USE OF DIFFERENTIAL CEPSTRA AS ACOUSTIC FEATURES IN HIDDEN TRAJECTORY MODELING FOR PHONETIC RECOGNITION

Li Deng and Dong Yu

Microsoft Research, Redmond, WA, 98052
{deng,dongyu}@microsoft.com

ABSTRACT

The earlier version of the hidden trajectory model (HTM) for speech dynamics which predicts the "static" cepstra as the observed acoustic feature is generalized to one which predicts joint static cepstra and their temporal differentials (i.e., delta cepstra). The formulation of this generalized HTM is presented in the generative-modeling framework, enabling efficient computation of the joint likelihood for both static and delta cepstral sequences as the acoustic features given the model. The parameter estimation techniques for the new model are developed and presented, giving closed-form estimation formulas after the use of vector Taylor series approximation. We show principled generalization from the earlier static-cepstra HTM to the new static/delta-cepstra HTM not only in terms of model formulations but also in terms of their respective analytical forms in (monophone) parameter estimation. Experimental results on the standard TIMIT phonetic recognition task demonstrate recognition accuracy improvement over the earlier best HTM system, both significantly better than state-of-the-art triphone HMM systems.

Index Terms— phonetic recognition, hidden trajectory modeling, delta cepstra, joint static/dynamic feature, generative modeling

1. INTRODUCTION

In recent years, we have been pursuing a research direction in speech modeling and recognition where the dynamic structure associated with human speech generation mechanisms is exploited for the purpose of providing a more accurate and parsimonious characterization of the speech process than the traditional hidden Markov model (HMM) [1, 4, 21]. One particular type of the statistical models that we have been focusing on developing is the hidden trajectory model (HTM) [4, 5, 21], which uses target-directed, cross-unit continuous movement of vocal tract resonances (VTR) as the basis for modeling the underlying dynamic speech structure and for predicting the acoustic features (i.e., observation data) in the form of "static" cepstra. In the work presented in this paper, we generalize the earlier HTM by predicting not only the "static" cepstra but also the frame-differential cepstra (also known as dynamic or delta or regression features [7, 10, 22]).

The importance of modeling speech dynamics for speech processing applications has been well known for many years [7]. The early approaches exploited frame-differential acoustic features as a simplistic representation of speech dynamics, and fed them into standard pattern recognition systems with weak dynamic modeling capabilities. It is widely known that the use of these "dynamic" features is problematic and inconsistent within the traditional pattern recognition frameworks (e.g., HMM). Numerous analyses on and empirical remedy of the inconsistency have appeared in the literature (e.g., [20, 22]), demonstrating the usefulness of the dynamic features such as the delta cepstra.

Later approaches to exploiting speech dynamics made use of the statistical "segment" models that represent correlations of observed acoustic features across frames (e.g., [15, 2]). The most recent approaches, exemplified by our previous HTM [4, 21], represented speech dynamics not on the observed acoustic domain, but on the unobserved or "hidden" domain (including VTR or articulatory domain), and the observed acoustic dynamics (in the form of sequences of "static" cepstral vectors) becomes a natural consequence of the modeled "hidden dynamics". Within this generative-modeling framework, it is natural to extend our earlier HTM so that it accounts for not only sequences of "static" vectors but also sequences of "delta" ones. Part of the motivation of this extension is the demonstrated usefulness of the delta features in HMM-based speech processing applications.

The incorporation of delta features in the HMM is straightforward — via direct training of additional means and variances associated with the delta-feature component in the HMM using the identical training algorithm (e.g., Baum-Welch algorithm) to that designed for the static-feature component in the HMM. In contrast, the use of delta features in the HTM requires an additional prediction stage for these new features, the main topic of this paper. This stage is bypassed in the HMM, responsible for the creation of apparent inconsistency between the uses of static and dynamic features in the overall generative modeling framework.

