Computational Models for Speech Production

Li Deng

Department of Electrical and Computer Engineering
University of Waterloo, Waterloo, Ontario, Canada N2L 3G1
email: deng@crg6.uwaterloo.ca

Summary. Major speech production models from speech science literature and a number of popular statistical "generative" models of speech used in speech technology are surveyed. Strengths and weaknesses of these two styles of speech models are analyzed, pointing to the need to integrate the respective strengths while eliminating the respective weaknesses. As an example, a statistical task-dynamic model of speech production is described, motivated by the original deterministic version of the model and targeted for integrated-multilingual speech recognition applications. Methods for model parameter learning (training) and for likelihood computation (recognition) are described based on statistical optimization principles integrated in neural network and dynamic system theories.

1. Introduction

In the past thirty years or so, the same physical entity of human speech has been studied and modeled using drastically different approaches undertaken by largely distinct communities of speech scientists and speech engineers. Models for how speech is generated in human speech production system developed by speech scientists typically have rich model structures [17, 24]. The structures have embodied detailed multi-level architectures which transform the high-level symbolic phonological construct to acoustic streams via intermediate stages of phonetic task specification, motor command generation, and articulation. However, these models are often underspecified due to 1) deterministic nature which does not accommodate random variabilities of speech and only weakly accommodates systematic variabilities; 2) lack of comprehensiveness in covering all classes of speech sounds (with some exceptions); and 3) lack of strong computational formalisms allowing for automatic model learning from data and for optimal choice of decision variables necessary for high-accuracy speech classifications.

On the other hand, models for how speech patterns are characterized by statistical generative mechanisms, which have been developed in speech technology, notably by speech recognition researchers, typically contain weak and poor structures. These models often simplistically assume direct, albeit statistical, correlates between phonological constructs and the surface acoustic properties. This causes recognizers built from these models to perform poorly for unconstrained tasks, and to break down easily when porting from one task domain to another, from one speaking mode to another, and from one language to another. Empirical system tuning and ever-more increasing data appear to be the only options for making the systems behave reasonably if no fundamental changes are made to the speech models underlying the recognizers. However, a distinct strength associated with these models is that they are equipped with powerful statistical formalisms, based solidly on mathematical optimization principles , and are suitable for implementation with flexible, integrated architecture. The precise mathematical framework, despite poor and simplistic model structure, gives rise to ease of automatic parameter learning (training) and to optimal decision rules for speech-class discrimination.

Logically, there are strong reasons to expect that a combination of the strengths of the above two styles of models, free from the respective weaknesses, will ultimately lead to superior speech recognition. In this paper, some versions of a statistical dynamic speech production model[1], with theoretical motivation, mathematical formulation, and procedures for parameter learning are described, aiming at achievement of such superiority.

2. Speech production models in science/technology literatures

In this section I will briefly survey major speech production models in science and technology literatures, with emphasis on drawing parallels and contrasts between these two styles of models developed by largely separate research communities. The purpose of scientific speech production models is to provide adequate representations to account for the conversion of a linguistic (mainly phonological) message to the actions of the production system and to the consequent articulatory and acoustic outputs. Critical issues addressed by these models are the serial-order problem, the degrees-of-freedom problem, and the related context-sensitivity or coarticulation problem in both articulatory and acoustic domains. The models can be roughly classied into categories of global and component models. Within the category of the global production models, the major classes (modified from the classification of [17] where a large number of references are listed) are: 1) Feedback-feedforward models, which enable both predictive and adaptive controls to operate, ensuring achievement of articulatory movement goals; 2) Motor program and generalized motor program (schema) models, which use preassembled forms of speech movement (called goals or targets) to regulate the production system; 3) Integrated motor program and feedback models, which combine the characteristics and modeling assumptions of 1) and 2); 4) Dynamic system and gestural patterning models, which employ coordinative structures with a small number of degrees of freedom to create families of functionally equivalent speech movement patterns, and use dynamic vocal tract constriction variables (in the "task" space) to directly define speech movement tasks or goals; 5) Models based on equilibrium point hypothesis, which use shifts of virtual target trajectories, arising from interactions among central neural commands, muscle properties, and external loads, in the articulatory space ("body" rather than "task" space) to control and produce speech movement; and finally 6) Connectionist models, which establish nonlinear units interconnected into a large network to

[1] Based mainly on the task-dynamic model originally developed by speech scientists [28] and on our earlier work on overlapping articulatory features [8].

functionally account for a number of prominent speech behaviors including serial order and coarticulation.

Within the category of the component or subsystem models of speech production are the models for respiratory subsystem, laryngeal subsystem, and supralaryngeal (vocal tract) subsystem. In addition, composite models have also been developed to integrate multiple subsystem models operating in parallel or in series [17, 24].

