Structured Speech Modeling

Li Deng, *Fellow, IEEE*, Dong Yu, *Member, IEEE*, and Alex Acero, *Fellow, IEEE*

***Abstract*—Modeling dynamic structure of speech is a novel paradigm in speech recognition research within the generative modeling framework, and it offers a potential to overcome limitations of the current hidden Markov modeling approach. Analogous to structured language models where syntactic structure is exploited to represent long-distance relationships among words [5], the structured speech model described in this paper makes use of the dynamic structure in the hidden vocal tract resonance space to characterize long-span contextual influence among phonetic units. A general overview is provided first on hierarchically classified types of dynamic speech models in the literature. A detailed account is then given for a specific model type called the hidden trajectory model, and we describe detailed steps of model construction and the parameter estimation algorithms. We show how the use of resonance target parameters and their temporal filtering enables joint modeling of long-span coarticulation and phonetic reduction effects. Experiments on phonetic recognition evaluation demonstrate superior recognizer performance over a modern hidden Markov model-based system. Error analysis shows that the greatest performance gain occurs within the sonorant speech class.**

***Index Terms*—Hidden dynamics, hidden trajectory, long span modeling, maximum-likelihood, nonlinear prediction, parameter learning, structured modeling, vocal tract resonance.**

I. Introduction

OVER THE past two decades, significant progress has been made in advancing speech recognition technology providing a key modality for human–machine interaction. The technology is improving at a steady pace, and is becoming increasingly usable and useful. Despite this progress, some fundamental and practical limitations in the technology have hindered its widespread use. There has been a large performance gap between human and machine speech recognition. Mainstream adoption of speech recognition would not be possible without the underlying recognition technology that can deliver a sufficiently robust and low-error performance. Reducing speech recognizers' error rates under all deployment environments remains the greatest challenge to making speech mainstream. Many leading researchers in the field understand the fragile nature of the current speech recognition system design, and have advocated that new, serious research is needed to overcome some fundamental limitations of the current speech recognition technology (e.g., [3], [6], [9], [29], [34], [40], [42], [49], [53], [59]).

One clear limitation of this kind pertains to the naturalness of speech on the part of the speaker interacting with automatic recognizers. While the ultimate goal for speech recognition is to make it indistinguishable with human–human verbal interaction, at present, when users interact with any existing speech recognition system, they have to be fully aware of the fact that their conversation "partner" is a machine. The machine would easily break if the users were to speak in a casual and natural style as if they were talking with a friend. In order to enable mainstream use of speech recognition, naturalness or the free style of speaking should not produce so many recognition errors, making the recognition systems practically unusable as it is the case today. This serious inadequacy of the current speech recognition technology forms the motivation of our research on structured speech modeling and on its application to speech recognition.

The research reported in this paper was conducted within the "Novel-Approach" component of the DARPA's EARS program during 2002–2005. Our overall research program has been developed based on principles of speech and language science and on computational models for the essential aspects of the human verbal communication process that is responsible for the generation of naturally uttered human speech signals with an unconstrained style. Quantitative and structured speech models have been developed in statistical terms, so that advanced algorithms can be developed to automatically and optimally determine the physically meaningful parameters in the models from a representative set of training data. The speech recognizer architecture designed in this approach is significantly different from that based on the conventional hidden Markov models (HMMs). Some detailed stages in the human speech generation chain from the distinctive feature-based linguistic units to speech acoustics are represented explicitly. The main advantage of representing such a detailed structure in the human speech process is that a highly compact set of parameters can now be used to capture phonetic context and speaking rate/style variations in a common framework. Using this framework, many important subjects in speech science and those in speech recognition that were previously studied separately by different communities of speech researchers can now be investigated in a unified fashion.

While our overall research program has been aimed at a comprehensive framework, where a detailed hierarchy in speech generation is exploited (see some descriptions in [9], [13], [16]), the approach presented in this paper is a significantly simplified version. The simplification is adopted to facilitate the development of learning algorithms and the implementation of the recognizer for evaluation. However, the essential aspect of the comprehensive framework—the dynamic structure of speech—remains unchanged in the current recognizer implementation. One key insight gained from scientific studies (e.g., [13], [52], [58]) is the importance of the highly regular dynamic patterns associated with natural human speech in the underlying structure that reflects both contextual influences (coarticulation) and incomplete articulation (reduction) as a common physical cause. It is our belief that making use of this dynamic structure can benefit

Manuscript received October 14, 2005; revised January 27, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Geoffrey Zweig.

The authors are with Microsoft Research, Redmond, WA 98052 USA (e-mail: deng@microsoft.com; dongyu@microsoft.com; alexac@microsoft.com).

Digital Object Identifier 10.1109/TASL.2006.878265

1558-7916/$20.00 © 2006 IEEE

automatic speech recognizers in its ability to handle free-style speech and to reduce the recognizers' reliance on large amounts of training data. And functional, quantitative modeling of this dynamic structure, simplistic as it may be, will make it feasible to construct an effective speech recognizer. In this paper, we will focus on a specific implementation of a vocal-tract-resonance (VTR)-based dynamic model and on a speech recognizer built on this model.

The organization of this paper is as follows. In Section II, we provide a general overview of a rich body of literature on functional dynamic models for speech features, including some of our earlier work. This sets up the background for a specific type of the model that we have spent most of the effort in developing during our participation in the EARS program. This model type, which we call hidden trajectory model (HTM), is introduced in Section III, based on the statistical generative process for the speech observations. Details of the HTM parameterization and parameter estimation are presented in Section IV. This is followed by experimental evaluation of the HTM in Section V.

II. Overview of Dynamic Speech Models

A. *Multiple Levels of Dynamics in Human Speech Generation*

As a linguistic and physical abstraction, human speech generation can be functionally represented at four distinctive but correlated levels of dynamics. The top level of the dynamics is symbolic or phonological. The linear sequence of speech units in linear phonology or the nonlinear (or multilinear) sequence of the units in autosegmental or articulatory phonology demonstrate the discrete, time-varying nature of the speech dynamics at the mental motor-planning level of speech production (cf. [13, Chap. 8, 9]). The next level of the dynamics is continuous valued and associated with the functional, "task" variables in speech production. At this level, the goal or "task" of speech generation is defined, which may be either the acoustic goal such as vocal tract resonances or formants, or the articulatory goal such as vocal-tract constrictions, or their combination (cf. [13, Chap. 7, 10]). This task level can be considered as the interface between phonology and phonetics, since it is at this level that each symbolic phonological unit is mapped to a unique set of the phonetic parameters. These parameters are often called the correlates of the phonological units. The third level of the dynamics occurs at the physiological articulators. Such articulatory dynamics are a nonlinear transformation of the task dynamics [47]. Finally, the last level of the dynamics is the acoustic one, where speech "observations" are computed in speech recognition applications.

Several different types of computational dynamic models for speech generation presented in this section will be organized in view of the above functional levels of the dynamics. To make this overview most relevant to the specific HTM implementation during our EARS work, we will classify the models into two main categories. In the first category are the models focusing on the lowest, acoustic level of dynamics. This class of models is often called the stochastic segment models as are well known through the earlier review paper [41]. The second category consists of what is called the *hidden dynamic model* where the task dynamic and articulatory dynamic levels are functionally grouped into a functional single-level dynamics. In contrast to the acoustic-dynamic model which represents coarticulation at the surface, observational level, the hidden dynamic model explores a deeper, unobserved (hence "hidden") level of the speech dynamic structure that regulates coarticulation and phonetic reduction.

B. *Category-I: Acoustic Dynamic Models*

Hidden Markov model (HMM) is the simplest kind of the acoustic dynamic model in this category. Stochastic segment models are a broad class of statistical models that generalize from the HMM and that intend to overcome some shortcomings of the HMM such as the conditional independent assumption and its consequences. (This assumption is grossly unrealistic and restricts the ability of the HMM as an accurate generative model.) The generalization is in the following sense: In an HMM, one frame of speech acoustics is generated by visiting each HMM state, while a variable-length sequence of speech frames is generated by visiting each "state" of a stochastic segment model. Each state is associated with a random sequence length.

Similar to an HMM, a stochastic segment model can be viewed as a generative process for the observation data sequences. It is intended to model the acoustic feature trajectories and temporal correlations that have been inadequately represented by an HMM. This is accomplished by introducing new parameters to characterize the trajectories and the temporal correlations.

A convenient way to understand a variety of stochastic segment models and their relationships is to establish a hierarchy showing how the HMM is generalized by gradually relaxing the modeling assumptions. Starting with a conventional HMM in this hierarchy, there are two main classes of its extended or generalized models. Each of these classes further contains subclasses of models. We describe this hierarchy below.

