A LONG-CONTEXTUAL-SPAN MODEL OF RESONANCE DYNAMICS FOR SPEECH RECOGNITION: PARAMETER LEARNING AND RECOGNIZER EVALUATION

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

Microsoft Research
One Microsoft Way, Redmond, WA 98052, USA
{deng,dongyu,t-xiaoli,alexac}@microsoft.com

ABSTRACT

We present a structured speech model that is equipped with the capability of jointly representing incomplete articulation and long-span co-articulation in natural human speech. Central to this model is compact statistical parameterization of the highly regular dynamic patterns (exhibited in the hidden vocal-tract-resonance domain) that are driven by the stochastic segmental targets. We provide a rigorous mathematical description of this model, and present novel algorithms for learning the full set of model parameters using the cepstral data of speech. In particular, the gradient ascend techniques for learning variance parameters (for both resonance targets and cepstral prediction residuals) are described in detail. Phonetic recognition experiments are carried out using two paradigms of N-best rescoring and lattice search. Both sets of results demonstrate higher recognition accuracy achieved by the new model compared with the best HMM system. The higher accuracy is consistently observed, with and without combining HMM scores, and with and without including the references in the N-best lists and lattices. Further, the new model with rich parameter-free structure uses only the context-independent, single-modal Gaussian parameters, which are fewer than one percent of the parameters in the context-dependent HMM system with mixture distributions.

1. INTRODUCTION

Natural human speech exhibits dynamic acoustic patterns that reflect contextual influences known as coarticulation as well as incomplete articulation known as reduction. We have in recent years been developing a statistical generative model with long-span contextual dependency to explicitly take into account these two closely related effects. Our model employs cross-phone temporal filtering of vocal tract resonance (VTR) targets as the basis for joint characterization of long-span coarticulation and reduction. Since this long-contextual-span model treats the VTR trajectories (constructed in a non-recursive manner) as an unobserved stochastic process, we call it hidden trajectory model (HTM). One specific prediction of the HTM is "static" speech-class confusion. That is, the VTR frequencies taken at fixed, mid-points in the phones, as well as any related acoustic parameters such as cepstra, associated with different phones tend to move closer to each other as speech utterances become more casual [3]. Recent acoustic measurements on recorded speech with a range of speech style and speaking rate variations show such reduced vowel formant frequency separations, rendering a support for our model [8].

Similar motivations to ours for modeling contextual and reduction effects have appeared in other earlier work. For example, an empirical predictive relationship between reduced and non-reduced spectra with the same underlying phones was modeled in [1] based on psychoacoustic mechanisms. The prediction in [1] is deterministic, and follows the direction from reduced speech to non-reduced one in the observed (non-hidden) domain. In contrast, our HTM provides statistical prediction of the reduction effect in the reversed direction and in the hidden domain. Another related work to our model is temporal decomposition [2], where the co-articulated speech observations are modeled as a time-varying linear sum of a set of pre-fixed deterministic vectors in the same domain as the speech observations. Our HTM extends this concept of coarticulation modeling in three ways: 1) The pre-fixed deterministic vectors are extended to be segmental random vectors (which we call segmental random targets) where all distributional parameters are learned via maximum likelihood (ML); 2) Co-articulation as a linear sum of targets is represented in the hidden VTR domain, distinct from the observed acoustic domain in [2] and with explicit statistical relations provided between the two domains; and 3) The linear weights that are used to implement coarticulation are carefully constrained so as to produce realistic VTR trajectories under all speaking conditions (with or without reduction).

Several aspects of the HTM and its preliminary evaluation in phonetic recognition have been presented in our earlier publications [4, 5, 10], including the training algorithms for a partial set of model parameters. In this paper, we provide the training algorithms for the entire set of model parameters, and in particular, we present details of gradient descent techniques for learning the covariance matrices for both the residuals and targets. Further, the current paper presents more comprehensive evaluation results on the HTM than in the earlier work. The work in [4, 5] showed the effectiveness of the HTM for phonetic recognition mainly when the reference transcripts are included in the N-best lists where $N = 1000$. This success was more recently extended to lattice search with an equivalent of much larger lists with N in the order of billions to trillions [10]. The evaluation results presented in this paper demonstrate improvement of phonetic recognition performance over a high-quality HMM recognizer in more rigorous tests when the reference transcripts are not contained in the N-best lists and lattices that are generated from the HMM recognizer.

2. MODEL OVERVIEW

The HTM presented in this paper is a structured generative model, from the top level of phonetic specification to the bottom level of

0-7803-9479-8/05/$20.00 © 2005 IEEE

acoustic observations. We now provide an overview of this generative chain and the statistical characterizations for the various levels in the chain.

