SWITCHING DYNAMIC SYSTEM MODELS FOR SPEECH ARTICULATION AND ACOUSTICS

LI DENG*

Abstract. A statistical generative model for the speech process is described that embeds a substantially richer structure than the HMM currently in predominant use for automatic speech recognition. This switching dynamic-system model generalizes and integrates the HMM and the piece-wise stationary nonlinear dynamic system (state-space) model. Depending on the level and the nature of the switching in the model design, various key properties of the speech dynamics can be naturally represented in the model. Such properties include the temporal structure of the speech acoustics, its causal articulatory movements, and the control of such movements by the multidimensional targets correlated with the phonological (symbolic) units of speech in terms of overlapping articulatory features.

One main challenge of using this multi-level switching dynamic-system model for successful speech recognition is the computationally intractable inference (decoding with confidence measure) on the posterior probabilities of the hidden states. This leads to computationally intractable optimal parameter learning (training) also. Several versions of BayesNets have been devised with detailed dependency implementation specified to represent the switching dynamic-system model of speech. We discuss the variational technique developed for general Bayesian networks as an efficient approximate algorithm for the decoding and learning problems. Some common operations of estimating phonological states' switching times have been shared between the variational technique and the human auditory function that uses neural transient responses to detect temporal landmarks associated with phonological features. This suggests that the variation-style learning may be related to human speech perception under an encoding-decoding theory of speech communication, which highlights the critical roles of modeling articulatory dynamics for speech recognition and which forms a main motivation for the switching dynamic system model for speech articulation and acoustics described in this chapter.

Key words. State-space model, Dynamic system, Bayesian network, Probabilistic inference, Speech articulation, Speech acoustics, Auditory function, Speech recognition.

AMS(MOS) subject classifications. Primary 68T10.

1. Introduction. Speech recognition technology has made dramatic progress in recent years (cf. [30, 28]), attributed to the use of powerful statistical paradigms, availability of increasing quantities of speech data corpus, and to the development of powerful algorithms for model learning from the data. However, the methodology underlying the current technology has been founded on weak scientific principles. Not only does the current methodology require prohibitively large amounts of training data and lack robustness under mismatch conditions, its performance also falls at least one order of magnitude short of that of human speech recognition on many comparable tasks (cf. [32, 43]). For example, the best recognizers

*Microsoft Research, One Microsoft Way, Redmond, WA 98052 (deng@microsoft.com). The author wishes to thank many useful discussions with and suggestions for improving the paper presentation by David Heckerman, Mari Ostendorf, Ken Stevens, B. Frey, H. Attias, G. Ramsay, J. Ma, L. Lee, Sam Roweis, and J. Bilmes.

today still produce errors in more than one quarter of the words in natural conversational speech in spite of many hours of speech material used as training data. The current methodology has been primarily founded on the principle of statistical "ignorance" modeling. This fundamental philosophy is unlikely to bridge the performance gap between human and machine speech recognition. A potentially promising approach is to build into the statistical speech model most crucial mechanisms in human speech communication for use in machine speech recognition. Since speech recognition or perception in humans is one integrative component in the entire closed-loop speech communication chain, the mechanisms to be modeled need to be sufficiently broad — including mechanisms in both speech production and auditory perception as well as in their interactions.

Some recent work on speech recognition have been pursued along this direction [6, 18, 13, 17, 46, 47]. The approaches proposed and described in [1, 5, 49] have incorporated the mechanisms in the human auditory process in speech recognizer design. The approaches reported in [18, 21, 19, 44, 3, 54] have advocated the use of the articulatory feature-based phonological units which control human speech production and are typical of human lexical representation, breaking away from the prevailing use of the phone-sized, "beads-on-a-string" linear phonological units in the current speech recognition technology. The approaches outlined in [35, 12, 11, 14, 13] have emphasized the functional significance of the abstract, "task" dynamics in speech production and recognition. The task variables in the task dynamics are the quantities (such as vocal tract constriction locations and degrees) that are closely linked to the goal of speech production, and are nonlinearly related to the physical variables in speech production. Work reported and surveyed in [10, 15, 38, 47] have also focused on the dynamic aspects in the speech process, but the dynamic object being modeled is in the space of speech acoustics, rather than in the space of the production-affiliated variables.

Although dynamic modeling has been a central focus of much recent work in speech recognition, the dynamic object being modeled either in the space of "task" variables or of acoustic variables does not and may not be potentially able to directly take into account the many important properties in *true* articulatory dynamics. Some earlier work used [16, 22] either quantized articulatory features or articulatory data to design speech recognizers, employing highly simplistic models for the underlying articulatory dynamics. Some other earlier proposals and empirical methods exploited pseudo-articulatory dynamics or abstract hidden dynamics for the purpose of speech recognition [2, 4, 23, 45], where the dynamics of a set of pseudo-articulators is realized either by FIR filtering from sequentially placed, phoneme-specific target positions or by applying trajectory-smoothness constraints. Such approaches relied on simplistic nature in the use of the pseudo-articulators. As a result, compensatory articulation, which is a key property of human speech production and which requires modeling correlations among a set of articulators, could not be taken into account. This has drastically diminished the power of such models for potentially successful use in speech recognition.

To incorporate crucial properties in human articulatory dynamics — including compensatory articulation, target behavior, and relatively constrained dynamics (due to biomechanical properties of the articulatory organs) — in a statistical model of speech, it appears necessary to use true, multidimensional articulators, rather than the pseudo-articulators attempted in the past. Given that much of the acoustic variation observed in speech that makes speech recognition difficult can be attributed to articulatory phenomena, and given that articulation is one key component in the closed-loop human speech communication chain, it is reasonable to expect that incorporating a faithful and explicit articulatory dynamic model in the statistical structure of automatic speech recognizer will contribute to bridging the performance gap between human and machine speech recognition. Based on this motivation, a general framework for speech recognition using a statistical description of the speech articulation and acoustic processes is developed and outlined in this chapter. Central to this framework is a switching dynamic system model used to characterize the speech articulation (with its control) and the related acoustic processes, and the Bayesian network (BayesNet) representation of this model. Before presenting some details of this model, we first introduce an encoding-decoding theory of human speech perception which formalizes key roles of modeling speech articulation.

