ELSEVIER

Speech Communication 22 (1997) 93–111

SPEECH COMMUNICATION

Production models as a structural basis for automatic speech recognition

L. Deng [a,*], G. Ramsay [a], D. Sun [b]

[a] *Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada*
[b] *Bell Laboratories, Murray Hill, NJ, USA*

Received 4 October 1996; revised 17 March 1997; accepted 28 April 1997

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 present two probabilistic speech recognition models with the structure designed based on approximations to human speech production mechanisms, and conclude by suggesting that many of the advantages to be gained from interaction between speech production and speech recognition communities will develop from integrating production models with the probabilistic analysis-by-synthesis strategy currently used by the technology community. © 1997 Elsevier Science B.V.

Résumé

Dans cet article, nous suggérons que des modèles de production de la parole fortement structurés pourront contribuer significativement à la réussite future des modèles de reconnaissance automatique de la parole, 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 des modèles phonétiques, et présentons deux modèles statistiques de reconnaissance de la parole basés sur des approximations des mécanismes de production de la parole. Nous suggérons en conclusion que l'interaction entre les domaines de la production et de la reconnaissance de la parole peut être particulièrement efficace si l'on intègre les modèles de production dans la stratégie d'analyse–synthèse probabiliste, utilisée déjà depuis longtemps en reconnaissance de la parole. © 1997 Elsevier Science B.V.

Keywords: Speech production; Speech recognition; Analysis by synthesis; Stochastic modeling; Nonlinear phonology; Phonetic interface; Articulatory features; Articulatory dynamics; Stochastic target model

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 or HMMs) have dominated such efforts (e.g., [82,55]). On the other hand, work from speech production researchers has centered on elaborating detailed models and

* Corresponding author. E-mail: deng@crg5.uwaterloo.ca or ldeng@itl.atr.co.jp.

PII S0167-6393(97)00018-6

theories intended to account for the nature of the transformation from discrete phonological symbols to continuous acoustic streams via motor control strategies and articulatory dynamics (e.g., [74,48,64]). Nowhere has this division been more evident than in the differing views of 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" [10] – 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 $\boldsymbol{O} = O_1, O_2, \ldots, O_T$ be a sequence of observable acoustic data of speech, which can either be speech waveforms [76], or continuous-valued acoustic vectors [82], or discrete-valued vector-quantized codes [55], or any other type of general acoustic measurements, and let $W = w_1, w_2, \ldots, w_n$ be the sequence of words intended by the speaker who produces the acoustic record $\boldsymbol{O}$ above. The goal of a speech recognizer is to "guess" the most likely word sequence $\hat{W}$ given the acoustic data $\boldsymbol{O}$. 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 $P(W|\boldsymbol{O})$:

$$\begin{aligned}\hat{W} &= \arg\max_W P(W|\boldsymbol{O}) \\ &= \arg\max_W \frac{P(\boldsymbol{O}|W)P(W)}{P(\boldsymbol{O})} \\ &= \arg\max_W P(\boldsymbol{O}|W)P(W), \end{aligned} \tag{1}$$

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 $P(\boldsymbol{O}|W)$ is the probability that the speaker produces (or the microphone of the speech recognizer receives) the acoustic data $\boldsymbol{O}$ 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 $P(\boldsymbol{O}|W)$? [1] and (2) a search for the word sequence W (the operation $\arg\max_W$ in Eq. (1)) that provides the optimal value of the posterior probability. Note that in Eq. (1) $P(\boldsymbol{O})$ is constant with respect to maximization over W and hence can be dropped during the recognition step. [2]

Eq. (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 $\boldsymbol{O}$ are produced from word sequence W (hence the necessity and possibility to evaluate the production probability $P(\boldsymbol{O}|W)$), and the analysis phase involves finding the solution $\hat{W}$ which best matches (in a maximum-likelihood sense) the outcome of the production. From this analysis-by-synthesis interpretation of Eq. (1), we will see that good speech production theories, when meeting the computational requirements implied in Eq. (1), should have much to contribute to advancing speech recognition technology. [3]

3. Critical review of HMMs: lessons learned from technologists

There is no doubt that Hidden Markov Model (HMM) is currently the most successful technology in many (heavily) constrained speech recognition applications; see [65,82] 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 Eq. (1). Implicit in Eq. (1) are the need to efficiently compute a production probability $P(\boldsymbol{O}|W)$ and the need to learn "production model" parameters so as to achieve high accuracy in evaluating $P(\boldsymbol{O}|W)$. HMMs are amenable to efficient computation and parameter learning thanks to Baum's work [6], and thus would fit naturally into the probabilistic analysis-by-synthesis framework of Eq. (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 $\boldsymbol{O}$. One simple way to view an HMM as a speech production or synthesis device is to run Monte Carlo simulation on the HMM and regard the outcome of the simulation as the synthetic speech (cf. [67]). Filter-based approaches have also been proposed for HMM speech synthesis [79].

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 $P(\boldsymbol{O}|W)$ which appeared in Eq. (1) into two factors:

$$P(\boldsymbol{O}|W) = \sum_{\mathscr{P}} P(\boldsymbol{O}|\mathscr{P})\,P(\mathscr{P}|W) \approx \max_{\mathscr{P}} P(\boldsymbol{O}|\mathscr{P})\,P(\mathscr{P}|W), \tag{2}$$

where $\mathscr{P}$ is a discrete-valued *phonological model* and specifies, according to probability $P(\mathscr{P}|W)$, how words and word sequences W can be expressed in terms of a particular organization of a small set of "atomic" phonological units; $P(\boldsymbol{O}|\mathscr{P})$ is the probability that a particular organization $\mathscr{P}$ of phonological units produces the acoustic data for the given 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 Eq. (2), state-of-the-art speech recognizers [55,82] based on phonetic HMMs can be analyzed as follows. The phonological model $\mathscr{P}$ is essentially a linearly-organized multiple-state phonetic sequence governed by a left-

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

[2] During the parameter estimation or training step, however, $P(\boldsymbol{O})$ can be an important quantity, especially in using discrimination motivated models [10,13].

[3] 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 [80,35].

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:

$$P(\mathscr{P}|W) = \prod_{t=0}^{T-1} P(s_{t+1}|s_t, W);$$

$$P(\boldsymbol{O}|\mathscr{P}) = \prod_{t=1}^{T} b_{s_t}(\boldsymbol{O}_t), \tag{3}$$

where $P(s_{t+1}|s_t)$ and $b_{s_t}(\boldsymbol{O}_t)$ 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 [11,49,14], 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 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 or 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 (Eq. (1)) and to the use of automatic learning which is an implicit component of the framework. 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 switch-board database demonstrate poor performance of state-of-the-art technology based on conventional HMMs [15,33]. The superior performance achieved by discrimination-motivated HMM parameter learning over the maximum-likelihood learning [12,13,35] further attests to the poor quality of conventional HMMs as a generative model for use in speech recognition. This is so because if the generative model describes true statistics of the model-generated data, then given a sufficient amount of training data, the consistency property of the maximum-likelihood estimate would give true model parameters. This would guarantee minimal recognition rates according to the decision rule specified by Eq. (1) and thus would eliminate the need for discriminative training.