All of the scientifically motivated speech production models briefly surveyed above have focused mainly on explanatory power for speech behaviors (including articulatory movements and its relations to speech acoustics), and paid relatively minor attention to computation issues. Further, in developing and evaluating these models, comprehensiveness in covering speech classes is often seriously limited (e.g. CV, CVC, VCV sequences only). In stark contrast, speech models developed by technologists usually cover all classes of speech sounds, and computation issues are given a high priority of consideration with no exception. Another key character of the technology-motivated speech models is the rigorous statistical frameworks for model formulation which permit automatic learning of model parameters from realistic acoustic data of speech. On the negative side, however, the structures of these models tend to be oversimplified, often deviating significantly from the true stages in the human speech generation mechanisms which the scientific speech production models are aiming to account for. To show this, let us view the HMM (which forms the theoretical basis of the modern speech recognition technology) as a primitive (very inaccurate) generative model of speech. To show such inaccuracy, we simply note that the unimodal Gaussian HMM generates its sample paths which are piecewise constant trajectories embedded in temporally uncorrelated Gaussian noise, and, since variances of the noise are estimated from time-independent (except with the HMM state-bound) speech data from all speakers, the model could freely allow generation of speech from different speakers over as short as every 10 msec. [2]

A simplified hierarchy of statistical generative or production models of speech developed in speech technology is briefly reviewed here. Under the root node of conventional HMM there are two main classes of its extended or generalized models: 1) nonstationary-state HMM[3] whose sample paths are piecewise, explicitly defined stochastic trajectories (e.g. [3, 14, 11]); 2) multi-region dynamic system model whose sample paths are piecewise, recursively defined stochastic trajectories (e.g. [10]). [4] The parametric form of Class-1 models typically uses polynomials to constrain the trajectories, with the standard HMM as a special case when the polyno-

[2] For the mixture Gaussian HMM, the sample paths are erratic and highly irregular due to lack of temporal constraints forcing fixed mixture component within each HMM state; this deviates significantly from speech data coming from heterogeneous sources such as from multiple speakers and multiple speech collection channels.

[3] also called trended HMM, segmental HMM, stochastic trajectory model, etc., with slight variations in technical detail according to whether the parameters defining the trend functions are random or not.

[4] Intimate relations and implementational differences between the explicitly defined and recursively defined time series models are discussed in [19].

mial order is set to zero. This model can be further divided according to whether the polynomial trajectories or trends can be observed directly [5] or the trends are hidden due to assumed randomness in the polynomial coefficients. For the latter, further classification gives discrete and continuous mixtures of trends depending on the assumed discrete or continuous nature of the model parameter distribution[6, 16, 12]. For Class-2, recursively defined trajectory models, the earlier linear model aiming at dynamic modeling at the acoustic level [10, 23] has been generalized to nonlinear models taking into account detailed mechanisms of speech production. Subclasses of such nonlinear dynamic models are 1) articulatory dynamic model; and 2) task-dynamic model. They differ from each other by distinct objects of dynamic modeling, one at the level of biomechanic articulators (body space) and the other at the level of more abstract task variables (task space)[6] [1, 2, 9, 4, 5, 25]. Depending on the assumptions about whether the dynamic model parameters are deterministic or random, and whether these parameters are allowed to change within phonological state boundaries, further subclasses can be categorized (see details in the following sections).

3. Derivation of discrete-time version of statistical task-dynamic model

In this section, a discrete-time statistical task-dynamic model , which is implementable and trainable for use in speech recognition, is derived from the original deterministic, continuous-time task-dynamic model well established in speech science literature [28]. Starting with the original model but incorporating random noise $\mathbf{w}(t)$:

$$\frac{d^2\mathbf{z}(t)}{dt^2} + 2\mathbf{S}(t)\frac{d\mathbf{z}(t)}{dt} + \mathbf{S}^2(t)(\mathbf{z}(t) - \mathbf{Z}^0(t)) = \mathbf{w}(t),$$

where $\mathbf{S}^2$ is normalized, gesture-dependent stiffness parameter (which controls fast or slow movement of tract variable $\mathbf{z}(t)$), and Z^0 is gesture-dependent point-attractor parameter of the dynamical system (which controls the target and hence direction of the movement). Here, for generality, we assume that the model parameters are (slowly) time-varying.

Rewrite the above into a canonical form (where $\dot{\mathbf{z}}(t) = \frac{d\mathbf{z}(t)}{dt}$):

$$\frac{d}{dt}\begin{pmatrix} \mathbf{z}(t) \\ \dot{\mathbf{z}}(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\mathbf{S}^2(t) & -2\mathbf{S}(t) \end{pmatrix} \begin{pmatrix} \mathbf{z}(t) \\ \dot{\mathbf{z}}(t) \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbf{S}^2(t)\mathbf{Z}^0(t) \end{pmatrix} + \begin{pmatrix} 0 \\ \mathbf{w}(t) \end{pmatrix}$$

This, in matrix form, is:

[5] This happens when the model parameters (i.e. polynomial coefficients) are deterministic.

[6] They also have distinct origins, one from scientific speech production models based on equilibrium point hypothesis, the other from the deterministic version of the task-dynamic model.

$$\frac{d}{dt}Z(t) = F(t)Z(t) - F(t)T(t) + W(t),$$

where composite state is defined by

$$Z(t) \equiv \begin{pmatrix} \mathbf{z}(t) \\ \dot{\mathbf{z}}(t) \end{pmatrix},$$

system matrix by

$$F(t) \equiv \begin{pmatrix} 0 & 1 \\ -\mathbf{S}^2(t) & -2\mathbf{S}(t) \end{pmatrix},$$

and attractor vector for the system dynamics by

$$T(t) \equiv -F^{-1}(t) \begin{pmatrix} 0 \\ \mathbf{S}^2(t)\mathbf{Z}^0(t) \end{pmatrix}.$$

