NOVEL ACOUSTIC MODELING WITH STRUCTURED HIDDEN DYNAMICS FOR SPEECH COARTICULATION AND REDUCTION

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 report in this paper our recent progress on the new development, implementation, and evaluation of the structured speech model with statistically characterized hidden trajectories. Uni-directionality in coarticulation modeling in such hidden trajectory models as presented in previous EARS workshops has been extended to bi-directionality (forward as well as backward in the temporal dimension), offering significantly more power in parsimonious modeling of long-span context dependency. This new type of model, when appropriately implemented, also simultaneously exhibits the property of contextually assimilated phonetic reduction or phonetic target undershooting that is prevalent in casual, fluent speech (e.g., conversational speech). Experiments on large-scale N-best rescoring (N=1000) have demonstrated substantially lower phone recognition errors achieved by the model compared with a context-dependent (triphone) HMM system built with HTK. When the "error propagation" effect of the long-span acoustic model is artificially removed in the N-best rescoring paradigm (via adding the reference hypotheses into the 1000-best list), the error rate is further cut down in a dramatic manner.

1. INTRODUCTION

Modeling hidden dynamics in the temporal structure of human speech has been a salient theme in recent speech recognition research, and a growing body of literature on this theme is emerging (e.g., [1, 2, 3, 5, 7, 9, 8, 11, 15, 14, 18, 19, 20]). Hidden dynamic modeling provides a potential to overcome fundamental limitations of the HMM, especially those related to recognizing casual, spontaneous, or conversational speech with a high degree of phonetic reduction. 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, 11]). 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, 14, 19]. 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 [19] for handling bi-directional coarticulation within the framework of uni-directional, target-directed hidden trajectory modeling. (This also provides greater implementational simplicity compared with the bi-directional coarticulation model via infinite-impulse response (FIR) filtering proposed in [2].) Second, compared with the HTM of [19] 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 [6], 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 (e.g., [10, 12, 13]). 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. 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 the 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. In the mean time, the filtering process automatically exhibits contextually assimilated reduction when the segment's duration is reasonably short. especially when the filter's "stiffness" parameter is close to one. Reduction is defined in this paper as VTR target undershooting, i.e., the physically realized VTR value being away from the VTR target. When reduction is controlled by the targets of contextual (left and right) segments, we say that the reduction is contextually assimilated.

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 [1]

$$h_s(k) = \begin{cases} c\gamma_{s(k)}^{-k} & -D < k < 0 \\ c & k = 0 \\ c\gamma_{s(k)}^{k} & 0 < k < D \end{cases} \tag{1}$$

where k represents time frame, and $\gamma_{s(k)}$ is the segment-dependent "stiffness" parameter vector, one component for each resonance. Each component is positive and real-valued, ranging between zero and one. In this paper, γ is treated as a deterministic quantity for simplicity purposes. The subscript $s(k)$ in $\gamma_{s(k)}$ indicates that the stiffness parameter is dependent on the segment state $s(k)$ which varies over time. D in (1) is the unidirectional length of the impulse response, representing the temporal extent of coarticulation in one temporal direction, assumed for simplicity to be equal in length for the forward direction (anticipatory coarticulation) and the backward direction (regressive 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

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

where the input target vector's value and the filter's stiffness vector's value typically take not only those associated with the current home segment, but also those associated with the adjacent segments. The latter case happens when the time τ in (**??**) goes beyond the home segment's boundaries; i.e., when the segment $s(\tau)$ occupied at time τ switches from the home segment to an adjacent one.

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 $\boldsymbol{o}(k)$, which is typically less smooth, reflecting quantal properties in speech production [16, 17]. The mapping, as has been implemented, is in a memoryless fashion (i.e., no temporal smoothing), and is statistical rather than deterministic.

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

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

where v_s is a subsegment-dependent,[2]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 $\mathcal{F}(z_s)$. A subsegment of a phone is defined to be a consecutive temporal portion of the phone segment. Linear concatenation of several subsegments constitutes a phone segment.

In (3), the output of the mapping function $\mathcal{F}(z)$ has the following parameter-free, analytical form [6] for its n-th vector component (i.e., n-th order cepstrum):

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

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 [6]. 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:

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

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

[1] For simplicity purposes, we use scalar instead of vector notations to describe the model construction as well as model implementations throughout this paper.

