A HIDDEN TRAJECTORY MODEL WITH BI-DIRECTIONAL TARGET-FILTERING: CASCADED VS. INTEGRATED IMPLEMENTATION FOR PHONETIC RECOGNITION

Li Deng, Xiang Li, Dong Yu, and Alex Acero

Microsoft Research, One Microsoft Way, Redmond WA 98052, USA

{deng, t-xli, dongyu, alexac}@microsoft.com

ABSTRACT

We present a novel acoustic model of speech, based on statistical hidden trajectory modeling (HTM) with bi-directional vocal tract resonance (VTR) target filtering, for speech recognition. The HTM consists of two stages of the generative process of speech: from the phone sequence to VTR dynamics and then to the cepstrum-based acoustic observation. Two types of model implementation are detailed, one with straightforward two-stage cascading, and another which integrates over the statistical distribution of VTR in model construction and in computing acoustic likelihood. With the use of first-order Taylor series approximation to the nonlinearity in the VTR-to-cepstrum prediction component of HTM, the acoustic likelihood is established in an analytical form. It is a Gaussian with the time-varying mean that gives structured long-span context dependence over the entire utterance, and with the dynamically adjusted variance proportional to the squared "local slope" in the nonlinear mapping function from VTR to cepstrum. When the HTM parameters are trained via maximizing this "integrated" likelihood, dramatic reduction of an upper error bound is achieved in the standard TIMIT phonetic recognition task using a large-scale N-best rescoring paradigm.

1. INTRODUCTION

Modeling hidden dynamics in the temporal structure of human speech has been a salient theme in recent speech recognition research (e.g., [1, 2, 3, 6, 8, 7, 9, 10, 11]) providing a potential to overcome fundamental limitations of the HMM, especially those related to recognizing highly reduced, spontaneous speech. One specific type of such modeling approaches is exemplified by the *hidden trajectory model* (HTM), where the hidden dynamics take parametric forms of temporal functions defined in a non-recursive manner. This offers implementational advantages over the recursive forms of hidden dynamic models (e.g., [2, 3, 9]). In the earlier work on HTM, various parametric forms of temporal functions with the properties of target-directedness and of uni-directional coarticulation have been proposed and positively evaluated [3, 11]. Two significant extensions of the earlier HTM have been recently developed and will be reported in this paper. First, the uni-directional coarticulation model in the vocal tract resonance (VTR) hidden space is extended to the bi-directional model via finite-impulse response (FIR) filtering of both forward and backward VTR targets. This overcomes the heuristic boundary-shift rule used in [11] for handling bi-directional coarticulation within the framework of uni-directional, target-directed hidden trajectory modeling. Second, compared with the HTM of [11] where the mapping function from the hidden VTR space to the observed acoustic space was implemented via a mixture of linear functions with a large number of trainable parameters, the new model presented in this paper exploits an analytical nonlinear mapping function developed in our recent work of [5], offering more precise and yet more parsimonious account for the speech dynamics in the observed acoustic (cepstral) domain.

Some detailed analyses of coarticulatory properties, including phonetic reduction, as exhibited by the bi-directional target-filtering HTM were presented recently in [4], where scientific evidence supporting the underlying concept of the model was provided. The focus of the current paper is on ways of implementing this HTM for the purpose of automatic speech recognition using the measured cepstral features. We have implemented two versions of the model, one with straightforward cascading of two stages in the model; i.e., passing the output of the VTR trajectory model (stage I) directly as input to the VTR-to-cepstrum mapping function (stage II), and another which integrates the two stages of the model in computing the likelihood of acoustic observations.

The organization of this paper is as follows. In Sec. 2, the HTM consisting of two stages of the speech generative process is outlined. Two ways of model implementation, cascaded one and integrated one, are presented in Secs. 3 and 4, respectively. HTM parameter training is described in Sec. 5. We provide experimental results in Sec. 6 on a standard TIMIT phonetic recognition task based on N-best rescoring, which demonstrates significant advantages of the integrated HTM implementation.

