ETRW-96 paper selected to be published in Speech Communication

PRODUCTION MODELS AS A STRUCTURAL BASIS FOR AUTOMATIC SPEECH RECOGNITION

L. Deng[1], *G. Ramsay*[1], *and D. Sun*[2]

[1] *Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada*

[2]*Bell Laboratories, Murray Hill, NJ, U.S.*

ABSTRACT

We postulate in this paper that highly structured speech production models will have much to contribute to the ultimate success of speech recognition in view of the weaknesses of the theoretical foundation underpinning current technology. These weaknesses are analyzed in terms of phonological modeling and of phonetic-interface modeling. We conclude by suggesting that many of the advantages to be gained from interaction between speech production and speech recognition communities will develop from integrating models from the production community with the probabilistic analysis-by-synthesis strategy currently used by the technology community.

RÉSUMÉE

Dans cet article, nous proposons que les modèles de production de la parole contribueront beaucoup à la réussite eventuelle des modèles de reconnaissance automatique, limités en ce moment par les faiblesses de la base théorique de la technologie actuelle. Nous analysons ces faiblesses au niveau des modèles phonologiques et modèles phonétiques, et suggérons un lien avantageux entre la production et la reconnaissance, fondé sur l'intégration des deux genres de modèles dans le cadre d'une stratégie d'analyse-synthèse probabilistique utilisée déjà depuis longtemps dans le domaine de la reconnaissance de la parole.

Key words: Speech production, speech recognition, analysis by synthesis, stochastic modeling, nonlinear phonology, phonetic interface, articulatory feature and articulatory dynamics.

---

* Send all correspondences to Li Deng, deng@crg5.uwaterloo.ca.

1. Introduction

The past quarter of a century has witnessed a conspicuous division between the research efforts of speech technologists and speech scientists. In speech technology, in particular speech recognition, the development and use of largely unstructured statistical models (e.g. hidden Markov models) have dominated such efforts (e.g., [72, 47]). On the other hand, work from speech production researchers has centered on elaborating detailed models and theories intended to account for the nature of the transformation from discrete phonological symbols to continuous acoustic streams via motor control strategies/articulatory dynamics (e.g., [64, 41]). Nowhere has this division been more evident than in viewing what constitutes the "atomic" units of speech. In linguistic and speech production theories, sub-phonemic or sub-segmental entities such as features, gestures, and motor commands have been a central focus permeating much research. In contrast, virtually all speech recognition systems have been built on speech units of the size of phonemes or larger, with limited exceptions.

In the light of this division, one purpose of writing this tutorial paper is to contribute to bridging this gap by arguing that there are advantages to be gained from interaction between the speech science and speech technology communities. The benefits of this interaction will lie both in deeper understanding of the nature of the human speech communication process and in making such understanding useful in industrial applications. On the one hand, state-of-the-art speech recognition technology has (arguably) reached a "local optimum" [9] — with the global optimum defined as machine performance indistinguishable from human performance on natural speech. In order to escape from this "local optimum", speech recognition needs new concepts, and we believe one key source of ideas should come from global speech production models which are capable of simulating key mechanisms of the human speech communication process, but at the same time remain amenable to computation. On the other hand, for speech scientists interested in making their models useful for speech recognition, we will argue that conventional deterministic approaches to modeling should eventually be replaced by the statistical Bayesian-theoretic approach (which may be viewed as probabilistic analysis/synthesis) already taken for granted by most technologists. Beginning in speech recognition, we have been conducting research over the past few years that has involved ideas from both speech production and recognition fields. In this tutorial, we will give an extensive review of background work and then describe our own experience and some results of our research. It is our hope that this tutorial will serve the purpose of demonstrating the benefit of integrating research in speech production and recognition, fields which have unfortunately been divided for so many years.

2. "Fundamental equation" of speech recognition

In order to present a convincing case that production-oriented models are truly useful for speech recognition, and that the probabilistic approaches emerging from the speech recognition community may be useful for speech production modeling, we need to give a brief description of the statistical framework that underlies much of modern speech recognition research and system development.

Let $\mathbf{O} = O_1, O_2, ..., O_T$ be a sequence of observable acoustic data of speech, which can either be speech waveforms [66], or continuous-valued acoustic vectors [72], or discrete-valued vector-quantized codes [47], or any other type of general acoustic measurements, and let $W = w_1, w_2, ..., w_n$ be the sequence of words intended by the speaker who produces the acoustic record $\mathbf{O}$ above. The goal of a speech recognizer is to "guess" the most likely word sequence $\hat{W}$ given the acoustic data $\mathbf{O}$. Bayesian decision theory provides a minimum Bayes-risk solution to the above "guessing game", and the minimum Bayes risk can be made equivalent to minimum probability of error if the risk is assigned values of one or zero for incorrect and correct guesses, respectively. According to Bayesian decision theory, speech recognition is formulated as a top-down search problem over the allowable word sequences W based on the posterior probability $P(W|\mathbf{O})$:

$$\hat{W} = \arg\max_{W} P(W|\mathbf{O}) = \arg\max_{W} P(\mathbf{O}|W)P(W), \tag{1}$$

where $P(W)$ is the prior probability that the speaker utters W, which is independent of the acoustic data (and hence of relatively minor interest to speech production researchers) and is determined by the language model, and $P(\mathbf{O}|W)$ is the probability that the speaker produces (or the microphone of the speech recognizer receives) the acoustic data $\mathbf{O}$ if W is the intended word sequence by the speaker. Disregarding the issue of language modeling, the above formulation, or fundamental equation (1), of the speech recognition problem can be reduced to two issues: 1) speech generation or production from word sequence to acoustic streams — how to accurately compute the probability $P(\mathbf{O}|W)$[1]? and 2) a search for the word sequence W (the operation $\arg\max_W$ in Eqn. 1) that provides the optimal value of the posterior probability.