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

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:

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

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:

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

where

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

and

$$\sigma^2_z(k) = \sum_{\tau=k-D}^{k+D} c^2(\gamma_{s(\tau)})\sigma^2_{T_{s(\tau)}}\gamma_{s(\tau)}^{2|k-\tau|}. \tag{7}$$

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

4.2. Linearization of nonlinear 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

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

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:

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

where the components of the Jacobian above (n-th order cepstrum's derivative with respect to VTR z) are

$$\mathcal{F}'_n[f_p(k)] = -\frac{4\pi}{f_s}e^{-\pi n \frac{b_p(k)}{f_s}} \sin(2\pi n \frac{f_p(k)}{f_s}) \tag{10}$$

for the VTR frequency components of z, and

$$\mathcal{F}'_n[b_p(k)] = -\frac{2\pi}{f_s}e^{-\pi n \frac{b_p(k)}{f_s}} \cos(2\pi n \frac{f_p(k)}{f_s}) \tag{11}$$

for the VTR bandwidth components of z. In the current implementation, the expansion point $z_0(k)$ in (9) is fixed to be the output of stage-I of the model, rather than being iteratively updated.

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:

$$p(o(k)|z(k), s) \approx \mathcal{N}[o(k); \mu_{o_s}(k), \sigma^2_{r_s}], \tag{12}$$

where

$$\begin{aligned} \mu_{o_s}(k) &= \mathcal{F}[z(k)] + \mu_{r_s} \\ &= \mathcal{F}'[z_0(k)]z(k) + B_s, \end{aligned} \tag{13}$$

where

$$B_s = \mathcal{F}[z_0(k)] + \mu_{r_s} - \mathcal{F}'[z_0(k)]z_0(k). \tag{14}$$

This then permits a closed-form solution for acoustic likelihood computation, which we derive now.

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:

$$\begin{aligned} & p(o(k)|s) = \int p(o(k)|z(k), s)p(z(k)|s)dz \\ \approx & \int \mathcal{N}[o(k); \mu_{o_s}, \sigma^2_{r_s}] \times \mathcal{N}[z(k); \mu_z(k), \sigma^2_z(k)]dz \\ = & \int \mathcal{N}\Big[o(k); \mathcal{F}'[z_0(k)]z(k) + B_s, \sigma^2_{r_s}\Big] \times \mathcal{N}[z(k); \mu_z(k), \sigma^2_z(k)]dz \\ = & \int \mathcal{N}\Big[\mathcal{F}'[z_0(k)]z(k); o(k) - B_s, \sigma^2_{r_s}\Big] \times \mathcal{N}[z(k); \mu_z(k), \sigma^2_z(k)]dz \\ = & \mathcal{N}\Big[o(k) - B_s; \mathcal{F}'[z_0(k)] \times \mu_z(k), \sigma^2_{r_s} + (\mathcal{F}'[z_0(k)])^2\sigma^2_z(k)\Big] \\ = & \frac{(2\pi)^{-0.5}}{\sqrt{\sigma^2_{r_s} + (\mathcal{F}'[z_0(k)])^2\sigma^2_z(k)}} \exp\Big\{-\frac{(o(k) - B_s - \mathcal{F}'[z_0(k)]\mu_z(k))^2}{2[\sigma^2_{r_s} + (\mathcal{F}'[z_0(k)])^2\sigma^2_z(k)]}\Big\} \\ = & \frac{(2\pi)^{-0.5}}{\sqrt{\sigma^2_{r_s} + (\mathcal{F}'[z_0(k)])^2\sigma^2_z(k)}} \exp\Big\{-\frac{(o(k) - \bar{\mu}_{o_s}(k))^2}{2[\sigma^2_{r_s} + (\mathcal{F}'[z_0(k)])^2\sigma^2_z(k)]}\Big\} \end{aligned} \tag{15}$$

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

$$\bar{\mu}_{o_s}(k) = \mu_{o_s}\ |_{z(k)=\mu_z(k)}= \mathcal{F}'[z_0(k)]\mu_z(k) + B_s \tag{16}$$