1) Nonstationary-State HMMs: This model class has also been called the trended HMM, constrained mean trajectory model, segmental HMM, or stochastic trajectory model, etc., with minor variations according to whether the parameters defining the trend functions are random or not and how their temporal properties are constrained. Given the HMM state s, the sample paths of most of these model types are piecewise, explicitly defined acoustic feature trajectories[1]

$$\mathbf{o}(k) = \mathbf{g}_k(\mathbf{\Lambda}_s) + \mathbf{r}_s(k) \tag{1}$$

where $\mathbf{g}_k(\mathbf{\Lambda}_s)$ is the deterministic function of time frame k, parameterized by state-specific $\mathbf{\Lambda}_s$, which can be either deterministic or random, and $\mathbf{r}_s(k)$ is a state-specific stationary residual signal. Further classification of nonstationary-state HMMs is as follows.

- Polynomial trended HMM:
  — Observable trend functions: This is the simplest trended HMM where there is no uncertainty in polynomial coefficients $\mathbf{\Lambda}_s$ [18], [21], [22], [27], [35].
  — Random trend functions: The trend functions $\mathbf{g}_k(\mathbf{\Lambda}_s)$ are stochastic due to the uncertainty in polynomial co-

[1] When no temporal recursion is involved in characterizing the time-varying (dynamic) function, we call this specific type of the dynamic function as a "trajectory," or a kinematic function.

efficients $\mathbf{\Lambda}_s$. $\mathbf{\Lambda}_s$ are random vectors in one of two ways: 1) $\mathbf{\Lambda}_s$ has a discrete distribution [19], [31]; and 2) $\mathbf{\Lambda}_s$ has a continuous distribution. In the latter case, the model has been called the segmental HMM, where the earlier versions have a polynomial order of zero [25] and the later versions have an order of one [30] or two [45].

- Nonparametric trended HMM: The trend function is determined by the training data after performing dynamic time warping [28], rather than by any parametric form.
- The trend function derived by differential (delta) features [38], [56].
- The trend function derived by a nonlinear function based on posterior probability computation [51].

2) Multiregion Recursive Linear Models: Common to this model class is the linear recursive form in dynamic modeling of the region-dependent time-varying acoustic feature vectors, where the "region" is often associated with a phonetic unit. The most typical recursion is of the form

$$\mathbf{o}(k) = \mathbf{\Lambda}_s(1)\mathbf{o}(k-1) + \cdots + \mathbf{\Lambda}_s(p)\mathbf{o}(k-p) + \mathbf{r}_s(k) \quad (2)$$

and the starting point of the recursion for each state s comes usually from the previous state's ending history.

The model expressed in (2) provides clear contrast to the trajectory or trended models where the time-varying acoustic features are approximated as an explicit temporal function of time. The sample paths of the model (2), on the other hand, are piecewise, recursively defined stochastic time-varying functions. Further classification of this model class is as follws.

- Autoregressive HMM: The time-varying function associated with each region (HMM state) is defined by linear prediction (i.e., recursively defined autoregression function) for the acoustic features [14], [33] or waveforms [44], [48].
- Switching linear dynamic system model: The model not only uses the autoregression function to recursively define "state" dynamics, but also involves an additional noisy observation function. The actual effect of autoregression is to smooth the observed acoustic feature vectors ([23]).

C. Category-II: Hidden Dynamic Models

The many types of acoustic dynamic or stochastic segment models described previously generalize the HMM by generating a variable-length sequence of speech frames in each state, overcoming the HMM's assumption of local conditional independence. Yet the inconsistency between the HMM assumptions and the properties of the realistic dynamic speech process goes beyond this limitation. In stochastic segment models, the speech frames assigned to the same segment/state have been modeled to be temporally correlated and the model parameters been time varying. However, the lengths of such segments are typically short. Longer-term correlation across phonetic units, which provides dynamic structure responsible for coarticulation and phonetic reduction, in a full utterance has not been captured.

In contrast, a more advanced class of dynamic speech models, *hidden dynamic models*, or structured speech models, explicitly captures the long-contextual-span properties over the phonetic units by imposing continuity constraints on the hidden dynamic variables internal to the acoustic observation data. This constraint is motivated by physical properties of speech generation. The constraint captures some key coarticulation and reduction properties in speech, and makes the model parameterization more parsimonious than the acoustic dynamic models where the coarticulation model requires a large number of free parameters. Since the underlying speech structure represented by the hidden dynamic model links a sequence of segments via continuity in the hidden dynamic variables, it can also be appropriately termed as the *super-segmental* model.

Differing from the acoustic dynamic models, the hidden dynamic models represent speech structure in the hidden dynamic variables. Depending on the nature of these dynamic variables in light of multilevel speech dynamics discussed earlier, the hidden dynamic models can be broadly classified into 1) articulatory dynamic model, 2) task-dynamic model, and 3) vocal tract resonance (VTR) dynamic model. The VTR dynamics are a special type of task dynamics, with the acoustic goal or "task" of speech production in the VTR domain. Key advantages of using VTRs as the "task" are their direct correlate to the acoustic information, and the lower dimensionality in the VTR vector compared with the counterpart hidden vectors either in the articulatory dynamic model or in the task-dynamic model with articulatorily defined goals or "tasks" such as vocal tract constriction properties.

As an alternative classification scheme, the hidden dynamic models, like acoustic dynamic models, can also be classified, from the computational perspective, according to whether the hidden dynamics are represented mathematically with temporal recursion or not. These two types of the models are briefly reviewed here.

1) Multiregion Nonlinear Dynamic System Models: The hidden dynamic models in this first class use the temporal recursion (k-recursion via the predictive function $\mathbf{g}_k$ in (3) below) to define the hidden dynamics $\mathbf{z}(k)$. Each region s of such dynamics is characterized by the s-dependent parameter set $\mathbf{\Lambda}_s$, with the "state noise" denoted by $\mathbf{w}_s(k)$. The memoryless nonlinear mapping function is exploited to link the hidden dynamic vector $\mathbf{z}(k)$ to the observed acoustic feature vector $\mathbf{o}(k)$, with the "observation noise" denoted by $\mathbf{v}_s(k)$, and parameterized also by region dependent parameters. The combined "state equation" (3) and "observation equation" (4) form a general multiregion nonlinear dynamic system model:

$$\begin{aligned}\mathbf{z}(k+1) &= \mathbf{g}_k[\mathbf{z}(k), \mathbf{\Lambda}_s] + \mathbf{w}_s(k) &(3)\\ \mathbf{o}(k') &= \mathbf{h}_{k'}[\mathbf{z}(k'), \mathbf{\Omega}_{s'}] + \mathbf{v}_{s'}(k'). &(4)\end{aligned}$$

where subscripts k and k' indicate that the functions $\mathbf{g}[\cdot]$ and $\mathbf{h}[\cdot]$ are time varying and may be asynchronous with each other. s or s' denotes the dynamic region correlated with phonetic categories.

Various simplified implementations of the aforementioned generic nonlinear system model have appeared in the literature (e.g., [4], [15], [17], [20], [24], [26], [37]). Most of these implementations reduce the predictive function $\mathbf{g}_k$ in the state equation (3) into a linear form and use the concept of phonetic targets as part of the parameters. This gives rise to linear target filtering (by infinite impulse response or IIR filters) as a model for the hidden dynamics. Also, many of these implementations use neural networks as the nonlinear mapping function $\mathbf{h}_k[\mathbf{z}(k), \mathbf{\Omega}_s]$ in the observation equation (4).

2) Hidden Trajectory Models: The second type of the hidden dynamic models use trajectories (i.e., explicit functions of time, with no recursion) to represent the temporal evolution of the hidden dynamic variables (e.g., VTR or articulatory vectors). This *hidden* trajectory model (HTM) differs conceptually from the acoustic dynamic or trajectory model in that the articulatory-like constraints and structure can be captured in the HTM via the continuous valued hidden variables that run across the phonetic units. Importantly, the polynomial trajectories, which were shown to fit well to the temporal properties of cepstral features [18], [35], are not appropriate for the hidden dynamics since they do not generally satisfy at least two physical constraints required for realistic articulation-like dynamics. One of these constraints is "segment-bound monotonicity," meaning that for a duration of about a segment's length, the hidden dynamic variables move in a monotonic manner to reflect the behavior of articulatory inertia. The second constraint violated by the polynomial trajectory is "target-directedness." That is, the movements of articulators or hidden dynamic variables are directed or attracted toward segment-bound targets. They may not reach the targets, but the direction of the movements are strongly constrained in this way.

