Articulatory Features and Associated Production Models in Statistical Speech Recognition

Li Deng

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

Summary. A statistical approach to speech recognition is outlined which draws close parallel with closed-loop human speech communication schematized as a joint process of encoding and decoding of linguistic messages. The encoder consists of the symbolically-valued overlapping articulatory feature model and of its interface to a nonlinear task-dynamic model of speech production. A general speech recognizer architecture based on optimal decoding strategy incorporating encoder-decoder interactions is described and discussed.

1. Introduction

The general concept of closed-loop speech chain underlying human speech communication has been known for many years [2]. However, engineering construction of automatic speech recognition machines, which have been known to perform orders of magnitude worse than human, so far has hardly been able to capitalize on any significant properties of the closed-loop human speech communication . This situation arises due to a number of important factors including 1) (justifiable) desires for short-term engineering success in limited tasks; 2) lack of interactions between scientific and technological research communities and hence lack of integration of the respective research accomplishments; 3) fragmental and incomplete nature of our understanding of the closed-loop human speech communication process; and 4) lack of suitable computational formalisms which would allow the scientific understanding to be readily useful in computation-intensive speech technology applications.

The purpose of this tutorial paper is to describe the general nature of the closed-loop human speech chain as an encoding-decoding process (analogous to information-theoretic design of engineering communication systems), and to show how within this framework computational formalisms can be established enabling graceful integration of engineering modeling-decoding techniques with scientific models and theories intended to faithfully describe the human speech process.

2. Functional description of human speech communication as an encoding-decoding process

At the global and functional level, human speech communication can be viewed as an encoding-decoding process, where the decoding process or perception is an ac-

tive process consisting of auditory reception followed by phonetic/linguistic interpretation. As an encoder implemented by the speech production system, the speaker uses knowledges of meanings of words (or phrases), of grammar in a language, and of the sound representations for the intended linguistic message. Such knowledges can be made analogous to the keys used in engineering communication systems. The phonetic plan, derived from the semantic, syntactic, and phonological processes, is then executed through the motor-articulatory system to produce speech waveforms.

As a decoder which aims to accomplish speech perception, the listener uses a key, or the internal "generative" model , which must be compatible with (may not be identical to) the key used by the speaker to interpret the speech signal received and transformed by the auditory system. This enables the listener to reconstruct, via (probabilistic) analysis-by-synthesis strategies, the linguistic message intended by the speaker. Such an encoding-decoding view of human speech communication, where the observable speech acoustics plays the role of carrier of deep, linguistically meaningful messages , is strikingly similar to the modulation-demodulation scheme in electronic digital communication and to the encryption-decryption scheme in secure electronic communication.

Since the nature of the key used in the phonetic-linguistic information decoding or speech perception/understanding lies in the strategies used in the production or encoding process, speech production and perception are intimately linked in the closed-loop speech chain. The implication of such a link for speech recognition technology is the need to develop functional and computational models of human speech production for use as an "internal model" in the decoding process by machines.

3. Overview of theories of speech perception

With respect to the above encoding-decoding review of human speech communication which advocates intimate links between speech production and perception, a number of popular theories and models of speech perception are reviewed here.

Motor theory of speech perception, addressing the issue of ubiquitous acoustic variability of speech, experienced two main stages of development, both emphasizing a specialized phonetic module mediating speech production and perception. The early version of the theory asserts existence of phonetic invariance at the levels of articulatory gesture or motor command [13]. Due to the failure of finding such invariance experimentally, this earlier version was modified to move the proposed phonetic invariance to higher, vaguely specified levels of speech production [14]. The abstract nature of the modified motor theory renders it practically useless for possible speech recognition applications.

Closely related to motor theory, the analysis-by-synthesis model [11] of speech perception adopted a more tangible, hypothesis-and-test approach to phonetic decoding by human. Elements of this model include the proposal of active internal synthesis of comparison signals, use of generative rules to convert lexical items into phonetic parameters (which describe the behavior of structures controlling the

vocal-tract configuration and vocal-cords activities), and rules to convert these phonetic parameters into time-varying speech spectra.

Sharing partial views with motor theory, direct-realist theory of speech perception proposes that listener directly perceives the articulatory gestures of the speaker via the structure that the gestures pass on to the common acoustic medium between listener and speaker. The theory does not require specialized phonetic module [10].

Contrary to motor theory, acoustic-auditory theory of speech perception asserts existence of phonetic invariance not in any internal levels of speech production, but in the acoustic-auditory domain, which is the outcome of speech production and determines the object of speech perception. In this theory, speech production and perception are indirectly linked by virtue of common acoustic goals or targets [20, 12, 9].