2.1. Stochastic segmental targets and phonetic units

The HTM presented in this paper assumes that each phonetic unit is associated with a multivariate distribution of the VTR targets (with the exception of several compound phonetic units where two distributions are used). Each phone-dependent target vector, $\boldsymbol{t}_s$, consists of four low-order resonance frequencies appended by their corresponding bandwidths, where s denotes the segmental phone unit. The target vector is a random vector — hence stochastic target — whose distribution is assumed to be a (gender-dependent) Gaussian:

$$p(\boldsymbol{t}|s) = \mathcal{N}(\boldsymbol{t}; \boldsymbol{\mu}_{T_s}, \boldsymbol{\Sigma}_{T_s}). \tag{1}$$

2.2. Target filtering for modeling coarticulation and reduction

The generative process in the HTM starts by temporal filtering the stochastic targets and it results in a time-varying pattern of stochastic hidden VTR vectors $\boldsymbol{z}_s(k)$. The filter is constrained so that the smooth temporal function of $\boldsymbol{z}_s(k)$ moves segment-by-segment towards the respective target vector $\boldsymbol{t}_s$ but it may or may not reach the target depending on the degree of reduction. These phonetic targets are segmental in that they do not change over the phone segment once the sample is taken, and they are assumed to be largely context independent. In the current implementation, the generation of the VTR trajectories from the segmental targets is by a bi-directional finite impulse response (FIR) filtering:

$$\boldsymbol{z}_s(k) = h_{s(k)} * \boldsymbol{t}(k) = \sum_{\tau=k-D}^{k+D} c_\gamma \gamma_{s(\tau)}^{|k-\tau|} \boldsymbol{t}_{s(\tau)}, \tag{2}$$

where c_γ is the normalization factor, which is needed to produce VTR target undershooting, instead of overshooting, for casually uttered speech. Parameter γ_s controls the *spatial* extent of coarticulation and is correlated with speaking effort. The length of the filter's impulse response $h_{s(k)}$, $2D+1$, determines the *temporal* extent of coarticulation.

The linearity between $\boldsymbol{z}$ and $\boldsymbol{t}$ as in Eq.(2) and Gaussianity of the target $\boldsymbol{t}$ make the VTR vector $\boldsymbol{z}(k)$ (at each frame k) a Gaussian as well. We now discuss the parameterization of this Gaussian process:

$$p(\boldsymbol{z}(k)|s) = \mathcal{N}[\boldsymbol{z}(k); \boldsymbol{\mu}_{z(k)}, \boldsymbol{\Sigma}_{z(k)}]. \tag{3}$$

The mean vector above is determined by the filtering function:

$$\boldsymbol{\mu}_{z(k)} = \sum_{\tau=k-D}^{k+D} c_\gamma \gamma_{s(\tau)}^{|k-\tau|} \boldsymbol{\mu}_{T_{s(\tau)}} = \mathbf{a}_k \cdot \boldsymbol{\mu}_T. \tag{4}$$

Each f-th component of vector $\boldsymbol{\mu}_{z(k)}$ is

$$\mu_{z(k)}(f) = \sum_{l=1}^{L} a_k(l) \mu_T(l, f), \tag{5}$$

where L is the total number of phone-like HTM units as indexed by l, and f=1,..., 8 for 4 VTR frequencies and 4 corresponding bandwidths.

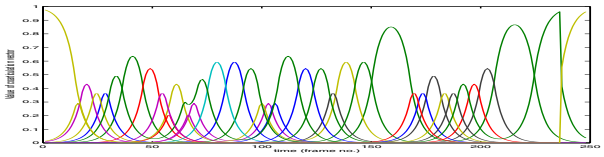

Fig. 1. *Numerical values of time-varying co-articulatory vectors $\mathbf{a}_k$'s for a TIMIT utterance.*

The covariance matrix in (3) can be similarly derived to be

$$\boldsymbol{\Sigma}_{z(k)} = \sum_{\tau=k-D}^{k+D} c_\gamma^2 \gamma_{s(\tau)}^{2|k-\tau|} \boldsymbol{\Sigma}_{T_{s(\tau)}}.$$

Approximating the covariance matrix by a diagonal one for each phone unit l, we represent its diagonal elements as a vector:

$$\boldsymbol{\sigma}_{z(k)}^2 = \boldsymbol{v}_k \cdot \boldsymbol{\sigma}_T^2. \tag{6}$$

where the target covariance matrix is also approximated as diagonal:

$$\boldsymbol{\Sigma}_T(l) \approx \begin{bmatrix} \sigma_T^2(l,1) & 0 & \cdots & 0 \\ 0 & \sigma_T^2(l,2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_T^2(l,8) \end{bmatrix}$$