In pursuing the development of high-quality global speech production models that 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 pa-

rameters from any available training data. [4] 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 $P(O|W)$ in terms of the phonological and interface models (Eq. (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. [5]

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 (cf. [55]) and of sub-phonemic units constructed systematically by training from acoustic data (e.g., [43]) has achieved promising results, there are reasons to believe that more challenging, unconstrained tasks with performance approaching human capability will require sub-phonemic models grounded solidly on modern phonological theories. The reasons are based partly on our understanding of the nature of the atomic units for lexical representation in our own human brains [14,30], and partly on our understanding of the fundamental limitations of the current speech recognition technology arising from use of heuristic phonological units [33].

Some of the sub-phonemic phonological models used with varying degrees of success in speech recognition include the microsegment model [25], the locus model focusing on CV and VC transitions [20], and that which directly takes Chomsky–Halle binary distinctive features as the recognition object [31,58]. A phonological model based on an articulation-based feature-geometric theory is reported in [7,52,78] 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 [4,5,8], 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) [70,26,32,47]. 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 [22,24] treats each articulatory feature as a symbolic entity (i.e. with no partial ordering) embodying phonological contrasts together with acoustic and possibly auditory correlates. [6] The overlapping nature of the articulatory features in this model is directly motivated by representations in autosegmental phonology [37] and by the way gestural scores are constructed in articulatory phonology [11]. Some details of this overlapping feature based model will be given in Section 5.

Unlike the phonological models which mostly originated from linguistic research, most of the inter-

[4] 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 the current phonetic-HMM based speech recognition technology, however, random and systematic variabilities are largely undistinguished and a combination of them leads to the demand of a large amount of training data.

[5] A review of potential roles of speech production modeling in speech recognition from a somewhat different viewpoint recently appeared in [72,57].

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

face 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 and acoustic observations of speech. [7] 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 [52,78,70] 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 [59,58,27]. Further, as we discussed in Section 3, the HMM-style interface model is the most common interface model in the current speech recognition technology. In addition to the conventional, stationary-state HMM, [8] various versions of (segment-based) nonstationary-state or trended HMMs [18,36,19,28,40] and several other types of the segment-based models [60] including the linear dynamical system model [29] have more recently been developed as theoretically superior interface models from phonological units to acoustics. 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 [5] where FIR filters are employed to functionally interface phoneme-specific targets (phonological entity) with motions of "abstract" or pseudo-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 phonological units such as articulatory features and acoustics. We will describe this latter model with some detail in Section 6.

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 [46,74,56,73,44]. The task-dynamic model differs conceptually from the explicit articulatory-dynamic model in that the dynamics 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 [17]. A non-statistical version of the model has been described in [56]. The main difficulty we have found in 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 [51]:

[...] 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

[7] 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.

[8] The recognizers described in [82,7,22] employed essentially an identical interface model (i.e., the stationary-state HMM) but with different phonological models.

number of articulatory configurations or (HMM) states. [...].

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 appears 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 [82].

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 for articulatory measurement data. The recognition process involves 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 [9] 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 [22] into a Markov state transition graph, which is summaried below.

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

Seg.	Lips (L)	T. Bld. (B)	T. Dors. (D)	Velm (V)	Larynx (X)
b	$L\alpha$	U	U	$V\alpha$	$X\alpha$
d	U	$B\alpha$	U	$V\alpha$	$X\alpha$
g	U	U	$D\alpha$	$V\alpha$	$X\alpha$
p	$L\alpha$	U	U	$V\alpha$	$X\beta$
t	U	$B\alpha$	U	$V\alpha$	$X\beta$
k	U	U	$D\alpha$	$V\alpha$	$X\beta$
s	U	$B\beta$	U	$V\alpha$	$X\beta$
f	$L\beta$	U	U	$V\alpha$	$X\beta$
v	$L\beta$	U	U	$V\alpha$	$X\alpha$
m	$L\alpha$	U	U	$V\beta$	$X\alpha$
n	U	$B\alpha$	U	$V\beta$	$X\alpha$
ng	U	U	$D\alpha$	$V\beta$	$X\alpha$
r	$L\gamma$	$B\gamma$	$D\beta$	$V\alpha$	$X\alpha$
iy	U	U	$D\gamma$	$V\alpha$	$X\alpha$
aa	U	U	$D\delta$	$V\alpha$	$X\alpha$
ao	$L\delta$	U	$D\epsilon$	$V\alpha$	$X\alpha$
ih	U	U	$D\zeta$	$V\alpha$	$X\alpha$
eh	U	U	$D\eta$	$V\alpha$	$X\alpha$
ae	U	U	$D\theta$	$V\alpha$	$X\alpha$

5.1. 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 [10] is shown in Table 1, where symbolic feature value 'U' denotes feature underspecification. The precise meaning of underspecification used here is somewhat different from that used in the phonology and phonetics literature. Underspecification used here denotes the fact that in actual utterances the features

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

[10] Throughout this paper, we will use TIMIT labels as computerized alphabet, instead of IPA symbols, to denote phonetic units. Use of the TIMIT labels has been wide spread among speech recognition researchers, and the correspondence between the TIMIT labels to IPA symbols can be found in [71].

implemented will be determined by the specified features in the segments adjacent to the current segment. This strategy is common among all five feature dimensions and hence we omit the feature label (L, B or D) in writing the underspecified features. Some of the choices made in Table 1 are briefly described here. All vowels are specified in the Tongue Dorsum (D) dimension. Vowels with different degrees of lip rounding are also specified with different Lips (L) features. Lip spreading in vowels (e.g., vowel /iy/) is not specified in the L (Lips) feature dimension because its effects on acoustics are negligible (compared with lip rounding). Articulatory correlates of some of the features in Table 1 are as follows. Feature $V\alpha$ is correlated with the velum-closing gesture implemented in all oral sounds, and feature $V\beta$ with the velum-opening gesture. Feature $X\beta$ is correlated with the wide glottal opening or aspiration gesture, and feature $X\alpha$ to the glottal-closing gesture responsible for voicing. Feature $L\alpha$ is correlated with the bilabial closing gesture, and feature $L\beta$ with the dental-labial constriction gesture, etc.

To help understand the procedure we developed to automatically construct the articulatory HMM states, we provide a step-by-step explanation below using the example phrase /t eh n k ae t s/ (*ten cats*).