2. THE HIDDEN TRAJECTORY MODEL WITH BI-DIRECTIONAL TARGET FILTERING

2.1. Model stage I

Stage I of the novel HTM presented in this paper is responsible for converting a sequence of VTR targets with discrete jumps at the phone segments' boundaries into a smooth dynamic pattern (i.e., trajectory) across all these boundaries. Forward as well as backward coarticulation occurs when the bi-directional filtering and smoothing process makes the VTR value at each time dependent on not only the VTR target at the current phone, but also the VTR targets from the adjacent phones.

The HTM developed in this work gives quantitative prediction of the magnitude of contextually assimilated reduction. It is constructed using a slowly time-varying, FIR filter characterized by the non-causal, vector-valued, impulse response function of

$$h_s(k) = c\gamma_{s(k)}^{k}, \qquad \text{for } 0 < k < D \tag{1}$$

Xiang Li (xiangl@cs.cmu.edu) was a summer student intern at Microsoft Research from Carnegie Mellon University.

and symmetric for $-D < k < 0$, where c is a normalization constant. $\gamma_{s(k)}$ is the segment-dependent "stiffness" parameter vector, one component for each resonance. D in (1) is the unidirectional length of the impulse response, representing the temporal extent of coarticulation.

Given the filter's impulse response and the input to the filter as the segmental VTR target sequence $T(k)$, the filter's output as the model's prediction for the VTR trajectories is the convolution between these two signals. The result of the convolution within the boundaries of the home segment s is

$$z_s(k) = h_{s(k)} * T(k) = \sum_{\tau=k-D}^{k+D} cT_{s(\tau)}\gamma_{s(\tau)}^{|k-\tau|}, \qquad (2)$$

where the input target vector's value and the filter's stiffness vector's value typically take account not only those associated with the current home segment, but also those associated with the adjacent segments.

2.2. Model stage II

Stage II of the HTM converts the VTR vector $z(k)$ at each time frame k into a corresponding vector of LPC cepstra $\boldsymbol{o}(k)$. Thus, the smooth dynamic pattern of $z(k)$ as the output from Stage I is mapped to a dynamic pattern of $\boldsymbol{o}(k)$.

To describe this mapping function, we decompose the VTR vector into a set of K resonant frequencies $f = (f_1, f_2, ..., f_K)'$ and bandwidth $b = (b_1, b_2, ..., b_K)'$, and let $z = (f \;\; b)'$. Then the statistical mapping from VTR to cepstrum, which constitutes Stage II of the model, is represented by

$$o(k) = \mathcal{F}(z_s(k)) + \mu_{r_s} + v_s(k), \qquad (3)$$

where v_s is a subsegment-dependent,[1] zero-mean Gaussian random vector: $v_s \sim \mathcal{N}(v; 0, \sigma_{r_s}^2)$, and μ_{r_s} is a subsegment-dependent bias vector for the nonlinear predictive function $\mathcal{F}(z_s)$.

In (3), the output of the mapping function $\mathcal{F}(z)$ has the following parameter-free, analytical form [5] for its n-th vector component (i.e., n-th order cepstrum):

$$\mathcal{F}_n(k) = \frac{2}{n}\sum_{p=1}^{P} e^{-\pi n \frac{b_p(k)}{f_s}} \cos(2\pi n \frac{f_p(k)}{f_s}), \qquad (4)$$

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

3. CASCADED IMPLEMENTATION

In this implementation, we assume that given the segment s, there is no variability in the VTR targets for a fixed speaker and consequently there is no variability in the VTR variable z in each frame within the segment. (Such variability is absorbed into the random component in model stage II.) That is, both z and T are treated as deterministic instead of random variables. Hence we have $p(z|s) = 1$ for $z = z_{max}$ as generated from the FIR filter, and $p(z|s) = 0$ otherwise. Segment-dependent and speaker-specific targets T_s in the training data are obtained by an iterative adaptative algorithm that adjusts T_s so that the FIR output from (2) matches, with minimal errors, the automatically tracked VTR produced from the algorithm described in [5]. For the test data, targets are estimated using an algorithm similar to vocal tract length normalization techniques.