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 (15) is intuitive. For example, when the Taylor series expansion point is set at $z_0(k) = \mu_z(k)$, (16) 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 (15) 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 (15), 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 (15), 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 (15), and K denotes the total duration of subsegment s in the training data. This gives

$$\begin{aligned} 0 &= \sum_{k=1}^{K} \Big[o(k) - \bar{\mu}_{o_s} \Big] \\ &\approx \sum_{k=1}^{K} \Big[o(k) - \mathcal{F}'[z_0(k)]\mu_z(k) - B_s \Big] \\ &= \sum_{k=1}^{K} \Big[o(k) - \mathcal{F}'[z_0(k)]\mu_z(k) - \\ &\quad - \underbrace{\{\mathcal{F}[z_0(k)] + \mu_{r_s} - \mathcal{F}'[z_0(k)]z_0(k)\}}_{B_s} \Big] \qquad (17) \end{aligned}$$

This gives the estimation formula:

$$\hat{\mu}_{r_s} = \frac{\sum_{k=1}^{K} \Big[o(k) - \mathcal{F}[z_0(k)] - \mathcal{F}'[z_0(k)]\mu_z(k) + \mathcal{F}'[z_0(k)]z_0(k) \Big]}{K}. \qquad (18)$$

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 $z_0(k) = \mu_z(k)$, (18) is simplified to:[3]

$$\hat{\mu}_{r_s} = \frac{\sum_{k=1}^{K} \Big[o(k) - \mathcal{F}[z_0(k)] \Big]}{K}. \qquad (19)$$

For training the (static) base residual variances in the integrated implementation, we set

$$\frac{\partial \log \prod_{k=1}^{K} p(o(k)|s)}{\partial \sigma^2_{r_s}} = 0.$$

This gives

$$\begin{aligned} &\sum_k \Big[\frac{-1}{[\sigma^2_{r_s} + (\mathcal{F}'[z_0(k)])^2 \sigma_z^2(k)]} \Big] \\ + &\sum_k \Big[\frac{(o(k) - \bar{\mu}_{o_s})^2}{[\sigma^2_{r_s} + (\mathcal{F}'[z_0(k)])^2 \sigma_z^2(k)]^2} \Big] = 0, \qquad (20) \end{aligned}$$

or

$$\sum_k \Big[\frac{\sigma^2_{r_s} + (\mathcal{F}'[z_0(k)])^2 \sigma_z^2(k) - (o(k) - \bar{\mu}_{o_s})^2}{[\sigma^2_{r_s} + (\mathcal{F}'[z_0(k)])^2 \sigma_z^2(k)]^2} \Big] = 0, \qquad (21)$$

Assuming the dependency of the denominator term, $(\mathcal{F}'[z_0(k)])^2 \sigma_z^2(k)$, on time k is not very strong, we obtain the approximate estimation formula:

$$\hat{\sigma}^2_{r_s} \approx \frac{\sum_k \Big\{ (o(k) - \bar{\mu}_{o_s})^2 - (\mathcal{F}'[\mu_z(k)])^2 \sigma_z^2(k) \Big\}}{K}. \qquad (22)$$

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, in training the residual means and variances. The expansion includes the use of two separate targets for diphthongs and affricates, assuming one target following another. 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. A bi-phone language model is used to generate this N-best list in order to improve the quality of the list as much as possible. Mel-frequency cepstral coefficients (with delta and acceleration features), which are known to be better than LPC cepstra, are used in generating this N-best list for improving its quality also.

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. In contrast, the HTM system dramatically increases 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 (19) and (22). Noticeable performance improvement is obtained after the new training. Third, Integrated-I system uses (15) 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 (15) for

[3] 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	101-Best		1001-Best		1000-Best	
of the	(with ref.)		(with ref.)		(no ref.)	
HTM	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" or no language model is used. Acoustic features for all systems are the same LPC cepstral vectors.

scoring, with the parameters $\hat{\mu}_{r_s}$ and $\hat{\sigma}^2_{r_s}$ trained using (19) and (22). 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 ONGOING RESEARCH