One parametric form of the hidden trajectory constructed to satisfy both of these constraints is the following explicit temporal function of time k:

$$\mathbf{g}_k(\mathbf{\Lambda}_s) = \mathbf{t}_s + (\mathbf{g}_0 - \mathbf{t}_s)(1 + \mathbf{d}_s \times k)\exp(-\boldsymbol{\gamma}_s \times k) \quad (5)$$

as proposed and analyzed in [15]. It can be shown that when two such trajectories corresponding to two adjacent phones are concatenated, the temporal derivative of at the connection point is smooth. (This would not be the case for the exponential trajectory.) In (5), the parameter set $\mathbf{\Lambda}_s$ consists of $[\mathbf{t}_s, \mathbf{d}_s, \boldsymbol{\gamma}_s]$, where $\mathbf{t}_s$ is the segmental target vector which directs the movement in the time-varying hidden dynamic vector, and $\mathbf{d}_s$ and $\boldsymbol{\gamma}_s$ jointly control the shape of the trajectory. $\mathbf{g}_0$ is a "nuisance" parameter, taking the value of the hidden trajectory vector associated with the preceding unit at the boundary with the currect unit denoted by s. An implementation and evaluation of this model for speech recognition can be found in [50] and [57], where a conversion of this model to a second-order recursive dynamic model was made in the implementation.

Another parametric form of the hidden trajectory, which also satisfies these two constraints has been developed more recently [8], [10], [11] based on finite-impulse response (FIR) filtering of VTR target sequences. No temporal recursions on the hidden VTR vectors are used for defining the dynamics. Compared with the trajctories parameterized by (5), the trajectories constructed by FIR filtering give more flexibility to control the switching point in the hidden dynamics from one segment to another. The development of this latter parametric form of the HTM has constituted our major effort in the EARS program. In the remainder of this paper, we provide a systematic account of this model, synthesizing and expanding our earlier descriptions of this work in [10] and [11] but with more detailed experimental evaluation results.

III. Hidden Trajectory Modeling With Target Filtering

As a special type of the hidden dynamic model, the HTM presented in this section is a structured generative model, from the top level of phonetic specification to the bottom level of acoustic observations via the intermediate level of (nonrecursive) FIR-based target filtering that generates hidden VTR trajectories. One advantage of the FIR filtering is its natural handling of the two constraints (segment-bound monotonicity and target-directedness discussed earlier) that often require asynchronous segment boundaries for the VTR dynamics and for the acoustic observations. This asynchrony can be explained as follows. On the one hand, the segment boundaries for acoustic observations are most appropriately determined by the switch between a change of the phonetic feature for manner of articulation. On the other hand, with the same "vocalic" manner of articulation (e.g., a vowel segment) VTRs or formants often curve in the middle of the segment (e.g., [58]). To apply the trajectory model such as (5) and to demand segment-bound monotonicity, the boundaries for hidden VTR dynamics would need to be defined somewhere inside the vocalic segment, making it asynchronous with the segment boundaries for acoustic observations. The trajectories generated by FIR filtering allow nonmonotonic VTR movements within the segment boundaries for acoustic observations, while satisfying the constraint of segment-bound monotonicity with the asynchronous segment-boundaries associated with the hidden VTR.

This section is devoted to the mathematical formulation of the HTM as a statistical generative model. Parameterization of the model is detailed here, with consistent notations set up to facilitate the derivation and description of algorithmic learning of the model parameters presented in Section IV.

A. Generating Stochastic Hidden VTR Trajectories

The HTM assumes that each phonetic unit is associated with a multivariate distribution of the VTR targets.[2] 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

$$p(\boldsymbol{t}|s) = \mathcal{N}(\boldsymbol{t}; \boldsymbol{\mu}_{T_s}, \boldsymbol{\Sigma}_{T_s}). \quad (6)$$

The generative process in the HTM starts by temporal filtering the stochastic targets. This results in a time-varying pattern of stochastic hidden VTR vectors $\boldsymbol{z}(k)$. The filter is constrained so that the smooth temporal function of $\boldsymbol{z}(k)$ moves segment-by-segment toward the respective target vector $\boldsymbol{t}_s$, but it may or may not reach the target depending on the degree of phonetic reduction.

These phonetic targets are segmental in that they do not change over the phone segment once the sample is taken, and

[2] There are exceptions for several compound phonetic units, including diphthongs and affricates, where two distributions are used.

they are assumed to be largely context independent. In our HTM implementation, the generation of the VTR trajectories from the segmental targets is through a bidirectional FIR filtering. The impulse response of this noncausal filter is

$$h_s(k) = \begin{cases} c_{\boldsymbol{\gamma}_s}\boldsymbol{\gamma}_{s(k)}^{-k}, & -D < k < 0 \\ c_{\boldsymbol{\gamma}_s}, & k = 0 \\ c_{\boldsymbol{\gamma}_s}\boldsymbol{\gamma}_{s(k)}^{k}, & 0 < k < D \end{cases} \tag{7}$$

where k represents time frame (typically with a length of 10 ms each), and $\gamma_{s(k)}$ is the segment-dependent "stiffness" scaler parameter. It is positive and real-valued, ranging between zero and one. $c_{\boldsymbol{\gamma}_s}$ in (7) is a normalization "constant" (a function of state s via the dependency on $\boldsymbol{\gamma}$), ensuring that $h_s(k)$ sums to one over all time frames k. 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 (7) 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 $\boldsymbol{t}(k)$, 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 home segment s is

$$\boldsymbol{z}(k) = h_{s(k)} * \boldsymbol{t}(k) = \sum_{\tau=k-D}^{k+D} c_{\boldsymbol{\gamma}_s}\boldsymbol{\gamma}_{s(\tau)}^{|k-\tau|}\boldsymbol{t}_{s(\tau)} \tag{8}$$

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 (8) 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.

The linearity between $\boldsymbol{z}$ and $\boldsymbol{t}$ as in (8) 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 trajectory

$$p(\boldsymbol{z}(k)|s) = \mathcal{N}[\boldsymbol{z}(k); \boldsymbol{\mu}_{z(k)}, \boldsymbol{\Sigma}_{z(k)}]. \tag{9}$$

The aforementioned mean vector is determined by the filtering function

$$\boldsymbol{\mu}_{z(k)} = \sum_{\tau=k-D}^{k+D} c_{\boldsymbol{\gamma}}\boldsymbol{\gamma}_{s(\tau)}^{|k-\tau|}\boldsymbol{\mu}_{T_{s(\tau)}} = \mathbf{a}_k \cdot \boldsymbol{\mu}_T. \tag{10}$$

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

$$\mu_{z(k)}(f) = \sum_{l=1}^{L} a_k(l)\mu_T(l,f) \tag{11}$$

where L is the total number of phone-like HTM units as indexed by l, and $f = 1, \ldots, 8$ for four VTR frequencies and four corresponding bandwidths. Note that $\mathbf{a}_k$ cannot be easily written in a concise mathematical form. It is defined by a constructive process which will be described shortly.

The covariance matrix in (9) 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

$$\boldsymbol{\sigma}_{z(k)}^2 = \boldsymbol{v}_k \cdot \boldsymbol{\sigma}_T^2 \tag{12}$$

and the target covariance matrix is also approximated as diagonal

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

The fth element of the vector in (12) is

$$\sigma_{z(k)}^2(f) = \sum_{l=1}^{L} v_k(l)\sigma_T^2(l,f). \tag{13}$$

In (10) and (12), $\mathbf{a}_k$ 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 unit beyond the $2D+1$ window contributes a zero value to these "coarticulation" vectors' elements. Both $\mathbf{a}_k$ and $\boldsymbol{v}_k$ are a function of the phones' identities and temporal orders in the utterance, and are independent of the VTR dimension f. As an illustration, we show in Fig. 1 the $\mathbf{a}_k$ values for a TIMIT utterance. The values are separated out for each of the L phones. 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 automatically [due to the normalization factor in the FIR impulse response function (7)]. 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].

B. Generating Acoustic Data

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 linear prediction coefficient cepstra (LPCC) (and their frequency-warped version) in this paper.