The f-th element of the vector in (6) is

$$\sigma_{z(k)}^2(f) = \sum_{l=1}^{L} v_k(l) \sigma_T^2(l, f). \tag{7}$$

In (5) and (6), $\mathbf{a}_k$ and $\boldsymbol{v}_k$ are frame (k)-dependent vectors. They are constructed for any given phone sequence and phone boundaries within the coarticulation range ($2D+1$ frames) centered at frame k. Any phone beyond the $2D+1$ window contributes a zero value to these "co-articulation" vectors' elements. They are both a function of the phones' identities and temporal orders in the utterance, and are independent of the VTR dimension f. As an illustration, we plot in Fig. 1 the $\mathbf{a}_k$ values for a TIMIT utterance. At each time frame k, the values of the vector (L components in total) represent the coarticulatory effect quantified as how much adjacent phones contribute to the current phone at frame k in its VTR value. The sum of such contributions over all phones is constrained to be one. And as shown in Fig. 1, the temporally closer phones exert greater coarticulatory effects than the phones farther away. We note that these time-varying vectors $\mathbf{a}_k$ play a similar role to the linear weighting parameters in temporal decomposition [2].

2.3. Generating acoustic observations

The next generative process in the HTM provides a forward probabilistic mapping or prediction from the stochastic VTR trajectory $\boldsymbol{z}(k)$ to the stochastic observation trajectory $\boldsymbol{o}(k)$. The observation takes the form of LPC cepstra or LPCC (and their frequency-warped version) in this paper. An analytical form of the nonlinear prediction function $\mathcal{F}[\boldsymbol{z}(k)]$ was presented in [4] and summarized as:

$$\mathcal{F}_j(k) = \frac{2}{j} \sum_{p=1}^{P} e^{-\pi j \frac{b_p(k)}{f_s}} \cos(2\pi j \frac{f_p(k)}{f_s}), \tag{8}$$

where f_s is the sampling frequency, P is the highest VTR order ($P = 4$ in this work), and j is the cepstral order.

We now introduce the cepstral prediction's *residual* vector:

$$\mathbf{r}_s(k) = \boldsymbol{o}(k) - \mathcal{F}[\boldsymbol{z}(k)].$$

We model this residual vector as a Gaussian parameterized by residual mean vector $\boldsymbol{\mu}_{r_{s(k)}}$ and covariance matrix $\boldsymbol{\Sigma}_{r_{s(k)}}$:

$$p(\mathbf{r}_s(k)|\boldsymbol{z}(k), s) = \mathcal{N}\Big[\mathbf{r}_s(k); \boldsymbol{\mu}_{r_{s(k)}}, \boldsymbol{\Sigma}_{r_{s(k)}}\Big]. \quad (9)$$

Then the conditional distribution of the observation becomes:

$$p(\boldsymbol{o}(k)|\boldsymbol{z}(k), s) = \mathcal{N}\Big[\boldsymbol{o}(k); \mathcal{F}[\boldsymbol{z}(k)] + \boldsymbol{\mu}_{r_{s(k)}}, \boldsymbol{\Sigma}_{r_{s(k)}}\Big]. \quad (10)$$

2.4. Linearization of the cepstral prediction function

In order to compute the acoustic observation likelihood (see next section), we 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 $p(\boldsymbol{o}|\boldsymbol{z}, s)$. We take the simplest approach to linearize the nonlinear mean function of $\mathcal{F}[\boldsymbol{z}(k)]$ in (10) by using the first-order Taylor series approximation:

$$\mathcal{F}[\boldsymbol{z}(k)] \approx \mathcal{F}[\boldsymbol{z}_0(k)] + \mathcal{F}'[\boldsymbol{z}_0(k)](\boldsymbol{z}(k) - \boldsymbol{z}_0(k)), \quad (11)$$

where the components of Jacobian matrix $\mathcal{F}'[\cdot]$ can be computed in a closed form of

$$\mathcal{F}'_j[f_p(k)] = -\frac{4\pi}{f_s} e^{-\pi j \frac{b_p(k)}{f_s}} \sin(2\pi j \frac{f_p(k)}{f_s}) \quad (12)$$

for the VTR frequency components of $\boldsymbol{z}$, and

$$\mathcal{F}'_j[b_p(k)] = -\frac{2\pi}{f_s} e^{-\pi j \frac{b_p(k)}{f_s}} \cos(2\pi j \frac{f_p(k)}{f_s}) \quad (13)$$

for the VTR bandwidth components of $\boldsymbol{z}$.