The paper is organized as follows. In Section 2, we review the earlier HTM which predicts static cepstral sequences. A generalized HTM which predicts joint static and delta cepstral sequences is presented in Section 3. The learning algorithm for the generalized HTM is described in Section 4. We present experimental results on phonetic recognition, showing recognition accuracy improvement over the earlier version of the HTM in Section 5.

2. HIDDEN TRAJECTORY MODEL WITH CEPSTRA AS ACOUSTIC FEATURES

In this section, we provide a brief overview of the earlier version of the HTM detailed in [4], where "static" cepstra are used as the acoustic features that are predicted by the model. The notation in [4] has been slightly modified so that the generalization of the HTM described in the next section can be more easily identified.

Given the VTR hidden trajectory, $\boldsymbol{z}_k$, which is a function of time frame k, the conditional distribution of cepstra is assumed to be Gaussian:

$$p(\boldsymbol{o}_k|\boldsymbol{z}_k, s) = \mathcal{N}\Big[\boldsymbol{o}_k; \mathcal{F}[\boldsymbol{z}_k] + \boldsymbol{\mu}_{s(k)}, \boldsymbol{\Sigma}_{s(k)}\Big]. \quad (1)$$

where $\mathcal{F}[\boldsymbol{z}_k]$ is a fixed, parameter-free nonlinear function, with $\boldsymbol{\mu}_{s(k)}$

and $\mathbf{\Sigma}_{s(k)}$ being the model parameters (related to the cepstral-prediction residual component) subject to optimization from the cepstral data $\boldsymbol{o}_k$. (The model parameters related to the stochastic VTR targets have been presented in [5, 4] and will not be discussed in this paper.)

We further assume that the prior distribution of $\boldsymbol{z}_k$ for each time k, given the phonetic unit or state s is a Gaussian:

$$p(\boldsymbol{z}_k|s) = \mathcal{N}\Big(\boldsymbol{z}_k; \boldsymbol{\mu}_{z_k}, \mathbf{\Psi}_{z_k}\Big) \tag{2}$$

where the mean vector $\boldsymbol{\mu}_{z_k}$ and covariance matrix $\mathbf{\Psi}_{z_k}$ are dependent on the underlying model parameters representing the phonetic targets and on the coarticulatory properties of the stochastic target-directed "hidden" speech dynamics (see details in [4, 1]).

To compute the acoustic likelihood, we marginalize the hidden variable $\boldsymbol{z}_k$ to obtain

$$\begin{aligned} p(\boldsymbol{o}_k|s) &= \int p(\boldsymbol{o}_k|\boldsymbol{z}_k, s)p(\boldsymbol{z}_k|s)d\boldsymbol{z}_k \\ &= \int \mathcal{N}(\boldsymbol{o}_k; \mathcal{F}[\boldsymbol{z}_k] + \boldsymbol{\mu}_{s(k)}, \mathbf{\Sigma}_{s(k)})\mathcal{N}(\boldsymbol{z}_k; \boldsymbol{\mu}_{z_k}, \mathbf{\Psi}_{z_k})d\boldsymbol{z}_k \\ &\approx \mathcal{N}\Big(\boldsymbol{o}_k; \mathcal{F}'[\boldsymbol{z}_k^0]\boldsymbol{\mu}_{z_k} + \mathbf{b}_k, \mathbf{\Omega}_k\Big) \end{aligned} \tag{3}$$

where

$$\mathbf{b}_k = \mathcal{F}[\boldsymbol{z}_k^0] + \boldsymbol{\mu}_{s(k)} - \mathcal{F}'[\boldsymbol{z}_k^0]\boldsymbol{z}_k^0 \tag{4}$$

$$\mathbf{\Omega}_k = \mathbf{\Sigma}_{s(k)} + \mathcal{F}'[\boldsymbol{z}_k^0]\mathbf{\Psi}_{z_k}(\mathcal{F}'[\boldsymbol{z}_k^0])^T \tag{5}$$

The approximation in (3) was due to the use of first-order vector Taylor series expansion:

$$\mathcal{F}[\boldsymbol{z}_k] \approx \mathcal{F}[\boldsymbol{z}_k^0] + \mathcal{F}'[\boldsymbol{z}_k^0](\boldsymbol{z}_k - \boldsymbol{z}_k^0) \tag{6}$$

where $\boldsymbol{z}_k^0$ is the Taylor series expansion point, which, in our current modeling implementation, is obtained by a high-quality VTR or formant tracker [3, 6].