Eqn.(1) is essentially a re-formulation of analysis-by-synthesis cast in a consistent probabilistic framework: the synthesis phase is embedded in the assumption that acoustic data $\mathbf{O}$ is produced from word sequence W (hence the necessity and possibility to evaluate the production probability $P(\mathbf{O}|W)$), and the analysis phase involves finding the solution $\hat{W}$

[1] Bayesian decision guarantees minimal Bayes risk only if the probabilities involved are estimated correctly.

which best matches (in a maximum-likelihood sense) the outcome of the production. From this analysis-by-synthesis interpretation of Eqn.(1), we will see that good speech production theories, when meeting the computational requirements implied in Eqn.(1), should have much to contribute to advancing speech recognition technology[2].

3. Critical review of HMMs: Lessons learned from technologists

There is no doubt that HMMs are currently the most successful technology in many (heavily) constrained speech recognition applications; see [56, 72] for an overview of this technology. In our view, this success is not so much due to the mathematical formulation of the HMM itself as due to its conformity to the probabilistic analysis-by-synthesis formulation epitomized in Eqn.(1). Implicit in Eqn.(1) are the need to efficiently compute a production probability $P(\mathbf{O}|W)$ and the need to learn "production model" parameters so as to achieve high accuracy in evaluating $P(\mathbf{O}|W)$. HMMs are amenable to efficient computation and parameter learning thanks to Baum's work [5], and thus would fit naturally into the probabilistic analysis-by-synthesis framework of Eqn.(1). This is entirely consistent with the qualification of an HMM as a speech generator or production model, because embedded in the HMM there is a mechanism for converting a word sequence W directly into acoustic data $\mathbf{O}$. One simple way to view an HMM as a speech production ot synthesis device is to run Monte Carlo simulation on the HMM and regard the outcome of the simulation as the synthetic speech (cf. [58]). Filter-based approaches have also been proposed for HMM speech synthesis [69].

The theoretical treatment of the HMM as a production model is one thing; how reasonably and effectively it behaves as a production model is another thing. To examine this latter issue, let us first examine the production probability $P(\mathbf{O}|W)$ which appeared in Eqn.(1) into two factors:

$$P(\mathbf{O}|W) \;=\; \sum_{\mathcal{P}} P(\mathbf{O}|\mathcal{P})P(\mathcal{P}|W) \approx \max_{\mathcal{P}} P(\mathbf{O}|\mathcal{P})P(\mathcal{P}|W), \tag{2}$$

where $\mathcal{P}$ is a discrete-valued *phonological model* and specifies, according to probability $P(\mathcal{P}|W)$, how words and word sequences W can be expressed in terms of a particular organization of a small set of "atomic" phonological units; $P(\mathbf{O}|\mathcal{P})$ is the probability that a particular organization $\mathcal{P}$ of phonological units produces the acoustic data for the given

[2] In fact, the power of this probabilistic analysis-by-synthesis view is already emerging from the area of robust speech recognition, where both speech and noise are treated as the simultaneous result produced from a "composite" HMM generator [70, 29].

word sequence W. We shall call this latter mapping device from phonological organization to speech acoustics the *interface model*.

In view of the factorization in Eqn.(2), state-of-the-art speech recognizers [47, 72] based on phonetic HMMs can be analyzed as follows. The phonological model $\mathcal{P}$ is essentially a linearly-organized multiple-state phonetic sequence governed by a left-to-right Markov chain, and the interface model is simply a temporally independent random sampling from a set of (trainable) acoustic distributions associated with the states in the Markov chain. Therefore, the following straightforward decomposition in computing the probabilities associated with the phonological model and with the interface model is possible:

$$P(\mathcal{P}|W) = \prod_{t=0}^{T-1} P(s_{t+1}|s_t); \quad P(\mathbf{O}|\mathcal{P}) = \prod_{t=1}^{T} b_{s_t}(\mathbf{O}_t),$$

where $P(s_{t+1}|s_t)$ and $b_{s_t}(\mathbf{O}_t)$ are the transition probabilities of the Markov chain and the state-dependent output distribution of the HMM, respectively.

It is obvious from this discussion that the conventional phonetic HMM outlined above is a poor and naive speech generator or production model: the use of linearly-organized units as the phonological model is outdated, and ignores developments in modern phonological theory, whereas the use of an independent and identically distributed (i.i.d.) stochastic process (conditioned on the HMM state sequence) as the acoustic interface model discards many of the key temporal correlation properties in the acoustic signal resulting from relatively smooth motion of the articulatory structures. There are clearly many opportunities for improving both phonological and interface components of the HMM.

Now, given our rejection of the phonetic HMM as a *good* generator/production model, one may ask why it has achieved so much success in present speech recognition technology. Does such success imply that good production models have no role to play in speech recognition? Our answer is just the opposite. In our view, the current (limited) success of the HMM is to a large degree due to the many constraints (such as benign application environments, limited speaker variability and vocabulary size, and unnatural speaking style) imposed on the recognition tasks. These constraints create an artificially sparse phonetic space and limit many possible phonetic confusions. Task constraints are often so strong that even a simplistic HMM is able to do a reasonably good job in disambiguating different phonetic classes in many practically useful speech recognition tasks. The important point to note is that such limited success can be largely attributed to the probabilistic "analysis by synthesis" framework expressed in the fundamental equation (Eqn.1) and to the use of automatic learning which is an implicit component of the framework. Early deterministic approaches, such as expert systems, failed to achieve similar success in comparable tasks. Here we see the compelling need for developing production models superior to conventional phonetic HMMs for greater success in speech recognition technology with fewer constraints on the task domain. Recent evaluation experiments on real-world speech recognition tasks using a telephone switchboard database demonstrate the poor performance of state-of-the-art technology based on conventional HMMs. The superior performance achieved by discrimination-motivated HMM parameter learning over the maximum-likelihood learning [11, 12, 29] further attests to the poor quality of conventional HMMs as a generative model for use in speech recognition.