Explicit solution to the above task-dynamic equation is [21]:

$$Z(t) = \Phi(t, t_0)Z(t_0) + \int_{t_0}^{t} \Phi(t, \tau)[-F(\tau)T(\tau) + W(\tau)]d\tau,$$

where $\Phi(t, t_0)$ (state transition matrix) is the solution to the following matrix homogeneous differential equation: [7]

$$\dot{\Phi}(t, \tau) = F(t)\Phi(t, \tau); \quad init.\ cond. : \Phi(t, t) = I.$$

Setting $t_0 = t_k, t = t_{k+1}$, we have

$$Z(t_{k+1}) \approx \Phi(t_{k+1}, t_k)Z(t_k) - [\int_{t_k}^{t_{k+1}} \Phi(t_{k+1}, \tau)F(\tau)d\tau]T(t_k) + \int_{t_k}^{t_{k+1}} \Phi(t_{k+1}, \tau)W(\tau)d\tau,$$

which leads to the discrete-time form of the task-dynamic state equation:

$$Z(k+1) = \Phi(k)Z(k) + \Psi(k)T(k) + W_d(k), \tag{1}$$

where

$$\begin{aligned}
\Phi(k) &\approx \Phi(t_{k+1}, t_k) = exp(F_k \Delta t), \quad \Delta t \equiv t_{k+1} - t_k \\
\Psi(k) &\approx -[\int_{t_k}^{t_{k+1}} \Phi(t_{k+1}, \tau)F(\tau)d\tau] \approx -\int_{t_k}^{t_{k+1}} exp[F_k(t_{k+1} - \tau)]F(\tau)d\tau \\
&\approx -F_k\ exp(F_k\ t_{k+1}) \int_{t_k}^{t_{k+1}} exp[-F_k\tau]d\tau = I - exp(F_k \Delta t) = I - \Phi(k),
\end{aligned}$$

and $W_d(k)$ is discrete-time white Gaussian sequence which is statistically equivalent through its first and second moments to $\int_{t_k}^{t_{k+1}} \Phi(t_{k+1}, \tau)W(\tau)d\tau$.

[7] The solution can be written in matrix exponential form: $\Phi(t, \tau) = exp[(t - \tau)F_k]$ if $F(t) = F_k$ for $t_k \le t \le t_{k+1}$)

For a speech recognizer which has only acoustic data sequences at its disposal, the dynamic associated with task variables $Z(k)$ described in Eqn.(1) is a hidden or unobservable process. Following the treatment of task-dynamic model which uses intermediate model-articulator variables $\mathbf{x}(k)$ to link the task variables $Z(k)$ to acoustic variables $O(k)$ via static nonlinear functions, $O = \mathcal{O}(\mathbf{x})$ and $Z = \mathcal{Z}(\mathbf{x})$, we treat the hidden task-variable dynamic as observed through noisy (i.i.d. noise $V(k)$) nonlinear relation $\mathbf{h}(\cdot)$ between task-variable $Z(k)$ and acoustic observation $O(k)$:

$$O(k) = \mathcal{O}(\mathbf{x}(k)) + V(k) = \mathcal{O}[\mathcal{Z}^{-1}(Z(k))] + V(k) = \mathbf{h}[Z(k)] + V(k). \qquad (2)$$

The above global nonlinearity has been implemented numerically in the deterministic version of task-dynamic model [28, 20] by geometric relationships in an improved version of Mermelstein-type articulatory model ($Z = \mathcal{Z}(\mathbf{x})$) together with a configurable articulatory synthesizer ($O = \mathcal{O}(\mathbf{x})$) [27]. For intended use in statistical speech recognition, we need to parameterize this nonlinearity with trainable sets of parameters and with the numerical simulation only serving as parameter initialization. While many possibilities exist for the parameterization based on well established statistical and neural-network techniques, in this tutorial I will describe a straightforward method based on use of Multi-Layer Perceptron (MLP) neural network which functions as a universal multidimensional functional approximation device [15]. Let i denote the element index of the acoustic-variable vector O as output of MLP, and l index of the task-variable vector Z as input to MLP. Then, each vector component of the MLP output can be parameterized by [8]

$$h_i[Z] = \sum_j W_{ij} x_j = \sum_j W_{ij} g(\sum_l w_{jl} g(Z_l)), \quad i = 1, 2, \ldots, I. \qquad (3)$$

The observation equation (2) can then be written in the parameterized form:

$$O_i(k) = \sum_j W_{ij} g(\sum_l w_{jl} g(Z_l)) + V_i(k), \quad i = 1, 2, \ldots, I. \qquad (4)$$

4. Algorithms for learning task-dynamic model parameters and for likelihood computation

The deterministic version of the task-dynamic model [28], although well developed and tested for use as an effective research tool in accounting for and in understanding dynamic behaviors of the speech process, is unlikely to be directly useful for

[8] In this tutorial, the MLP is further simplified to contain only one hidden layer. The output (acoustic variables $O_1, O_2, ...$) layer is linear with weights W_{ij}; input (task variables $Z_1, Z_2, ...$) and hidden layers are sigmoid nonlinear: $g(v) = 1/[1 + exp(-v)]$, with weights w_{jl}. In actual implementation, since the hidden layer $\mathbf{x}$ is intended to represent model-articulator variables, and since the relationship between $\mathbf{x}$ and Z and that between $\mathbf{x}$ and O are known to be strongly nonlinear, two more hidden layers are placed between the three layers illustrated here.

engineering applications in speech recognition. Its lack of statistical structure does not allow the model to effectively capture variabilities, either systematic or random, in the observed speech data. Within the modeling framework of [28], there also seem to be no principled ways to devise an optimal decision rule for speech classification by matching the model's output to speech data. In contrast, the statistical version of the task-dynamic model derived in the previous section permits the use of computable likelihoods to construct the optimal decision rule with the optimality guaranteed by Bayesian decision theory .