2. Roles of articulation in encoding-decoding theory of speech perception. At a global and functional level, human speech communication can be viewed as an encoding-decoding process, where the decoding process or perception is an active process consisting of auditory reception followed by phonetic/linguistic interpretation. As an encoder implemented by the speech production system, the speaker uses knowledge of meanings of words (or phrases), of grammar in a language, and of the sound representations for the intended linguistic message. Such knowledge can be made analogous to the keys used in engineering communication systems. The phonetic plan, derived from the semantic, syntactic, and phonological processes, is then executed through the motor-articulatory system to produce speech waveforms.

As a decoder which aims to accomplish speech perception, the listener uses a key, or the internal "generative" model, which is compatible with the key used by the speaker to interpret the speech signal received and transformed by the peripheral auditory system. This would enable the listener to reconstruct, via (probabilistic) analysis-by-synthesis strategies, the linguistic message intended by the speaker.[1] This encoding-decoding theory of

[1] While it is not universally accepted that listeners actually do analysis-by-synthesis in speech perception, it would be useful to use such a framework to interpret the roles

human speech communication, where the observable speech acoustics plays the role of the carrier of deep, linguistically meaningful messages, may be likened to the modulation-demodulation scheme in electronic digital communication and to the encryption-decryption scheme in secure electronic communication. Since the nature of the key used in the phonetic-linguistic information decoding or speech perception lies in the strategies used in the production or encoding process, speech production and perception are intimately linked in the closed-loop speech chain. The implication of such a link for speech recognition technology is the need to develop functional and computational models of human speech production for use as an "internal model" in the decoding process by machines. Fig. 1 is a schematic diagram showing speaker-listener interactions in human speech communication and showing the several components in the encoding-decoding theory.

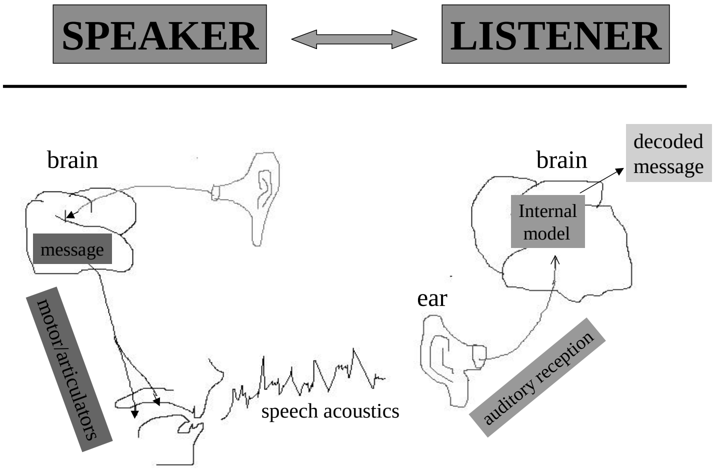

FIG. 1. *Speaker-listener interactions in the encoding-decoding theory of speech perception.*

The encoding-decoding theory of speech perception outlined above highlights crucial roles of speech articulation for speech perception. In summary, the theory consists of three basic, integrated elements: 1) approximate motor-encoding — the symbolic phonological process interfaced with dynamic phonetic process in speech production; 2) robust auditory reception — speech signal transformation prior to the cognitive process;

of articulation in speech perception.

3) cognitive decoding — optimal (by statistical criteria) matching of the auditory transformed signal with the "internal" model derived from a set of motor encoders distinct for separate speech classes. In this theory, the "internal" model in the brain of the listener is hypothesized to have been "approximately" established during the childhood speech acquisition process (or during the process of learning foreign languages in adulthood).

The speech production process as the approximate motor encoder in the above encoding-decoding theory consists of the control strategy of speech articulation, the actual realized speech articulation, and the acoustic signal as the output of the speech articulation system. On the other hand, the auditory process plays two other key roles. First, it transforms the acoustic signal of speech to make it robust against environmental variations. This provides the modified information to the decoder to make its job easier than otherwise. Second, many transient and dynamic properties in the auditory system's responses to speech help create temporal landmarks in the stream of speech to guide the decoding process [50, 53, 54]. (See more detailed discussions on the temporal landmarks in Section 4.3). As will be shown in this chapter, the optimal decoding using the switching dynamic system model as the encoder incurs exponentially growing computation. Use of the temporal landmarks generated from the auditory system's responses may successfully overcome such computational difficulties, hence providing an elegant approximate solution to the otherwise formidable computational problem in the decoding.

In addition to accounting for much of the existing human speech perception data, the computational nature of this theory, with some details described in the remaining of this chapter with special focus on statistical modeling of the dynamic speech articulation and acoustic processes, enables it to be used as the basic underpinning of computer speech recognition systems.

3. Switching state space model for multi-level speech dynamics. In this section, we outline each component of the multi-level speech dynamic model. The model serves as a computational device for the approximate encoder in the encoding-decoding theory of speech perception outlined above. We provide motivations for the construction of each model component from principles of speech science, present a mathematical description of each model component, and justify assumptions made to the mathematical description. The components in the overall model consists of a phonological model, a model for the segmental target, a model for the articulatory dynamics, and a model for the mapping from articulation to acoustics. We start with the phonological-model component.

3.1. Phonological construct. Phonology is concerned with sound patterns of speech and the nature of discrete or symbolic units that form such patterns. Traditional theories of phonology differ in the choice and interpretation of the phonological units. Early distinctive feature based theory [8] and subsequent autosegmental, feature-geometry theory [9] assumed a rather direct link between phonological features and their phonetic correlates in the articulatory or acoustic domain. Phonological rules for modifying features represented changes not only in the linguistic structure of the speech utterance, but also in the phonetic realization of this structure. This weakness has been recognized by more recent theories, e.g., articulatory phonology [7], which emphasize the importance of accounting for phonetic levels of variation as distinct from those at the phonological levels.