To compute the acoustic likelihood required for scoring in recognition, the above stage-I output $z = z_{max}$, as the deterministic signal, is passed to model's stage-II to produce the cepstral prediction $\mathcal{F}(z_{max}(k))$ on a frame-by-frame basis. For each frame of the observed cepstral vector $o(k)$ within each segment s (or subsegment), we have the following approximate likelihood score:

$$\begin{aligned} p(o(k)|s) &\approx \max_z p(o(k)|z(k), s)p(z(k)|s) \\ &\approx p(o(k)|z_{max}(k), s)p(z_{max}(k)|s) = p(o(k)|z_{max}(k), s) \\ &= \mathcal{N}\Big[o(k); \mathcal{F}(z_{max}(k)) + \mu_{r_s}, \sigma_{r_s}^2\Big]. \end{aligned} \qquad (5)$$

This Gaussian likelihood computation is done directly using the HTK's forced-alignment tool (Hvite) for the N-best rescoring experiments (to be presented in Sec. 6).

Training of model parameters $(\mu_{r_s}, \sigma_{r_s}^2)$ is carried out in a similar way, using the same assumption and approximation as above. This is also easily accomplished using the HTK tool for training monophone HMMs on the cepstral residuals after the model prediction is subtracted from the cepstral data.

4. INTEGRATED IMPLEMENTATION

This more elaborate implementation removes the assumption in the above cascaded implementation that the VTR target T or VTR z is deterministic and that the optimal VTR vector z_{max} is not a function of the acoustic observation $o(k)$. Instead, we incorporate uncertainty in T (or equivalently in z) in the formal model construction and in computing the acoustic likelihood. This likelihood scoring is essential for speech recognition, and is accomplished by marginalizing (integrating) over the statistical distribution of VTR variables.

4.1. Characterizing VTR uncertainty in model stage-I

In order to perform the marginalization, we first need to characterize the VTR uncertainty in terms of its statistical distribution. In the current implementation, for each gender (not denoted here for simplicity) and for each segment s, we assume a separate Gaussian distribution for the target:

$$p(T|s) = \mathcal{N}(T; \mu_{T_s}, \sigma_{T_s}^2).$$

Given a sampled target sequence $T_{s(k)}$ from this distribution, we have the random VTR trajectory $z(k)$ in the form of (2). Hence we have the Gaussian distribution (gender-specific) for VTR:

$$p(z(k)|s) = \mathcal{N}[z(k); \mu_z(k), \sigma_z^2(k)] \qquad (6)$$

where

$$\mu_z(k) = \sum_{\tau=k-D}^{k+D} c(\gamma_{s(\tau)})\mu_{T_{s(\tau)}}\gamma_{s(\tau)}^{|k-\tau|}$$

and

$$\sigma_z^2(k) = \sum_{\tau=k-D}^{k+D} c^2(\gamma_{s(\tau)})\sigma_{T_{s(\tau)}}^2\gamma_{s(\tau)}^{2|k-\tau|}. \qquad (7)$$

In our implementation, VTR target means μ_{T_s} and variances $\sigma_{T_s}^2$ above are estimated using sample statistics for the empirically estimated VTR targets for each of the speakers in the training set.

[1] For notational simplicity, we use the same label s to denote a segment as well as for a subsegment.

4.2. Linearizing cepstral prediction in model stage-II

In order to perform the marginalization, we also need to characterize the cepstrum uncertainty in terms of its conditional distribution on the VTR, and to simplify the distribution to a computationally tractable form. That is, we need to specify and approximate $p(o|z,s)$.

For the simplest case where Gaussianity is assumed for subsegment-dependent cepstral prediction residuals as in the current implementation, we have

$$p(o(k)|z(k),s) = \mathcal{N}\Big[o(k); \mathcal{F}[z(k)] + \mu_{r_s}, \sigma^2_{r_s}\Big]. \qquad (8)$$

For computational tractability in marginalization (next subsection), we need to linearize the nonlinear mean function of $\mathcal{F}[z(k)]$ in (8). To do this, we use the following first-order Taylor series approximation to the nonlinear mean function:

$$\mathcal{F}[z(k)] \approx \mathcal{F}[z_0(k)] + \mathcal{F}'[z_0(k)](z(k) - z_0(k)), \qquad (9)$$

where the components of the Jacobian $\mathcal{F}'[\cdot]$ can easily computed in a closed form using (4).