A novel acoustic model of speech, based on hidden trajectory modeling of structured dynamics with bi-directional VTR target filtering, is developed in this work and is presented in this paper for the purpose of automatic speech recognition. The HTM consists of two stages of a simplified generative process of human speech: 1) from the phone sequence to VTR dynamics and 2) from the VTR dynamics to the cepstrum-based acoustic observation sequence. Two types of model implementation are detailed, one with straightforward two-stage cascading (that does not take into account randomness in the VTR dynamics), and another which integrates over the statistical distribution of VTR in model construction and in computing acoustic likelihood. With the use of the first-order Taylor series approximation to the nonlinearity in the VTR-to-cepstrum prediction component of the 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 speech utterance, and with the dynamically adjusted variance proportional to the squared "local slope" in the nonlinear mapping function from VTR to cepstrum. When the key HTM parameters are trained via maximizing this "integrated" likelihood, dramatic error reduction is achieved in the standard TIMIT phonetic recognition task using a large-scale N-best (N=1000) rescoring paradigm.

The recognition results presented in this paper are from a large-scale N-best rescoring experiment where N=1000. Since the oracle phone accuracy (TIMIT) of the entire 1000-best list is only 82.4% (and oracle sentence accuracy is 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 has been 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 been 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 a full decoder using highly conservative pruning strategies (to ensure that minimal errors occur during the search process). Our preliminary work also found that extending the current acoustic features from the LPC cepstrum to its Mel-warped version gives substantial performance gain. This requires that the HTM be able to predict the warped cepstrum's dynamics based on the model's structure with reasonable accuracy. Finally, we currently use TIMIT database mainly for the purpose of convenient model development and software debugging. While the reduction phenomenon is not manifestly strong for the TIMIT-style read speech, we nevertheless observed good performance advantages with the use of the HTM. For the conversation-style speech with much stronger reduction that our model is intended for, greater performance gain is expected. We are currently pursuing the structured acoustic modeling research in these promising directions.

8. REFERENCES

[1] J. Bilmes. "Graphical models and automatic speech recognition," in M. Johnson, M. Ostendorf, S. Khudanpur, and R. Rosenfeld (eds.): *Mathematical Foundations of Speech and Language Processing,* Springer-Verlag, New York, 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 for the 1998 Workshop on Language Engineering, Center for Language and Speech Processing at Johns Hopkins University, 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, G. Ramsay, and D. Sun. "Production models as a structural basis for automatic speech recognition," *Speech Communication,* Vol. 22, No. 2, 1997, pp. 93-111.

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

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

[8] W. Holmes. "Segmental HMMs: Modeling dynamics and underlying structure in speech," in M. Johnson et. al.(eds.): in M. Johnson, M. Ostendorf, S. Khudanpur, and R. Rosenfeld (eds.): *Mathematical Foundations of Speech and Language Processing,* Springer-Verlag, New York, 2004, pp. 135-156.

[9] 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.

[10] B. Lindblom. “Spectrographic study of vowel reduction,” J. Acoust. Soc. Am., Vol. 35, 1963, pp. 1773-1781.

[11] 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.

[12] S. Moon and B. Lindblom. “Interaction between duration, context, and speaking style in English stressed vowels,” J. Acoust. Soc. Am., Vol. 96, 1994, pp. 40-55.

[13] M. Pitermann. “Effect of speaking rate and contrastive stress on formant dynamics and vowel perception,” J. Acoust. Soc. Am., Vol. 107, 2000, pp. 3425-3437.

[14] F. Seide, J. Zhou, and L. Deng. “Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM — MAP decoding and evaluation,” *Proc. ICASSP*, 2003, pp. 748-751.

[15] T. Stephenson, H. Bourlard, S. Bengio, and A. Morris, “ASR using dynamic Bayesian networks with acoustic and articulatory variables,” *Proc. ICSLP*, Vol. 1, 2000.

[16] K. Stevens. “On the quantal nature of speech,” *Journal of Phonetics*, 17, 1989, 3-45.

[17] K. Stevens, *Acoustic Phonetics*, The MIT Press, Cambridge, MA, 1998.

[18] J. Sun and L. Deng. “An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition,” *J. Acoust. Soc. Am.*, Vol. 111, No. 2, February 2002, pp. 1086-1101.

[19] 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.

[20] G. Zweig. “Bayesian network structures and inference techniques for automatic speech recognition,” *Computer Speech and Language*, Vol. 17, 2003, pp. 173-193.