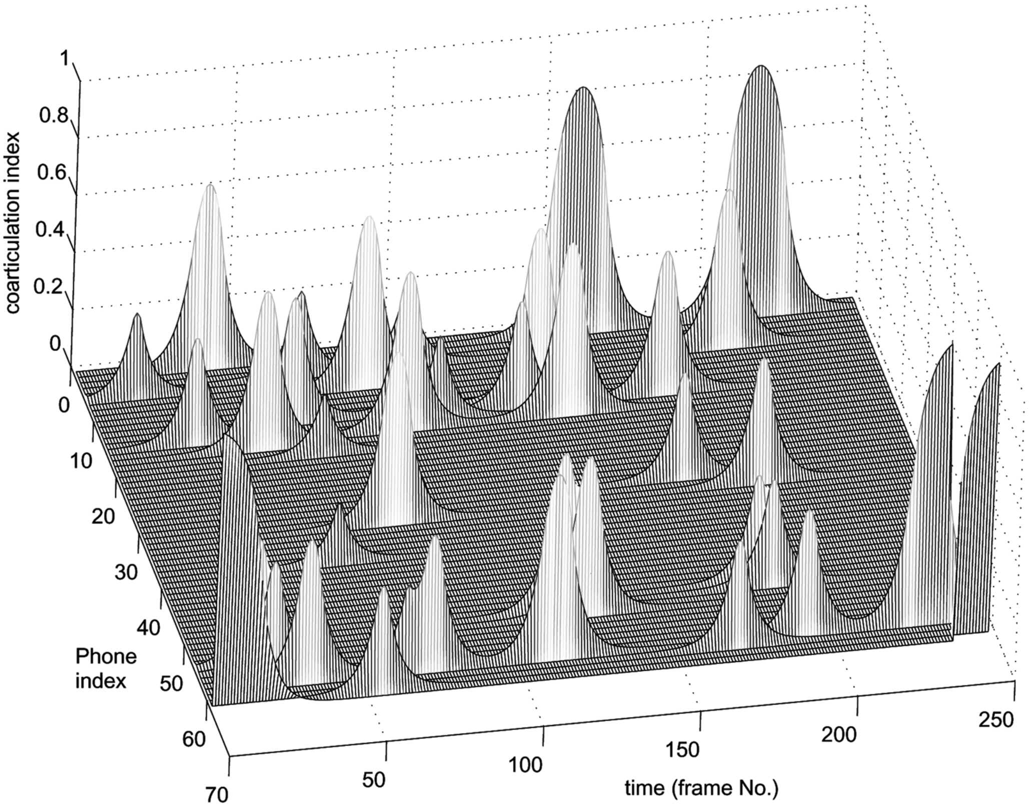

Fig. 1. Illustration of time-varying coarticulatory vectors $\mathbf{a}_k$'s for a TIMIT utterance. See text for detailed explanations.

An analytical form of the nonlinear prediction function $\mathcal{F}[\boldsymbol{z}(k)]$ was presented in [7] and is summarized here. The nth component of the vector-valued function output $\mathcal{F}[\boldsymbol{z}(k)]$ has the form

$$\mathcal{F}_n(k) = \frac{2}{n}\sum_{p=1}^{P} e^{-\pi n (b_p(k)/f_s)} \cos\left(2\pi n \frac{f_p(k)}{f_s}\right) \qquad (14)$$

where f_s is the sampling frequency, P is the highest VTR order ($P = 4$ in this work), and n is the cepstral order. And f_p is the pth order VTR frequency, whose corresponding VTR bandwidth is b_p. Four f_p's and four b_p's constitute the VTR vector as the input argument $\boldsymbol{z}(k)$ in the nonlinear function $\mathcal{F}[\boldsymbol{z}(k)]$. A detailed derivation of (14) provided in [7] has been based on an all-pole model of the speech waveform with no additional assumptions.

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 $\boldsymbol{\Sigma}_{r_{s(k)}}$

$$p(\mathbf{r}_s(k)|\boldsymbol{z}(k), s) = \mathcal{N}\left[\mathbf{r}_s(k); \boldsymbol{\mu}_{r_{s(k)}}, \boldsymbol{\Sigma}_{r_{s(k)}}\right]. \qquad (15)$$

Then, the conditional distribution of the observation becomes

$$p(\boldsymbol{o}(k)|\boldsymbol{z}(k), s) = \mathcal{N}\left[\boldsymbol{o}(k); \mathcal{F}[\boldsymbol{z}(k)] + \boldsymbol{\mu}_{r_{s(k)}}, \boldsymbol{\Sigma}_{r_{s(k)}}\right]. \qquad (16)$$

An alternative form of the distribution in (16) is the following "observation equation:"

$$\boldsymbol{o}(k) = \mathcal{F}[\boldsymbol{z}(k)] + \boldsymbol{\mu}_{r_{s(k)}} + \mathbf{w}_s(k)$$

with the Gaussian observation noise $\mathbf{w}_s(k) \sim \mathcal{N}(\boldsymbol{v}_s; \mathbf{0}, \boldsymbol{\Sigma}_{r_{s(k)}})$.

C. Linearizing Cepstral Prediction Function

To facilitate computing the acoustic observation (LPCC) likelihood (Section III-D), it is important to characterize the LPCC 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 (16) by using the first-order Taylor series approximation

$$\mathcal{F}[\boldsymbol{z}(k)] \approx \mathcal{F}[\boldsymbol{z}_0(k)] + \mathcal{F}'[\boldsymbol{z}_0(k)](\boldsymbol{z}(k) - \boldsymbol{z}_0(k)) \qquad (17)$$

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

$$\mathcal{F}'_n[f_p(k)] = -\frac{4\pi}{f_s} e^{-\pi n (b_p(k)/f_s)} \sin\left(2\pi n \frac{f_p(k)}{f_s}\right) \qquad (18)$$

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

$$\mathcal{F}'_n[b_p(k)] = -\frac{2\pi}{f_s} e^{-\pi n (b_p(k)/f_s)} \cos\left(2\pi n \frac{f_p(k)}{f_s}\right) \qquad (19)$$

for the VTR bandwidth components of $\boldsymbol{z}$. In the current implementation, the Taylor series expansion point $\boldsymbol{z}_0(k)$ in (17) is taken as the tracked VTR values based on the HTM.[3]

Substituting (17) into (16), 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 $\boldsymbol{z}$

$$p(\boldsymbol{o}(k)|\boldsymbol{z}(k), s) \approx \mathcal{N}(\boldsymbol{o}(k); \boldsymbol{\mu}_{o_{s(k)}}, \boldsymbol{\Sigma}_{r_{s(k)}}) \tag{20}$$

where

$$\boldsymbol{\mu}_{o_{s(k)}} = \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{z}(k) + \Big[\mathcal{F}[\boldsymbol{z}_0(k)] - \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{z}_0(k) + \boldsymbol{\mu}_{r_{s(k)}}\Big]. \tag{21}$$

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

D. Computing Acoustic Likelihood by Marginalizing Over VTR Uncertainty

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 LPCC 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. Some detailed steps of derivation give

$$\begin{aligned} p(\boldsymbol{o}(k)|s) &= \int p[\boldsymbol{o}(k)|\boldsymbol{z}(k), s]p[\boldsymbol{z}(k)|s]dz \\ &\approx \int \mathcal{N}[\boldsymbol{o}(k); \boldsymbol{\mu}_{o_{s(k)}}, \boldsymbol{\Sigma}_{r_{s(k)}}] \\ &\quad \times \mathcal{N}[\boldsymbol{z}(k); \boldsymbol{\mu}_{z(k)}, \boldsymbol{\Sigma}_{z(k)}]dz \\ &= \mathcal{N}\left\{\boldsymbol{o}(k); \bar{\boldsymbol{\mu}}_{o_s(k)}, \bar{\boldsymbol{\Sigma}}_{o_s(k)}\right\} \end{aligned} \tag{22}$$

where the time (k)-varying mean vector is

$$\bar{\boldsymbol{\mu}}_{o_s}(k) = \mathcal{F}[\boldsymbol{z}_0(k)] + \mathcal{F}'[\boldsymbol{z}_0(k)][\mathbf{a}_k \cdot \boldsymbol{\mu}_T - \boldsymbol{z}_0(k)] + \boldsymbol{\mu}_{r_{s(k)}} \tag{23}$$

and the time-varying covariance matrix is

$$\bar{\boldsymbol{\Sigma}}_{o_s}(k) = \boldsymbol{\Sigma}_{r_{s(k)}} + \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{\Sigma}_{z(k)}(\mathcal{F}'[\boldsymbol{z}_0(k)])^{\mathrm{Tr}}. \tag{24}$$

1) Interpretation and Analysis: The final result of (22)–(24) are quite intuitive. For instance, when the Taylor series expansion point is set at $\boldsymbol{z}_0(k) = \boldsymbol{\mu}_{z(k)} = \mathbf{a}_k \cdot \boldsymbol{\mu}_T$, (23) is simplified to $\bar{\boldsymbol{\mu}}_{o_s}(k) = \mathcal{F}[\boldsymbol{\mu}_{z(k)}] + \boldsymbol{\mu}_{r_s}$, which is the noise-free part of cepstral prediction. Also, the covariance matrix in (22) is increased by the quantity $\mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{\Sigma}_{z(k)}(\mathcal{F}'[\boldsymbol{z}_0(k)])^{\mathrm{Tr}}$ over the covariance matrix for the cepstral residual term $\boldsymbol{\Sigma}_{r_{s(k)}}$ only. This magnitude of increase reflects the newly introduced uncertainty in the hidden variable, measured by $\boldsymbol{\Sigma}_{z(k)}$. The variance amplification factor $\mathcal{F}'[\boldsymbol{z}_0(k)]$ results from the local "slope" in the nonlinear function $\mathcal{F}[\boldsymbol{z}]$ which maps from the VTR vector $\boldsymbol{z}(k)$ to cepstral vector $\boldsymbol{o}(k)$.