In this section, algorithms for learning parameters of three versions of statistical task-dynamic model, with increasing complexity in the model construct consistent with differing assumptions invoked by various phonetic theories, are outlined. Some main steps of algorithm derivation, based on statistical optimization principles integrated in neural network and dynamic system theories, are included also.

4..1 Model with deterministic, time-invariant parameters

This is the simplest version of statistical task-dynamic model where deterministic, unconstrained, time-invariant parameters are assumed in the state equation Eqn.(1), rewritten as

$$Z(k+1) = \Phi Z(k) + (I - \Phi)T + W_d(k).$$

The nonlinearity in the observation equation is parameterized by a form of MLP according to Eqn.(4).

With use of chain rule twice and use of $\frac{d}{dv} g(v) = g(v)(1 - g(v))$, the Jacobian matrix (needed for parameter learning) of the MLP-parameterized nonlinear mapping (Eqn.(3)) can be computeded in an analytical form:

$$\mathbf{H}_z(Z) \equiv \frac{d}{dZ}\mathbf{h}(Z) = [H_{il}(Z)] = \begin{pmatrix} \frac{\partial h_1}{\partial Z_1} & \frac{\partial h_1}{\partial Z_2} & \cdots & \frac{\partial h_1}{\partial Z_L} \\ \frac{\partial h_2}{\partial Z_1} & \frac{\partial h_2}{\partial Z_2} & \cdots & \frac{\partial h_2}{\partial Z_L} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial h_I}{\partial Z_1} & \frac{\partial h_I}{\partial Z_2} & \cdots & \frac{\partial h_I}{\partial Z_L} \end{pmatrix}$$

where

$$H_{il}(Z) = \sum_j W_{ij} g[\sum_m w_{jm} g(Z_m)][1 - g(\sum_m w_{jm} g(Z_m))] w_{jl}\ g(Z_l)[1 - g(Z_l))].$$

In developing a parameter-learning procedure, the joint log-likelihood for acoustic observation sequence $O = [O(1), O(2), \ldots, O(N)]$ and hidden task-variable sequence $Z = [Z(1), Z(2), \ldots, Z(N)]$ is first written out as

$$\begin{aligned} &\log L(Z, O, \Theta) = \\ &-\frac{1}{2}\sum_{k=1}^{N-1}\{\log Q + [Z(k+1) - \Phi Z(k) - (I - \Phi)T]^T Q^{-1} [Z(k+1) - \Phi Z(k) - (I - \Phi)T]\} \\ &-\frac{1}{2}\sum_{k=1}^{N}\{\log R + [O(k) - \mathbf{h}(Z(k))]^T R^{-1} [O(k) - \mathbf{h}(Z(k))]\} + const. \end{aligned}$$

Then a pseudo-EM algorithm is used for learning model parameters including those in task-dynamics and those in MLP nonlinear mapping: Θ=$\{T_{\mathcal{F}}, \Phi_{\mathcal{F}}, W_{ij}, w_{jl}, i = 1, 2, \ldots, I; j = 1, 2, \ldots, J; l = 1, 2, \ldots, L\}$.

E-step of the EM algorithm involves computation of the following conditional expectation: [9]

$$Q(Z,O,\Theta) = E\{\log L(Z,O)|O,\Theta\} = -\frac{N-1}{2}\log Q - \frac{N}{2}\log R - \frac{1}{2}\sum_{k=1}^{N}$$
$$E_N\{[Z(k+1) - \Phi Z(k) - (I-\Phi)T]^T Q^{-1}[Z(k+1) - \Phi Z(k) - (I-\Phi)T]|O,\Theta\}$$
$$-\frac{1}{2}\sum_{k=1}^{N-1} E_N\{[O(k) - \mathbf{h}(Z(k))]^T R^{-1}[O(k) - \mathbf{h}(Z(k))]|O,\Theta\}.$$

This can be simplified by standard algebraic manipulations to

$$\begin{aligned} Q(Z,O,\Theta) &= Q_1(Z,O,\Phi,T) + Q_2(Z,O,W_{ij},w_{jl}) \qquad (5)\\ &= -\frac{N-1}{2}\log\{\frac{1}{N-1}\sum_{k=1}^{N-1} E_N\{[Z(k+1) - \Phi Z(k) - (I-\Phi)T]^2|O,\Theta\}\}\\ &\quad -\frac{N}{2}\log\{\frac{1}{N}\sum_{k=1}^{N} E_N\{[O(k) - \mathbf{h}(Z(k))]^2|O,\Theta\}\} + const. \end{aligned}$$

Note that the task-dynamic parameters (Φ, T) contained in Q_1 only and the MLP weight parameters (W_{ij}, w_{jl}) of the observation equation contained in Q_2 only can be optimized independently in the subsequent M-step which is discussed now.