The learning algorithm for the cepstral prediction residual parameters $\boldsymbol{\mu}_s$ and $\mathbf{\Sigma}_s$ (as well as the model parameters related to stochastic targets) can be found in [4] and will not be reviewed here.

3. GENERALIZATION: JOINT CEPSTRA/DELTA-CEPSTRA AS ACOUSTIC FEATURES

We now generalize the above HTM with static cepstra as the observation vectors to one that accounts for joint static cepstra and its temporal differentials. The differential or delta cepstra is defined by

$$\boldsymbol{d}_k = \frac{\boldsymbol{o}_{k+\theta} - \boldsymbol{o}_{k-\theta}}{2\theta} \tag{7}$$

Using (7) and (1), we obtain conditional pdf for delta-cepstra:

$$\begin{aligned} &p[\boldsymbol{d}_k|\boldsymbol{z}_k, \boldsymbol{z}_{k+\theta}, \boldsymbol{z}_{k-\theta}, s] = \\ &\mathcal{N}\Big[\boldsymbol{d}_k; \frac{\mathcal{F}[\boldsymbol{z}_{k+\theta}] - \mathcal{F}[\boldsymbol{z}_{k-\theta}]}{2\theta} + \boldsymbol{\delta}_{s(k)}, \mathbf{\Gamma}_{s(k)}\Big]. \end{aligned} \tag{8}$$

where $\boldsymbol{\delta}_{s(k)}$ and $\mathbf{\Gamma}_{s(k)}$ are related to $\boldsymbol{\mu}_{s(k+\theta)}, \boldsymbol{\mu}_{s(k-\theta)}, \mathbf{\Sigma}_{s(k+\theta)}$, and $\mathbf{\Sigma}_{s(k-\theta)}$. In the current model implementation, however, we treat $\boldsymbol{\delta}_{s(k)}$ and $\mathbf{\Gamma}_{s(k)}$ as new parameters in the training for simplicity purposes. This treatment avoids otherwise more complex constrained optimization problem where the constraints based on the relations are imposed. (Solving this difficult constrained optimization problem is our planned future research work.)

Then, the conditional distribution of the joint cepstral/delta-cepstral features becomes:

$$p(\boldsymbol{o}_k, \boldsymbol{d}_k|\boldsymbol{z}_k, \boldsymbol{z}_{k-\theta}, \boldsymbol{z}_{k+\theta}, s) = \tag{9}$$
$$\mathcal{N}\Big(\begin{bmatrix} \boldsymbol{o}_k \\ \boldsymbol{d}_k \end{bmatrix}; \begin{bmatrix} \mathcal{F}[\boldsymbol{z}_k] + \boldsymbol{\mu}_{s(k)} \\ \frac{\mathcal{F}[\boldsymbol{z}_{k+\theta}] - \mathcal{F}[\boldsymbol{z}_{k-\theta}]}{2\theta} + \boldsymbol{\delta}_{s(k)} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{s(k)} & 0 \\ 0 & \mathbf{\Gamma}_{s(k)} \end{bmatrix}\Big)$$