Substituting (9) into (8), we obtain the approximate conditional acoustic observation probability where the mean μ_{o_s} is expressed as a linear function of the VTR variable z:

$$p(o(k)|z(k),s) \approx \mathcal{N}[o(k); \mu_{o_s}(k), \sigma^2_{r_s}], \qquad (10)$$

where

$$\mu_{o_s}(k) = \mathcal{F}'[z_0(k)]z(k) + \underbrace{\{\mathcal{F}[z_0(k)] + \mu_{r_s} - \mathcal{F}'[z_0(k)]z_0(k)\}}_{B_s}.$$

4.3. Marginalizing VTR uncertainty

Given the results above, the marginalization over the random VTR variable z in computing the acoustic likelihood can be proceeded analytically as follows:

$$\begin{aligned} p(o(k)|s) &= \int p(o(k)|z(k),s)p(z(k)|s)dz \\ &\approx \int \mathcal{N}[o(k); \mu_{o_s}, \sigma^2_{r_s}] \times \mathcal{N}[z(k); \mu_z(k), \sigma^2_z(k)]dz \qquad (11) \\ &= \int \mathcal{N}\Big[o(k); \mathcal{F}'[z_0(k)]z(k) + B_s, \sigma^2_{r_s}\Big] \times \mathcal{N}[z(k); \mu_z(k), \sigma^2_z(k)]dz \\ &= \mathcal{N}\Big[o(k) - B_s; \mathcal{F}'[z_0(k)] \times \mu_z(k), \sigma^2_{r_s} + (\mathcal{F}'[z_0(k)])^2\sigma^2_z(k)\Big] \\ &= \frac{(2\pi)^{-0.5}}{\sqrt{\sigma^2_{r_s} + (\mathcal{F}'[z_0(k)])^2\sigma^2_z(k)}} \exp\Big\{-\frac{(o(k) - \bar{\mu}_{o_s}(k))^2}{2[\sigma^2_{r_s} + (\mathcal{F}'[z_0(k)])^2\sigma^2_z(k)]}\Big\} \end{aligned}$$

where the (time-varying) mean of this Gaussian distribution

$$\bar{\mu}_{o_s}(k) = \mu_{o_s}\,|_{z(k)=\mu_z(k)} = \mathcal{F}'[z_0(k)]\mu_z(k) + B_s \qquad (12)$$

is the expectation of $\mu_{o_s}(k)$ over $z(k)$ (i.e., when the VTR random variable $z(k)$ is replaced by its mean $\mu_z(k)$). The final result of (11) is intuitive. For example, when the Taylor series expansion point is set at $z_0(k) = \mu_z(k)$, (12) is simplified to $\bar{\mu}_{o_s}(k) = \mathcal{F}[\mu_z(k)] + \mu_{r_s}$, as the noise-free part of prediction. Also, the variance in (11) is increased by a quantity of $(\mathcal{F}'[z_0(k)])^2\sigma^2_z(k)$ compared with the corresponding variance $\sigma^2_{r_s}$ in the cascaded implementation. This magnitude of increase reflects the newly introduced uncertainty in the hidden variable, measured by $\sigma^2_z(k)$ as computed from (7). The variance amplification factor $(\mathcal{F}'[z_0(k)])^2$ results from the local "slope" in the nonlinear function $\mathcal{F}[z]$ which maps from VTR $z(k)$ to cepstrum $o(k)$. Note that in (11), the variance changes dynamically as a function of time frame, instead of as a function of segment as in the conventional HMM.

5. ML TRAINING OF RESIDUAL PARAMETERS

In the cascaded implementation, the parameters of the cepstral prediction residuals, μ_s and $\sigma^2_{r_s}$, are trained using the standard Baum-Welch algorithm (HTK tool for monophones) on the prediction residual signals. It can be easily shown that this gives maximum-likelihood (ML) parameter estimates for the likelihood function of (5). However, in the integrated implementation, where the likelihood function is in the form of (11), a new training technique is required, which we have developed and are describing now.