In pursuing the development of high-quality global speech production models that can theoretically guarantee superiority in speech recognition tasks as argued above, two key modeling requirements must be emphasized. First, the models must take into account critical mechanisms in the human speech communication process that describe systematic variability in speech acoustics as a necessary and natural means to convey phonologically meaningful information from speaker to listener; much of such systematic variability has been detrimental to the current HMM-based speech technology. Second, the current success of the HMM in technology has taught us the lesson that any good speech production model for use in speech recognition should be compatible with the computational requirements imposed by the probabilistic analysis-by-synthesis framework. This would include the ability of the model to allow for efficient top-down search through sentence-level hypotheses and the possibility for automatic learning of model parameters from any available training data[3]. What appear to be incompatible with the probabilistic top-down approach are the schemes stemming from the idea of deterministic bottom-up "inversion" (from acoustics to articulation, from articulation to motor control commands, and from control commands to phonological units, etc.) frequently adopted by speech production researchers interested in recognition.

4. Phonological and interface models: A review

The decomposition of the likelihood $P(\mathbf{O}|W)$ in terms of the phonological and interface models (Eqn.2) serves as a conceptually convenient way to review and classify many known approaches in speech recognition. We will conduct such a review from a viewpoint that treats most of the approaches as based on either an implicit or an explicit speech production model.[4]

[3]The training is required here because of the structural simplicity of the model assumed for computation reasons and of the existence of random variability in speech data. In contrast, in the current phonetic-HMM based speech recognition technology, random and systematic variabilities are largely undistinguished and a combination of them leads to the demand of a large amount of training data.

[4]A review of potential roles of speech production modeling in speech recognition from a somewhat different

A host of phonological models used in speech recognition include word, syllable, demisyllable, diphone, phoneme, context-dependent allophone, and sub-phoneme models. While for reasonably large tasks, use of context-dependent allophones as in [47] and of sub-phonemic units constructed systematically by training from acoustic data (e.g., [36]) has achieved promising results, there are compelling reasons to believe that more challenging, unconstrained tasks with performance approaching human capability will require sub-phonemic models grounded solidly on modern phonological theories.

Some of the sub-phonemic phonological models used with varying degrees of success in speech recognition include the microsegment model [22], the locus model focusing on CV and VC transitions [17], and that which directly takes Chomsky-Halle binary distinctive features as the recognition object [27, 49]. A phonological model based on an articulation-based feature-geometric theory is reported in [6, 44, 68] which provides preliminary evidence for its value in classification of limited phonetic classes.

Although no positive results have been reported yet, another interesting phonological model is the one used in the Bakis-type speech recognizer [3, 4, 7], where abstract phoneme-specific control commands, or targets, serve as the phonological construct. Yet another phonological model, in a rather different spirit than the previous ones, is based on the idea of quantizing articulatory variables (called (pseudo)-articulatory or multi-valued phonetic features) [61, 23, 28, 40]. Our experience showed that this type of model can be made effective for classification of limited phonetic classes but it is difficult to extend this effectiveness to broader classes of speech sounds. This difficulty arises because the strict and precise ordering of the quantized features is not flexible enough to describe the compensatory effects associated with production of a wide class of speech sounds.

Overcoming the above weakness and again motivated by the articulatory organization of speech, the overlapping articulatory feature model reported in [14, 19, 21] treats each articulatory feature as a symbolic entity (i.e. with no partial ordering) embodying phonological contrasts together with acoustic and possibly auditory correlates [5]. The overlapping nature of the articulatory features in this model is directly motivated by representations in autosegmental phonology [31] and by the way gestural scores are constructed in articulatory phonology [10]. Some details of this overlapping feature based model will be given in Section 5.

---

viewpoint recently appeared in [62].

[5]Capitalization on such correlates at the structural level (e.g., static or dynamic acoustic patterns), however, is the job of the phonetic interface model.

Unlike the phonological models which mostly originated from linguistic research, most of the interface models used in speech recognition have been developed by speech technologists or statisticians. As mentioned in Section 2, the role of an interface model is to provide perceptually significant linkage between phonological units to acoustic observations of speech[6]. An interface model can be viewed as a "forward" or production model from the production viewpoint, and equivalently, as we argued in Section 2, as an "inverse" or recognition model from the analysis-by-synthesis viewpoint.

Several notable approaches which play the role of the interface model used in speech recognition are reviewed briefly here. The most straightforward, largely deterministic approach is exemplified by early rule-based systems; the methods reported in [44, 68, 61] also can be classified in this broad category. In the second category of the interface models is the neural-network based approach, which is characterized by direct mapping between phonological symbols and speech acoustics [50, 49, 24]. Further, as we discussed in Section 3, HMMs are the most common interface model in the current speech recognition technology. In addition to the conventional, stationary-state HMM[7], various versions of (segment-based) nonstationary-state HMMs [15, 30, 16, 25, 33] and several other types of the segment-based models [51] including the linear dynamical system model [26] have more recently been developed as theoretically superior interface models from phonological units to acosutics. The above stochastic dynamical-system or trajectory models can be viewed as the acoustic-dynamic model.

It is worth pointing out that the interface models mentioned above have lacked a level of explicit representation of articulatory dynamics in mapping from phonological units to acoustics. Two recent models have provided such an explicit representation, and thereby can be called the articulatory-dynamic model. One is the model in [4] where FIR filters are employed to functionally interface phoneme-specific targets (phonological entity) with motions of "abstract" or psudo-articulators, and nonlinear neural networks are used to interface the articulator motions with acoustics. The other is the model we have recently developed where a linear dynamical system and articulatory synthesizer are integrated into a stochastic framework which serves as an interface between abstract phonological units and acoustics. We will describe this latter model with some detail in Section 6.

[6]In a broader sense, this is the phonology-phonetics interface discussed much in the phonology and phonetics literature. It would include the interface to articulation if articulatory measurements are available to the speech recognizer, and even to motor control commands if they can be estimated accurately enough.

[7]The recognizers described in [72, 6, 19] employed essentially an identical interface model but with different phonological models.

Another general type of dynamic-interface model, which employs an abstract and implicit rather than explicit articulatory representation, is the task-dynamic model developed over the past years largely at Haskins Laboratories [39, 64, 48, 63, 37]. The task-dynamic model differs conceptually from the explicit articulatory-dynamic model in that the dymamics is described in the "task" space (vocal-tract constriction degrees and locations) rather than at the biomechanic articulatory level (positions of tongue, lips, and jaw, etc.) A statistical and computational framework for interfacing overlapping articulatory features as the phonological model to the task-dynamic model for potential use in speech recognition was presented in [14]. The main difficulty for successful applications has been the formidable computational burden required to characterize the highly nonlinear relationship between the task-space variables and the biomechanic articulatory variables.