In the framework described here, it will be assumed that the linguistic function of phonological units is to maintain linguistic contrasts and is separate from phonetic implementation. It is further assumed that the phonological unit sequence can be described mathematically by a discrete-time, discrete-state homogeneous Markov chain. This Markov chain is characterized by its state transition matrix $\mathbf{A} = [a_{ij}]$ where $a_{ij} = P(s_k = j | s_{k-1} = i)$.

How to construct sequences of symbolic phonological units for any arbitrary speech utterance and how to built them into an appropriate Markov state (i.e., phonological state) structure will not be dealt with here. We merely mention that for effective use of the current framework in speech recognition, the symbolic units must be of multiple dimensions that overlap with each other temporally, overcoming beads-on-a-string limitations. We refer the readers to some earlier work for ways of constructing such overlapping units, either by rules or by automatic learning, which have proved effective in the HMM-like speech recognition framework [21, 19, 18, 56].

3.2. Articulatory control and targets. After a phonological model is constructed, the processes for converting abstract phonological units into their phonetic realization need to be specified. This is a central issue in speech production. It concerns the nature of invariance and variability in the processes interfacing phonology and phonetics, and specifically, whether the invariance is more naturally expressed in the articulatory or acoustic/auditory domains. Early proposals assumed a direct link between abstract phonological units and physical measurements. The "quantal theory" [53] proposed that phonological features possessed invariant acoustic correlates that could be measured directly from the speech signal. The "motor theory" [31] proposed instead that articulatory properties are associated with phonological symbols. No conclusive evidence supporting either hypothesis has been found without controversy, however.

In the current framework, a commonly held view in the phonetics literature is adopted that discrete phonological units are associated with a temporal segmental sequence of phonetic targets or goals [34, 29, 40, 41, 42]. The function of the articulatory motor control system is to achieve such targets or goals by manipulating the articulatory organs according to some control principles subject to the articulatory inertia and possibly minimal-energy constraints.

Compensatory articulation has been widely documented in the phonetics literature where trade-offs between different articulators and non-uniqueness in the articulatory-acoustic mapping allow for the possibilities that many different articulatory target configurations may be able to realize the same underlying goal, and that speakers typically choose a range of possible targets depending on external environments and their interactions with listeners [29]. In order to account for compensatory articulation, a complex phonetic control strategy need be adopted. The key modeling assumptions adopted regarding such a strategy is as follows. First, each phonological unit is associated with a number of phonetic parameters that are described by a state-dependent distribution. These measurable parameters may be acoustic, articulatory or auditory in nature, and they can be computed from some physical models for the articulatory and auditory systems. Further, the region determined by the phonetic correlates for each phonological unit can be mapped onto an articulatory parameter space. Hence the target distribution in the articulatory space can be determined simply by stating what the phonetic correlates (formants, articulatory positions, auditory responses, etc.) are for each of the phonological units (many examples are provided in [55]), and by running simulations in suitably-detailed articulatory and auditory models.

A convenient mathematical representation for the distribution of the articulatory target vector $\mathbf{t}$ is a multivariate Gaussian distribution, denoted by

$$\mathbf{t} \sim \mathcal{N}(\mathbf{t}; \mathbf{m}(s), \mathbf{\Sigma}(s)).$$

Since the target distribution is conditioned on a specific phonological unit (such as a bundle of overlapped features represented by an HMM state s) and since the target does not switch until the phonological unit changes, the statistics for the temporal sequence of the target process follows that of a segmental HMM. A most recent review of the segmental HMM can found in [26].

3.3. Articulatory dynamics. At the present state of knowledge, it is difficult to speculate how the conversion of higher-level motor control into articulator movement takes place. Ideally, modeling of articulatory dynamics and control would require detailed neuromuscular and biomechanical models of the vocal tract, as well as an explicit model of the control objectives and strategies. This is clearly too complicated to implement. A reasonable, simplifying assumption would be that the combined (non-linear) control system and articulatory mechanism behave, at a functional level, as a linear dynamic system that attempts to track the control input equivalently represented by the articulatory target in the articulatory parameter space. Articulatory dynamics can then be approximated as the response of a dynamic vocal tract model driven by a random target sequence (as a segmental HMM). (The output of the vocal tract model then produces a time-varying tract shape which modulates the acoustic properties of the speech signal as observed data.)

This simplifying assumption then reduces the generic nonlinear state equation:

$$\mathbf{z}(k+1) = \mathbf{g}_s[\mathbf{z}(k), \mathbf{t}_s, \mathbf{w}(k)]$$

into a mathematically tractable linear one:

$$\mathbf{z}(k+1) = \mathbf{\Phi}_s \mathbf{z}(k) + (\mathbf{I} - \mathbf{\Phi}_s)\mathbf{t}_s + \mathbf{w}(k), \tag{3.1}$$

where $\mathbf{z} \in \mathcal{R}^n$ is the articulatory-parameter vector, $\mathbf{I}$ is the identity matrix, $\mathbf{w}$ is the IID and Gaussian system noise ($\mathbf{w}(k) \sim \mathcal{N}[\mathbf{w}(k); \mathbf{0}, \mathbf{Q}_{s_k}]$), $\mathbf{t}_s$ is the HMM-state dependent, target vector (expressed in the articulatory domain), and $\mathbf{\Phi}_s$ is the HMM-state-dependent system matrix. The dependence of the $\mathbf{t}_s$ and $\mathbf{\Phi}_s$ parameters of the above dynamic system on the phonological state is justified by the fact that the functional behavior of an articulator depends on the particular goal it is trying to implement, and on the other articulators with which it is cooperating in order to produce compensatory articulation.