It is also interesting to interpret the likelihood score (22) as probabilistic characterization of a temporally varying Gaussian process, where the time-varying mean vector is expressed in (23) and the time-varying covariance matrix is expressed in (24). This may make the HTM look ostensibly like a nonstationary-state HMM (within the acoustic dynamic model category). However, the key difference is that in HTM the dynamic structure represented by the hidden VTR trajectory enters into the time-varying mean vector (23) in two ways: 1) as the argument $\boldsymbol{z}_0(k)$ in the nonlinear function $\mathcal{F}[\boldsymbol{z}_0(k)]$; and 2) as the term $\mathbf{a}_k \cdot \boldsymbol{\mu}_T = \boldsymbol{\mu}_{z(k)}$ in (23). Being closely related to the VTR tracks, they both capture long-span contextual dependency, yet with mere context-independent VTR target parameters. Similar properties apply to the time-varying covariance matrices in (24). In contrast, the time-varying *acoustic* dynamic models do not have these desirable properties. For example, the polynomial trajectory model [18], [27], [35] does regression fitting directly on the cepstral data, exploiting no underlying speech structure and hence requiring context-dependent polynomial coefficients for representing coarticulation. Likewise, the more recent trajectory model [51] also relies on a very large number of free model parameters to capture acoustic feature variations.

IV. Parameter Estimation for Hidden Trajectory Model

In this section, we will present in detail a novel parameter estimation algorithm we have developed and implemented for the HTM described in the preceding section, using the LPCCs as the acoustic observation data in the training set. The criterion used for this training is to maximize the acoustic observation likelihood in (22). The full set of the HTM parameters consists of those characterizing the LPCC residual distributions and those characterizing the VTR target distributions. We present their estimation separately below, assuming that all phone boundaries are given (as in the TIMIT training data set in our experiments reported in Section V).[4]

A. 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 subphone segmental unit s.

[3] Due to space limitation, we omit presenting the VTR tracking algorithm in this paper. Use of the tracked VTR as $\boldsymbol{z}_0(k)$ has been shown to give better recognition results than using $\boldsymbol{\mu}_{z(k)}$ as $\boldsymbol{z}_0(k)$ that we attempted in the past.

[4] If the phone boundaries are not given in the database, then the training algorithms in this section will be applied based on predefined phone segmentation obtained from a baseline HMM system. Then the segmentation will be refined using the trained HTM and the process iterates itself. The segmentation algorithm for the HTM is currently under development and will not be described in this paper.

1) Mean Vectors: To find the ML (maximum likelihood) estimate of parameters $\boldsymbol{\mu}_{r_s}$, we set

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

where $p(\boldsymbol{o}(k)|s)$ is given by (22), and K_s denotes the total duration of subphone s in the training data. This gives

$$\sum_{k=1}^{K_s} \left[\boldsymbol{o}(k) - \bar{\boldsymbol{\mu}}_{o_s}\right] = 0 \text{ or} \tag{25}$$

$$\sum_{k=1}^{K_s} \Big[\boldsymbol{o}(k) - \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{\mu}_{z(k)} - \left\{\mathcal{F}[\boldsymbol{z}_0(k)] + \boldsymbol{\mu}_{r_s} - \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{z}_0(k)\right\}\Big] = 0. \tag{26}$$

Solving for $\boldsymbol{\mu}_{r_s}$, we have the estimation formula of

$$\hat{\boldsymbol{\mu}}_{r_s} = \frac{\sum_k \left[\boldsymbol{o}(k) - \mathcal{F}[\boldsymbol{z}_0(k)] - \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{\mu}_{z(k)} + \mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{z}_0(k)\right]}{K_s}. \tag{27}$$

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

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

which gives

$$\sum_{k=1}^{K_s} \left[\frac{\boldsymbol{\sigma}^2_{r_s} + \mathbf{q}(k) - (\boldsymbol{o}(k) - \bar{\boldsymbol{\mu}}_{\boldsymbol{o}_s})^2}{\left[\boldsymbol{\sigma}^2_{r_s} + \mathbf{q}(k)\right]^2}\right] = 0 \tag{28}$$

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

$$\mathbf{q}(k) = \text{diag}\left[\mathcal{F}'[\boldsymbol{z}_0(k)]\boldsymbol{\Sigma}_{z(k)}(\mathcal{F}'[\boldsymbol{z}_0(k)])^{\text{Tr}}\right]. \tag{29}$$

Due to the frame (k) dependency in the denominator in (28), no simple closed-form solution is available for solving $\boldsymbol{\sigma}^2_{r_s}$ from (28). 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 (28) can be cancelled, yielding the approximate closed-form estimate of

$$\hat{\boldsymbol{\sigma}}^2_{r_s} \approx \frac{\sum_{k=1}^{K_s} \left\{(\boldsymbol{o}(k) - \bar{\boldsymbol{\mu}}_{o_s})^2 - \mathbf{q}(k)\right\}}{K_s}. \tag{30}$$

2) **Direct gradient ascent:** Make no assumption of the above, and take the left-hand-side of (28) as the gradient ∇L of log-likelihood of the data in the standard gradient-ascent algorithm

$$\boldsymbol{\sigma}^2_{r_s}(t+1) = \boldsymbol{\sigma}^2_{r_s}(t) + \epsilon_t \nabla L \left(\boldsymbol{o}_1^{K_s} \middle| \boldsymbol{\sigma}^2_{r_s}(t)\right)$$

where ϵ_t is a heuristically chosen positive constant controlling the learning rate at the tth iteration.

3) **Constrained gradient ascent:** Add to the previous standard gradient ascent technique the constraint that the variance estimate be 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 $\boldsymbol{\sigma}^2_{r_s}$

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

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

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

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

Among the three techniques above, the first one is the fastest but gives a slightly lower performance than the other two techniques which are computationally more expensive. The second technique occasionally causes poor training when the variance estimate becomes negative. To avoid the negative variance estimates requires careful setting of the learning rate. The third technique is robust against the above problem but is much slower than the first technique while comparable to the second one in computation.

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

We note that the quality of the estimates for the residual parameters discussed above plays a crucial role in phonetic recognition performance. These parameters provide an important mechanism for distinguishing speech sounds that belong to different manners of articulation. This is attributed to the fact that nonlinear cepstral prediction from VTRs has different accuracy for these different classes of sounds. Within the same manner class, the phonetic separation is largely accomplished by distinct VTR targets, which typically induce significantly different cepstral prediction values via the "amplification" mechanism provided by the Jacobian matrix $\mathcal{F}'[z]$.

B. 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 are also conditioned on phone segment s and not on subphone segment.

1) Mean Vectors: To obtain a closed-form estimation solution, we assume diagonality (as used in Section IV-A previously) of the prediction cepstral residual's covariance matrix $\boldsymbol{\Sigma}_{r_s}$. Denoting its jth component by $\boldsymbol{\sigma}^2_r(j)$ $(j = 1, 2, \ldots, J)$, we decompose the multivariate Gaussian of (22) element-by-element into

$$p(\boldsymbol{o}(k)|s(k)) = \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma^2_{o_{s(k)}}(j)}} \times \exp\left\{-\frac{(o_k(j)-\bar{\mu}_{o_{s(k)}}(j))^2}{2\sigma^2_{o_{s(k)}}(j)}\right\}, \quad (31)$$

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

The log-likelihood function for a training data sequence ($k = 1, 2, \ldots, K_s$) relevant to the VTR mean vector μ_{T_s} becomes

$$P = \sum_{k=1}^{K_s}\sum_{j=1}^{J}\left\{-\frac{(o_k(j)-\bar{\mu}_{o_{s(k)}}(j))^2}{\sigma^2_{o_{s(k)}}(j)}\right\}$$
$$= \sum_{k=1}^{K_s}\sum_{j=1}^{J}\left\{\frac{\left[\sum_f \mathcal{F}'[z_0(k),j,f]\sum_l a_k(l)\mu_T(l,f) - d_k(j)\right]^2}{\sigma^2_{o_{s(k)}}(j)}\right\} \quad (32)$$

where l and f are indices to phone and to VTR component, respectively, and

$$d_k(j) = o_k(j) - F[z_0(k),j] + \sum_f \mathcal{F}'[z_0(k),j,f]z_0(k,f) - \mu_{r_{s(k)}}(j).$$

While the acoustic feature's distribution is Gaussian for both HTM and HMM given the state s, the key difference is that the mean and variance in HTM as in (22) are both time varying functions (hence trajectory model). These functions provide context dependency (and possible target undershooting) via the smoothing of targets across phonetic units in the utterance. This smoothing is explicitly represented in the weighted sum over all phones in the utterance (i.e., $\sum_l$) in (32).