M-step of the EM algorithm aims at optimizing the Q function in Eqn.(5) with respect to model parameters Θ=$\{T, \Phi, W_{ij}, w_{jl}\}$. For the model at hand, it seeks solutions for

$$\frac{\partial Q_1}{\partial \Phi} \propto \sum_{k=1}^{N-1} E_N[\frac{\partial}{\partial \Phi}\{[Z(k+1) - \Phi Z(k) - (I-\Phi)T]^2|O,\Theta\} = 0 \qquad (6)$$

$$\frac{\partial Q_1}{\partial T} \propto \sum_{k=1}^{N-1} E_N[\frac{\partial}{\partial T}\{[Z(k+1) - \Phi Z(k) - (I-\Phi)T]^2|O,\Theta\} = 0 \qquad (7)$$

$$\frac{\partial Q_2}{\partial W_{ij}} \propto \sum_{k=1}^{N} E_N[\frac{\partial}{\partial W_{ij}}\{[O(k) - \mathbf{h}(Z(k))]^2|O,\Theta\} = 0 \qquad (8)$$

$$\frac{\partial Q_2}{\partial w_{jl}} \propto \sum_{k=1}^{N} E_N[\frac{\partial}{\partial w_{jl}}\{[O(k) - \mathbf{h}(Z(k))]^2|O,\Theta\} = 0. \qquad (9)$$

Eqns.(6) and (7) are third-order nonlinear algebraic equations (in Φ and T):

[9] Together with a set of related sufficient statistics needed to complete evaluation of the conditional expectation .

$$N\Phi T^2 - 2\{\sum_k E_N[Z(k)|O]\}\Phi T - NT^2 + \{\sum_k E_N[Z^2(k)|O]\}\Phi +$$
$$\{\sum_k E_N[Z(k)+Z(k+1)|O]\}T - \{\sum_k E_N[Z(k+1)Z(k)|O]\} = 0,$$
$$N\Phi^2 T - 2N\Phi T - \{\sum_k E_N[Z(k)|O]\}\Phi^2 +$$
$$\{\sum_k E_N[Z(k)+Z(k+1)|O]\}\Phi + NT - E_N[Z(k+1)|O]\} = 0.$$

The coefficients in the above algebraic equations constitute the sufficient statistics, which can be obtained by the standard technique of Iterated Extended Kalman Filtering (IEKF) with fixed-interval smoothing [21, 10], and the equations are solved for (Φ, T) by numerical methods. Alternatively, optimization of Q_1 can be found using gradient decent with explicit expressions of gradients given by Eqns.(6) and (7).

Solutions to Eqns.(8) and (9) for finding (W_{ij}, w_{jl}) to maximize Q_2 in Eqn.(5) have to rely on approximation (due to the complexity in the $\mathbf{h}(.)$ function). The approximation involves first finding smoothed estimates of hidden variables $Z(k)$, $Z(k|N)$, via the IEKF fixed-interval smoother. Given such estimates, the conditional expectations can be approximated (pseudo-EM) to give

$$\frac{\partial Q_2}{\partial W_{ij}} \propto \sum_k [O(k) - \mathbf{h}(Z(k|N))]^T \frac{\partial \mathbf{h}(Z(k|N))}{\partial W_{ij}}$$
$$\frac{\partial Q_2}{\partial w_{jl}} \propto \sum_k [O(k) - \mathbf{h}(Z(k|N))]^T \frac{\partial \mathbf{h}(Z(k|N))}{\partial w_{jl}},$$

where the partial derivatives can be evaluated using the MLP structure to give

$$\frac{\partial \mathbf{h}(Z(k|N))}{\partial W_{ij}} = \begin{pmatrix} 0 \\ \vdots \\ g[\sum_m w_{jm} g(Z_m(k))] \\ \vdots \\ 0 \end{pmatrix} \leftarrow (i^{th}\ element\ non-zero)$$

$$\frac{\partial \mathbf{h}(Z(k|N))}{\partial w_{jl}} = g[\sum_m w_{jm} g(Z_m(k))][1 - g(\sum_m w_{jm} g(Z_m(k)))]g(Z_l(k))] \begin{pmatrix} W_{1j} \\ W_{2j} \\ \vdots \\ W_{Ij} \end{pmatrix}.$$

Given the explicit expressions for $\frac{\partial Q_2}{\partial W_{ij}}, \frac{\partial Q_2}{\partial w_{jl}}$ derived above, gradient-decent algorithm is effectively used to obtain optimized parameters (W_{ij}, w_{jl}) in the M-step.

4..2 Model with random, time-invariant parameters

Statistical motivation for developing a model with random parameters (contrasting the deterministic model parameters discussed in the previous section) is a Bayesian one [21]; that is, we, as modelers, desire to achieve robustness in model parameter estimation and have some prior knowledge to use about the model parameters. Phonetic motivation for allowing for random parameters in task-dynamic model is that different classes of speech sounds have well known, systematic variations in their production strategies (which can be directly quantified in terms of the task-dynamic model's parameter variations), and that speakers tend to use a great degree of (constrained) freedom in choosing their production strategies (plasticity of phonetic gestures advocated in H&H theory).

For example, since parameters $T = (T_1, ..., T_l, ..., T_L)$ (l is index to utterance token) in task-dynamic model represents attractor constriction properties (degree and location) of the vocal tract, it is possible to use well established speech production knowledge (e.g. articulation-acoustics relationships described in quantal theory of speech [29]) to construct the prior in the form of inverse Gaussian distribution (non-negatively valued):

$$f(T_l; \mu_l, \lambda_l) = \sqrt{\frac{\lambda_l}{2\pi}} \times T_l^{-\frac{3}{2}} \times exp[-\frac{\lambda_l (T_l - \mu_l)^2}{2\mu_l^2 T_l}], \quad T_l > 0$$

Note that various broad classes of speech sounds have systematically different hyperparameters μ's and λ's, which are largely predictable, in the above inverse Gaussian distribution . Hence, such prior information can be effectively used to initialize these parameters in subsequent automatic MAP training.