3.4. Acoustic model. While a truly consistent framework we are striving for based on explicit knowledge of speech production and perception ideally should include detailed high-order state-space models of the physical mechanisms involved, this becomes unfeasible due to excessive computational requirements. The simplifying assumption adopted is that the articulatory and acoustic state of the vocal tract can be adequately described by low-order vectors of variables representing respectively the relative positions of the major articulators, and the corresponding time-averaged spectral parameters derived from the acoustic signal (or other parameters computed from auditory models). Given further that an appropriate time scale is chosen, it will also be assumed that the relationship between articulatory and acoustic representations can be modeled by a static memoryless transformation, converting a vector of articulatory parameters into a vector of acoustic (or auditory) measurements.

This noisy static memoryless transformation can be mathematically represented by the following observation equation in the state-space model:

$$\mathbf{o}(k) = \mathbf{h}[\mathbf{z}(k)] + \mathbf{v}(k). \tag{3.2}$$

where $\mathbf{o} \in \mathcal{R}^m$ is the observation vector, $\mathbf{v}$ is the IID observation noise vector ($\mathbf{v}(k) \sim \mathcal{N}[\mathbf{v}(k); \mathbf{0}, \mathbf{R}]$) uncorrelated with the state noise $\mathbf{w}$, and $\mathbf{h}[.]$ is the static memoryless transformation from the articulatory vector to its corresponding acoustic observation vector.

There are many ways of choosing the static nonlinear function for $\mathbf{h}[\mathbf{z}]$. Let us take an example of multi-layer perceptron (MLP) with three layers (input, hidden and output). Let w_{jl} be the MLP weights from input to hidden units and W_{ij} be the MLP weights from hidden to output units, where l is the input node index, j the hidden node index and i the output node index. Then the output signal at node i can be expressed as a (nonlinear) function $\mathbf{h}(.)$ of all the input nodes (making up the input vector) according to

$$h_i(\mathbf{z}) = \sum_{j=1}^{J} W_{ij} \cdot s\left(\sum_{l=1}^{L} w_{jl} \cdot z_l \right), \quad 1 \le i \le I, \tag{3.3}$$

where I, J and L are the numbers of nodes at the output, hidden and input layers, respectively. $s(.)$ is the hidden unit's nonlinear activation function, taken as the standard sigmoid function of

$$s(z) = \frac{1}{1+\exp(-z)}. \tag{3.4}$$

The derivative of this sigmoid function has the following concise form:

$$s'(z) = s(z)(1 - s(z)), \tag{3.5}$$

making it convenient for use in many computations.

Typically, the analytical forms of nonlinear functions, such as the MLP, make the associated nonlinear dynamic systems difficult to analyze and make the estimation problems difficult to solve. Approximations are frequently used to gain computational simplifications while sacrificing accuracy for approximating the nonlinear functions.

One most commonly used technique for the approximation is the truncated (vector) Taylor series expansion. If all the Taylor series terms of order two and higher are truncated, then we have the linear Taylor series approximation that is characterized by the Jacobian matrix $\mathbf{J}$ and the point of Taylor series expansion $\mathbf{z}_0$:

$$\mathbf{h}(\mathbf{z}) \approx \mathbf{h}(\mathbf{z}_0) + \mathbf{J}(\mathbf{z}_0)(\mathbf{z} - \mathbf{z}_0). \tag{3.6}$$

Each element of the Jacobian matrix $\mathbf{J}$ is partial derivative of each vector component of the nonlinear output with respect to each of the input vector components. That is,

$$\mathbf{J}(\mathbf{z}_0) = \frac{\partial \mathbf{h}}{\partial \mathbf{z}_0} = \begin{bmatrix} \frac{\partial h_1(\mathbf{z}_0)}{\partial z_1} & \frac{\partial h_1(\mathbf{z}_0)}{\partial z_2} \cdots \frac{\partial h_1(\mathbf{z}_0)}{\partial z_n} \\ \frac{\partial h_2(\mathbf{z}_0)}{\partial z_1} & \frac{\partial h_2(\mathbf{z}_0)}{\partial z_2} \cdots \frac{\partial h_2(\mathbf{z}_0)}{\partial z_n} \\ \vdots & \vdots \\ \frac{\partial h_m(\mathbf{z}_0)}{\partial z_1} & \frac{\partial h_m(\mathbf{z}_0)}{\partial z_2} \cdots \frac{\partial h_m(\mathbf{z}_0)}{\partial z_n} \end{bmatrix}. \tag{3.7}$$

As an example, for the MLP nonlinearity of Eqn. 3.3, the (i, l)-th element of the Jacobian matrix is

$$\sum_{j=1}^{J} W_{ij} \cdot s_j(y) \cdot (1 - s_j(y)) \cdot w_{jl}, \qquad 1 \leq i \leq I, \quad 1 \leq l \leq L, \tag{3.8}$$

where $y = \sum_{l'=1}^{L} W_{jl'} z_{l'}$.

Use of the radial basis function as the nonlinearity in the general nonlinear dynamic system model, as an alternative to the MLP described above, can be found in [24].

3.5. Switching state space model. Eqns. 3.1 and 3.2 form a special version of the switching state-space model appropriate for describing multi-level speech dynamics. The top-level dynamics occurs at the discrete-state phonology, represented by the state transitions of s with a relatively long time scale. The next level is the target ($\mathbf{t}$) dynamics; it has the same time scale and provides systematic randomness at the segmental level. At the level of articulatory dynamics, the time scale is significantly shortened. This is continuous-state dynamics driven by the target process as input, which follows HMM statistics. The state equation 3.1 explicitly describes this dynamics in $\mathbf{z}$, with index of s (which takes discrete values) implicitly representing the switching process. At the lowest level of acoustic dynamics, there is no switching process. Since the observation equation 3.2 is static, this simplifying speech model assumes that acoustic dynamics results solely from articulatory dynamics.