Setting

$$\frac{\partial P}{\partial \mu_T(l_0,f_0)} = 0$$

and grouping terms involving unknown $\mu_T(l,f)$ on the left and the remaining terms on the right, we obtain

$$\sum_f\sum_l A(l,f;l_0,f_0)\mu_T(l,f) = \sum_k\left\{\sum_j \frac{\mathcal{F}'[z_0(k),j,f_0]}{\sigma^2_{o_{s(k)}}(j)}d_k(j)\right\}a_k(l_0) \quad (33)$$

with $f_0 = 1, 2, \ldots, 8$ for each VTR dimension, and with $l_0 = 1, 2, \ldots 58$ for each phone unit. In (33)

$$A(l,f;l_0,f_0) = \sum_{k,j}\frac{\mathcal{F}'[z_0(k),j,f]\mathcal{F}'[z_0(k),j,f_0]}{\sigma^2_{o_{s(k)}}(j)} \times a_k(l_0)a_k(l). \quad (34)$$

Equation (33) is a 464×464 full-rank linear system of equations.[5] Matrix inversion gives an ML estimate of the complete set of target mean parameters: a 464-dimensional vector formed by concatenating all eight VTR components (four frequencies and four bandwidths) of the 58 phone units in TIMIT.

In implementing (33) for the ML solution to target mean vectors, we kept other model parameters constant. The estimation of the target and residual parameters was carried out in an iterative manner. Initialization of the parameters $\mu_T(l,f)$ was provided by the values described in [13].

2) Diagonal Covariance Matrices: To establish the objective function for optimization, we take the sum of the logarithm of the likelihood function (31) (over K_s frames) to obtain

$$L_T \propto -\sum_{k=1}^{K_s}\sum_{j=1}^{J}\left\{\frac{(o_k(j)-\bar{\mu}_{o_{s(k)}}(j))^2}{\sigma^2_{r_s}(j)+q(k,j)} + \log\left[\sigma^2_{r_s}(j)+q(k,j)\right]\right\} \quad (35)$$

where $q(k,j)$ is the jth element of the vector $\mathbf{q}(k)$ as defined in (29). When $\boldsymbol{\Sigma}_{z(k)}$ is diagonal, it can be shown that

$$q(k,j) = \sum_f \sigma^2_{z(k)}(f)\left(F'_{jf}\right)^2 = \sum_f\sum_l v_k(l)\sigma^2_T(l,f)\left(F'_{jf}\right)^2, \quad (36)$$

where F'_{jf} is the (j,f) element of Jacobian matrix $\mathcal{F}'[\cdot]$ in (29), and the second equality in the above is due to (13).

Using chain rule to compute the gradient, we obtain

$$\nabla L_T(l,f) = \frac{\partial L_T}{\partial \sigma^2_T(l,f)} = \sum_{k=1}^{K_s}\sum_{j=1}^{J}\left\{\frac{(o_k(j)-\bar{\mu}_{o_{s(k)}}(j))^2\left(F'_{jf}\right)^2 v_k(l)}{\left[\sigma^2_{r_s}(j)+q(k,j)\right]^2} - \frac{\left(F'_{jf}\right)^2 v_k(l)}{\sigma^2_{r_s}(j)+q(k,j)}\right\}. \quad (37)$$

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.

V. Phonetic Recognition Experiments

We have carried out 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 [13]) in training the HTM parameters using the standard training utterances. Phonetic recognition errors are tabulated using the commonly

[5] The dimension $464 = 58 \times 8$ where we have a total of 58 phones in the TIMIT database after decomposing each diphthong into two "phones," and 8 is the VTR vector dimension.

TABLE I
TIMIT PHONETIC RECOGNITION PERFORMANCE COMPARISONS BETWEEN AN HMM SYSTEM AND THREE VERSIONS OF THE HTM SYSTEM. HTM-1: N-BEST RESCORING WITH HTM SCORES ONLY; HTM-2: N-BEST RESCORING WITH WEIGHTED HTM, HMM, AND LM SCORES; HTM-3: LATTICE-CONSTRAINED A^* SEARCH WITH WEIGHTED HTM, HMM, AND LM SCORES. IDENTICAL ACOUSTIC FEATURES (FREQUENCY-WARPED LPCCS) ARE USED

	Acc %	Corr %	Sub %	Del %	Ins %
HMM	**71.43**	73.64	17.14	9.22	2.21
HTM-1	**74.31**	77.76	16.23	6.01	3.45
HTM-2	**74.59**	77.73	15.61	6.65	3.14
HTM-3	**75.07**	78.28	15.94	5.78	3.20

TABLE II
COMPARISONS OF HMM AND HTM PERFORMANCES (PERCENT CORRECT) WITHIN EACH OF FOUR BROAD PHONE CLASSES

	Sonorants	Stops	Fricatives	Closures
Occurrences	3814	889	1252	1578
HMM	64.05	72.10	75.64	88.72
HTM	72.42	76.27	75.74	90.94

adopted 39 labels after the label folding. The results are reported on the standard core test set of 192 utterances by 24 speakers [29].

Due to the high implementation and computational complexity for the full-fledged HTM decoder (currently under development), we have restricted the results in this paper only to those obtained by N-best rescoring and lattice-constrained search. For each of the core test utterances, a standard decision-tree-based triphone HMM with a bigram language model is used to generate a large N-best list ($N = 1000$) and a large lattice.[6] These N-best lists and lattices are used for the rescoring experiments with the HTM. The range of error rates in the N-best list generated by the HMM system is from 60% (oracle worst) to 82% (oracle best), giving a sufficient room for the new HTM system to improve the performance. The HTM system is trained using the algorithms presented in the preceding section. Learning rates in the gradient ascent techniques have been tuned empirically.

In Table I, we show phonetic recognition performance comparisons between the HMM system described previously and three evaluation versions of the HTM system. The HTM-1 version uses the HTM likelihood computed from (22) to rescore the 1000-best lists, and no HMM score and language model (LM) score attached in the 1000-best list are exploited. The HTM-2 version improves the HTM-1 version slightly by linearly weighting the log-likelihoods of the HTM, the HMM, and the (bigram) LM, based on the same 1000-best lists.[7] The HTM-3 version replaces the 1000-best lists by the lattices, and carries out A^* search, constrained by the lattices and with linearly weighted HTM-HMM-LM scores, to decode phonetic sequences.[8] Notable performance improvement is obtained as shown in the final row of Table I. For all the systems, the performance is measured by percent phone recognition accuracy (i.e., including insertion errors) averaged over the core test-set sentences (numbers in bolds in column two). The percent-correctness performance (i.e., excluding insertion errors) is listed in column three. The substitution, deletion, and insertion error rates are shown in the remaining columns.

The performance results in Table I are obtained using the identical acoustic features of frequency-warped LPCCs for all the systems. Frequency warping of LPCCs [39] has been implemented by a linear matrix-multiplication technique on both acoustic features and the observation-prediction component of the HTM. The warping gives slight performance improvement for both HMM and HTM systems by a similar amount. Overall, the lattice-based HTM system (75.07% accuracy) gives 13% fewer errors than the HMM system (71.43% accuracy, with the use of the same bigram LM as for the HTM system).[9] This performance is better than any HMM system on the same task as summarized in [29], and is approaching the best-ever result (75.6% accuracy) obtained by using many heterogeneous classifiers as reported in [29] also.

Error analysis has been carried out to examine whether the performance improvement by the HTM system is restricted to certain classes of phones or it is spread over all classes. The analysis results are shown in Table II where performance comparisons are made within each of the four broad sound classes:

1) sonorants (vowels, semivowels, nasals);
2) stop consonants;
3) fricative consonants and affricates;
4) stop closures and silence segments.

We note that the improvements are most significant in the sonorant class, followed by the stop-consonant class. No improvement is observed in the fricative-consonant class. This is in accord with our expectation and suggests that the target-filtering component of the HTM has been better designed than the acoustic residual component where a single Gaussian per state is used for the modeling. The improvement for the stop class is likely due to the better modeling of vocalic portions of formant transitions from stop to vowel and from vowel to stop.