Using first-order Taylor series approximation for nonlinear function $\mathcal{F}[\boldsymbol{z}_k]$ according to (6), we approximate the conditional pdf of (9) by

$$\mathcal{N}\Big(\begin{bmatrix} \boldsymbol{o}_k \\ \boldsymbol{d}_k \end{bmatrix}; \mathbf{A}_k \begin{bmatrix} \boldsymbol{z}_k \\ \boldsymbol{z}_{k+\theta} \\ \boldsymbol{z}_{k-\theta} \end{bmatrix} + \mathbf{b}_k, \begin{bmatrix} \mathbf{\Sigma}_{s(k)} & 0 \\ 0 & \mathbf{\Gamma}_{s(k)} \end{bmatrix}\Big) \tag{10}$$

where

$$\mathbf{A}_k = \begin{bmatrix} \mathcal{F}'[\boldsymbol{z}_k^0] & 0 & 0 \\ 0 & \frac{\mathcal{F}'[\boldsymbol{z}_{k+\theta}^0]}{2\theta} & -\frac{\mathcal{F}'[\boldsymbol{z}_{k-\theta}^0]}{2\theta} \end{bmatrix} \tag{11}$$

and

$$\mathbf{b}_k = \begin{bmatrix} \mathcal{F}[\boldsymbol{z}_k^0] + \boldsymbol{\mu}_{s(k)} - \mathcal{F}'[\boldsymbol{z}_k^0]\boldsymbol{z}_k^0 \\ \frac{\mathcal{F}[\boldsymbol{z}_{k+\theta}^0] - \mathcal{F}'[\boldsymbol{z}_{k+\theta}^0]\boldsymbol{z}_{k+\theta}^0 - \mathcal{F}[\boldsymbol{z}_{k-\theta}^0] + \mathcal{F}'[\boldsymbol{z}_{k-\theta}^0]\boldsymbol{z}_{k-\theta}^0}{2\theta} + \boldsymbol{\delta}_{s(k)} \end{bmatrix} \tag{12}$$

We assume a block-diagonal Gaussian distribution for the joint hidden trajectory vector, $p(\boldsymbol{z}_k, \boldsymbol{z}_{k+\theta}, \boldsymbol{z}_{k-\theta}|s)$. That is,

$$\mathcal{N}\Big(\begin{bmatrix} \boldsymbol{z}_k \\ \boldsymbol{z}_{k+\theta} \\ \boldsymbol{z}_{k-\theta} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_{z_k} \\ \boldsymbol{\mu}_{z_{k+\theta}} \\ \boldsymbol{\mu}_{z_{k-\theta}} \end{bmatrix}, \begin{bmatrix} \mathbf{\Psi}_{z_k} & 0 & 0 \\ 0 & \mathbf{\Psi}_{z_{k+\theta}} & 0 \\ 0 & 0 & \mathbf{\Psi}_{z_{k-\theta}} \end{bmatrix}\Big) \tag{13}$$

where $\mathbf{\Psi}_{z_k}$, $\mathbf{\Psi}_{z_{k+\theta}}$, and $\mathbf{\Psi}_{z_{k-\theta}}$ are the covariances of the hidden trajectory vectors at three different frames. They are determined by the VTR-targets' covariance parameters of the HTM and by the coarticulation properties of the VTR dynamics as elaborated in [4], and are considered fixed for the discussion of the cepstral prediction residual parameters as the focus of this paper.

Now we compute the acoustic likelihood, $p(\boldsymbol{o}_k, \boldsymbol{d}_k|s)$, by marginalizing the hidden trajectory variables:

$$\int p(\boldsymbol{o}_k, \boldsymbol{d}_k|\boldsymbol{z}_k, \boldsymbol{z}_{k-\theta}, \boldsymbol{z}_{k+\theta}, s)p(\boldsymbol{z}_k, \boldsymbol{z}_{k+\theta}, \boldsymbol{z}_{k-\theta}|s)d\boldsymbol{z}_k d\boldsymbol{z}_{k+\theta} d\boldsymbol{z}_{k-\theta} \tag{14}$$