4. BayesNet representation of the segmental switching dynamic speech model. Developed traditionally by machine-learning researchers, BayesNets have found many useful applications. A BayesNet is a graphical model that describes dependencies in the probability distributions defined over a set of variables. A most interesting class of the BayesNet, as relevant to speech modeling, is dynamic BayesNets that are specifically aimed at modeling time series statistics. For time series data such as speech vector sequences, there are causal dependencies between random variables in time. The causal dependencies give some specific, left-to-right BayesNet structures. Such specific structures either permit development of highly efficient algorithms (e.g., for the HMM) for the probabilistic inference (i.e., computation of conditional probabilities for hidden variables) and for learning (i.e., model parameter estimation), or enable the use of approximate techniques (such as variational techniques) to solve the inference and learning problems.

Both the HMM and the stationary (i.e., no switching) dynamic system model are two of the simplest examples of a dynamic BayesNet, for which the efficient algorithms developed already in statistics and in speech processing [51, 38, 20] turn out to be identical to those based on the more general principles of BayesNet theory applied to the special network structures associated with these models. However, for the more complex speech model such as the switching dynamic system model described above, no exact solutions for inference and learning are available without exponentially growing computation with the size of the data. Approximate solutions have been provided for some simple versions of the the switching dynamic system model in literatures of statistics [52], speech processing [33], and of neural computation and BayesNet [25, 39]. The BayesNet framework allows us to take a fresh view on the complex computational issues for such a model, and provides guidance and insights to the algorithm development as well as model refinement.

4.1. Basic BayesNet model. We now discuss how the particular multi-component speech model described in Section 3 can be represented and implemented by BayesNets. Fig. 2 shows one type of dependency structure (indicated by the direction of arrows) of the model, where (discrete) time index runs from left to right. The top-row random variables $s(k)$ take discrete values over the set of phonological states (overlapped feature bundles), and the remaining random variables for the targets, articulators, and acoustic vectors are continuously valued for each time index.

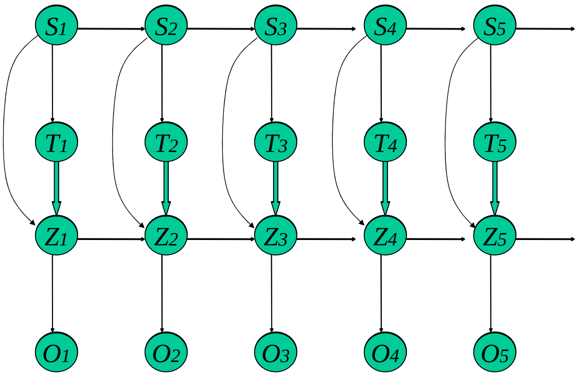

FIG. 2. *Dynamic BayesNet for a basic version of the switching dynamic system model of speech. The random variables on Row 1 are discrete, hidden linguistic states with the Markov-chain temporal structure. Those on Row 2 are continuous, hidden articulatory targets as ideal articulation. Those on Row 3 are continuous, hidden states representing physical articulation with the Markov temporal structure also. Those on Row 4 are continuous, observed acoustic/auditory vectors.*

Each dependency in the above BayesNet can be implemented by specifying the associated conditional probability. In the speech model presented in Section 3, the horizontal (temporal) dependency for the phonological (discrete) states is specified by the Markov chain transition probabilities:

$$P(s_k = j | s_{k-1} = i) = a_{ij}. \tag{4.1}$$

The vertical (level)[2] dependency for the target random variables is specified by the following conditional density function:

$$p(\mathbf{t}(k)|s_k) = \mathcal{N}(\mathbf{t}(k); \mathbf{m}(s_k), \mathbf{\Sigma}(s_k)). \tag{4.2}$$

Possible structures in the covariance matrix $\mathbf{\Sigma}(s_k)$ in the above target distribution can be explored using physical interpretations of the targets as idealized articulation. For example, the velum component is largely uncorrelated with other components; so is the glottal component. On the other hand, tongue components are correlated with each other and with the jaw component. For some linguistic units (/u/ for instance), some tongue components are correlated with the lip components. Therefore, the covariance matrix $\mathbf{\Sigma}(s_k)$ has a block diagonal structure. If we represent each component in the target vector in the BayesNet, then each target node in Fig. 2 will contain a sub-network.

The joint horizontal and vertical dependency for the articulatory (continuous) state is specified, based on state equation 3.1, by the conditional density function:

$$\begin{aligned} p_{\mathbf{z}}[\mathbf{z}(k+1)|\mathbf{z}(k), \mathbf{t}(k), s_k] &= p_{\mathbf{w}}[\mathbf{z}(k+1) - \mathbf{\Phi}_{s_k}\mathbf{z}(k) - (\mathbf{I} - \mathbf{\Phi}_{s_k})\mathbf{t}(k)] \\ &= \mathcal{N}[\mathbf{z}(k+1); \mathbf{\Phi}_{s_k}\mathbf{z}(k) + (\mathbf{I} - \mathbf{\Phi}_{s_k})\mathbf{t}(k), \mathbf{Q}_{s_k}]. \end{aligned} \tag{4.3}$$

The vertical dependency for the observation random variables is specified, based on observation equation 3.2, by the conditional density function:

$$\begin{aligned} p_{\mathbf{o}}[\mathbf{o}(k)|\mathbf{z}(k)] &= p_{\mathbf{v}}[\mathbf{o}(k) - \mathbf{h}(\mathbf{z}(k))] \\ &= \mathcal{N}[\mathbf{o}(k)); \mathbf{h}(\mathbf{z}(k)), \mathbf{R}]. \end{aligned} \tag{4.4}$$

Eqns. 4.1, 4.2, 4.4, and 4.5 then completely specify the switching dynamic model in Fig. 2 since they define all possible dependencies in its BayesNet representation. Note that while the phonological state s_k and its associated target $\mathbf{t}(k)$ in principle are at a different time scale than the phonetic variables $\mathbf{z}(k)$ and $\mathbf{o}(k)$, for simplicity purposes and as one possible implementation, Eqns. 4.1-4.5 have placed them at the same time scale.