VI. DISCUSSION AND CONCLUSION

As pointed out by several authors (e.g., [13], [41], [42]), the HMM as a generative model of speech suffers from a number of obvious handicaps. First, the assumption of conditional independence of successive observations is grossly unrealistic. Second, because a typical HMM does not represent continuous dynamics in the phonetic structure of speech, it has to rely on a large amount of training data to (partially) capture, by contextual enumeration, speech variability due to coarticulation and other factors. Third, as a generative model, the HMM inherently requires the

[6] The summarizing statistics of the 192 lattices for the core test set are as follows: the average numbers of the lattice nodes and links are 1289 and 8276, respectively. If expanded to an N-best list, the corresponding N is calculated to be as large as 1.003 281 449 186 39E+36.

[7] We had expected greater improvement after the use of HMM scores and LM scores. This did not occur, likely due to the fact that the selection of the N-best hypotheses by the HMM and LM already embeds much of the HMM- and LM-based discriminative information. Hence, additional information from the weighted HMM scores in the HTM provides understandably only a minor contribution to the final performance.

[8] We refer the interested readers to a detailed technical description of this A^*-based search algorithm in [55].

[9] Somewhat better HMM performance, 73.04% accuracy, has been achieved using the MFCC acoustic features, instead of warped LPCCs with 71.43% accuracy. However, the observation-prediction component of the HTM would become more complex when changing from the (warped) LPCC features to the MFCC features, which is currently under investigation.

use of uniform acoustic features across all speech categories. The structurally compact HTM presented in this paper has been developed to aim at mitigating the first two handicaps while retaining the same generative modeling framework where uniform cepstral features are used. In the closely related field of language modeling, a similar structured modeling approach has been developed to overcome limitations of the traditional N-gram statistical model [5], [54], also within the generative modeling framework. Syntactic structure is exploited for characterizing the long-distance relationship among words in [5], [54]. In a similar manner, the dynamic structure in the hidden VTR space is exploited in the HTM presented in this paper for characterizing the long-span contextual influence among phonetic units.

There have been numerous scientific and modeling studies (e.g., [12], [32], [36], [43], [46]) offering direct support to the basic premise of our HTM that VTR dynamics by target filtering can adequately account for the contextually assimilated phonetic reduction that increases "static" phonetic confusion. The increased phonetic confusion is particularly strong for casual and fast speech. It is very difficult to capture this type of phonetic variability by the HMM due to its lack of structural representation of speech dynamics. The HTM presented in this paper provides a rigorous mathematical framework to accomplish such structured modeling, with promising phonetic recognition results obtained. We note that 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 nonreduced spectra with the same underlying phones was modeled in [1], pointing to the same difficulty we faced in the reduction-induced increase of phonetic confusion. Another related work to our model is temporal decomposition [2], where the coarticulated speech observations are modeled as a time-varying linear sum of a set of prefixed deterministic vectors in the same domain as the speech observations. Our HTM extends this concept of coarticulation modeling in three significant 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 ML.
2) Coarticulation 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;
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).

The presentation of the HTM in this paper has been made in the context of larger classes of structured hidden dynamic models. We provide a general overview of the main classes of such models, and discuss their differences and relationships. While the specific HTM presented in this paper has made a number of approximations (as detailed in Sections III and IV) to facilitate its implementation, we note that using more general machine learning tools (e.g., [3], [59]) may relax these approximations and possibly enable implementation and evaluation of more general classes of the hidden dynamic models. Our future work will involve improved HTM design (e.g., adaptive learning of the "stiffness" parameters and exploitation of differential input features, etc.) and more comprehensive evaluation of the model especially for larger recognition tasks with more casual speaking style than TIMIT. We will also investigate producing novel acoustic-phonetic features from the HTM and from other classes of hidden dynamic models. Within the discriminative framework, this will permit the use of nonuniform input features, including the HTM-induced features for vocalic speech sounds in particular, for more effective classification and recognition than the traditional use of uniform input features as done in current HTM and HMM systems.

ACKNOWLEDGMENT

The authors would like to thank A. Gunawardana and M. Seltzer for their help in generating high-quality lattices from an HMM system. They would also like to thank also thank X. Li for implementing an initial version of the LPCC warping function in the HTM system.

REFERENCES

[1] M. Akagi, "Modeling of contextual effects based on spectral peak interaction," *J. Acoust. Soc. Amer.*, vol. 93, no. 2, pp. 1076–1086, 1993.

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

[3] J. Bilmes and C. Bartels, "Graphical model architectures for speech recognition," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 89–100, Sep. 2005.

[4] J. Bridle *et al.*, "An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition," in *Final Report for the 1998 Workshop on Language Engineering, Center for Language and Speech Processing*, 1998, pp. 1–61.

[5] C. Chelba and F. Jelinek, "Structured language modeling," *Comput. Speech Lang.*, pp. 283–332, Oct. 2000.

[6] L. Deng, K. Wang, and W. Chou, "Speech technology and systems in human–machine communication—Guest editors' editorial," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 12–14, Sep. 2005.

[7] L. Deng, A. Acero, and I. Bazzi, "Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 2, pp. 425–434, Mar. 2006.

[8] L. Deng, D. Yu, and A. Acero, "A bidirectional target-filtering model of speech coarticulation and reduction: Two-stage implementation for phonetic recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 256–265, Jan. 2006.

[9] L. Deng and X. D. Huang, "Challenges in adopting speech recognition," *Commun. ACM*, vol. 47, no. 1, pp. 69–75, Jan. 2004.

[10] 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. ICASSP*, Philadelphia, PA, Mar. 2005, pp. 337–340.

[11] 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*, Lisbon, Portugal, Sep. 2005, pp. 1097–1100.

[12] ——, "A quantitative model for formant dynamics and contextually assimilated reduction in fluent speech," in *Proc. ICSLP*, Jeju Island, Korea, 2004, pp. 719–722.

[13] L. Deng and D. O'Shaughnessy, *Speech Processing—A Dynamic and Optimization-Oriented Approach*. New York: Marcel Dekker, 2003.

[14] L. Deng and C. Rathinavalu, "A Markov model containing state-conditioned second-order nonstationarity: Application to speech recognition," *Comput. Speech Lang.*, vol. 9, pp. 63–86, 1995.

[15] L. Deng, "A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition," *Speech Commun.*, vol. 24, no. 4, pp. 299–323, 1998.

[16] ——, "Computational models for speech production," in *Computational Models of Speech Pattern Processing*, K. Ponting, Ed. Berlin, Gemany: Springer, 1999, pp. 199–213.

[17] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for vocal-tract-resonance dynamics," *J. Acoust. Soc. Amer.*, vol. 108, pp. 3036–3048, 2000.

[18] L. Deng, M. Aksmanovic, D. Sun, and J. Wu, "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 507–520, Oct. 1994.

[19] L. Deng and M. Aksmanovic, "Speaker-independent phonetic classification using hidden Markov models with mixtures of trend functions," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 4, pp. 319–324, Jul. 1997.

[20] L. Deng, G. Ramsay, and D. Sun, "Production models as a structural basis for automatic speech recognition," *Speech Commun.*, vol. 22, pp. 93–111, 1997.

[21] L. Deng, "Speech modeling and recognition using a time series model containing trend functions with Markov modulated parameters," in *Proc. IEEE Workshop Automatic Speech Recognition*, 1991, pp. 24–26.

[22] ——, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal, signal processing," *Signal Process.*, vol. 27, pp. 65–78, 1992.

[23] V. Digalakis, J. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 4, pp. 431–442, Oct. 1993.

[24] J. Frankel and S. King, "ASR—Articulatory speech recognition," *Proc. Eurospeech*, vol. 1, pp. 599–602, 2001.

[25] M. Gales and S. Young, "Segmental HMMs for speech recognition," *Proc. Eurospeech*, pp. 1579–1582, 1993.

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

[27] H. Gish and K. Ng, "A segmental speech model with applications to word spotting," in *Proc. ICASSP*, vol. 1, 1993, pp. 447–450.

[28] O. Ghitza and M. Sondhi, "Hidden Markov models with templates as nonstationary states: An application to speech recognition," *Comput. Speech Lang.*, vol. 7, pp. 101–119, 1993.

[29] J. Glass, "A probabilistic framework for segment-based speech recognition," *Comput. Speech Lang.*, vol. 17, no. 2–3, pp. 137–152.

[30] W. Holmes and M. Russell, "Probabilistic-trajectory segmental HMMs," *Comput. Speech Lang.*, vol. 13, pp. 3–27, 1999.