5. Speech recognition using overlapping articulatory features

One principal motivation of the articulatory feature model described in this section comes from our recognition of the weakness of the conventional phoneme-sized HMM viewed as a speech production model. The following is a quote from an early seminal paper which significantly contributed to the popularity of the HMM in speech recognition: "[...] It is quite natural to think of the speech signal as being generated by such a (HMM) process. We can imagine the vocal tract as being in one of a finite number of articulatory configurations or (HMM) states. [...]" [43]. It has become apparent nowadays that the mechanism described above that associates HMM states to articulatory configurations has been highly superficial. The phoneme-sized HMM is essentially a flexible piece-wise data-fitting device and describes mere surface phenomena of speech acoustics rather than any underlying mechanisms of the speech process. This is the reason why an increase in size of the HMM state-space and in the amount of training data appear to be the only possibility for more accurate representation of speech if one is to build more robust speech recognizers for tasks with fewer constraints [72].

The overlapping articulatory feature model aims at constructing a multi-dimensional HMM whose states can be made to directly correspond to the symbolically-coded, phonologically-contrastive articulatory structure responsible for generating acoustic observations from the states. The very nature of multiple dimensionalities, separate for each phonologically significant articulatory gesture tier, of the HMM allows embodiment of the asynchronous articulatory feature/gesture overlaps (coarticulation) in a natural way.

The overall design of the speech recognizer is cast in the probabilistic analysis-by-synthesis framework; no direct inversion operation is necessary and to perform speech recognition there is no requirement (although desirable) for articulatory measurement data. The recognition process involves top-down hypothesizing sentence-level solutions W (either by search techniques or by N-best inputs obtained from conventional recognizers), together with scoring (matching) each hypothesis with the acoustic data using the assumption that the data are produced from a sequence of multi-dimensional articulatory HMM states. The articulatory states are constructed in advance (see below) using a phonetic transcription[8] for each sentence-level hypothesis and can be retrieved instantaneously during recognition. At the heart of the recognizer is our algorithm for automatic conversion of any probabilistic and fractional articulatory feature overlap pattern [19] into a Markov state transition graph, which is summaried below.

Construction of articulatory HMM states

Two key components are required for constructing the articulatory HMM states: 1) an articulatory feature specification system; and 2) constraints on feature overlaps and spreads. A portion of the feature specification system for American English is shown in Table 1, where symbolic feature value *'U'* denotes feature underspecification.

Notations

To describe the procedure for the articulatory state construction, we first introduce the following notation. Let

$$\mathbf{\Phi} = (\phi_1, \cdots, \phi_m)$$

be the phonetic transcription of a sentence, where m is the number of phonetic segments and ϕ_i takes a discrete value of phonetic symbols. Let

$$\mathbf{f}(\phi_i) = (f_1(\phi_i), \cdots, f_D(\phi_i))^T$$

be the vector of articulatory features of target segment ϕ_i in the phonetic transcription. For example, $\mathbf{f}(/b/) = (L\alpha, U, U, V\alpha, X\alpha)$ shown in Table 1. Similarly, we let

$$\mathbf{g}(\mathbf{f}(\phi_{i+\delta}), \phi_i)$$

be the vector of contextual articulatory features of target segment ϕ_i assimilated by the features at segment $i+\delta$, where δ takes integer values ($\delta > 0$ for anticipatory coarticulation and $\delta < 0$ for carry-over coarticulation). Obviously, $\mathbf{g}(\mathbf{f}(\phi_i), \phi_i) \triangleq \mathbf{f}(\phi_i)$ when $\delta = 0$.

Algorithm Description

[8] Research on incorporating prosodic information and syllabic structure in the state construction, intended for multilingual speech recognition, is currently underway.

Input:

phonetic transcription of a given utterance (part or whole sentence): $\mathbf{\Phi} = (\phi_1, \cdots, \phi_m)$.

Output:

articulatory state transition graph.

Algorithm:

1. Attaching D-tuple features to each target segment ($\mathbf{f}(\phi_i)$) according to the feature specification system (D=5).

2. Feature specification of contextual segments ($\mathbf{g}(\mathbf{f}(\phi_{i+\delta}), \phi_i)$):

   With the notation introduced above, the articulatory feature specification of each target segment ϕ_i *in context* can be written as:

$\cdots$	$g_1(\mathbf{f}(\phi_{i-1}), \phi_i)$	$f_1(\phi_i)$	$g_1(\mathbf{f}(\phi_{i+1}), \phi_i)$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\cdots$	$g_D(\mathbf{f}(\phi_{i-1}), \phi_i)$	$f_D(\phi_i)$	$g_D(\mathbf{f}(\phi_{i+1}), \phi_i)$	$\cdots$

   The values of the contextual features of the segment ϕ_i are determined by the values of $\mathbf{f}(\phi_i), \cdots, \mathbf{f}(\phi_{i+\delta})$ together with feature overlap and spread rules.

   Initially, we set

   $$\mathbf{g}(\mathbf{f}(\phi_{i+\delta}), \phi_i) = \mathbf{f}(\phi_{i+\delta}).$$

   Then, the values of $\mathbf{g}(\mathbf{f}(\phi_{i+\delta}), \phi_i)$ are modified according to a set of rules including:

   (a) Maximum feature spread constraint:

   $g_d(\mathbf{f}(\phi_{i+\delta}), \phi_i) =$ 'U' for $d = 1, \cdots, D$, and $|\delta| \leq \Delta$, where Δ is a constant indicating the maximum amount of spread.
   This rule specifies the constraint on the maximum span of feature spreading.

   (b) Discontinuity with underspecified features:

   If $g_d(\mathbf{f}(\phi_{i+\delta}), \phi_i) =$ 'U', then $g_d(\mathbf{f}(\phi_{i+\delta'}), \phi_i) =$ 'U', for $|\delta'| \geq |\delta|$.
   This rule prevents feature spreading across some segments with underspecified features.