The EM algorithm similar to the earlier deterministic-parameter case applies here for parameter estimation, except now three additional terms are needed in the auxiliary function Q_1 of the E-step due to use of the additional prior distribution:

$$\begin{aligned} Q_1(Z, O, \Phi, T) \quad &\propto \quad -\{\frac{1}{N-1} \sum_{k=1}^{N-1} E_N\{[Z(k+1) - \Phi Z(k) - (I - \Phi)T]^2 | O, \Theta\}\} \\ &+ \quad \sum_{l=1}^{L} \{\frac{1}{2}\, log\lambda_l - \frac{3}{2}\, logT_l - \lambda_l \frac{(T_l - \mu_l)^2}{2\mu_l^2 T_l}\} \end{aligned}$$

M-step of the EM algorithm then gives MAP (empirical Bayes) estimates of both hyper and random model parameters by solving

$$\frac{\partial Q_1}{\partial T} = 0 \quad and \quad \frac{\partial Q_1}{\partial \Phi} = 0,$$

where the second equation above is identical to that in the earlier deterministic-parameter case (Eqn.(7)). Solutions for the first one require that hyper parameters be given, and can be obtained by jointly or iteratively solving $\frac{\partial Q_1}{\partial \mu} = 0$ and $\frac{\partial Q_1}{\partial \lambda} = 0$, which gives optimal estimates for the hyper-parameters. Alternatively, M-step can be accomplished by gradient-decent methods.

4..3 Model with random, smoothly time-varying parameters

This further extension of the statistical task-dynamic model is again motivated by H&H theory of speech gesture plasticity: speakers have significant freedom in articulation, only to be constrained by tradeoffs between the speech-economy principle and by the listener's demand for clarity or sufficient perceptual contrast. In addition to using random parameters, the speaker's freedom in articulation can be further quantitatively represented in the task-dynamic model by allowing for time-varying parameters (within phonological-unit boundaries). On the other hand, the speech-economy principle can be simultaneously quantified by smoothness prior constraints imposed on possible sample paths of these random, time-varying parameters.

For example, the task dynamic with time-varying state transition matrix can be described by

$$Z(k+1) = \Phi(k)Z(k) + [I - \Phi(k)]T(k) + W_d(k), \tag{10}$$

where $\Phi(k)$ is a random parameter (matrix) and is constrained to change slowly over time. A stochastically perturbed difference-equation (order r) model is used to quantitatively provide smoothness prior constraints for random time variation of $\Phi(k)$:

$$\nabla^r \Phi(k) = v(k), \quad v(k) \sim \mathcal{N}(0, \sigma^2).$$

When $r = 1$, $\Phi(k)$ has locally constant trend (i.e. random walk); when $r = 2$ and 3, $\Phi(k)$ has locally linear and quadratic trends, respectively, etc.

For the special case of $T = 0$ in Eqn.(10), a constrained least-square solution to parameter estimation of $\Phi(k)$ has been proposed in [18] as an optimization problem for the following objective function:

$$\sum_k [Z(k+1) - \Phi(k)Z(k)]^2 + \lambda^2 \sum_k [\nabla^r \Phi(k)]^2,$$

where λ^2 is the tradeoff parameter which balances the infidelity of the model to the data $Z(k)$ and the infidelity of the model to the smoothness constraint. [10]

Difficulties in solving the above least-square problem have prompted statisticians to devise more elegant solutions. Results contained in [19, 13] have shown that the smoothness-constraint problem on polynomial parameter trajectories can be equivalently treated as optimal smoothing problem using state-space model formulation. To see this, the smoothness polynomial constraint $\nabla^r \Phi(k) = v(k)$ is rewritten as an equivalent state-space (Gauss-Markov) system:

$$\begin{array}{ll}
r = 1: & \Phi(k+1) = \Phi(k) + v(k) \\
r = 2: & \Phi(k+1) = 2\Phi(k) - \Phi(k-1) + v(k) \\
r = 3: & \Phi(k+1) = 3\Phi(k) - 3\Phi(k-1) + \Phi(k-2) + v(k) \\
\quad \ldots & \ldots
\end{array}$$

This, via the state-augmentation technique, can be equivalently written as a time-invariant linear system:

$$\tilde{\Phi}(k+1) = G\tilde{\Phi}(k) + H\tilde{v}(k),$$

[10] Bayesian interpretation of the above least-square problem is also given in [18].

where

$$\begin{array}{lcl}
for\ \ r=1: & G=I & (constant\ trajectory) \\
for\ \ r=2: & G=\begin{pmatrix} 2I & -I \\ I & 0 \end{pmatrix} & (linear\ trajectory) \\
for\ \ r=3: & G=\begin{pmatrix} 3I & -3I & I \\ I & 0 & 0 \\ 0 & I & 0 \end{pmatrix} & (quadratic\ trajectory) \\
\cdots & \cdots &
\end{array}$$

and

$$H=\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \tilde{\Phi}(k)=\begin{pmatrix} \Phi(k) \\ \Phi(k-1) \\ \vdots \\ \Phi(k-r+1) \end{pmatrix}.$$

This state-space formulation for time variation of parameter $\Phi(k)$, together with the task dynamic state equation, allows all the EM estimation results for the model with time-invariant parameters (discussed earlier) apply to the current time-varying parameter case.