Note also that in Eqn. 4.5 the "forward" conditional probability for the observation vector (when the corresponding articulatory vector $\mathbf{z}(k)$ is known) is Gaussian, as is the measurement noise vector's distribution. The

[2] This refers to the level of the speech production chain as the "encoder".

mean of the Gaussian is the prediction of the nonlinear function $\mathbf{h}(\mathbf{z}(k))$. However, the "inverse" or "inference" conditional probability $p[\mathbf{z}(k)|\mathbf{o}(k)]$ will not be Gaussian due to the nonlinearity of $\mathbf{h}(.)$ as well as the switching process that controls the dynamics in $\mathbf{z}(k)$. The fact that the conditional distribution for $\mathbf{z}(k)$ is not Gaussian is one major source of difficulty for the inference and learning problems associated with the nonlinear switching dynamic system model.

4.2. Extended BayesNet model. One modification and extension of the basic BayesNet model of Fig. 2 is to explicitly represent parallel streams of the overlapping phonological features and their associated articulatory dimensions. As discussed in Section 3.1, the phonological construct of the model consists of multidimensional symbols (feature bundles) overlapping in time. The BayesNet for this expanded model is shown in Fig. 3, where the individual components of the articulator vector from the parallel overlapping streams are ultimately combined to generate the acoustic vectors.

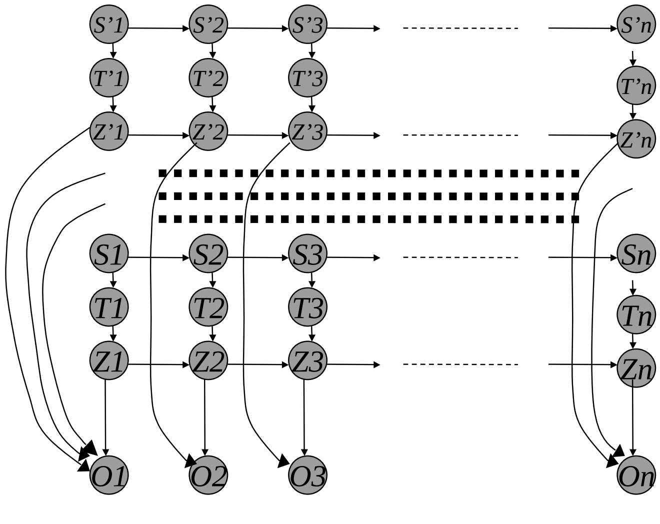

FIG. 3. *Dynamic BayesNet for an expanded version of the switching dynamic system model of speech. Parallel streams of the overlapping phonological features and their associated articulatory dimensions are explicitly represented. The articulators from the parallel streams are ultimately combined to jointly determine the acoustic vectors.*

Another modification of the basic Bayesian-net model of Fig. 2 is to incorporate the segmental constraint on the switching process for the dynamics of the target random vector $\mathbf{t}(k)$. That is, while random, $\mathbf{t}(k)$ remains fixed until the phonological state s_k switches. The switching of target $\mathbf{t}(k)$ is synchronous with that of the phonological state, and only at the time of switching, $\mathbf{t}(k)$ is allowed to take a new value according to its probability density function. This segmental constraint can be described mathematically by the following conditional probability density function:

$$p[\mathbf{t}(k)|s_k, s_{k-1}, \mathbf{t}(k-1)] = \begin{cases} \delta[\mathbf{t}(k) - \mathbf{t}(k-1)] & \textit{if} \quad s_k = s_{k-1}, \\ \mathcal{N}(\mathbf{t}(k); \mathbf{m}(s_k), \mathbf{\Sigma}(s_k)) & \textit{otherwise}. \end{cases}$$

This adds the new dependency of random vector of $\mathbf{t}(k)$ on s_{k-1} and $\mathbf{t}(k-1)$, in addition to the existing s_k as in Fig. 2. The modified BayesNet incorporating this new dependency is shown in Fig. 4.

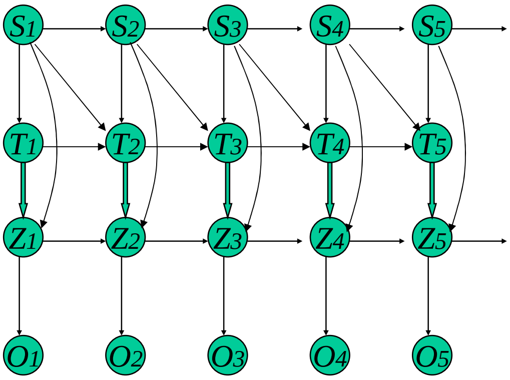

FIG. 4. *Dynamic BayesNet for the switching dynamic system model of speech incorporating the segmental constraint for the target random variables.*

4.3. Discussions. Given the BayesNet representations of switching dynamic system models for speech, rich tools for approximate inference and learning can be exploited and further developed. Since the exact inference is impossible, at least in theory, the success of applying such a model to speech recognition crucially depends on the accuracy of the approximate algorithms.

It is worth noting that while the exact optimal inference for the phonological states (the speech recognition problem) has exponential complexity in computation, once the approximate times of the switching in the phonological states become known, computational complexity can be substantially reduced. With the application of the variational technique (e.g., [27]) developed for BayesNet inference and learning to some generic, unstructured versions of the switching state-space model [39, 25]), one can separate the discrete states from the remaining portion of the network. (Recent research [37] also provides evidence that approximate methods such variational learning work well for a speech model called loosely-coupled HMM.) For the structured switching state-space model of speech dynamics as presented in this paper, this allows one to iteratively estimate the posterior distributions of the discrete phonological and continuous articulatory states. Inference on the phonological states becomes essentially a search for the state switching times with soft decisions. For example, when one uses the Gaussian mixture distribution to approximate the true posteriors in the speech model discussed so far, the E-step (needed for the recognizer's MAP decoding procedure) in the variational EM algorithm can be shown to be a solution to a set of algebraic nonlinear system of equations. Achieving efficient and accurate solutions to these closely coupled equations for the purpose of decoding the optimal phonological state sequence can be greatly facilitated when some crude estimates (e.g., within the range of several frames) of the phonological state boundaries, which we call the *landmarks*, are made available.