A drastically different theory of speech perception from all the above ones proposes that it is the interactions of speaker and listener based on balances between speaker's efforts and listener's contrastive perceptual goals , not the phonetic invariance at any levels of the speech chain, which are essential properties of speech perception. This theory is called Hyper-Hypo or H&H theory [15, 16]. H&H theory proposes that the distal object of speech perception is the speaker's intention (shared with motor theory), but that such an intention has both articulatory-gesture production component and contrastive perceptual component, and is determined by short-term, dynamic interactions of the two components. An essential concept of the theory is plasticity of phonetic gestures — speakers adaptively tune phonetic gestures to the needs of speaking situations under motor and perceptual constraints, and phonetic gestures are not invariant but are adaptations to constraints on production mechanisms for least "efforts" (or speech economy , low-cost behavior, or "hypo" speech) and on perceptual mechanisms for achieving sufficient contrast ("hyper" speech). These mechanisms are language independent and not special to speech; "invariance" must be defined according to the global purpose of speech communication (e.g. lexical access and speech comprehension).

One supporting evidence of H&H theory is the phenomenon of compensatory articulation where speakers are capable of re-organizing articulation to reach fixed acoustic and perceptual goals under both artificial bite-block condition and natural loud, clear, fast or spontaneous speaking conditions. Another evidence comes from the formation of the phonetic system with "quantal" properties which can be shown as being driven by a demand for sufficient perceptual contrast. Speech communication system is established via constant interaction between speaker and listener: the listener force the speaker to make sufficient phonetic distinctions (negative control), and the speaker tries to use least "efforts" but is simultaneously constrained by the listener's demand.

4. A general framework of statistical speech recognition

The Bayesian framework is adopted as a general framework for intended incorporation of scientifically motivated speech models in statistical speech recognition. Let

$\mathbf{O} = O_1, O_2, ..., O_T$ be a sequence of observable acoustic data of speech, and let $W = w_1, w_2, ..., w_n$ be the sequence of words intended by the speaker who produces the acoustic record $\mathbf{O}$. The goal of a speech recognizer is to "guess" the most likely word sequence $\hat{W}$ given the acoustic data $\mathbf{O}$. The problem can be formulated as a top-down search problem over the allowable word sequences:

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

Decomposition of the word-to-acoustics probability $P(\mathbf{O}|\mathbf{W})$ above is accomplished by using law of total probability:

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

where $\mathcal{F}$ is a discrete-valued *phonological* construct (or "pronunciation" model), which specifies, according to probability $P(\mathcal{F}|W)$, how words and word sequences W can be expressed in terms of a particular organization of a small set of fundamental phonological units; $P(\mathbf{O}|\mathcal{F})$ is the probability that a particular organization $\mathcal{F}$ of phonological units produces the acoustic data $\mathbf{O}$. This probability is determined by the phonetic interface model .

According to phonetic theories, the interface model ideally should consist of at least three hierarchical levels of mapping: from phonological symbols ($\mathcal{F}$) to motor commands ($\mathcal{M}$), from motor commands to articulation ($\mathcal{A}$), and from articulation to acoustics ($\mathbf{O}$). That is, one can further decompose the probability $P(\mathbf{O}|\mathcal{F})$ associated with the global interface model into:

$$P(\mathbf{O}|\mathcal{F}) = \sum_{\mathcal{M},\mathcal{A}} P(\mathbf{O}|\mathcal{A})P(\mathcal{A}|\mathcal{M})P(\mathcal{M}|\mathcal{F}) \approx \max_{\mathcal{M},\mathcal{A}} P(\mathbf{O}|\mathcal{A})P(\mathcal{A}|\mathcal{M})P(\mathcal{M}|\mathcal{F}). \tag{3}$$

For efficient engineering construction of speech recognizers, an approximation to the above layered, multi-level mapping is necessary and can be made by one-level or two-level mappings from the phonological level $\mathcal{F}$ to the acoustic level $\mathbf{O}$. Any approximation must faithfully retain the dynamic character of the speech production process. [1]

5. Brief analysis of weaknesses of current speech recognition technology

Despite some success in highly constrained recognition tasks, the current HMM-based, data-driven speech recognition technology is fundamentally limited in its

[1] Some work done in our research group included three types of approximation (differing by three distinct levels at which lies the object of dynamic modeling): 1) Acoustic-dynamic model based on nonstationary-state or trended HMM [8]; 2) Articulatory-dynamic or stochastic target model [7, 18]; and 3) Task-dynamic model [3, 4, 5].

ability to achieve human-like speech recognition. Such a limitation stems from its weak theoretical foundations from both phonological and phonetic perspectives. First, nearly all currently popular speech recognition strategies use more or less the same set of phone-like phonological speech units (e.g. triphones) arranged in strictly linear sequences, like "beads-on-a-string". This, however, is not how human language faculty organizes its phonological primitives. Second, the weak theoretical foundation of the current speech recognition technology from phonetic perspective is reflected in the weak structure of the HMM in use and in the simplistic strategy of surface data fitting to the observable acoustics (equipped with virtually no underlying data generation mechanisms). A consequence of this weakness is that the sample paths of the HMM as a nonstationary stochastic process deviate significantly from true speech data trajectories.