5.2. Notations

Some notations are introduced here to describe the procedure for the articulatory state construction. Let

$$\boldsymbol{\Phi} = (\phi_1, \ldots, \phi_m)$$

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. For example, $\phi_1 = /t/$ in the example phrase. Let

$$f(\phi_i) = (f_1(\phi_i), \ldots, f_D(\phi_i))^{\mathrm{T}}$$

be the vector of articulatory features of target segment ϕ_i in the phonetic transcription (which can be found in Table 1). For example, $f(/t/) = (U, B\alpha, U, V\alpha, X\beta)^{\mathrm{T}}$. (The dimensionality of the vector $D = 5$.) Similarly, let

$$g(\phi_i|\phi_{i+\delta})$$

be the vector of contextual articulatory features of target segment ϕ_i assimilated by the features of segment $\phi_{i+\delta}$, where δ takes an integer value ($\delta > 0$ for anticipatory coarticulation and $\delta < 0$ for carry-over coarticulation). Obviously, $g(\phi_i|\phi_i) \triangleq f(\phi_i)$ when $\delta = 0$. In general, the value of $g(\phi_i|\phi_{i+\delta})$ is determined by $f(\phi_i)$, ..., $f(\phi_{i+\delta})$, and by a set of rules controlling feature overlaps and spreads as described in Step 2 of the algorithm below.

5.3. Algorithm description

Input:

A phonetic transcription of a given utterance (word, phrase or sentence): $\boldsymbol{\Phi} = (\phi_1, \ldots, \phi_m)$. In our example, $\boldsymbol{\Phi} =$ (/t/, /eh/, /n/, /k/, /ae/, /t/, /s/).

Output:

An articulatory feature-based HMM state transition graph.

Algorithm:

(1) For each phonetic unit ϕ_i in the context of $\boldsymbol{\Phi}$, find its feature specification $f(\phi_i)$ from Table 1. For example, taking $i = 2$ in the phrase /t eh n k ae t s/, we have

$$f(\phi_2) = f(/\mathrm{eh}/) = (U, U, D\eta, V\alpha, X\alpha)^{\mathrm{T}}.$$

(2) Specify features for contextual segments, $g(\phi_i|\phi_{i+\delta})$. The purpose of this step is to apply phonological rules to the sequence of feature vectors. Initially, set

$$g(\phi_i|\phi_{i+\delta}) \triangleq f(\phi_{i+\delta}).$$

In the example phrase,

$$g(\phi_2|\phi_1) = g(/\mathrm{eh}/|/\mathrm{t}/) = (U, B\alpha, U, V\alpha, X\beta)^{\mathrm{T}},$$

$$g(\phi_2|\phi_3) = g(/\mathrm{eh}/|/\mathrm{n}/) = (U, B\alpha, U, V\beta, X\alpha)^{\mathrm{T}}.$$

Then the value of $g(\phi_i|\phi_{i+\delta})$ is modified by a set of constraints on feature overlaps and spreads. For simplicity, we only describe here the basic rules for illustration purposes:

(a) Set $g_d(\phi_i|\phi_{i+\delta}) =$ 'U' for $d = 1, \ldots, D$, and $|\delta| > \Delta$, where Δ is a constant indicating the maxi-

mal amount of feature spread. This rule specifies the constraint on the maximum span of feature spreading. In the case of $\Delta = 1$, we assume the influence on the target segment is only from its immediate left and right context. (By allowing $\Delta > 1$, we can increase the flexibility of modeling long range contextual influences.)

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

(c) If $g_d(\phi_i|\phi_{i+\delta}) = f_d(\phi_i)$, set $g_d(\phi_i|\phi_{i+\delta})$ to 'U'. This rule removes influences from neighboring phonetic segments in some feature dimensions if they have the same feature values in these dimensions.

In our example phrase,

$$\boldsymbol{g}(\phi_2|\phi_1) = \boldsymbol{g}(/\text{eh}/|/\text{t}/) = (U, B\alpha, U, U, X\beta)^{\mathrm{T}},$$
$$\boldsymbol{g}(\phi_2|\phi_3) = \boldsymbol{g}(/\text{eh}/|/\text{n}/) = (U, B\alpha, U, V\beta, U)^{\mathrm{T}}.$$

Finally, the feature values are modified by appending a symbol of 'l' or 'r' to distinguish the left from the right contextual influence. In the example phrase,

$$\boldsymbol{g}(\phi_2|\phi_1) = \boldsymbol{g}(/\text{eh}/|/\text{t}/) = (U, lB\alpha, U, U, lX\beta)^{\mathrm{T}},$$
$$\boldsymbol{g}(\phi_2|\phi_3) = \boldsymbol{g}(/\text{eh}/|/\text{n}/) = (U, rB\alpha, U, rV\beta, U)^{\mathrm{T}}.$$

Based on $\boldsymbol{f}(\phi_i)$ and $\boldsymbol{g}(\phi_i|\phi_{i+\delta})$, we arrive at the following feature specification array, $F(\phi_i|\phi_{i-\Delta}, \ldots, \phi_{i+\Delta})$, for ϕ_i with contextual influences:

$g_1(\phi_i\|\phi_{i-\Delta})$	$\cdots$	$g_1(\phi_i\|\phi_{i-1})$	$f_1(\phi_i)$	$g_1(\phi_i\|\phi_{i+1})$	$\cdots$	$g_1(\phi_i\|\phi_{i+\Delta})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$g_D(\phi_i\|\phi_{i-\Delta})$	$\cdots$	$g_D(\phi_i\|\phi_{i-1})$	$f_D(\phi_i)$	$g_D(\phi_i\|\phi_{i+1})$	$\cdots$	$g_D(\phi_i\|\phi_{i+\Delta})$

In our example phrase /t eh n k ae t s/, ϕ_2 = /eh/ with $\Delta = 1$, and the contextual feature array is

$$F(/eh/|/t/, /eh/, /n/) = \begin{array}{|c|c|c|}\hline U & U & U \\ lB\alpha & U & rB\alpha \\ U & D\eta & U \\ U & V\alpha & rV\beta \\ lX\beta & X\alpha & U \\ \hline \end{array}$$

(3) Construct the feature-based states S. According to the result from Step 2 which gives the feature specifications for a phonetic segment in context, enumerate all distinct D-tuple feature bundles (states):

$$S \triangleq \left(g_1(\phi_i|\phi_{i+\delta_1}), \ldots, g_D(\phi_i|\phi_{i+\delta_D})\right)^{\mathrm{T}},$$

which satisfy the following two conditions: [11]

(a) $\delta_1, \ldots, \delta_D$ are either all positive or all negative;

(b) $g_d(\phi_i|\phi_{i+\delta_d}) \neq 'U'$ when $\delta_d \neq 0$.