[31] H. Hon and K. Wang, "Unified frame and segment based models for automatic speech recognition," in *Proc. ICASSP*, vol. 2, 2000, pp. 1017–1020.

[32] J. Krause and L. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *J. Acoust. Soc. Amer.*, vol. 115, no. 1, pp. 362–378, 2004.

[33] P. Kenny, M. Lennig, and P. Mermelstein, "A linear predictive HMM for vector-valued observations with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 2, pp. 220–225, Feb. 1990.

[34] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next-generation automatic speech recognition," in *Proc. ICSLP*, Korea, Oct. 2004, pp. 109–111.

[35] C. Li and M. Siu, "An efficient incremental likelihood evaluation for polynomial trajectory model with application to model training and recognition," in *Proc. ICASSP*, 2003, pp. 756–759.

[36] B. Lindblom, "Spectrographic study of vowel reduction," *J. Acoust. Soc. Amer.*, vol. 35, pp. 1773–1781, 1963.

[37] J. Ma and L. Deng, "Target-directed mixture linear dynamic models for spontaneous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 47–58, Jan. 2004.

[38] Y. Minami, E. McDermott, A. Nakamura, and S. Katagiri, "Recognition method with parametric trajectory generated from mixture distribution HMMs," in *Proc. ICASSP*, 2003, pp. 124–127.

[39] A. Oppenheim and D. Johnson, "Discrete representation of signals," *Proc. IEEE*, vol. 60, no. 6, pp. 681–691, Jun. 1972.

[40] N. Morgan *et al.*, "Pushing the envelope—Aside," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 81–88, Sep. 2005.

[41] M. Ostendorf, V. Digalakis, and J. Rohlicek, "From HMMs to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 360–378, Sep. 1996.

[42] F. Pereira, "Linear models for structure prediction," in *Proc. Interspeech*, Lisbon, Sep. 2005, pp. 717–720.

[43] M. Pitermann, "Effect of speaking rate and contrastive stress on formant dynamics and vowel perception," *J. Acoust. Soc. Amer.*, vol. 107, pp. 3425–3437, 2000.

[44] A. Poritz, "Hidden Markov models: A guided tour," in *Proc. ICASSP*, vol. 1, 1988, pp. 7–13.

[45] C. Rathinavelu and L. Deng, "A maximum *a posteriori* approach to speaker adaptation using the trended hidden Markov model," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 549–557, Jul. 2001.

[46] R. Rose, J. Schroeter, and M. Sondhi, "The potential role of speech production models in automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 99, pp. 1699–1709, 1996.

[47] E. Saltzman and K. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, vol. 1, pp. 333–382.

[48] H. Sheikhazed and L. Deng, "Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 80–91, Jan. 1994.

[49] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 1781–1784.

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

[51] K. C. Sim and M. Gales, "Temporally varying model parameters for large vocabulary continuous speech recognition," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 2137–2140.

[52] K. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 1998.

[53] ——, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Amer.*, vol. 111, pp. 1872–1891, Apr. 2002.

[54] W. Wang, A. Stolcke, and M. Harper, "The use of a linguistically motivated language model in conversational speech recognition," in *Proc. ICASSP*, vol. 1, Montreal, QC, Canada, 2004, pp. 261–264.

[55] 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, Portugal, Sep. 2005, pp. 553–556.

[56] H. Zen, K. Tokuda, and T. Kitamura, "A Viterbi algorithm for a trajectory model derived from HMM with explicit relationship between static and dynamic features," in *Proc. ICASSP*, Montreal, QC, Canada, 2004, pp. 837–840.

[57] J. Zhou, F. Seide, and L. Deng, "Coarticulation modeling by embedding a targetdirected hidden trajectory model into HMM—Modeling and training," in *Proc. ICASSP*, vol. I, 2003, pp. 744–747.

[58] V. Zue, *Lecture Notes on Spectrogram Reading*. Cambridge, MA: MIT Press, 1991.

[59] G. Zweig, "Bayesian network structures and inference techniques for automatic speech recognition," *Comput. Speech Lang.*, vol. 17, no. 2–3, pp. 173–193, 2003.

Li Deng (M'86–SM'91–F'05) received the Ph.D. degree in electrical engineering from the University of Wisconsin, Madison, in 1986.

In 1989, he joined the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, as an Assistant Professor, where he became a Full Professor in 1996. From 1992 to 1993, he conducted sabbatical research at the Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, and from 1997–1998, at the ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. In 1999, he joined Microsoft Research, Redmond, WA, as Senior Researcher, where he is currently Principal Researcher. He also has been an Affiliate Professor in electrical engineering at the University of Washington, Seattle, since 2000. His research interests include acoustic–phonetic modeling of speech, speech and speaker recognition, speech synthesis and enhancement, speech production and perception, auditory speech processing, noise robust speech processing, statistical methods and machine learning, nonlinear signal processing, spoken language systems, multimedia signal processing, and multimodal human–computer interaction. In these areas, he has published over 250 refereed papers in leading international conferences and journals, 12 book chapters, and has given keynotes, tutorials, and lectures worldwide. He has been granted over a dozen U.S. and international patents in acoustics, speech/language technology, and signal processing. He authored two books: *Speech Processing—A Dynamic and Optimization-Oriented Approach* (Marcel Dekker, 2003), and *Dynamic Speech Models—Theory, Algorithms, and Applications* (Morgan & Claypool, 2006).

Dr Deng served on the Education Committee and Speech Processing Technical Committees of the IEEE Signal Processing Society (1996–2000) and was Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (2002–2005). He is currently a member of the Society's Multimedia Signal Processing Technical Committee. He also serves on the editorial board of IEEE *Signal Processing Magazine* and is the Magazine's Area Editor. He was a Technical Chair of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP04), and is the General Chair of the IEEE Workshop on Multimedia Signal Processing, 2006. He is Fellow of the Acoustical Society of America.

Dong Yu (M'97–SM'06) received the B.S. degree (with honors) in electrical engineering from Zhejiang University, Hangzhou, China, the M.S. degree in computer science from Indiana University, Bloomington, the M.S. degree in electrical engineering from the Chinese Academy of Sciences, Beijing, and the Ph.D. degree in computer science from the University of Idaho, Moscow.

He joined the Microsoft Corporation, Redmond, WA, in 1998 and Microsoft Speech Research Group, Redmond, in 2002. Prior to joining Microsoft, he was an Assistant Researcher in the Institute of Automation, Chinese Academy of Sciences, from 1994 to 1995. His research interests are in speech recognition, multimodal interaction, dialog systems, and computer security. He has published dozens of refereed journal and conference papers in the above areas.

Alex Acero (S'85–M'90–SM'00–F'04) received the M.S. degree from the Polytechnic University of Madrid, Madrid, Spain, in 1985, the M.S. degree from Rice University, Houston, TX, in 1987, and the Ph.D. degree from Carnegie Mellon University, Pittsburgh, PA, in 1990, all in electrical engineering.

He worked in Apple Computer's Advanced Technology Group from1990 to 1991. In 1992, he joined Telefonica I+D, Madrid, as a Manager of the Speech Technology Group. In 1994, he joined Microsoft Research, Redmond, WA, where he became a Senior Researcher in 1996 and Manager of the Speech Research Group in 2000. Since 2005, he has been a Research Area Manager overseeing speech, natural language, communication, and collaboration. He is currently an affiliate Professor of Electrical Engineering at the University of Washington, Seattle. He is the author of the books *Acoustical and Environmental Robustness in Automatic Speech Recognition* (Kluwer, 1993) and *Spoken Language Processing* (Prentice Hall, 2001), has written invited chapters in three edited books and over 120 technical papers, and has given keynotes, tutorials, and other invited lectures worldwide. He holds 19 U.S. patents. His research interests include speech recognition, synthesis and enhancement, speech denoising, language modeling, spoken language systems, statistical methods and machine learning, multimedia signal processing, and multimodal human–computer interaction.

Dr. Acero served on the Speech Technical Committee of the IEEE Signal Processing Society between 1996 and 2002, chairing the committee in 2000–2002. He was the Publications Chair of ICASSP98, Sponsorship Chair of the 1999 IEEE Workshop on Automatic Speech Recognition and Understanding, and General Co-Chair of the 2001 IEEE Workshop on Automatic Speech Recognition and Understanding. He has served as Associate Editor for SIGNAL PROCESSING LETTERS and is currently an Associate Editor for IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and member of the editorial board of *Computer Speech and Language*. He was a member of the board of governors of the IEEE Signal Processing Society between and 2003 and 2005. He is a 2006 Distinguished Lecturer for the IEEE Signal Processing Society.