Substituting (11) into (10), we obtain the approximate conditional acoustic observation probability where the mean vector $\boldsymbol{\mu}_{o_s}$ is expressed as a linear function of the VTR vector $\boldsymbol{z}$:

$$p(\boldsymbol{o}(k)|\boldsymbol{z}(k), s) \approx \mathcal{N}(\boldsymbol{o}(k); \boldsymbol{\mu}_{o_{s(k)}}, \boldsymbol{\Sigma}_{r_{s(k)}}), \quad (14)$$

where

$$\boldsymbol{\mu}_{o_{s(k)}} = \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{z}(k) + [\mathcal{F}[\boldsymbol{z}_0(k)] - \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{z}_0(k) + \boldsymbol{\mu}_{r_{s(k)}}]. \quad (15)$$

3. LIKELIHOOD COMPUTATION

An essential aspect of the HTM is its ability to provide the likelihood value for any sequence of acoustic observation vectors $\boldsymbol{o}(k)$ in the form of cepstral parameters. The efficiently computed likelihood provides a natural scoring mechanism comparing different linguistic hypotheses as needed in speech recognition. No VTR values $\boldsymbol{z}(k)$ are needed in this computation as they are treated as the hidden variables. They are marginalized (i.e., integrated over) in the likelihood computation. Given the model construction and the approximation described in the preceding section, the HTM likelihood computation by marginalization can be carried out in a closed form. The final result of the computation is as follows:

$$\begin{aligned} p(\boldsymbol{o}(k)|s) &= \int p[\boldsymbol{o}(k)|\boldsymbol{z}(k), s] p[\boldsymbol{z}(k)|s] d\boldsymbol{z} \\ &= \mathcal{N}\Big\{\boldsymbol{o}(k); \bar{\boldsymbol{\mu}}_{o_{s(k)}}, \bar{\boldsymbol{\Sigma}}_{o_{s(k)}}\Big\} \end{aligned} \quad (16)$$

where the time-varying mean vector is

$$\bar{\boldsymbol{\mu}}_{o_s}(k) = \mathcal{F}[\boldsymbol{z}_0(k)] + \mathcal{F}'[\boldsymbol{z}_0(k)][\mathbf{a}_k \cdot \boldsymbol{\mu}_T - \boldsymbol{z}_0(k)] + \boldsymbol{\mu}_{r_{s(k)}}$$

and the time-varying covariance matrix is

$$\bar{\boldsymbol{\Sigma}}_{o_s}(k) = \boldsymbol{\Sigma}_{r_{s(k)}} + \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{\Sigma}_z(k)(\mathcal{F}'[\boldsymbol{z}_0(k)])^{\mathrm{Tr}}. \quad (17)$$

To facilitate the development of the parameter learning algorithms for VTR targets' distributional parameters, we assume diagonality of the prediction cepstral residual's covariance matrix $\boldsymbol{\Sigma}_{r_s}$. Denoting its j-th component by $\boldsymbol{\sigma}^2_r(j)$ ($j = 1, 2, ..., J$), we decompose the multivariate Gaussian of (16) element-by-element into

$$p(\boldsymbol{o}(k)|s(k)) = \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma^2_{o_{s(k)}}(j)}} \exp\Big\{-\frac{(o_k(j) - \bar{\mu}_{o_{s(k)}}(j))^2}{2\sigma^2_{o_{s(k)}}(j)}\Big\}, \quad (18)$$

where $o_k(j)$ denotes the j-th component (i.e., j-th order) of the cepstral observation vector at frame k.

4. PARAMETER LEARNING

We describe the parameter learning algorithms for the HTM using the cepstral observation data as the training set. The criterion used for this training is to maximize the observation likelihood in (18). We describe the algorithms for the full set of model parameters, including an outline of the algorithm derivation and with more detail given to the covariance matrices' parameter estimation (for both the residuals and targets) that has not been described in our earlier papers.

4.1. Learning cepstral residuals' distributional parameters

This subset of the HTM parameters consists of 1) the mean vectors $\boldsymbol{\mu}_{r_s}$ and 2) the diagonal elements $\boldsymbol{\sigma}^2_{r_s}$ in the covariance matrices of the cepstral prediction residuals. Both of them are conditioned on phone or sub-phone s.