3. Construction of states S:

According to Step-2 results which give the feature specifications for all the segments *in context*), enumerate all D-tuple feature bundles (states):

$$S \triangleq (g_1(\mathbf{f}(\phi_{i+\delta_1}), \phi_i), \cdots, g_D(\mathbf{f}(\phi_{i+\delta_D}), \phi_i))^T$$

which satisfy

(a) $\delta_1, \cdots, \delta_D$ are either all positive or all negative.

(b) $g_d(\mathbf{f}(\phi_{i+\delta_d}), \phi_i) \neq$ 'U' for $\delta_d \neq 0$.

4. Determination of state transitions:

Enumerate all transitions from state (a)

$$S(a) = (g_1(\mathbf{f}(\phi_{i+\delta_1(a)}), \phi_i), \cdots, g_D(\mathbf{f}(\phi_{i+\delta_D(a)}), \phi_i))^T$$

to state (b)

$$S(b) = (g_1(\mathbf{f}(\phi_{i+\delta_1(b)}), \phi_i), \cdots, g_D(\mathbf{f}(\phi_{i+\delta_D(b)}), \phi_i))^T$$

which satisfy

$$\delta_d(a) \leq \delta_d(b), \quad \text{for } d = 1, \cdots, D$$

where at least one inequality holds strictly, and

$$\delta_d(b) - \delta_d(a) \leq \Delta_{\text{skip}} \quad \text{for } d = 1, \cdots, D$$

A speech recognizer which uses the above method of constructing articulatory HMM states has been implemented and evaluated on standard tasks commonly used in speech recognition research. Our experience suggests that a great deal of care needs to be taken in incorporating feature-overlap constraint rules, with the degree of rule relaxation made dependent on the amount of training data. Once this care is taken, our results have demonstrated consistently superior performance of the recognizer in comparison with several benchmark systems [19, 18, 21, 20].

6. A stochastic target model for potential use in speech recognition

The previous section described a symbolic, knowledge-based approach to the incorporation of speech production information in the phonological component of a statistical speech recognizer. The current section outlines a model for the phonetic interface linking discrete-state phonological representations to their continuous-state articulatory and acoustic counterparts for potential use in speech recognition.

Determining the relationship between the many different levels of representation involved in generating and perceiving the speech signal is a central and long-standing problem in speech production research, but has largely been ignored in main-stream speech recognition, which typically assumes only two representations – phonemic and acoustic, neither of which is explicitly linked to the behavior of the vocal tract. Given that much of the variation in speech that makes speech recognition difficult is due to articulatory phenomena, it is reasonable to assume that incorporating an accurate, explicit articulatory model in the structure of the recognizer should lead to improved performance, or at least to a more compact parameterization of certain types of variability.

Stop epenthesis after nasals, for example, occurs when an additional silence and burst is introduced into the acoustic signal due to differences in timing between velic and oral closure. In current recognition systems, the only way to deal with this is to assume an extra lexical entry (including a stop phoneme) for the word in question, or to broaden the acoustic model of the nasal to include the characteristics of the epenthetic stop. Neither of these generalizes correctly, since the explanation lies in the articulatory domain and is neither acoustic nor phonological in origin. Modifying the lexicon or the acoustic model does not capture the *reason* for the change which would allow extrapolation to unseen contexts without explicitly enumerating all possibilities.

The central concern for speech recognition is to provide a mechanism which can describe all possible observed variation succinctly without overgenerating, learn the parameters of this description from a finite set of training data, and use this information to discriminate between possible lexical alternatives, even when examples of particular contexts are not present in the training data.

Unfortunately, there appears to be no simple direct correspondence between underlying phonological structures and their surface phonetic realizations, and it is difficult to see how this aim might be accomplished without eventually including some prior knowledge of the underlying physics which constrain human production and perception.

As a first step towards incorporating an explicit production model into speech recognition, we have developed a *stochastic target model* for recognition and synthesis [57, 58, 59, 60] which is capable of accounting for some of the systematic articulatory variation observed in speech, within a recognition framework. The present model is summarized below, to provide an illustration of the kind of approach that may prove successful in integrating speech production and speech recognition.

We note first of all that a number of important alternative attempts have already been made to include articulatory representations in speech recognition. Most of these have centered on direct deterministic approaches to the inversion problem, or on functional approximation of the articulatory-acoustic mapping, and do not fit easily into the statistical HMM-based framework with which we are concerned. A comprehensive review of this work has already appeared in Schroeter and Sondhi [65] and will not be discussed further here. Examples of similar recognition models which also contain an intermediate articulatory representation appeared in [67, 71, 4, 7].

In the model, utterances are assumed to consist, at the phonological level, of abstract sequences of overlapping symbols drawn from a finite alphabet. These might be phonemes, or feature bundles, or abstract gestures, depending on the phonological framework and lexical representation adopted for recognition. A higher-level phonological model is assumed to be available to generate hypothesized symbol sequences and their probability of occurrence.

Each phonological symbol is taken to correspond to a family of physical correlates, which may be articulatory, acoustic, or perceptual, not all of which need necessarily be realized in any particular instance or context. The class of correlates for each symbol is described statistically by a probability distribution over the appropriate measurement space in which the correlates are observed, and this can be constructed empirically by examining an ensemble of realizations from real or modeled data. The choice of correlates must come from phonetic knowledge, (though the parameters of the distributions can be trained from data), and it is quite possible that different specifications may give rise to equivalent distributions. For example, the correlates for $/u/$ might be defined by lip-rounding and velar constriction, or by a low $F1/F2$ pattern, and it is not immediately clear which is the better description, or whether one automatically entails the other.

The key modeling assumption is that all probability distributions on any number of measurement spaces can be projected onto a single equivalent spatial distribution of targets on a space of articulatory parameters describing the state of an articulatory model. The parameters may describe simple kinematic properties (jaw angle, tongue elevation), or could eventually represent muscle activations etc., depending on the complexity of the available model, and the target distribution for each symbol represents the probability that any particular point in the model space will be used to realize the specified phonetic correlates. The "control strategy" is therefore represented in purely articulatory terms, though it may well be constructed according to its acoustic consequences.