Interestingly, such an important role of the phonological state boundary estimates fits closely with the encoding-decoding theory of speech perception outlined in Section 2. As we discussed in Section 2, one crucial role of auditory reception for human speech perception is to provide temporal landmarks for the phonological features via the many transient neural response properties in the auditory system [50, 53, 54]. Recall that in the switching dynamic system model of speech presented in this paper, the phonological units are represented not in terms of phones that consist of a bundle of synchronously aligned features, but in terms of individual features. Therefore, the temporal landmarks associated with the individual features that may be detected by transient neural responses in the auditory system have important functional roles to play in providing the crude boundary information to facilitate the decoding of phonological states (speech perception). This common operation performed by the auditory system and by one aspect of the variational technique suggests that the variational-style decoding algorithms may be closely related to human speech perception.

5. Summary and discussions. We outlined an encoding-decoding theory of speech perception in this chapter, which highlights the importance and critical role of modeling articulatory dynamics in speech recognition. This is an integrated motor-auditory theory where the motor or production system provides the internal model for the listener's speech decoding device, while the auditory system provides sharp temporal landmarks for phonological features to constrain the decoder's search space and to minimize possible loss of decoding accuracy.

Most of current speech systems are very fragile. For further progress in the field, the author believes that it is necessary to bring in human-like intelligence of speech perception into computer systems. The switching dynamic system models discussed in this chapter offer one powerful mathematical tool for implementing the encoding-decoding mechanism of human speech communication. We have shown that the BayesNet framework allows us to take a fresh view on the complex computational issues in inference (decoding) and in learning, and to provide guidance and insights to the algorithm development.

It is hoped that the framework presented here will help integrate results from speech production and advanced machine learning within the statistical paradigms for speech recognition. An important, long-term goal will involve development of computer systems to the extent that they can be evaluated efficiently on realistic, large speech databases, collected in a variety of speaking styles (conversational styles in particular) and for a large population of speakers.

The ultimate goal of the research, whose components are described in some detail in this chapter, is to develop high-performance systems for integrated speech analysis, coding, synthesis, and recognition within a consistent statistical framework. Such a development is guided by the encoding-decoding theory of human speech communication, and is based on computational models of speech production and perception. The switching dynamic system models of speech and their BayesNet representations presented are a significant extension of the current highly simplified statistical models used in speech recognition. Further advances in this research direction will require greater integration within a statistical framework of existing research in modeling speech production, speech recognition, and advanced machine learning.

REFERENCES

[1] J. Allen. "How do humans process and recognize speech," *IEEE Trans. Speech Audio Proc.*, Vol. **2**, 1994, pp. 567–577.

[2] R. Bakis. "Coarticulation modeling with continuous-state HMMs," *Proc. IEEE Workshop Automatic Speech Recognition*, Harriman, New York, 1991, pp. 20–21.

[3] N. Bitar and C. Espy-Wilson. "Speech parameterization based on phonetic features: Application to speech recognition," *Proc. Eurospeech*, Vol. **2**, 1995, pp. 1411–1414.

[4] C. Blackburn and S. Young. "Towards improved speech recognition using a speech production model," *Proc. Eurospeech*, Vol. **2**, 1995, pp. 1623–1626.

[5] H. Bourlard and S. Dupont. "A new ASR approach based on independent processing and recombination of partial frequency bands," *Proc. ICSLP*, 1996, pp. 426–429.

[6] H. Bourlard, H. Hermansky, and N. Morgan. "Towards increasing speech recognition error rates," *Speech Communication*, Vol. **18**, 1996, pp. 205–231.

[7] C. Browman and L. Goldstein. "Articulatory phonology: An overview," *Phonetica*, Vol. **49**, pp. 155–180, 1992.

[8] N. Chomsky and M. Halle. *The Sound Pattern of English*, New York: Harper and Row, 1968.

[9] N. Clements. "The geometry of phonological features," *Phonology Yearbook*, Vol. **2**, 1985, pp. 225–252.

[10] L. Deng. "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Processing*, Vol. **27**, 1992, pp. 65–78.

[11] L. Deng. "A computational model of the phonology-phonetics interface for automatic speech recognition," Summary Report, SLS-LCS, Massachusetts Institute of Technology, 1992–1993.

[12] L. Deng. "Design of a feature-based speech recognizer aiming at integration of auditory processing, signal modeling, and phonological structure of speech." *J. Acoust. Soc. Am.*, Vol. **93**, 1993, pp. 2318.

[13] L. Deng. "Computational models for speech production," in *Computational Models of Speech Pattern Processing* (NATO ASI), Springer-Verlag, 1999, pp. 67–77.

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

[15] 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 Proc.*, Vol. **2**, 1994, pp. 507–520.

[16] L. Deng, and K. Erler. "Structural design of a hidden Markov model based speech recognizer using multi-valued phonetic features: Comparison with segmental speech units," *J. Acoust. Soc. Am.*, Vol. **92**, 1992, pp. 3058–3067.

[17] L. Deng and Z. Ma. "Spontaneous speech recognition using a statistical coarticulatory model for the hidden vocal-tract-resonance dynamics," *J. Acoust. Soc. Am.*, Vol. **108**, No. 6, 2000, pp. 3036–3048.

[18] L. Deng, G. Ramsay, and D. Sun. "Production models as a structural basis for automatic speech recognition," *Speech Communication*, Vol. **22**, No. 2, 1997, pp. 93–111.

[19] L. Deng and H. Sameti. "Transitional speech units and their representation by the regressive Markov states: Applications to speech recognition," *IEEE Trans. Speech Audio Proc.*, Vol. **4**, No.4, July 1996, pp. 301–306.