4.1.1. Mean vectors

To find the ML estimate of parameters $\boldsymbol{\mu}_{r_s}$, we set

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

where $p(\boldsymbol{o}(k)|s)$ is given by (16), and K denotes the total duration of sub-phone s in the training data. This gives

$$\sum_{k=1}^{K} \Big[\boldsymbol{o}(k) - \bar{\boldsymbol{\mu}}_{o_s}\Big] = 0$$

This leads to the estimation formula of

$$\hat{\boldsymbol{\mu}}_{r_s} = \frac{\sum_k \Big[\boldsymbol{o}(k) - \mathcal{F}[\boldsymbol{z}_0(k)] - \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{\mu}_z(k) + \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{z}_0(k)\Big]}{K}. \quad (19)$$

4.1.2. Diagonal covariance matrices

Denote the diagonal elements of the covariance matrices for the residuals as a vector $\boldsymbol{\sigma}^2_{r_s}$. To derive the ML estimate, we set

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

This gives

$$\sum_{k=1}^{K}\Big[\frac{\boldsymbol{\sigma}^2_{r_s} + \mathbf{q}(k) - (\boldsymbol{o}(k) - \bar{\boldsymbol{\mu}}_{\boldsymbol{o}_s})^2}{[\boldsymbol{\sigma}^2_{r_s} + \mathbf{q}(k)]^2}\Big] = 0, \tag{20}$$

where vector squaring above is the element-wise operation, and

$$\mathbf{q}(k) = \text{diag}\Big[\mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{\Sigma}_z(k)(\mathcal{F}'[\boldsymbol{z}_0(k)])^{\text{Tr}}\Big]. \tag{21}$$

Due to frame (k) dependency of the denominator in (20), no simple closed-form solution is available for solving $\boldsymbol{\sigma}^2_{r_s}$ from (20). We have implemented three different techniques for seeking approximate ML estimates which we outline here.

1. Frame-independent approximation: Assume the dependency of $\mathbf{q}(k)$ on time frame k is mild, or $\mathbf{q}(k) \approx \bar{\mathbf{q}}$. Then the denominator in (20) can be cancelled, yielding the approximate closed-form estimate of

$$\hat{\boldsymbol{\sigma}}^2_{r_s} \approx \frac{\sum_{k=1}^{K}\Big\{(\boldsymbol{o}(k) - \bar{\boldsymbol{\mu}}_{o_s})^2 - \mathbf{q}(k)\Big\}}{K}. \tag{22}$$

2. Direct gradient ascent: Make no assumption of the above, and take the left-hand-side of (20) as the gradient ∇L of log-likelihood of the data in the standard gradient-ascent algorithm:

$$\boldsymbol{\sigma}^2_{r_s}(t+1) = \boldsymbol{\sigma}^2_{r_s}(t) + \epsilon_t \nabla L(\boldsymbol{o}_1^K | \boldsymbol{\sigma}^2_{r_s}(t)),$$

where ϵ_t is a heuristically chosen positive constant controlling the learning rate at the t-th iteration.

3. Constrained gradient ascent: This technique improves on the previous standard gradient ascent by imposing the constraint that the variance estimate is always positive. The constraint is established by the parameter transformation: $\tilde{\boldsymbol{\sigma}}^2_{r_s} = \log \boldsymbol{\sigma}^2_{r_s}$, and by performing gradient ascent for $\tilde{\boldsymbol{\sigma}}^2_{r_s}$ instead of for $\boldsymbol{\sigma}^2_{r_s}$:

$$\tilde{\boldsymbol{\sigma}}^2_{r_s}(t+1) = \tilde{\boldsymbol{\sigma}}^2_{r_s}(t) + \tilde{\epsilon}_t \nabla \tilde{L}(\boldsymbol{o}_1^K | \tilde{\boldsymbol{\sigma}}^2_{r_s}(t)),$$

Using chain rule, we show below that the new gradient $\nabla \tilde{L}$ is related to the gradient ∇L before parameter transformation in a simple manner:

$$\nabla \tilde{L} = \frac{\partial \tilde{L}}{\partial \tilde{\boldsymbol{\sigma}}^2_{r_s}} = \frac{\partial \tilde{L}}{\partial \boldsymbol{\sigma}^2_{r_s}} \frac{\partial \boldsymbol{\sigma}^2_{r_s}}{\partial \tilde{\boldsymbol{\sigma}}^2_{r_s}} = (\nabla L) \;\; \exp(\tilde{\boldsymbol{\sigma}}^2_{r_s}).$$

At the end of algorithm iterations, the parameters are transformed via $\boldsymbol{\sigma}^2_{r_s} = \exp(\tilde{\boldsymbol{\sigma}}^2_{r_s})$, which is guaranteed to be positive.

For efficiency purposes, parameter updating in the above gradient ascent techniques is carried out after each utterance in the training, rather than after the entire batch of all utterances.