Any hypothesized sequence of phonological units thus induces a succession of target distributions on the articulatory space, which are sampled randomly, as each new symbol appears, to construct a control trajectory for the articulators which lasts until the occurrence of a new symbol. It is possible that target distributions may need to be modified dynamically to account for external feedback, but the basic assumption is that control is open-loop over short periods of time, with adjustments occurring only at phonological boundaries.

The concept of a spatial target is originally due to MacNeilage [45], and reappears in various recent proposals; Keating [38] and Guenther [32] assume that phonemes correspond to underspecified target regions in a planning space; Perkell [52, 53] has suggested target regions corresponding to abstract "goals" which may be defined in articulatory, acoustic, or oro-sensory terms; Shirai and Honda [67] and Coker [13] define targets in terms of phoneme-specific articulatory target points; Honda and Kaburagi [34], and Bailly et al. [1, 2] suggest targets specified by prototype trajectories passing through constrained spatio-temporal regions; Saltzman and Munhall [64] assume a "task-dynamic" formulation for the control of a dynamical system, where targets are represented implicitly by attractors in an underlying task space; Laboissière et al. [42] and Perrier et al. [55] view targets as equilibrium-point trajectories. Experimental and modeling studies by Perkell [54], Maeda [46], Boë et al. [8] among others have shown that trade-offs due to various compensatory effects are indeed reflected in the shape of distributions of articulatory and acoustic measurements, so this appears to be a useful functional representation to adopt.

At present, it is difficult to speculate how the conversion of higher-level control signals into articulator movement takes place. A popular assumption, adopted here, which has partly been justified by experimental data [35], is that the combined (non-linear) control system and articulatory mechanism behave macroscopically as a linear system tracking the control input in the articulatory parameter space. Articulator motion then occurs as the response of the vocal tract model to a random control trajectory, and results in a time-varying tract shape modulating the acoustic properties of the speech signal.

The mapping between articulatory parameters and measurements can be simulated using an appropriate acoustic or mechanical model to generate the observed measurements. Any observation space may be used, as long as a model is available to represent the relationship between articulatory model parameters and the measurements to be exploited during recognition.

In mathematical terms, the phonological sequence is modeled by a semi-Markov chain (S,T), where $S = \{S_m : m \in \mathbb{N}\}$ is a Markov process with transition matrix Π and initial distribution π_0 taking values in a finite set of symbols $\mathcal{S} = \{s_i : i = 1 \ldots N\}$, and $T = \{T_m : m \in \mathbb{N}\}$ is a Poisson process describing state durations, distributed according to the Markov state S_m.

A Markov-modulated point process $U = \{U_m : m \in \mathbb{N}\}$ representing piecewise-constant target trajectories takes values in an articulatory space $\mathcal{X} = \mathbb{R}^m$, where the U_m are independent conditioned on S, and each target point U_m is drawn from one of a number of distributions, determined by the current Markov state S_m. For convenience, Gaussian mixtures can be used to approximate arbitrary continuous target distributions.

The articulatory process $X = \{X_n : n \in \mathbb{N}\}$ is assumed to admit a linear state-space representation driven by the marked point process (S, T, U) according to:

$$X_{n+1} = \sum_{j=1}^{d-1} A_j(S_{J(n)}) X_{n+1-j} + A_d(S_{J(n)}) U_{J(k)} + V_n, \tag{3}$$

$$Y_n = h(X_n) + W_n, \tag{4}$$

where $V = \{V_n : n \in \mathbb{N}\}$ and $W = \{W_n : n \in \mathbb{N}\}$ are zero-mean Gaussian white noise processes representing modeling error, and J is an appropriate index function. The system matrices $\mathcal{A} = \{A_j \in \mathbb{R}^{p \times p} : j = 1 \ldots d\}$ are selected by the Markov state S, and constrained so that the system relaxes asymptotically towards the current target input.

Any observation process $Y = \{Y_n : n \in \mathbb{N}\}$ evolving in a measurement space $\mathcal{Y} = \mathbb{R}^{q \times q}$ is generated from the articulatory trajectories by a static non-linear mapping $h : \mathcal{X} \to \mathcal{Y}$, which can be constructed from a codebook of sample points derived from model simulations. An explicit inverse model is not required, and in a stochastic framework we do not need to use complex gradient-descent-based optimization techniques to recover the underlying dynamics.

The important feature of this framework is that it consists of a general stochastic state-space representation for the relationship between observed and unobserved signals, which is constrained by a production model to mimic certain aspects of speech. Although the model is certainly very crude and over-simplistic, it at least provides a starting point for investigating the possibility of articulatory speech recognition.

In our work to date, we have developed algorithms for state and parameter estimation, using Kalman filtering and the EM algorithm, that can be applied to yield recognition and training techniques capable of recovering phonological state sequences, control trajectories, and articulatory trajectories from acoustic data alone. The model can also be used for random articulatory synthesis through Monte-Carlo simulation; unlike any other synthesis technique, the results incorporate a degree of random but systematic variability, even when the same input commands are used, reflecting what is commonly observed in repeated speech production experiments. Descriptions of this preliminary work can be found in [57, 58, 59, 60]. Evaluation of the model on a realistic recognition task is underway, and will be reported in a future publication.

7. Summary and conclusions

This paper is intended to address issues concerning the need for integrating computational speech production models into automatic speech recognition. We began by providing an introduction to the "fundamental equation" of speech recognition which epitomizes the probabilistic analysis-by-synthesis framework underpinning much of modern speech recognition research and development. This framework essentially treats recognition as a process of stochastically matching or searching the output (as random variables) of the modeled speech generator or production system – however crude an approximation it may be – with observable speech acoustics. The theory of HMMs as the mathematical backbone behind the current speech recognition technology was then critically reviewed, where we contended that HMMs can be viewed as a primitive speech generator or "production" model consistent with the probabilistic analysis-by-synthesis framework. We pointed out, however, that while use of HMMs in the probabilistic framework accounts for much of the (limited) success of the current recognition technology, the conventional phoneme-sized HMM viewed as a primitive speech generator suffers from strong theoretical weaknesses both from the *phonological* modeling and from the *interface* modeling standpoints. After a brief review of a variety of approaches to speech recognition based either implicitly or explicitly on some phonological and interface models beyond the conventional phoneme-sized HMMs, we presented two more elaborate phonological and interface models, both developed in our laboratory at Waterloo over the past few years.