For maximum-likelihood training of residual means, we set

$$\frac{\partial \log \prod_{k=1}^{K} p(o(k)|s)}{\partial \mu_{r_s}} = 0,$$

where $p(o(k)|s)$ is given by (11), and K denotes the total duration of subsegment s in the training data. This gives the estimation formula:

$$\hat{\mu}_{r_s} = \frac{\sum_{k=1}^{K}\Big[o(k) - \mathcal{F}[z_0(k)] - \mathcal{F}'[z_0(k)]\mu_z(k) + \mathcal{F}'[z_0(k)]z_0(k)\Big]}{K}. \qquad (13)$$

When the Taylor series expansion point is chosen to be the output of model stage-I with the target mean as the FIR filter's input, or $z_0(k) = \mu_z(k)$, (13) is simplified to:[2]

$$\hat{\mu}_{r_s} = \frac{\sum_{k=1}^{K}\Big[o(k) - \mathcal{F}[z_0(k)]\Big]}{K}. \qquad (14)$$

Similarly, variance estimation can be established as

$$\hat{\sigma}^2_{r_s} \approx \frac{\sum_k\Big\{(o(k) - \bar{\mu}_{o_s})^2 - (\mathcal{F}'[\mu_z(k)])^2\sigma^2_z(k)\Big\}}{K}. \qquad (15)$$

The above estimation formulas are applied iteratively since new boundaries of subsegments are obtained after the new updated parameters become available. The initial parameters used for the iteration are obtained using HTK for training monophone HMMs (with three left-to-right states for each phone).

6. EXPERIMENTS AND RESULTS

The phonetic recognition experiments which we have carried out to evaluate the bi-directional, target-filtering HTM with both cascaded and integrated implementations are based on the widely used TIMIT database. No language model is used in any HTM experiment. We build the acoustic models based on HTMs using the standard TIMIT label set, with slight expansion for diphthongs and affricates, in training the residual means and variances. Phonetic recognition errors are tabulated using the 39 labels adopted by many researchers to report recognition results. The results are reported on the standard core test set with a total of 192 utterances by 24 speakers.

We use the N-best rescoring paradigm to evaluate the HTM. For each of the core test utterances, we use a standard triphone HMM with a decision tree to generate a very large N-best list where N=1000. The average number (over the 192 utterances) of distinct phone sequences in this N=1000 list is 788, the remaining being due to variations in the phone segmentation in the same phone sequence.

[2] We have found in our empirical experiments that this simple way of setting Taylor series expansion points is more effective than other more elaborative ways.

Types of the HTM	101-Best (with ref.)		1001-Best (with ref.)		1000-Best (no ref.)	
	sent	phn	sent	phn	sent	phn
Cascaded I	26.6	78.8	16.2	**76.4**	0.0	71.8
Cascaded II	52.6	86.8	33.1	81.0	0.0	71.7
Integrated I	22.4	79.0	16.2	76.6	0.0	72.4
Integrated II	83.3	95.6	78.1	**94.3**	0.5	73.0
CD-HMM	0.0	64.0	0.0	**64.0**	0.0	64.0

Table 1. Performance comparison of four types of HTM implementation (all with context independent parameters) and of triphone HMM baseline (with context dependent parameters). Performance is measured by percent sentence and phone recognition accuracies (%) in the core test set defined in the TIMIT database. See text for details of the four types of HTM implementations. "Flat" language model is used. Acoustic features for all systems are the same LPC cepstral vectors.

With the use of a flat phone language model and of the LPC cepstra as features (the same conditions as the HTM), the phone recognition accuracy for the HTK-implemented standard triphone HMM in N-best list rescoring, with (N=1001) and without (N=1000) adding reference hypotheses, is 64.04%. The sentence recognition accuracy is 0.0% for HMM, even with references included. That is, the HMM system does not score the reference phone sequence higher than the N-best candidates for any of the 192 test sentences. The HTM systems dramatically increase both phone and sentence recognition accuracies, as shown in Table 6. We list the HTM performances for two types of cascaded and integrated implementations, respectively. First, the HTM with Cascaded-I implementation uses (5) for likelihood scoring, with the residual parameters $(\hat{\mu}_{r_s}, \hat{\sigma}^2_{r_s})$ trained by HTK based on the residual features computed as the difference between cepstral data and cepstral prediction. Second, Cascaded-II system uses (5) for scoring also, but with the parameters $\hat{\mu}_{r_s}$ and $\hat{\sigma}^2_{r_s}$ trained using (14) and (15). Noticeable performance improvement is obtained after the new training. Third, Integrated-I system uses (11) for scoring, with the parameters $\hat{\mu}_{r_s}$ and $\hat{\sigma}^2_{r_s}$ trained by HTK in the same way as for the Cascaded-I implementation. Rather poor performance is observed. Finally, Integrated-II system uses (11) for scoring, with the parameters $\hat{\mu}_{r_s}$ and $\hat{\sigma}^2_{r_s}$ trained using (14) and (15). The best performance, both in sentence and phone recognition accuracies, is achieved. The improvement is the greatest when references are added into the N-best list.