In our example of /eh/ above, a total of seven distinct feature bundles or states are enumerated:

S_1: $(U, lB\alpha, D\eta, V\alpha, lX\beta)^{\mathrm{T}}$,
S_2: $(U, U, D\eta, V\alpha, lX\beta)^{\mathrm{T}}$,
S_3: $(U, lB\alpha, D\eta, V\alpha, X\alpha)^{\mathrm{T}}$,
S_4: $(U, U, D\eta, V\alpha, X\alpha)^{\mathrm{T}}$,
S_5: $(U, rB\alpha, D\eta, V\alpha, X\alpha)^{\mathrm{T}}$,
S_6: $(U, U, D\eta, rV\beta, X\alpha)^{\mathrm{T}}$,
S_7: $(U, rB\alpha, D\eta, rV\beta, X\alpha)^{\mathrm{T}}$.

(4) Determine the state transitions. We build a connection (or transition) from state (a),

$$S(a) = \left(g_1(\phi_i|\phi_{i+\delta_1(a)}), \ldots, g_D(\phi_i|\phi_{i+\delta_D(a)})\right)^{\mathrm{T}},$$

to state (b),

$$S(b) = \left(g_1(\phi_i|\phi_{i+\delta_1(b)}), \ldots, g_D(\phi_i|\phi_{i+\delta_D(b)})\right)^{\mathrm{T}},$$

which satisfy $\delta_d(a) \leqslant \delta_d(b)$, for $d = 1, \ldots, D$ with at least one inequality held strictly.

The state transitions of our example /eh/ is given in Fig. 1. (Self loops and transitions with skips larger than one are not shown in Fig. 1 for simplicity in the presentation; these loops and skips are incor-

[11] The condition (a) below says that the spreading features from left and from right at different dimensions do not overlap in the home segment. In the example of /t eh n/ above, the left-spread feature $lB\alpha$ into the home segment /eh/ should not overlap with the right-spread feature $rV\beta$, since the formant transition in /eh/ induced from the $lB\alpha$ feature spread is rarely simultaneously nasalized (which would be an acoustic consequence of the $rV\beta$ feature spread). The condition (b) below says that the underspecified features in any dimension in the contextual (non-home) segment(s) do not contribute to forming the HMM states. Both of these constraining conditions have substantially reduced the complexity of the articulatory HMM states as the algorithm's output.

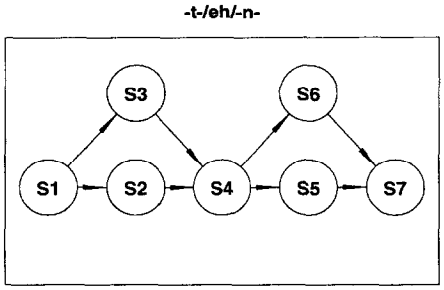


Fig. 1. State transition topology for allophone [eh] in context [t eh n].

porated in the actual implementation of the speech recognizer.)

(5) Form a complete state transition graph for the entire utterance. Once the state transition graph for each of the phonetic segments is constructed from above steps, these component graphs are connected from left to right to form the complete topology for the entire utterance.

The algorithm described above has been comprehensive in nature – capable of accepting any arbitrary string of phonetic labels to produce the corresponding articulatory HMM state topology. This output state topology reflects many flexible ways in articulation in implementing the articulation-based phonological features associated with the input phonetic string. The comprehensive nature of the algorithm may make it somewhat difficult to grasp at first. Therefore, in addition to the example phrase provided above in describing the algorithm, we use another example for input phonetic string /b iy m ih ng/, included in Appendix A, to illustrate some intermediate functions in the operation of the algorithm and the algorithm output.

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 [22,21,24,23].

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

The previous section described a largely symbolic, knowledge-based approach to the incorporation of speech production mechanism in the phonological component of a statistical speech recognizer. Information about articulatory constraints was embedded in an HMM structure, but an explicit model of the phonetic processes governing motion of the vocal tract and production of speech was not given. The relationship between phonology and phonetics was approximated instead by the output distributions on dynamically defined acoustic patterns associated with each phonological Markov chain state. The current section describes a second alternative model designed to account explicitly for the phonetic interface linking discrete-state phonological representations to their continuous-state articulatory and acoustic counterparts. (A third alternative, not described in this paper, would be to interface phonological representations to functionally defined task variables, rather than to articulatory variables.) The aim is to provide a consistent framework for incorporating current ideas about the phonetic processes governing articulatory control into a statistical framework for potential use in speech recognition and synthesis.

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. 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 are introduced into the acoustic signal due to differences in timing between adjacent velic and oral closures [1,34]. The word "princess" is often realized as [printses] or [prinses], depending on whether the velum is raised before or after the release of the alveolar stop. The continuous variation in synchronization observed in the articulatory data is partly due to deliberate control on the part of the speaker, and partly due to intrinsic biomechanical properties of the tongue, jaw and velum. The apparent categorical change observed in the acoustic domain is a result of the non-linear processes responsible for converting vocal tract movement into sound – formation of simultaneous oral and velic closures causes a build-up in pressure which, when released, appears as an intrusive stop. To represent stop epenthesis successfully, we therefore need to be able to describe the possible changes in vocal tract shape induced by small changes in timing between the two closures, as well as their acoustic consequences.

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 approaches generalizes correctly, since the explanation lies in the articulatory domain, and is neither purely acoustic nor purely 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. In view of the complexity of the physical processes involved in producing the stop, the only consistent way to account for this is to model the relevant aspects of articulatory motion directly.

As a first step towards incorporating an explicit production model into speech recognition, we have developed *a stochastic target model* for recognition and synthesis [66–69] which is capable of accounting for some of the systematic articulatory variation observed in speech, within a speech 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 [75] and will not be discussed further here. Examples of similar recognition models that also contain an intermediate articulatory representation appeared in [77,81,5,8].

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 of the kind discussed in previous sections 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 may 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 statistical 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. At present, control trajectories are assumed for simplicity to be piecewise-constant functions, representing essentially static spatial targets, but there is no reason why these should not be replaced in the future with more general parametric sample paths. The probability distribution of possible control trajectories on the articulatory space is intended to represent the statistical ensemble of idealized articulatory movements which would produce the required sequence of distributions of phonetic correlates for the phonological sequence in question. It is possible that the target distributions may need to be modified dynamically to account for proprioceptive or exteroceptive feedback, but the basic assumption is that control is open-loop over short periods of time, with adjustments in control strategy occurring only at phonological boundaries.