4..4 Discriminative learning of production models' parameters

Discriminative model learning, as opposed to the maximum likelihood one discussed so far, can be theoretically motivated by the argument expressed in H&H theory that the listener's perceptual contrast is the primary objective of human speech communication while employing speech economy in speech production. For possible speech recognition applications in the context of task-dynamic model discussed so far, such an objective can be quantitatively formulated as the problem of minimizing speech recognition errors subject to a tradeoff principle of "least effort" implemented by smoothness constraint on time-varying, random model parameters. Analogies can be made here between the above machine speech recognition strategy and human speech perception: the smoothness constraint on model parameter variations implemented in the recognizer is analogous to minimizing speaker's efforts of production, and the criterion of minimizing speech recognition errors is analogous to maximizing human perceptual contrasts across different phonetic or lexical classes.

The basis of computational formalisms for carrying out the above constrained optimization in the framework of task-dynamic model is already established by speech technologists. The major step involves computation of the gradient of a smoothed estimate of the empirical recognition error with respect to all parameters in the model. For efficiency of training, the gradient has to be expressed in an analytical form. The computation of the gradient is lengthy and laborious and is not included here, but the general spirit of such computation can be gleaned from the work published in [26] applied to a far less sophisticated speech model.

5. Other types of computational models of speech production

So far in this paper I have concentrated on a specific type of speech production model, i.e., task dynamic one. This is a functional model, with no direct representation of biomechanical properties of the vocal tract and with dynamic properties of the system residing only in the “controller”. One main virtue of this model is its uniform definition of the goal of speech production across all consonant and vowel classes in terms of vocal tract constriction properties. This has greatly facilitated algorithmic developments which enable implementation of the model for speech recognition applications. A number of other, non-task-dynamic types of computational models of speech production have been developed, intending to incorporate dynamics either at the biomechanical articulator level or more directly at the acoustic observation level. Within the former class or articulatory-dynamic models, an articulatory stochastic target model was developed which aims at accounting for detailed movement behaviors of biomechanical articulators guided by the highly complex, multi-dimensional target distributions defined in the biomechanical articulator coordinate [9, 25]. Correlations among subsets of articulators in such target distributions are essential because it is the articulator coordinate, rather than the task-variable coordinate, in which the targets are defined. Some more empirical methods used to model articulatory dynamics for purpose of speech recognition include those in [1, 2], where the dynamic of a set of pseudo-articulators is realized by FIR filtering from sequentially placed, phoneme-specific target positions or by applying trajectory-smoothness constraints.

Within the class of acoustic-dynamic models, the model which attempts to condition the properties of the dynamic directly on specific feature-coded speech production mechanisms is described in [7]. In that model, the underlying articulatory-feature based phonological units are used to determine dynamic or static trajectories (order of polynomials) that describe the acoustic correlates of the phonological units, and substantial phonetic recognition performance improvements have been demonstrated. An earlier version of this model, using piecewise-static trajectories (conventional HMM) to approximate continuous trajectories in speech acoustics, is described in [8].

Along the line of acoustic-dynamic model of speech production, there exists a further possibility of choosing more appropriate parametric forms than polynomials to describe production-correlated acoustic variables. For example, the polynomial trajectories (as used in many earlier segmental models) do not entail the concept of formant target since they do not have the asymptotic property which allows the trajectory to slowly and smoothly relax to an asymptotic value such as the formant target. Exponential form of trajectories, however, has such an asymptotic property; e.g. $f(t) \propto f^0 \times (1 - \alpha t \times exp[-\gamma \times t])$. But some serious difficulties would arise if this exponential form of the trajectory model were to be used for formants directly. Because many consonants do not show acoustically measurable formants (due to full or partial pole-zero cancellation in vocal-tract acoustics caused by supraglottal excitation sources), the trajectory model has to be generalized from that describing measurable formant trajectories (applicable only to vowels) to that describing hidden vocal-tract resonance dynamics (applicable to all types of speech sounds). Smoothness and continuity constraints can then be naturally applied to the hidden vocal-tract resonance trajectories through entire utterances, thus naturally producing speech undershoot phenomena characteristic of casual, fast speech (as observed in Switchboard data). Due to the hidden nature of vocal-tract resonances, especially for consonants, it will be appropriate to use MFCCs as speech observations and to empirically build noisy nonlinear mappings from vocal-tract resonances (poles of vocal-tract transfer function) to MFCCs.

6. Summary and discussions

In this tutorial, major classes of speech models developed by two largely separate, scientific and technological, communities are surveyed, compared, and analyzed. Similar comparisons and analyses from a more global perspective have been made earlier in [22]. A particular type of speech production model, task-dynamic one, is developed which integrates the strengths of the two previously separate styles of production models. This integration owes much to successful use of the smoothness-prior (Bayesian equivalent) technique, motivated by gesture-plasticity and movement-economy principles in human speech production, in establishing the statistical task-dynamic model. Both maximum likelihood (via EM) and minimum classification error (via gradient descent) criteria are used for model parameter learning, justified in terms of various versions of phonetic theories. In either case, optimization of the likelihood or empirical classification error rate for a small number of hyperparameters in the model permits robust modeling of true dynamic behaviors of human speech. Modeling such behaviors requires a complex structure, but the technique we adopted enables use of a most compact set of hyperparameters which encompass a large number of implicitly inferred parameters.