4.2. Learning VTR targets' distributional parameters

This subset of the HTM parameters consists of 1) the mean vectors $\boldsymbol{\mu}_{T_s}$ and 2) the diagonal elements $\boldsymbol{\sigma}^2_{T_s}$ in the covariance matrices of the stochastic segmental VTR targets. They also are conditioned on phone segment s.

4.2.1. Mean vectors

Optimizing the log likelihood function of Eq.(18) with respect to the joint parameter set $\boldsymbol{\mu}_T$ (i.e., including each phone indexed by l and each of the vector component indexed by f in $\mu_T(l,f)$) results in a large full-rank linear system of equations. The derivation and solution implementation of this system of equations have been described in detail in [5]. We now turn to the variance estimation problem where higher complexity arises and approximate solutions are needed.

4.2.2. Diagonal covariance matrices

To establish the objective function for optimization, we take logarithm on the sum of the likelihood function Eq.(18) (over K frames) to obtain

$$L_T \propto -\sum_{k=1}^{K}\sum_{j=1}^{J}\Big\{\frac{(o_k(j) - \bar{\mu}_{o_{s(k)}}(j))^2}{\sigma^2_{r_s}(j) + q(k,j)} + \log[\sigma^2_{r_s}(j) + q(k,j)]\Big\} \tag{23}$$

where $q(k,j)$ is the j-th element of the vector $\mathbf{q}(k)$ as defined in (21). When $\boldsymbol{\Sigma}_z(k)$ is diagonal, it can be shown that

$$q(k,j) = \sum_f \sigma^2_{z(k)}(f)(F'_{jf})^2 = \sum_f \sum_l v_k(l)\sigma^2_T(l,f)(F'_{jf})^2, \tag{24}$$

where F'_{jf} is the (j,f) element of Jacobian matrix $\mathcal{F}'[\cdot]$ in (21), and the second equality in the above is due to (7).

Using chain rule to compute the gradient, we obtain

$$\begin{aligned}\nabla L_T(l,f) &= \frac{\partial L_T}{\partial \sigma^2_T(l,f)} \qquad (25)\\ &= \sum_{k=1}^{K}\sum_{j=1}^{J}\Big\{\frac{(o_k(j) - \bar{\mu}_{o_{s(k)}}(j))^2 (F'_{jf})^2 v_k(l)}{[\sigma^2_{r_s}(j) + q(k,j)]^2} - \frac{(F'_{jf})^2 v_k(l)}{\sigma^2_{r_s}(j) + q(k,j)}\Big\}\end{aligned}$$

Gradient-ascend iterations then proceed as follows:

$$\boldsymbol{\sigma}^2_T(l,f) \leftarrow \boldsymbol{\sigma}^2_T(l,f) + \epsilon \nabla L_T(l,f),$$

for each phone l and for each element f in the diagonal VTR target covariance matrix.

5. RECOGNITION EXPERIMENTS

We have carried out a set of phonetic recognition experiments aimed at evaluating the HTM and the parameter learning algorithms described in this paper. The standard TIMIT phone set with 48 labels is expanded to 58 (as described in [6]) in training the HTM parameters using the standard 4620 training utterances. Phonetic recognition errors are tabulated using the commonly adopted 39 labels after the label folding. The results are reported on the standard core test set of 192 utterances by 24 speakers [7].

While the full decoder is currently under development for the HTM, we report in this paper the N-best rescoring and lattice rescoring results. For each of the core test utterances, a standard

decision-tree-based triphone HMM is used to generate a large N-best list ($N = 1000$) and a large lattice. These N-best lists and lattices are used for the rescoring experiments with the HTM.

The HTM system is trained using the algorithms described in Section 4. Learning rates in the gradient ascent techniques have been tuned empirically.

5.1. Results on N-best Rescoring

Table 1 shows N-best rescoring results with the use of two types of language models (LM) for the HTM and HMM systems, respectively. One is the standard bi-phone (or bi-gram) LM trained from the TIMIT training set, and another uses a zero-value LM score denoted as "Flat-LM". Phonetic recognizers' performance is measured by percent phone recognition accuracy (i.e., including deletion errors) for the core test set. Two types of acoustic observations are used. First, LPC cepstra (LPCC) as described in Section 2.3 give the most straightforward system implementation. Second, to overcome the lack of perceptual correlate of LPCC, we implemented an extended HTM system (not described in this paper) using frequency-warped LPCCs [9] for both acoustic features and the observation-prediction component of the HTM. The accuracy performance in Table 1 is achieved with no reference hypotheses included in the N-best lists, except for the much greater accuracy numbers in parentheses for the HTM system where references are included. (Note that inclusion of references does not change the HMM system's performance.) The HTM outperforms the HMM in all cases. Note that no HMM scores are used to combine with the scores of the HTM in Table 1. The sole role of HMM for the HTM system is to create the 1000-best lists on which rescoring is carried out.