The concept of a spatial target is originally due to MacNeilage [53], and reappears in various recent proposals; Keating [45] and Guenther [38] assume that phonemes correspond to underspecified target regions in a planning space; Perkell [61,62] has suggested target regions corresponding to abstract "goals" which may be defined in articulatory, acoustic or oro-sensory terms; Shirai and Honda [77] and Coker [16] define targets in terms of phoneme-specific articulatory target points; Honda and Kaburagi [41], Bailly et al. [2,3] and Hogden [39] suggest targets specified by prototype trajectories passing through constrained spatio-temporal regions; Saltzman and Munhall [74] 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. [50] and Perrier et al. [64] view targets as equilibrium-point trajectories. Experimental and modeling studies by Perkell [63], Maeda [54], Boë et al. [9] 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 trajectories into articulator movement takes place. Ideally, modeling of articulatory dynamics and control would require detailed neuromuscular and biomechanical models of the vocal tract, as well as an explicit model of the control objectives and strategies realized by a speaker's motor control system. This is clearly too complicated to implement at present; a popular assumption, adopted here, is that the combined (nonlinear) control system and articulatory mechanism behave macroscopically as a stable linear system that attempts to track the control input as closely as possible in the articulatory parameter space. Articulator motion can then be approximated as the response of a simple linear vocal tract model driven by a random control trajectory, producing a time-varying tract shape which modulates the acoustic properties of the speech signal. The use of a linear dynamical system driven by a control signal to approximate articulatory trajectories has partly been justified by data fitting experiments conducted on experimental data, as described for example by Houde [42].

The final stage involves accounting for the generation of the observed measurements, and the transformation between articulatory parameters and observations must be simulated using an appropriate acoustic or mechanical model. 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. Depending on the choice of rep-

resentation and time scale, it may be possible to approximate this using a static mapping, otherwise a full dynamic model will be needed.

We now have an empirical account of the major aspects of the processes responsible for converting an abstract phonological description of an utterance into continuous-state articulatory and acoustic trajectories. We need to describe how the model is implemented, and how it can be used in automatic speech recognition.

In mathematical terms, the phonological sequence can be 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 $\mathscr{S} = \{s_i : i = 1 \ldots N\}$, and $T = \{T_m : m \in \mathbb{N}\}$ is a random process describing state durations, where T_m is distributed according to the Markov state S_m. Poisson distributions are chosen for convenience.

A Markov-modulated point process $U = \{U_m : m \in \mathbb{N}\}$ representing piecewise-constant target trajectories takes values in an articulatory space $\mathscr{X} = \mathbb{R}^p$, where the U_m are independent conditioned on (S,T), 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 sample paths of the control signal U represent idealized spatial target trajectories for the articulators.

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

$$X_{n+1} = \sum_{j=1}^{d-1} A_j\big(S_{J(n)}\big) X_{n+1-j} + A_d\big(S_{J(n)}\big) U_{J(k)} + V_n, \tag{4}$$

where $V = \{V_n : n \in \mathbb{N}\}$ is a zero-mean Gaussian white-noise process representing modeling error, and J is an appropriate index function indicating which phonological state is active at any point in time. The system matrices $\mathscr{A} = A_j \in \mathbb{R}^{p \times p} : j = 1 \ldots d\}$ determining articulatory dynamics are selected by the phonological Markov state S, and constrained so that the system relaxes asymptotically towards the current target input U. In practice, any linear or non-linear state–space model could be used instead, provided that this is chosen so that the articulatory state X moves towards the target U; the example given in Eq. (4) is simply the most convenient approximation to use.

Any observation process $Y = \{Y_n : n \in \mathbb{N}\}$ evolving in a measurement space $\mathscr{Y} = \mathbb{R}^8$ is assumed to be generated from the articulatory trajectories by a static non-linear mapping $h : \mathscr{X} \to \mathscr{Y}$, according to Eq. (5), where $W = \{W_n : n \in \mathbb{N}\}$ is a zero-mean Gaussian white-noise process representing measurement error,

$$Y_n = h(X_n) + W_n. \tag{5}$$

The mapping $h(\cdot)$ can be constructed from a codebook of sample points derived from model simulations, or perhaps derived from a corpus of real measurement data.

This completes the formal description of the model structure; an illustration of the mechanism for converting a phonological symbol sequence into articulatory and acoustic representations is given in Fig. 2, showing typical sample paths of the S, T, U, X and Y processes. The important characteristic 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 at present, 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 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. 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 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 [66–69]. Evaluation of the model on a realistic recognition

