chance when exposed to the reference transcription. The VTR and HDM systems are fairly close in performance, and slightly inferior to the HMM systems when not exposed to the reference transcription.

We have also included in these evaluations an HMM system significantly different than the one used to generate the baseline results to increase the credibility of these results. Further, an HMM system, labeled "Small" in Table 1. This system was trained on exactly the same material as the HDMs — a small subset of the SWB training database consisting of one speaker for whom there was the most amount of data.

4. SUMMARY

We believe the HDM systems presented in this paper demonstrate superior performance when exposed to the reference transcription. This demonstrates the promise of the HDM approach, and underscores the importance of time realignment of the N-best hypotheses for the HDM system. Further research in this direction is planned as follow-on work to WS'98.

We are grateful to Dr. George R. Doddington for his valuable insights into the evaluation paradigm, and Professor Mari Ostendorf for her useful suggestions on improving the performance of the VTR model.

5. REFERENCES

[1] R. Bakis, "Coarticulation modeling with continuous-state HMMs," *Proc. of the IEEE Workshop on Automatic Speech Rec.*, pp. 20-21, Arden House, New York, November 1991.

[2] L. Deng, "A Dynamic, Feature-Based Approach to Speech Modeling and Recognition," *Proc. of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 107-114, Santa Barbara, CA, USA, Dec. 1997.

[3] N. Deshmukh, A. Ganapathiraju, J. Hamaker, and J. Picone, "Resegmentation of Switchboard," to be presented at the International Conference on Spoken Language Processing, Sydney, Australia, November 1998.

[4] H.B. Richards and J. S. Bridle, "The HDM: A Segmental Hidden Dynamic Model of Coarticulation," submitted to the 1999 Int. Conf. on Acoustics, Speech, and Signal Processing, Phoenix, Arizona, USA, March 1999.

[5] J. M. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications, and Control*, Prentice Hall, New Jersey, USA, pp. 561, 1995.

[6] A. Ganapathiraju, V. Goel, J. Picone, A. Corrada, G. Doddington, K. Kirchoff, M. Ordowski, and B. Wheatley, "Syllable - A Promising Recognition Unit for LVCSR," *Proc. of the IEEE Automatic Speech Rec. and Underst. Workshop*, pp. 207-214, Santa Barbara, California, USA, December 1997.