Table 1. *N-best rescoring results where N=1000. No HMM scores are used to combine with scores of the HTM.*

Acoustic Features	HMM Recognizer		**HTM** Recognizer	
	Bi-gram	Flat-LM	Bi-gram	Flat-LM
LPCC	68.2	64.0(64.0)	73.0	72.8 (95.3)
W-LPCC	71.4	68.1(68.1)	**74.3**	73.5 (96.0)

When the HMM scores are combined with those of the HTM, further performance improvement is observed, as shown in Table 2, for both accuracy and correctness measures (the latter does not count deletion as errors). We had expected greater improvement after the use of HMM scores. However, since the selection of the N-best hypotheses by the HMM already embeds much of the HMM-based discriminative information, the additional information from the weighted HMM scores provides understandably only a minor contribution to the final performance. Selected values of the relative score weights in this experiment are provided in Table 2.

5.2. Results on Lattice Search

The use of lattices provides much richer hypothesis candidates than N-best lists for evaluating the HTM, but due to the long-contextual-span property of the model, the search algorithm is complex. We refer the interested readers to the detailed technical description of this A*-based search algorithm we have successfully developed described in [10].

Table 2. *N-best rescoring results where N=1000. The HMM scores and bi-gram LM scores in the N-best lists are combined with the HTM scores using various weights as shown.*

HTM Wt.	HMM Wt.	LM Wt.	Accuracy/Correct(%)
1	0	1	73.60 / 76.20
1	0	5	74.26 / 77.08
1	1	5	74.23 / 77.12
1	5	5	74.04 / 77.23
1	1	1	74.53 / 77.70
1	1.5	1	**74.59** / 77.73

This search algorithm is applied to the warped-LPCC version of the HTM, where the warping factor for the testing is fixed at $\alpha = 0.48$. With adjustments of relative weights of various scores as well as the phone insertion penalty value, we obtained sizable performance improvement, as shown in Table 3, over the best result in N-best rescoring.

Table 3. *Lattice search results. The HMM scores and bi-gram LM scores in the lattices are combined with the HTM scores using various weights. Phone insertion penalty (IP) is also varied. Warping factor is fixed at* $\alpha = 0.480$.

HTM Wt.	HMM Wt.	LM Wt.	IP	Acc/Corr(%)
1	2	14	0	73.34 / 76.90
1	3	18	0	74.23 / 78.03
1	4	28	0	74.80 / 78.07
1	8	40	0	74.39 / 78.82
1	5	35	0	74.91 / 78.22
1	5	35	-0.5	74.99 / 78.25
1	5	35	-1	**75.02** / 78.25

Finally, we empirically optimized the warping factor in the warped-LPCC version of the HTM, with detailed results shown in Table 4. The best accuracy achieved so far is 75.1% (or phone error rate of 24.9%). This performance is better than any HMM system as summarized in [7], and is approaching the ever best result in the same task (phone error rate of 24.4%) obtained by using many heterogeneous classifiers reported in [7].

Table 4. *Lattice search results. Performance as a function of the warping factor* α. *The weights for the HTM score, HMM score, and bi-gram LM score are fixed at 1, 5, and 35, respectively. Insertion penalty is fixed at -1.*

α	Accuracy/Correct(%)
0.420	74.51 / 78.42
0.450	74.58 / 77.87
0.470	74.79 / 78.04
0.475	74.96 / 78.21
0.478	**75.07** / 78.28
0.480	75.02 / 78.25
0.482	74.92 / 78.17
0.490	74.62 / 78.00

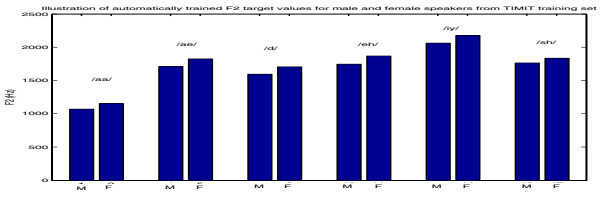

Fig. 2. *Male and female VTR-frequency means (F2 as shown) are separated apart after the training while having the same initial values before the training.*

5.3. Results on model learning

We show in this subsection limited results (due to the space constraint) demonstrating the effectiveness of the training algorithms presented in Section 4, as well as their correct implementation. First, a monotonic increase of the likelihood on the training data is observed over the training iterations. A typical training score sequence is shown in Table 5. Second, typical VTR target mean values (F2) for selected vowels and consonances before and after training are shown in Fig. 2, separated out for male and female speakers. Whereas male and female gender-dependent VTR target mean vectors are initialized with identical values (pp. 364 in [6]), they become well separated after the training. The estimated female's F2 are higher than the male's counterpart by an amount consistent with acoustic-phonetic intuition. Detailed analysis has been carried out to assess acoustic-phonetic reality of the training results and overwhelming consistency has been found.