This integration has a closed form, which gives

$$p(\boldsymbol{o}_k, \boldsymbol{d}_k|s) = \mathcal{N}\Big(\begin{bmatrix} \boldsymbol{o}_k \\ \boldsymbol{d}_k \end{bmatrix}; \mathbf{A}_k \begin{bmatrix} \boldsymbol{\mu}_{z_k} \\ \boldsymbol{\mu}_{z_{k+\theta}} \\ \boldsymbol{\mu}_{z_{k-\theta}} \end{bmatrix} + \mathbf{b}_k, \ \mathbf{\Omega}_k\Big) \tag{15}$$

where the covariance matrix can be shown to be

$$\mathbf{\Omega}_k = \begin{bmatrix} \mathbf{\Sigma}_{s(k)} & 0 \\ 0 & \mathbf{\Gamma}_{s(k)} \end{bmatrix} + \mathbf{A}_k \begin{bmatrix} \mathbf{\Psi}_{z_k} & 0 & 0 \\ 0 & \mathbf{\Psi}_{z_{k+\theta}} & 0 \\ 0 & 0 & \mathbf{\Psi}_{z_{k-\theta}} \end{bmatrix} \mathbf{A}_k^T \tag{16}$$

which is enlarged from that in the conditional likelihood of (9).

It is clear that the acoustic likelihood in (15), where joint cepstra and their temporal differentials are used as the acoustic features, is a principled generalization of that in (3) with only static cepstra as the acoustic features. The quantity $\mathbf{A}_k = \mathcal{F}'[\boldsymbol{z}_k^0]$ in (3) is generalized to (11) with a higher dimension, and the quantity $\mathbf{b}_k$ in (4) is also expanded to a higher dimension shown in (12). The same kind of expansion can be seen for the time-varying covariance matrix $\mathbf{\Omega}_k$, from (5) to (16). Note these quantities are not intrinsic model parameters but functions of them. Hence these quantities and the acoustic likelihood can be computed once the intrinsic model parameters are determined from the training data, which is presented next.

4. PARAMETER ESTIMATION

In this section, we present estimating formulas for the intrinsic model parameters — $\boldsymbol{\mu}_s$, $\boldsymbol{\Sigma}_s$, $\boldsymbol{\delta}_s$, and $\boldsymbol{\Gamma}_s$ — in the above generalized HTM using maximum likelihood. It is assumed that the boundaries of each phone unit, denoted by s, are given (either provided by the databases such as TIMIT, or computed from an HMM system in advance).

We first take derivatives of log products of (15), over all frames in the training data, with respect to the vector-valued parameters $\boldsymbol{\mu}_s$ and $\boldsymbol{\delta}_s$ (related to mean vectors of cepstral prediction residuals). Setting the derivatives to zero and solving for the parameters, we obtain the closed-form estimation formulas:

$$\hat{\boldsymbol{\mu}}_s = \frac{1}{K_s}\sum_{k=1}^{K_s}\Big(\boldsymbol{o}_k - \mathcal{F}[\boldsymbol{z}_k^0] - \mathcal{F}'[\boldsymbol{z}_k^0](\boldsymbol{\mu}_{z_k} - \boldsymbol{z}_k^0)\Big) \tag{17}$$

$$\begin{aligned}\hat{\boldsymbol{\delta}}_s \;=\; & \frac{1}{K_s}\sum_{k=1}^{K_s}\Big[\boldsymbol{d}_k - \frac{1}{2\theta}\Big(\mathcal{F}[\boldsymbol{z}_{k+\theta}^0] + \mathcal{F}'[\boldsymbol{z}_{k+\theta}^0](\boldsymbol{\mu}_{z_{k+\theta}} - \boldsymbol{z}_{k+\theta}^0) \\ & - \mathcal{F}[\boldsymbol{z}_{k-\theta}^0] - \mathcal{F}'[\boldsymbol{z}_{k-\theta}^0](\boldsymbol{\mu}_{z_{k-\theta}} - \boldsymbol{z}_{k-\theta}^0)\Big)\Big]\end{aligned} \tag{18}$$

where K_s is the total number of frames associated with phone unit s in the training data.