In conclusion, we suggest from our limited research experience that integration of high-quality global speech production models into the probabilistic analysis-by-synthesis strategy is a fruitful path towards ultimate success of human-like speech recognition. This integration must call for close interaction and collaboration between speech production and speech recognition communities, which we believe have been long overdue.

Acknowledgements

Valuable discussions with Ken Stevens, Joe Perkell, Victor Zue, Cathy Browman, L. Goldstein, E. Saltzman, R. McGowan, and with Chin Lee on issues of speech production modeling and on several views presented in this paper are gratefully acknowledged.

Seg	Lips (L)	T.Bld (B)	T.Dors (D)	Velm (V)	Larynx (X)
b	Lα	U	U	Vα	Xα
d	U	Bα	U	Vα	Xα
g	U	U	Dα	Vα	Xα
p	Lα	U	U	Vα	Xβ
t	U	Bα	U	Vα	Xβ
k	U	U	Dα	Vα	Xβ
s	U	Bβ	U	Vα	Xβ
f	Lβ	U	U	Vα	Xβ
v	Lβ	U	U	Vα	Xα
m	Lα	U	U	Vβ	Xα
n	U	Bα	U	Vβ	Xα
ng	U	U	Dα	Vβ	Xα
r	Lγ	Bγ	Dβ	Vα	Xα
iy	U	U	Dγ	Vα	Xα
aa	U	U	Dδ	Vα	Xα
ao	Lδ	U	Dϵ	Vα	Xα

Table 1: Five-tuple (L B D V X) articulatory feature specification for some common segments in American English.

References

[1] Bailly G., Laboissière R., and Schwartz J.L., "Formant trajectories as audible gestures: an alternative for speech synthesis," *J. Phonetics*, 19(1), 9-23.

[2] Bailly G. (1995), "Recovering place of articulation for occlusives in VCV's," *Proc. XIII'th ICPhS* vol. 2, 230-233.

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

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

[5] Baum L. (1972), "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, 3, 1-8.

[6] Bitar N., and Espy-Wilson C. (1995) "Speech parameterization based on phonetic features: Application to speech recognition," *Proc. Eurospeech*, vol.2, 1411-1414.

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

[8] Boë L-J., Perrier P., and Bailly G. (1992) "The geometric vocal tract variables controlled for vowel production : proposals for constraining acoustic-to-articulatory inversion," *J. Phonetics* 20, 27-38.

[9] Bourlard H. (1995), "Towards increasing speech recognition error rates," *Proc. Eurospeech*, vol. 2, 883-894.

[10] Browman C. and Goldstein L. (1992), "Articulatory phonology: An overview," *Phonetica*, 49, 155-180.

[11] Chow, W., Juang B. H., and Lee C. H. (1996), "A minimum error rate pattern recognition approach to speech recognition," *Int. J. Pattern Recog. Artificial Intel.,* vol. 8, pp. 5-31.

[12] Chow, W., Juang B. H., and Lee C. H. (1996) "Statistical and discriminative methods for speech recognition," in *Automatic Speech and Speaker Recognition – Advanced Topics,* C. Lee, F. Soong, and K. Paliwal (eds.), Kluwer Academic Publishers, pp. 109-132.

[13] Coker, C. (1976) "A model of articulatory dynamics and control," *Proc. IEEE*, 64, 452-460.

[14] Deng, L. (1993) "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(4) Pt.2, pp. 2318.

[15] 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.

[16] Deng L., Aksmanovic M., Sun D., and Wu J. (1994) "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states," *IEEE Trans. Speech Audio Proc.*, 2(4), 507-520.

[17] Deng L., and Braam D. (1994), "Context-dependent Markov model structured by locus equations: Application to phonetic classification," *J. Acoust. Soc. Am.*, 96, 2008–2025.

[18] Deng L., and Sameti H. (1994) "Speech recognition using dynamically defined speech units," *Proc. ICSLP*, vol. 4, 2167-2170.

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

[20] Deng L., Wu J., and Sameti H. (1995) "Improved speech modeling and recognition using multi-dimensional articulatory states as primitive speech units," *Proc. ICASSP*, vol. 1, 385-388.

[21] 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.

[22] Deng L., Lennig M., and Mermelstein P. (1990) "Modeling microsegments of stop consonants in a hidden Markov model based word recognizer," *J. Acoust. Soc. Am.*, 87, 2738-2747.

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

[24] Deng L., Hasanein K., and Elmasry M. (1994) "Analysis of correlation structure for a neural predictive model with application to speech recognition," *Neural Networks*, 7(2), 1994, pp. 331–339.

[25] Deng, L. and Rathinavalu, C. (1995) "A Markov model containing state-conditioned second-order nonstationarity: Application to speech recognition," *Computer Speech and Language*, Vol. 9, No. 1, pp. 63-86.

[26] 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 and Audio Proc.*, 1, 431-442.

[27] Eide E. (1993) "A linguistic feature representation of the speech waveform," *Proc. ICASSP*, vol. 2, 483-486.

[28] Erler K. and L. Deng (1993) "Hidden Markov model representation of quantized articulatory features for speech recognition," *Computer Speech and Language*, vol. 7(3), pp. 265-282.

[29] Furui S. (1995) "Flexible speech recognition," *Proc. Eurospeech*, vol. 3, 1595-1603.

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

[31] Goldsmith J.A. (1990) *Autosegmental and Metrical Phonology*, Blackwell: Oxford.

[32] Guenther F.H. (1995) "A modeling framework for speech motor development and kinematic articulator control," *Proc. XIII'th ICPhS*, vol. 2, 92-99.

[33] Holmes, W. and Russell, M. (1996) "Modeling speech variability with segmental HMMs," *Proc. Eurospeech*, vol. 1, pp. 447-450.