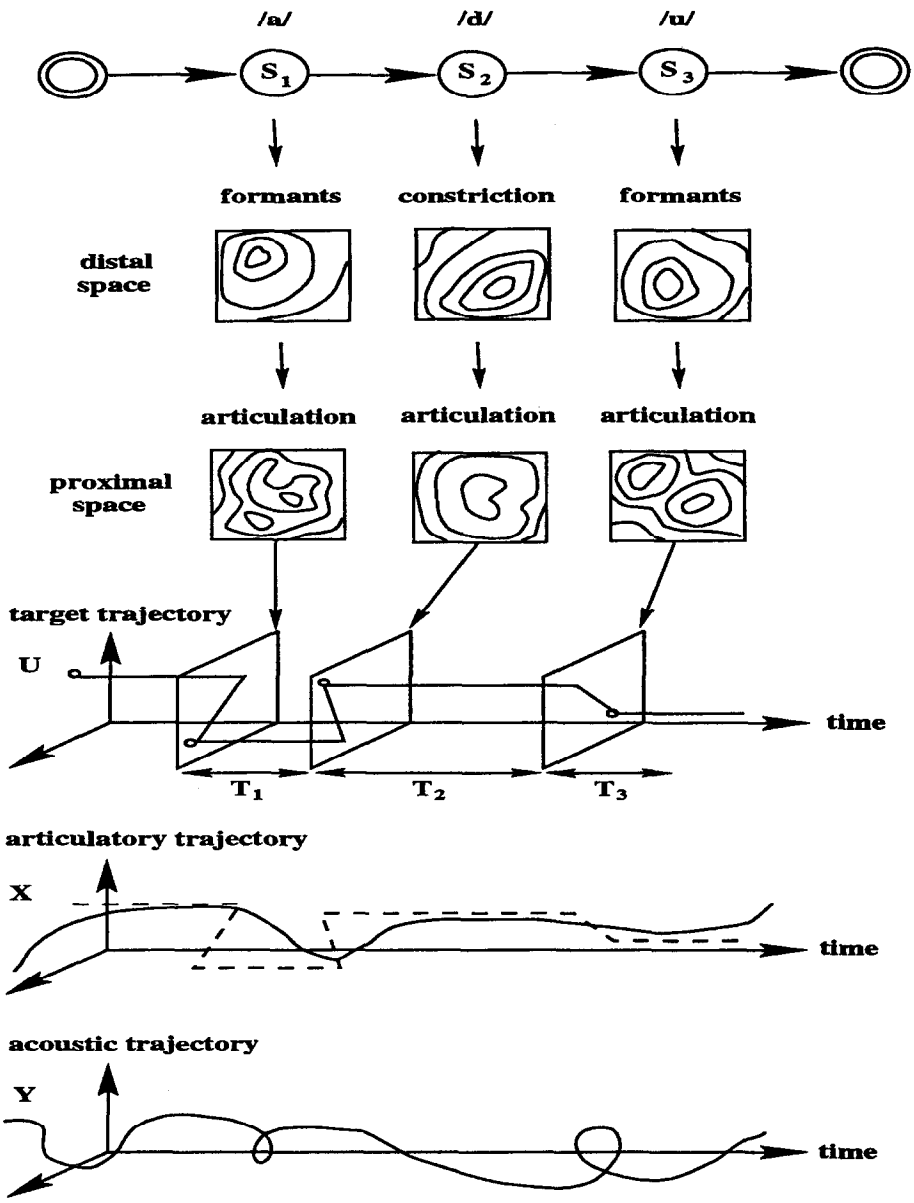


Fig. 2. A stochastic target model of speech production.

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 speech processing laboratory at University of 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

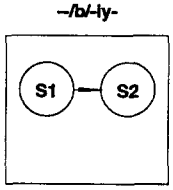
Valuable discussions with Ken Stevens, Joe Perkell, Victor Zue, Cathy Browman, Louis Goldstein, Elliot Saltzman, Richard McGowan, and with Chin Lee on issues of speech production modeling and on several views presented in this paper are gratefully acknowledged. We also thank John Hogden, Rafael Laboissière, Pascal Perrier, and an anonymous reviewer for their constructive comments which have improved the presentation of the ideas contained in this paper.

Appendix A. Example: automatic construction of articulatory HMM states

This appendix provides an example for some intermediate functions in the operation of the algorithm presented in Section 5 for automatic construction of articulatory HMM states. The example uses input phonetic string /b iy m ih ng/ (*beaming*) and uses constraint $\Delta = 1$. Each of the allophones in this phonetic string is mapped to its corresponding contexual HMM states. The sizes of these HMM states are 2, 5, 7, 7 and 4, respectively, for each of the five allophones. The contextual feature arrays and the HMM state topologies for the five allophones are shown below.

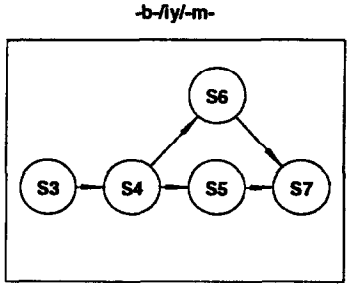
(1) /b/:

$$F(/b/|/b/,/iy/) = \begin{array}{|c|c|c|}\hline U & L\alpha & U \\ \hline U & U & U \\ \hline U & U & r D\gamma \\ \hline U & V\alpha & U \\ \hline U & X\alpha & U \\ \hline \end{array}$$



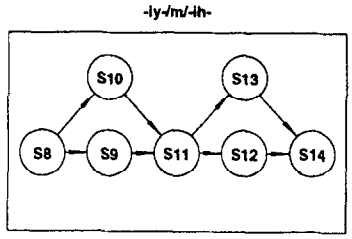
(2) /iy/:

$$F(/iy/|/b/,/iy/,/m/) = \begin{array}{|c|c|c|}\hline \ell L\alpha & U & r L\alpha \\ \hline U & U & U \\ \hline U & D\gamma & U \\ \hline U & V\alpha & V\beta \\ \hline U & X\alpha & U \\ \hline \end{array}$$



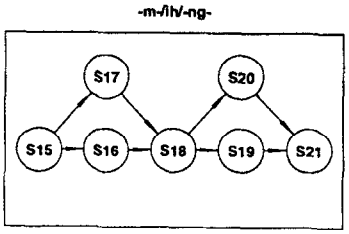
(3) /m/:

$$F(/m/|/iy/,/m/,/ih/) = \begin{array}{|c|c|c|}\hline U & L\alpha & U \\ U & U & U \\ \ell D\gamma & U & rD\zeta \\ \ell V\alpha & V\beta & rV\alpha \\ U & X\alpha & U \\ \hline\end{array}$$



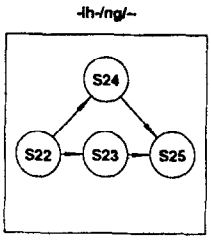
(4) /ih/:

$$F(/ih/|/m/,/ih/,/ng/) = \begin{array}{|c|c|c|}\hline \ell L\alpha & U & U \\ U & U & U \\ U & D\zeta & rD\alpha \\ \ell V\beta & V\alpha & rV\beta \\ U & X\alpha & U \\ \hline\end{array}$$



(5) /ng/:

$$F(/ng/|/ih/,/ng/) = \begin{array}{|c|c|c|}\hline U & U & U \\ U & U & U \\ \ell D\zeta & D\alpha & U \\ \ell V\alpha & V\beta & U \\ U & X\alpha & U \\ \hline\end{array}$$



All the states above are enumerated below with their five-tuple feature contents (features are specified in Table 1):

S_1: $(L\alpha,U,U,V\alpha,X\alpha)^T$,
S_2: $(L\alpha,U,rD\gamma,V\alpha,X\alpha)^T$,

S_3: $(lL\alpha,U,D\gamma,V\alpha,X\alpha)^T$,
S_4: $(U,U,D\gamma,V\alpha,X\alpha)^T$,
S_5: $(rL\alpha,U,D\gamma,V\alpha,X\alpha)^T$,
S_6: $(U,U,D\gamma,rV\beta,X\alpha)^T$,
S_7: $(rL\alpha,U,D\gamma,rV\beta,X\alpha)^T$,

S_8: $(L\alpha,U,lD\gamma,lV\alpha,X\alpha)^T$,
S_9: $(L\alpha,U,U,lV\alpha,X\alpha)^T$,
S_{10}: $(L\alpha,U,lD\gamma,V\beta,X\alpha)^T$,
S_{11}: $(L\alpha,U,U,V\beta,X\alpha)^T$,
S_{12}: $(L\alpha,U,rD\zeta,V\beta,X\alpha)^T$,
S_{13}: $(L\alpha,U,U,rV\alpha,X\alpha)^T$,
S_{14}: $(L\alpha,U,rD\zeta,rV\alpha,X\alpha)^T$,

S_{15}: $(lL\alpha,U,D\zeta,lV\beta,X\alpha)^T$,
S_{16}: $(U,U,D\zeta,lV\beta,X\alpha)^T$,
S_{17}: $(lL\alpha,U,D\zeta,V\alpha,X\alpha)^T$,
S_{18}: $(U,U,D\zeta,V\alpha,X\alpha)^T$,
S_{19}: $(U,U,rD\alpha,V\alpha,X\alpha)^T$,
S_{20}: $(U,U,D\zeta,rV\beta,X\alpha)^T$,
S_{21}: $(U,U,rD\alpha,rV\beta,X\alpha)^T$,

S_{22}: $(U,U,lD\zeta,lV\alpha,X\alpha)^{T}$,
S_{23}: $(U,U,D\alpha,lV\alpha,X\alpha)^{T}$,
S_{24}: $(U,U,lD\zeta,V\beta,X\alpha)^{T}$,
S_{25}: $(U,U,D\alpha,V\beta,X\alpha)^{T}$.