For estimating the parameters $\boldsymbol{\Sigma}_s$ and $\boldsymbol{\Gamma}_s$, covariance matrices of cepstral prediction residuals, no closed-form estimation formulas can be obtained. In the current model implementation, we simplify the problem by assuming diagonality of $\boldsymbol{\Sigma}_s$ and $\boldsymbol{\Gamma}_s$ and then estimate their diagonal elements only. Further, we use the "frame-independent approximation" (described in [4] in detail) and generalize the approximate estimation formula derived in [4] for cepstral features only to those with joint static and delta cepstra:

$$\begin{aligned}\mathrm{diag}(\hat{\boldsymbol{\Sigma}}_s) \;\approx\; & \frac{1}{K_s}\sum_{k=1}^{K_s}\Big[\Big(\boldsymbol{o}_k - \mathcal{F}[\boldsymbol{z}_k^0] \\ & - \mathcal{F}'[\boldsymbol{z}_k^0](\boldsymbol{\mu}_{z_k} - \boldsymbol{z}_k^0) - \hat{\boldsymbol{\mu}}_s\Big)^2 \\ & - \mathrm{diag}\Big(\mathcal{F}'[\boldsymbol{z}_k^0]\boldsymbol{\Psi}_{z_k}(\mathcal{F}'[\boldsymbol{z}_k^0])^T\Big)\Big]\end{aligned} \tag{19}$$

$$\begin{aligned}\mathrm{diag}(\hat{\boldsymbol{\Gamma}}_s) \;\approx\; & \frac{1}{K_s}\sum_{k=1}^{K_s}\Big\{\Big[\boldsymbol{d}_k - \frac{1}{2\theta}\Big(\mathcal{F}[\boldsymbol{z}_{k+\theta}^0] \\ & + \mathcal{F}'[\boldsymbol{z}_{k+\theta}^0](\boldsymbol{\mu}_{z_{k+\theta}} - \boldsymbol{z}_{k+\theta}^0) - \mathcal{F}[\boldsymbol{z}_{k-\theta}^0] \\ & - \mathcal{F}'[\boldsymbol{z}_{k-\theta}^0](\boldsymbol{\mu}_{z_{k-\theta}} - \boldsymbol{z}_{k-\theta}^0)\Big) - \hat{\boldsymbol{\delta}}_s\Big]^2 \\ & - \mathrm{diag}\Big(\mathbf{A}_k\begin{bmatrix}\boldsymbol{\Psi}_{z_k} & 0 & 0\\ 0 & \boldsymbol{\Psi}_{z_{k+\theta}} & 0\\ 0 & 0 & \boldsymbol{\Psi}_{z_{k-\theta}}\end{bmatrix}\mathbf{A}_k^T\Big)\Big\}\end{aligned} \tag{20}$$

where element-by-element vector square operations are used above.

5. PHONETIC RECOGNITION EXPERIMENTS

Phonetic recognition experiments are carried out, aimed at evaluating the effectiveness of adding the differential cepstra as acoustic features in the HTM and of the parameter learning algorithm described in this paper. The standard TIMIT phone set with 48 labels is expanded to 58 (as described in [2, 21]) in training the HTM parameters using standard 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 testset of 192 utterances (24 speakers), the same setup as that described in [8].

Phonetic recognition performance is measured from the phonetic decoding results obtained by an A* search on the phonetic lattice, which is generated by our baseline triphone HMM system with the bi-gram language model. (The same LM is used for evaluating the HTM recognizer.) The numbers of the lattice nodes and links per utterance are 1289 and 8276, respectively, averaged over 192 core test-set utterances. A detailed description of this lattice-constrained A* search algorithm can be found in [21] for the HTM with the use of static (frequency-warped) cepstra as the acoustic features, where the complete log likelihood score used in the decoding process uses a weighted sum of 1) log HTM likelihood, 2) log HMM likelihood, 3) LM score, and 4) insertion penalty. The basic structure of this decoder remains the same for the new HTM with the use of joint (frequency-warped) cepstra and their temporal differentials. The main change is to replace the old log HTM likelihood computed using Eq.(3) by the new one according to Eq.(15). In addition, the weights are re-adjusted.