[34] Honda M., and Kaburagi T. (1995) "A dynamical articulatory model using potential task representation," *Proc. ICSLP*, vol. 1, 179-182.

[35] Houde R.A. (1968) *SCRL Monograph No. 2: A study of tongue body motion during selected speech sounds*, Santa Barbara, California.

[36] Hwang M., Huang X., and Alleva F. (1993) "Predicting unseen triphones with senones," *Proc. ICASSP*, vol.1, 311-315.

[37] Kaburagi, T. and Honda, M. (1996) "A model of articulator trajectory formation based on the motor tasks of vocal-tract shapes," *J. Acoust. Soc. Am.*, vol. 99(5), pp. 3154-3170.

[38] Keating P. (1990) "The window model of coarticulation : articulatory evidence," in *Papers in Laboratory Phonology I*, J. Kingston and M. Beckman (eds), Cambridge University Press: Cambridge.

[39] Kelso, J., Saltzman, E., and Tuller, B. (1986) "The dynamical perspectives on speech production: data and theory," *J. Phonetics*, vol. 14, pp. 29-59.

[40] Kenny P., et al. (1991) "Articulatory Markov models," *Proc. IEEE Workshop Automatic Speech Recognition*, Harriman, New York, 22-23.

[41] Kent R., et al. (1995) "Models of speech production," in *Principles of Experimental Phonetics*, Ed. N. Lass, Mosby: London, 3-45.

[42] Laboissière R., Ostry D.J., and Perrier P. (1995) "A model of human jaw and hyoid motion and its implications for speech production," *Proc. XIII'th ICPhS*, vol. 2, 60-67.

[43] Levinson S., Rabiner L., and Sondhi M. (1983) "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell System Technical J.*, 62, 1035-1074.

[44] Liu S. (1995) *Landmark detection for distinctive feature-based speech recognition,* Ph.D. thesis, MIT.

[45] MacNeilage P. (1970) "Motor control of serial ordering in speech," *Psychological Review*, 77, 182-196.

[46] Maeda S. (1991) "On articulatory and acoustic variabilities," *J. Phonetics* 19, 321-331.

[47] Mahkoul J., and Schwartz R. (1995) "State of the art in continuous speech recognition," *Proc. Natl. Acad. Sci. USA*, 92, 9956-9963.

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

[49] Meng H. and Zue V. (1991) "Signal representation comparison for phonetic classification," *Proc. ICASSP*, vol. 1, 285-288.

[50] Morgan N. and Bourlard H. (1994) *Connectionist speech recognition - A hybrid approach*, Kluwer Academic Publishers.

[51] 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.

[52] Perkell J.S. (1980) "Phonetic features and the physiology of speech production," in *Language Production*, B. Butterworth (ed.), Academic Press: London.

[53] Perkell J.S., Matthies M.L., Svirsky M.A., and Jordan M.I. (1995) "Goal-based speech motor control: a theoretical framework and some preliminary data," *J. Phonetics*, 23, 23-35.

[54] Perkell J.S. (1996) "Properties of the tongue help to define vowel categories: hypotheses based on physiologically-oriented modeling," *J. Phonetics* 24, 3-22.

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

[56] Rabiner L., Juang B. H., and Lee C. H. (1996) "An overview of automatic speech recognition," in *Automatic Speech and Speaker Recognition – Advanced Topics,* C. Lee, F. Soong, and K. Paliwal (eds.), Kluwer Academic Publishers, pp. 1-30.

[57] Ramsay G. and Deng L. (1994) "A stochastic framework for articulatory speech recognition ," *J. Acoust. Soc. Am.* 95 (5) Pt.2 Abstract 2aSP19.

[58] Ramsay G. and Deng L. (1995) "Articulatory synthesis using a stochastic target model of speech production," *Proc. XIII'th ICPhS* vol. 2, 338-341.

[59] Ramsay G. and Deng L. (1995) "Maximum Likelihood Estimation for Articulatory Speech Recognition Using a Stochastic Target Model," *Proc. Eurospeech*, vol. 2, pp. 1401-1404.

[60] Ramsay G. and Deng L. (1996) "Optimal filtering and smoothing for speech recognition using a stochastic target model," *Proc. ICSLP-96* (to appear in October).

[61] Randolph M. (1994) "Speech analysis based on articulatory behavior," *J. Acoust. Soc. Am.*, 95 (5) Pt. 2, 1aSP15 (abstract).

[62] Rose, R., Schroeter, J., and Sondhi, M. (1996) "The potential role of speech production models in automatic speech recognition," *J. Acoust. Soc. Am.*, 99(3), pp. 1699-1709.

[63] Rubin, P., Saltzman, E., Goldstein, L., McGowan, R., M. Tiede, and Browman, C. (1996) "CASY and extensions to the task-dynamic model," *Proc. 4th European Speech Production Workshop*, Autrans, France, pp. 125-128.

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

[65] Schroeter J. and Sondhi M.M. (1994) "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Trans. Speech Audio Proc.*, 2(1), 133-150.

[66] Sheikhzadeh H. and Deng L. (1994) "Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization," *IEEE Trans. Speech Audio Proc.*, 2, 80-91.

[67] Shirai K. and Honda M. (1976) "Estimation of articulatory motion," in *Dynamic Aspects of Speech Production*, University of Tokyo Press: Tokyo.

[68] Stevens K., et al. (1992) "Implementation of a model for lexical access based on features," *Proc. ICSLP*, vol. 1, 499-502.

[69] Tokuda K., et al. (1995) "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," *Proc. Eurospeech*, vol.1, 757-760.

[70] Varga A. and Moore R. (1990) "HMM decomposition of speech and noise," *Proc. ICASSP*, 845-848.

[71] Wilhelms R. (1987) "Schätzung von artikulatorischen Bewegungen eines stylisierten Artikulatormodells aus dem Sprachsignal," *Ph.D. Dissertation*, Georg-August Univerität, Göttingen.

[72] Young S. (1995) "Large vocabulary continuous speech recognition: a review," *Proc. IEEE Workshop on Automatic Speech Recognition*, Snowbird, Utah, 3-28.