Acknowledgements

Over the past several years and on the subject matter of the three tutorial papers written by the author in this book, many discussions with or experimental contributions from the following individuals are gratefully acknowledged: K. Stevens, J. Perkell, V. Zue, R. McGowan, E. Saltzman, C. Browman, L. Goldstein, C. Lee, G. Chollet, G. Ramsay, D. Sun, H. Sameti, J. Wu, H. Sheikhzadeh, and I. Kheirallah.

7. References

[1] Bakis R. (1993), “An articulatory-like speech production model with controlled use of prior knowledge,” notes from *Frontiers in Speech Processing*, CD-ROM.

[2] Blackburn C., and Young. S. (1995), “Towards improved speech recognition using a speech production model,” *Proc. Eurospeech*, vol. 2, pp. 1623-1626.

[3] Deng L. (1992) “A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal,” *Signal Processing*, vol.27, pp. 65-78.

[4] Deng L. (1993) “Design of a feature-based speech recognizer aiming at integration of auditory processing, signal modeling, and phonological structure of speech.” *JASA*, vol. 93(4) Pt.2, pp. 2318.

[5] Deng L. (1992-1993) “A Computational Model of the Phonology-Phonetics Interface for Automatic Speech Recognition,” Summary Report of Research in Spoken Language Systems, Laboratory for Computer Science, MIT.

[6] Deng L. and Aksmanovic M. (1997) “Speaker-independent phonetic classification using hidden Markov models with mixtures of trend functions,” *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 319-324.

[7] Deng L. and Sameti H. (1996) “Transitional speech units and their representation by the regressive Markov states: Applications to speech recognition,” *IEEE Trans. Speech Audio Proc.*, vol. 4(4), pp. 301–306.

[8] Deng L. and Sun D. (1994), “A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features,” *JASA*, vol. 95, pp. 2702-2719.

[9] Deng L., Ramsay L., and Sun D. (1997) “Production models as a structural basis for automatic speech recognition,” *Speech Communication*, August issue.

[10] Digalakis V., Rohlicek J., and Ostendorf M., (1993) “ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition”, *IEEE Trans. Speech Audio Processing*, pp. 431-442.

[11] Ghitza O., and Sondhi M. (1993) “Hidden Markov models with templates as nonstationary states: an application to speech recognition,” *Computer Speech and Language*, vol. 7, pp. 101–119.

[12] Gales M. and Young S. (1993) “Segmental HMMs for speech recognition,” *Proc. Eurospeech*, pp. 1579-1582.

[13] Gersch W. (1992) “Smoothness priors,” in *New Directions in Time Series Analysis*, D. Brillinger et al. (eds.), Springer, New York, pp. 111-146.

[14] Gish H. and Ng K. (1993) “A segmental speech model with applications to word spotting,” *Proc. ICASSP*, pp. 447-450.

[15] Haykin S. (1994) *Neural Networks — A Comprehensive Foundation*, Maxwell Macmillan, Toronto.

[16] Holmes W. and Russell M. (1995) “Speech recognition using a linear dynamic segmental HMM,” *Proc. Eurospeech*, pp. 1611-1641.

[17] Kent R., Adams S. and Turner G. (1995) “Models of speech production,” in *Principles of Experimental Phonetics*, Ed. N. Lass, Mosby: London, pp. 3-45.

[18] Kitagawa G. and W. Gersch W. (1996) *Smoothness Priors Analysis of Time Series*, Springer, New York.

[19] Kohn R. and Ansley C. (1988) “Equivalence between Bayesian smoothness priors and optimal smoothing for function estimation,” in *Bayesian Analysis of Time Series and Dynamic Models*, J. Spall (ed.), Marcel Dekker, New York, pp. 393-430.

[20] McGowan R. (1994) “Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests,” *Speech Communication*, 14, pp. 19-48.

[21] Mendel J. (1995) *Lessons in Estimation Theory for Signal Processing, Communications, and Control*, Prentice Hall, New Jersey.

[22] Moore R. (1994) “Twenty things we still don’t know about speech,” *Proc. CRIM/FORWISS Workshop on Speech Research and Technology*, pp. 1-9.

[23] Ostendorf M. (1996) “From HMMs to segment models,” in *Automatic Speech and Speaker Recognition – Advanced Topics*, C. Lee, F. Soong, and K. Paliwal (eds.), Kluwer Academic Publishers, pp. 185-210.

[24] Perrier P. et al. (eds.) *Proceedings of the First ESCA Tutorial & Research Workshop on Speech Production Modeling* , Autrans, France, May 24-27, 1996.

[25] Ramsay G. and Deng L. (1996) “Optimal filtering and smoothing for speech recognition using a stochastic target model,” *Proc. ICSLP*, pp. 1113-1116.

[26] Rathinavalu C. and Deng L. (1997) “HMM-based speech recognition using state-dependent, discriminatively derived transforms on Mel-warped DFT features”, *IEEE Trans. Speech Audio Processing*, pp. 243-256.

[27] Rubin P. et al (1996) “CASY and extensions to the task-dynamic model,” *Proc. 4th European Speech Production Workshop*, Autrans, France, pp. 125-128.

[28] Saltzman E. and Munhall K. (1989) “A dynamical approach to gestural patterning in speech production,” *Ecological Psychology*, 1, 333-382.

[29] Stevens K. (1989) “On the quantal nature of speech,” *J. Phonetics*, vol.17, 1989, pp. 3–45.