In Table 1 we show phonetic recognition performance comparisons between the earlier version of the HTM [4] and the new version presented in this paper. In addition to the percent Accuracy and Correctness as the most common recognition performance measures, error types (Substitution, Deletion, and Insertion) are listed also for both versions of the HTM. We also show the performance of the baseline HMM using joint static and delta cepstra features in the final row of Table 1. The HMM system has a total of 1,170,000 parameters. In contrast, the two versions of the HTM have much fewer parameters — 6,272 and 11,056, respectively.

Table 1. *Phonetic recognition performance comparisons on the TIMIT core test set between two versions of HTM: HTM which predicts static (frequency-warped) cepstra and HTM which predicts joint static and delta cepstra. Lattice-constrained A* search [21] is used for phonetic decoding with weighted HTM, HMM, and LM scores. The final row lists the performance of the baseline HMM.*

	Acc %	Corr %	Sub %	Del %	Ins %
HTM (static cepstra)	**75.07**	78.28	15.94	5.78	3.20
HTM (static/ delta cepstra)	**75.17**	78.40	15.80	5.80	3.23
HMM (baseline)	**72.48**	75.70	17.74	6.76	3.22

The results in Table 1 show that both HTM systems perform significantly better than the HMM system, and the addition of delta cepstral features in the new version of HTM gives a moderate gain over the earlier version of HTM with static cepstral features alone. The magnitude of this gain for the HTM is smaller than the counterpart for the HMM, likely due to the fact that the HTM has already incorporated dynamic information in the model structure while the HMM does not. Nevertheless, since the performance is already at a high level, the moderate gain observed Table 1 indicates benefits of the expanded capability of the new HTM in predicting the delta cepstral features. To put such performance comparisons into a perspective, we summarize in Table 2 the accuracy performances on the same TIMIT phonetic recognition task reported in the literature using a wide variety of different techniques by other groups worldwide. The best-ever result on TIMIT phonetic recognition task with the core test set is one obtained by combining the results from a large number of classifiers each with different acoustic measurements. Without using such combinations, our HTM performs better than all other techniques in the literature, and is only 0.43% lower

than the best-ever recognition accuracy. In light of these high levels of the performance,[1] the 0.10% gain we have achieved as shown in Table 1 is meaningful and informative.

Table 2. *Summary of phonetic recognition accuracy on the TIMIT core test set in the literature using a wide range of techniques*

Technique	Phone Accuracy %
Triphone Discrete HMMs [12]	66.08
Triphone Continuous HMMs [11]	72.90
Conditional Random Field [13]	65.23
Recurrent Neural Nets [16]	73.90
Large-Margin GMM [18]	67.00
Monophone HTMs (this paper)	75.17
Anti-Phone, Heterogeneous Classifiers [9]	75.60

6. SUMMARY AND CONCLUSION

While the conventional technique for incorporating speech dynamics into speech recognizers involves the use of cross-frame differentials (deltas) of speech features, the realistic speech dynamics (as illustrated in [23, 14, 19], etc.) exhibit more intricate, linguistically correlated patterns far beyond what these simplistic differentials can characterize. Our previous version of the HTM [4] captures some realistic aspects of such speech dynamics in the model structure while predicting the static speech features (in the form of cepstra). In this paper, the HTM is extended so that its dynamic structure is used also to predict cross-frame differentials of speech features as well. We provide a rigorous mathematical formulation of this extended model, and present the newly developed parameter estimation technique.

The evaluation experiments on the standard TIMIT phonetic recognition task demonstrate benefits of adding the new component of the HTM in predicting delta cepstra, measured by moderate recognition accuracy improvement over the earlier version of the HTM. This improvement is on top of the best performance on this task without using many heterogeneous classifiers/recognizers with combined scores. Our future work involves design of more elaborate dynamic acoustic features and of new elements in the hidden structure of the HTM that can accurately predict such features. This enhanced generative modeling capability is expected to further increase phonetic and word recognition accuracy in our HTM-based recognizer.