7. SUMMARY AND DISCUSSION

The work described in this paper represents our recent effort and continuing research on structured generative modeling approaches to speech recognition. We present a new statistical HTM, improving upon an earlier version of the HTM by extending the coarticulation modeling from uni-directionality to bi-directionality. Two types of model implementation are presented and compared: cascaded vs. integrated ones. The latter integrates over the statistical distribution of VTR in model construction and in computing acoustic likelihoods, and with rigorous training, produces the best recognition results in a standard TIMIT phonetic recognition task.

The recognition results presented in this paper are from a large-scale N-best rescoring experiment where N=1000. Since the oracle phone accuracy of the entire 1000-best list is only 82.4% (and oracle sentence accuracy only 3%), the long-span coarticulatory HTM is negatively affected by a large number of incorrect phone hypotheses in the N-best list for virtually all test utterances. Although such an "error-propagation" effect does not hurt short-span context-dependent HMM nearly as much as it hurts long-span models such as our HTM, the results of Sec. 6 nevertheless show a much lower error rate for the HTM than for the HMM in the rigorous N-best rescoring experiment (see last column in Table 1). When the "error-propagation" effect is artificially removed by adding references into the N-best lists, further drastic error reduction is obtained. This illustrates that the desired behavior in the HTM design — for the model to accurately account for detailed acoustic dynamics (given the correct phone and the corresponding VTR target sequence) — has indeed be established in the integrated implementation. In order to achieve analogous dramatic performance gain with no reference information available, it would be necessary to carry out lattice rescoring using large, virtually error-free lattices, or to develop full decoders with highly conservative pruning strategies. We are currently pursuing the HTM research in this promising direction.

8. REFERENCES

[1] J. Bilmes. "Graphical models and ASR," in M. Johnson, et. al.(eds.): *Mathematical Foundations of Speech & Language Processing,* Springer, NY, 2004, pp. 135-186.

[2] J. Bridle, L. Deng, J. Picone, et. al. "An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition," Final Report, Johns Hopkins Univ. 1998, pp. 1-61.

[3] L. Deng. "A dynamic, feature-based approach to the interface between phonology and phonetics for speech recognition," *Speech Communication,* Vol. 24, 1998, pp. 299-323.

[4] L. Deng, D. Yu, and A. Acero. "A quantitative model for formant dynamics and contextually assimilated reduction in fluent speech," *Proc. ICSLP*, 2004, Jeju Island, Korea.

[5] L. Deng, I. Bazzi, and A. Acero. "Tracking vocal tract resonances using an analytical nonlinear predictor," *Proc. Eurospeech,* 2003, pp. 73-76.

[6] J. Frankel and S. King. "ASR — Articulatory speech recognition", *Proc. Eurospeech,* Vol. 1, 2001, pp. 599-602.

[7] W. Holmes. "Segmental HMMs: Modeling dynamics and underlying structure in speech," in M. Johnson et. al.(eds.): *Mathematical Foundations of Speech & Language Processing,* Springer, NY, 2004, pp. 135-156.

[8] Y. Gao, R. Bakis, J. Huang, and B. Zhang. "Multistage coarticulation model combining articulatory, formant, and cepstral features", *Proc. ICSLP*, Vol. 1, 2000, pp. 25-28.

[9] J. Ma and L. Deng. "Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state-space model for vocal-tract-resonance dynamics," *IEEE Trans. Speech and Audio Proc.*, Vol.11, 2003, pp. 590-602.

[10] T. Stephenson, H. Bourlard, S. Bengio, and A. Morris, "ASR using Dynamic Bayesian Networks with acoustic and articulatory variables," *Proc. ICSLP*, Vol. 1, 2000.

[11] J. Zhou, F. Seide, and L. Deng. "Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM," *IEEE Proc. ICASSP*, April 2003, Vol.I, pp. 744-747.