[20] L. Deng and X. Shen. "Maximum likelihood in statistical estimation of dynamic systems: Decomposition algorithm and simulation results", *Signal Processing*, Vol. **57**, 1997, pp. 65–79.

[21] L. Deng and D. Sun. "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *J. Acoust. Soc. Am.*, Vol. **95**, 1994, pp. 2702–2719.

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

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

[24] Z. Ghahramani and S. Roweis. "Learning nonlinear dynamic systems using an EM algorithm". *Advances in Neural Information Processing Systems*, Vol. **11**, 1999, 1–7.

[25] Z. Ghahramani and G. Hinton. "Variational learning for switching state-space model". *Neural Computation*, Vol. **12**, 2000, pp. 831–864.

[26] W. Holmes. "Segmental HMMs: Modeling dynamics and underlying structure in speech," in M. Ostendorf and S. Khudanpur (eds.) *Mathematical Foundations of Speech Recognition and Processing, Volume X in IMA Volumes in Mathematics and Its Applications,* Springer-Verlag, New York, 2002.

[27] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. "In introduction to variational methods for graphical models," in *Learning in Graphical Models* M. Jordon (ed.), The MIT Press, Cambridge, MA, 1999.

[28] F. Juang and S. Furui (eds.), Proc. of the IEEE (special issue), Vol. **88**, 2000.

[29] R. Kent, G. Adams, and G. Turner. "Models of speech production," in *Principles of Experimental Phonetics*, N. Lass (ed.), Mosby: London, 1995, pp. 3–45.

[30] C.-H. Lee, F. Soong, and K. Paliwal (eds.) *Automatic Speech and Speaker Recognition – Advanced Topics,* Kluwer Academic, 1996.

[31] A. Liberman and I. Mattingly. "The motor theory of speech perception revised" *Cognition*, Vol. **21**, 1985, pp. 1–36.

[32] R. Lippman. "Speech recognition by human and machines," *Speech Communication*, Vol. **22**, 1997, pp. 1–15.

[33] Z. Ma and L. Deng. "A path-stack algorithm for optimizing dynamic regimes in a statistical hidden dynamic model of speech," *Computer Speech and Language*, Vol. **14**, 2000, pp. 101–104.

[34] P. MacNeilage. "Motor control of serial ordering in speech," *Psychological Review*, Vol. **77**, 1970, pp. 182–196.

[35] R. McGowan. "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests," *Speech Communication*, Vol. **14**, 1994, pp. 19–48.

[36] R. McGowan and A. Faber. "Speech production parameters for automatic speech recognition," *J. Acoust. Soc. Am.*, Vol. **101**, 1997, pp. 28.

[37] H. Nock. *Techniques for Modeling Phonological Processes in Automatic Speech Recognition*, Ph.D. thesis, Cambridge University, 2001, Cambridge, U.K.

[38] 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 Proc.*, Vol. **4**, 1996, pp. 360–378.

[39] V. Pavlovic, B. Frey, and T. Huang. "Variational learning in mixed-state dynamic graphical models," *Proc. Annual Conf. in Uncertainty in Artificial Intelligence*, 1999, UAI-99.

[40] J. Perkell, M. Matthies, M. Svirsky, and M. Jordan. "Goal-based speech motor control: a theoretical framework and some preliminary data," *J. Phonetics*, Vol. **23**, 1995, pp. 23–35.

[41] J. Perkell. "Properties of the tongue help to define vowel categories: hypotheses based on physiologically-oriented modeling," *J. Phonetics* Vol. **24**, 1996, pp. 3–22.

[42] P. Perrier, D. Ostry, and R. Laboissière. "The equilibrium point hypothesis and its application to speech motor control," *J. Speech & Hearing Research,* Vol. **39**, 1996, pp. 365–378.

[43] L. Pols. "Flexible human speech recognition," *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding,* Santa Barbara, 1997, pp. 273–283.

[44] M. Randolph. “Speech analysis based on articulatory behavior,” *J. Acoust. Soc. Am.*, Vol. **95**, 1994, pp. 195.

[45] H. Richards, and J. Bridle. “The HDM: A segmental hidden dynamic model of coarticulation”, Proc. ICASSP, Vol. **1**, 1999, pp. 357–360.

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

[47] M. Russell. “Progress towards speech models that model speech,” *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding,* Santa Barbara, 1997, pp. 115–123.

[48] J. Schroeter and M. Sondhi. “Techniques for estimating vocal-tract shapes from the speech signal,” *IEEE Trans. Speech Audio Proc.*, Vol. **2**, 1994, pp. 133–150.

[49] H. Sheikhzadeh and L. Deng. “Speech analysis and recognition using interval statistics generated from a composite auditory model,” *IEEE Trans. Speech Audio Proc.*, Vol. **6**, 1998, pp. 50–54.

[50] H. Sheikhzadeh and L. Deng. “A layered neural network interfaced with a cochlear model for the study of speech encoding in the auditory system,” *Computer Speech and Language*, Vol. **13**, 1999, pp. 39–64.

[51] R. Shumway and D. Stoffer. “An approach to time series smoothing and forecasting using the EM algorithm,” *J. Time Series Analysis*, Vol. **3**, 1982, pp. 253–264.

[52] R. Shumway and D. Stoffer. “Dynamic linear models with switching”, *J. American Statistical Association*, Vol. **86**, 1991, pp. 763–769.

[53] K. Stevens. “On the quantal nature of speech,” *J. Phonetics,* Vol. **17**, 1989, pp. 3–45.

[54] K. Stevens. “From acoustic cues to segments, features and words,” *Proc. ICSLP*, Vol. **1**, 2000, pp. A1–A8.

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

[56] J. Sun, L. Deng, and X. Jing. “Data-driven model construction for continuous speech recognition using overlapping articulatory features,” *Proc. ICSLP*, Vol. **1**, 2000, pp. 437–440.