Table 5. *Log likelihood for the entire training set as a function of the training algorithm iteration number.*

Iteration Number	Log-Likelihood
1	10233229.0
2	57950989.4
3	60116655.8
4	60986576.2
5	61038928.8
6	61041398.8

6. DISCUSSION AND CONCLUSIONS

While the main motivation of the HTM is to capture the structure of the underlying speech dynamics to account for long-span coarticulation and phonetic reduction in an integrative manner, to facilitate implementation we evaluated the model thus far only on the TIMIT database as presented in this paper. It is known that the speech data in TIMIT suffer less from incomplete articulation and long-span contextual influences than those in free-style speech data such as in Switchboard databases. Hence, in our future work we expect greater advantages of the HTM for these difficult databases than the already demonstrated superiority for TIMIT as demonstrated in this paper.

In our earlier work on TIMIT [4, 5], we found that the oracle error rate for the N-best lists (N as large as 2000) produced by the HMM is as high as 18%. This accounts for the large difference between the N-best rescoring accuracies with and without including the reference hypotheses in the N-best lists. That is, the effect known as "error spreading" associated with any long-span model would hurt recognition when local errors occur. Use of the lattices reduces the oracle error rate for the phone identities. However, when phone segmentation is considered (TIMIT provides such information), the lattice oracle error rate is found to be still very large. Despite this, strong results are obtained on these lattices as we showed in Section 5.2. To further improve recognition accuracy, we are currently expanding the lattices for reducing the oracle errors related to phone segmentation so as to mitigate the error-spreading effect. We are also developing a time-synchronous decoder for the HTM free from the lattice constraint and thus are able to eliminate this effect. Further, we are continuing the research on improving the quality of the current HTM and on improving the efficiency of the search techniques that are specific to long-contextual-span models. Finally, since TIMIT exhibits relatively minor problems of phonetic reduction, we expect greater benefits of the HTM over the HMM for more challenging tasks of conversational speech in our future work.

7. ACKNOWLEDGEMENTS

We thank Bishnu Atal, B.-H. Juang, and C.-H. Lee for insightful discussions on the similar motivations of this work to those in [2, 1]. We also thank Xiang Li for implementing an initial version of the LPCC warping function. Finally, the help of Asela Gunawardana and Mike Seltzer for generating high-quality lattices from an HMM system and useful discussions are gratefully acknowledged.

8. REFERENCES

[1] M. Akagi. "Modeling of contextual effects based on spectral peak interaction," in *J. Acoust. Soc. Am.*, Vol. 93, No. 2, pp. 1076-1086, 1993.

[2] B. S. Atal. "Efficient coding of LPC parameters by temporal decomposition," in *Proc. IEEE ICASSP*, pp. 81–84, 1983.

[3] L. Deng, D. Yu, and A. Acero. "A quantitative model for formant dynamics and contextually assimilated reduction in fluent speech", in Proc. ICSLP, pp 719-722, Jeju, Korea, 2004.

[4] L. Deng, X. Li, D. Yu, and A. Acero,"A hidden trajectory model with bi-directional target-filtering: Cascaded vs. integrated implementation for phonetic recognition", in *Proc. IEEE ICASSP*, pp 337-340, March, 2005, Philadelphia.

[5] L. Deng, D. Yu, and A. Acero. "Learning statistically characterized resonance targets in a hidden trajectory model of speech coarticulation and reduction,", in *Proc. Interspeech 2005*, Lisbon, Sept 2005, pp. 1097-1100.

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

[7] J. Glass. "A probabilistic framework for segment-based speech recognition," in *Computer Speech and Language,* Vol. 17, 2003, pp. 137-152.

[8] J. Krause and L. Braida. "Acoustic properties of naturally produced clear speech at normal speaking rates," in *J. Acoust. Soc. Am.*, Vol. 115, No. 1, pp. 362-378, 2004.

[9] A. Oppenheim and D. Johnson. "Discrete representation of signals," in *Proc. IEEE,* Vol. 60, No. 6, 1972, pp. 681-691.

[10] D. Yu, L. Deng, and A. Acero. "Evaluation of a long-contextual-span trajectory model and phonetic recognizer using A* lattice search", in *Proc. Interspeech*, Lisbon, Sept 2005, pp. 553-556.