References

[1] S. Anderson, Nasal consonants and the internal structure of segments, Language 52 (1976) 326–344.

[2] G. Bailly, R. Laboissière, J.L. Schwartz, Formant trajectories as audible gestures: An alternative for speech synthesis, J. Phonetics 19 (1) (1991) 9–23.

[3] G. Bailly, Recovering place of articulation for occlusives in VCV's. In: Proc. XIII-th ICPhS, 1995, Vol. 2, pp. 230–233.

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

[5] R. Bakis, An articulatory-like speech production model with controlled use of prior knowledge. Notes from Frontiers in Speech Processing, CD-ROM, 1993.

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

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

[8] C. Blackburn, S. Young, Towards improved speech recognition using a speech production model. In: Proc. Eurospeech, 1995, Vol. 2, pp. 1623–1626.

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

[10] H. Bourlard, H. Hermansky, N. Morgan, Towards increasing speech recognition error rates, Speech Communication 18 (3) (1996) 205–231.

[11] C. Browman, L. Goldstein, Articulatory phonology: An overview, Phonetica 49 (1992) 155–180.

[12] W. Chow, B.H. Juang, C.H. Lee, A minimum error rate pattern recognition approach to speech recognition, Internat. J. Pattern Recog. Artif. Intell. 8 (1996) 5–31.

[13] W. Chow, B.H. Juang, C.H. Lee, Statistical and discriminative methods for speech recognition. In: C. Lee, F. Soong, K. Paliwal (Eds.), Automatic Speech and Speaker Recognition – Advanced Topics. Kluwer Academic Publishers, Boston, MA, 1996, pp. 109–132.

[14] N. Clements, E. Hume, The internal organization of speech sounds. In: J. Goldsmith (Ed.), The Handbook of Phonological Theory. Blackwell, Cambridge, 1995, pp. 206–244.

[15] J. Cohen, The summers of our discontent. In: Proc. Addendum ICSLP, 1996, pp. 9–10.

[16] C. Coker, A model of articulatory dynamics and control, Proc. IEEE 64 (1976) 452–460.

[17] L. Deng, Design of a feature-based speech recognizer aiming at integration of auditory processing, signal modeling, and phonological structure of speech. J. Acoust. Soc. Amer. 93 (4) Pt. 2 (1993) 2318.

[18] L. Deng, A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal, Signal Processing 27 (1) (1992) 65–78.

[19] L. Deng, M. Aksmanovic, D. Sun, J. Wu, Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states, IEEE Trans. Speech Audio Process. 2 (4) (1994) 507–520.

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

[21] L. Deng, H. Sameti, Speech recognition using dynamically defined speech units, Proc. ICSLP 4 (1994) 2167–2170.

[22] L. Deng, D. Sun, A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features, J. Acoust. Soc. Amer. 95 (1994) 2702–2719.

[23] L. Deng, J. Wu, H. Sameti, Improved speech modeling and recognition using multi-dimensional articulatory states as primitive speech units. In: Proc. ICASSP, Vol. 1, 1995, pp. 385–388.

[24] L. Deng, H. Sameti, Transitional speech units and their representation by the regressive Markov states: Applications to speech recognition, IEEE Trans. Speech Audio Process. 4 (4) (1996) 301–306.

[25] L. Deng, M. Lennig, P. Mermelstein, Modeling microsegments of stop consonants in a hidden Markov model based word recognizer, J. Acoust. Soc. Amer. 87 (1990) 2738–2747.

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

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

[28] L. Deng, C. Rathinavalu, A Markov model containing state-conditioned second-order nonstationarity: Application to speech recognition, Comput. Speech Language 9 (1) (1995) 63–86.

[29] V. Digalakis, J. Rohlicek, M. Ostendorf, ML estimation of a stochastic linear system with the *EM* algorithm and its application to speech recognition, IEEE Trans. Speech Audio Process. 1 (1993) 431–442.

[30] J. Durand, F. Katamba, Frontiers of Phonology: Atoms, Structures, Derivations. Longman, London, 1995.

[31] E. Eide, A linguistic feature representation of the speech waveform. In: Proc. ICASSP, Vol. 2, 1993, pp. 483–486.

[32] K. Erler, L. Deng, Hidden Markov model representation of quantized articulatory features for speech recognition, Comput. Speech Language 7 (3) (1993) 265–282.

[33] E. Foster et al., Automatic learning of word pronunciation from data, In: Proc. Addendum ICSLP, 1996, pp. 28–29.

[34] T. Fourakis, R. Port, Stop epenthesis in English, J. Phonetics 14 (1986) 197–221.