7. REFERENCES

[1] L. Deng. *DYNAMIC SPEECH MODELS — Theory, Algorithms, and Applications*, Morgan & Claypool Publishers, 2006.

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

[3] L. Deng, H. Attias, L. Lee, and A. Acero. "Adaptive Kalman smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model" *IEEE Trans. Audio, Speech & Language Processing*, Vol. 15, Jan. 2007, pp. 13-23.

[4] L. Deng, D. Yu, and A. Acero. "Structured speech modeling," *IEEE Trans. Audio, Speech & Language Processing*, Vol. 14, No. 5, Sept. 2006, pp. 1492-1504.

[1] We note recently Brno University researchers published much higher, 78.52%, accuracy using hierarchical neural nets [17]. But they merged a stop closure and burst into a single unit, deviating from the standard unit set [12, 8] commonly used in this task and making performance comparisons difficult.

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

[6] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan. "A database of vocal tract resonance trajectories for research in speech processing," *Proc. ICASSP*, May, 2006, Toulouse, France, pp. 60-63.

[7] S. Furui. "Speaker independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoustic, Speech & Sig. Proc.*, Vol. 34 (1), Feb. 1986, pp. 52-59.

[8] J. Glass. "A probabilistic framework for segment-based speech recognition," *Computer Speech and Language,* Vol. 17, No. 2-3, 2003, pp. 137-152.

[9] A. Halberstadt and J. Glass. "Heterogeneous measurements and multiple classifiers for speech recognition," *Proc. ICSLP*, 1998, Sydney Australia, pp. 995-998.

[10] X. Huang, A. Acero, and H. Hon. *SPOKEN LANGUAGE PROCESSING*, Prentice Hall, New York, 2001.

[11] L. Lamel and J. Gauvain. "High performance speaker-independent phone recognition using CDHMM," *Proc. EuroSpeech*, 1993, Berlin, Germany, pp. 121-124.

[12] K.F. Lee and H.W. Hon. "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, Vol. 37, 1989, pp. 1641-1648.

[13] J. Morris and E. Fosler-Lussier. "Combining phonetic attributes using conditional random fields," *Proc. Interspeech*, Pittsburgh, PA, Sept. 2006, pp. 597-600.

[14] J. Olive, A. Greenwood, and J. Coleman. *ACOUSTICS OF AMERICAN ENGLISH SPEECH — A Dynamic Approach*, Springer-Verlag, New York, 1993.

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

[16] A. Robinson. "An application to recurrent nets to phone probability estimation," *IEEE Trans. Neural Networks*, Vol. 5 (2), 1994, pp. 298-305.

[17] P. Schwarz, P. Matejka, and J. Cernocky. "Hierarchical structure of neural networks for phoneme recognition," *Proc. ICASSP*, May, 2006, Toulouse, France.

[18] F. Sha and L. Saul. "Large margin Gaussian mixture modeling for phonetic classification and recognition," *Proc. ICASSP*, Vol. 1, May, 2006, Toulouse, France, pp. 265-268.

[19] K. Stevens, *ACOUSTIC PHONETICS*, MIT Press, 1998.

[20] C. Williams. "How to pretend that correlated variables are independent by using difference observations," *Neural Computation*, Vol. 17, 2005, pp. 1-6.

[21] D. Yu, L. Deng, and A. Acero. "A lattice search technique for long-contextual-span hidden trajectory model of speech," *Speech Communication*, Vol. 48, 2006, pp. 1214-1226.

[22] H. Zen, K. Tokuda, T. Kitamura. "A Viterbi algorithm for trajectory model derived from HMM with explicit relation between static/dynamic features," *Proc. ICASSP*, 2004, pp. 837-840.

[23] V. Zue."NOTES ON SPEECH SPECTROGRAM READING," MIT, Cambridge MA, 1991.