[35] S. Furui, Flexible speech recognition. In: Proc. Eurospeech, Vol. 3, 1995, pp. 1595–1603.
[36] O. Ghitza, M. Sondhi, Hidden Markov models with templates as nonstationary states: an application to speech recognition, Comput. Speech Language 7 (1993) 101–119.
[37] J.A. Goldsmith, 1990. Autosegmental and Metrical Phonology. Blackwell, Oxford.
[38] F.H. Guenther, A modeling framework for speech motor development and kinematic articulator control. In: Proc. XIII-th ICPhS, Vol. 2, 1995, pp. 92–99.
[39] J. Hogden, A maximum likelihood approach to estimating speech articulator positions from speech acoustics. J. Acoust. Soc. Amer. 100 (4) Pt. 2 (1996) 2663–2664.
[40] W. Holmes, M. Russell, Modeling speech variability with segmental HMMs. In: Proc. Eurospeech, 1996, Vol. 1, pp. 447–450.
[41] M. Honda, T. Kaburagi, A dynamical articulatory model using potential task representation. In: Proc. ICSLP, 1995, Vol. 1, pp. 179–182.
[42] R.A. Houde, A study of tongue body motion during selected speech sounds. SCRL, Santa Barbara, CA, 1968, SCRL Monogr. No. 2.
[43] M. Hwang, X. Huang, F. Alleva, Predicting unseen triphones with senones. In: Proc. ICASSP, 1993, Vol. 1, pp. 311–315.
[44] T. Kaburagi, M. Honda, A model of articulator trajectory formation based on the motor tasks of vocal-tract shapes, J. Acoust. Soc. Amer. 99 (5) (1996) 3154–3170.
[45] P. Keating, The window model of coarticulation: articulatory evidence. In: J. Kingston, M. Beckman (Eds.), Papers in Laboratory Phonology, I. Cambridge University Press, Cambridge, 1990, pp. 451–470.
[46] J. Kelso, E. Saltzman, B. Tuller, The dynamical perspectives on speech production: Data and theory, J. Phonetics 14 (1986) 29–59.
[47] P. Kenny, M. Lennig, P. Mermelstein, Articulatory Markov models. In: Proc. IEEE Workshop Automatic Speech Recognition. Harriman, New York, 1991, pp. 22–23.
[48] R. Kent et al., 1995. Models of speech production. In: N. Lass (Ed.), Principles of Experimental Phonetics. Mosby, London, pp. 3–45.
[49] J. Keyser, K. Stevens, Feature geometry and the vocal tract, Phonology 11 (2) (1994) 207–236.
[50] R. Laboissière, D.J. Ostry, P. Perrier, A model of human jaw and hyoid motion and its implications for speech production. In: Proc. XIII-th ICPhS, 1995, Vol. 2, pp. 60–67.
[51] S. Levinson, L. Rabiner, M. Sondhi, An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition, Bell System Tech. J. 62 (1983) 1035–1074.
[52] S. Liu, Landmark detection for distinctive feature-based speech recognition. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1995.
[53] P. MacNeilage, Motor control of serial ordering in speech, Psychol. Rev. 77 (1970) 182–196.
[54] S. Maeda, On articulatory and acoustic variabilities, J. Phonetics 19 (1991) 321–331.
[55] J. Mahkoul, R. Schwartz, State of the art in continuous speech recognition, Proc. Nat. Acad. Sci. USA 92 (1995) 9956–9963.
[56] R. McGowan, Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests, Speech Communication 14 (1) (1994) 19–48.
[57] R. McGowan, A. Faber, Speech production parameters for automatic speech recognition, J. Acoust. Soc. Amer. 101 (1) (1997) 28.
[58] H. Meng, V. Zue, Signal representation comparison for phonetic classification. In: Proc. ICASSP, 1991, Vol. 1, pp. 285–288.
[59] N. Morgan, H. Bourlard, Connectionist Speech Recognition – A Hybrid Approach. Kluwer, Boston, MA, 1994.
[60] M. Ostendorf, From HMMs to segment models. In: C. Lee, F. Soong, K. Paliwal (Eds.), Automatic Speech and Speaker Recognition – Advanced Topics. Kluwer, Boston, MA, 1996, pp. 185–210.
[61] J.S. Perkell, Phonetic features and the physiology of speech production. In: B. Butterworth (Ed.), Language Production. Academic Press, London, 1980.
[62] J.S. Perkell, M.L. Matthies, M.A. Svirsky, M.I. Jordan, Goal-based speech motor control: A theoretical framework and some preliminary data, J. Phonetics 23 (1995) 23–35.
[63] J.S. Perkell, Properties of the tongue help to define vowel categories: Hypotheses based on physiologically-oriented modeling, J. Phonetics 24 (1996) 3–22.
[64] P. Perrier, D. Ostry, R. Laboissière, The equilibrium point hypothesis and its application to speech motor control, J. Speech Hearing Res. 39 (1996) 365–378.
[65] L. Rabiner, B.H. Juang, C.H. Lee, An overview of automatic speech recognition. In: C. Lee, F. Soong, K. Paliwal (Eds.), Automatic Speech and Speaker Recognition – Advanced Topics. Kluwer, Boston, MA, 1996, pp. 1–30.
[66] G. Ramsay, L. Deng, A stochastic framework for articulatory speech recognition. J. Acoust. Soc. Amer. 95 (5) Pt. 2 (1994) Abstract 2aSP19.
[67] G. Ramsay, L. Deng, Articulatory synthesis using a stochastic target model of speech production. In: Proc. XIII-th ICPhS, 1995, Vol. 2, pp. 338–341.
[68] G. Ramsay, L. Deng, Maximum likelihood estimation for articulatory speech recognition using a stochastic target model. In: Proc. Eurospeech, 1995, Vol. 2, pp. 1401–1404.
[69] G. Ramsay, L. Deng, Optimal filtering and smoothing for speech recognition using a stochastic target model. In: Proc. ICSLP, 1996, Vol. 2, pp. 1113–1116.
[70] M. Randolph, Speech analysis based on articulatory behavior. J. Acoust. Soc. Amer. 95 (5) Pt. 2 (1994) 1aSP15.
[71] A. Robinson, F. Fallside, A recurrent error propagation network speech recognition system, Comput. Speech Language 5 (1991) 259–274.
[72] R. Rose, J. Schroeter, M. Sondhi, The potential role of speech production models in automatic speech recognition, J. Acoust. Soc. Amer. 99 (3) (1996) 1699–1709.
[73] P. Rubin, E. Saltzman, L. Goldstein, R. McGowan, M.

Tiede, C. Browman, CASY and extensions to the task-dynamic model. In: Proc. 4th European Speech Production Workshop, Autrans, 1996, pp. 125–128.

[74] E. Saltzman, K. Munhall, A dynamical approach to gestural patterning in speech production, Ecol. Psychol. 1 (1989) 333–382.

[75] J. Schroeter, M.M. Sondhi, Techniques for estimating vocal-tract shapes from the speech signal, IEEE Trans. Speech Audio Process. 2 (1) (1994) 133–150.

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

[77] K. Shirai, M. Honda, Estimation of articulatory motion. In: Dynamic Aspects of Speech Production. University of Tokyo Press, Tokyo, 1976.

[78] K. Stevens et al., Implementation of a model for lexical access based on features. In: Proc. ICSLP, 1992, Vol. 1, pp. 499–502.

[79] K. Tokuda et al., An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. In: Proc. Eurospeech, 1995. Vol. 1, pp. 757–760.

[80] A. Varga, R. Moore, HMM decomposition of speech and noise. In: Proc. ICASSP, 1990, pp. 845–848.

[81] R. Wilhelms, Schätzung von artikulatorischen Bewegungen eines stylisierten Artikulatormodells aus dem Sprachsignal. Ph.D. Dissertation, Georg-August Universität, Göttingen, 1987.

[82] S. Young, Large vocabulary continuous speech recognition: A review. In: Proc. IEEE Workshop on Automatic Speech Recognition, Snowbird, UT, 1995, pp. 3–28.