The above weaknesses associated with the current speech recognition technology lead to speech recognizers which inherently lack robustness, and cannot generalize from training data to mismatched test data. The problem is particularly serious when little supervised adaptation data are available to recognizers, as in most real-world speech recognition applications. Such recognizers inevitably break down when moving from read or clear speech style to casual, fast and spontaneous speaking mode, switching from "sheep" speakers to "goat" speakers, or porting from one language to another or from one task to another. When new tasks or new languages are involved, re-design and re-training of the recognizers are undesirably needed.

It appears that the ultimate success of human-like speech recognition will require not only extensions of existing recognizer architectures, but fundamental changes in the statistical models of speech underlying speech recognizers. Such new models must at a functional level faithfully characterize essential properties of human behaviors in closed-loop speech communication (production and perception) and be equipped with effective computational formalisms and model learning strategies.

6. Phonological model: Overlapping articulatory features and related HMMs

Motivations of using vocal-tract constriction based articulatory features as the phonological primitive can be succinctly summaried by a quote from modern phonology literature [1]: "Phonetic Interpretation of the Feature Hierarchy: ...the basic organizing principle of the feature hierarchy is the *vocal tract constriction*.... The place features define constriction location and the articulator-free features define constriction degree. The notion "constriction" is central to many current theories of speech production, both acoustic and articulatory. It is therefore not surprising that phonological representations may be organized in terms of (vocal tract) constrictions as well."

In the work described briefly in [5], a compact set of universal phonological/articulatory features across world languages is designed. The resulting feature

specification systems, one for each language, share intensively among the component features. Through appropriate combinations of component features, new sounds or segments in new, target languages from the sounds in the source language(s) can be reliably predicted. The phonological model uses hierarchically organized articulatory features as the primitive phonological units motivated by articulatory phonology and feature-geometry theory.

The phonological model further entails a statistical scheme to allow probabilistic, asynchronous but constrained temporal overlapping among components (symbols) in the sequentially placed feature bundles. A set of feature overlapping and constraining rules are designed based on syllable structure and other prosodic factors . Examples of the rules for English are: 1) overlap among consonant clusters in syllable onset and coda, and in consonant sequence across connected words; 2) overlap between syllable onset and nucleus; 3) overlap between syllable nucleus and coda; 4) overlap of the tongue-dorsum feature between two adjacent syllable nuclei; 5) except for Lips and Velum features, no overlap between onset and coda within the same syllable.

The phonological model finally contains a crucial component which converts the above probabilistic feature overlap pattern to a finite state automaton (FSA) or HMM state topology. This FSA represents ensemble sequences of phonological units composed of the overlapped features, serving as the phonetic plan which controls lower (phonetic) levels of speech production resulting in dynamic patterns of speech acoustics.

7. Task-dynamic model of speech production

Before I discuss how the above overlapping-articulatory feature based phonological model can be interfaced to the phonetic variables (including ultimately speech acoustics), the (deterministic) task-dynamic model of speech production developed in speech science [19] is briefly reviewed. This is a most comprehensive speech production model, well developed and tested, based originally on a general model of skilled movement control. In this model, the control signal is derived from abstract gestural units defined in articulatory phonology; these gestural units are organized into utterance-specific “gestural scores”. Each gesture is correlated with two task variables , vocal-tract constriction degree and constriction location. At any point in time, only a small subset (fewer than 3 or 4 usually) of the gestures are co-occurring (overlapping or “blending”) during speech production. When blending occurs, competitive blending rules are used to determine the final values of the correlated task variables.

The intrinsic dynamics for each task variable is modeled as critically damped second order system, which is characterized by the gesture-dependent (normalized) stiffness and by the gesture-dependent point-attractor of the dynamical system. The

relation between the task variables and the model-articulators is characterized by a static and nonlinear function, which is constructed by vocal-tract geometry [2].

Using piecewise locally linear approximation of the nonlinear relation between task variables and model-articulators, the linear dynamics for the former is converted into quasi-linear dynamics in the latter. Jacobian transformation matrix derived from the nonlinear relation becomes a component of the linearized dynamics. Finally, given the time-varying model-articulator motions, a further static, nonlinear relation maps the model-articulators into speech acoustics using Haskins Lab's configurable articulatory synthesizer .

8. Interfacing overlapping features to task-dynamic model and a general architecture for speech recognition

Our computational approach to phonology-to-phonetic interface is based on a discrete-time task-dynamic model derived from the continuous-time model reviewed above, with a statistical structure imposed [4]. In this approach, each individual articulatory feature (as symbolic phonological unit) is made to associate with the task-dynamic model parameters including stiffness and the point or region of the attractor (continuous phonetic variables). The stiffness and attractor parameters corresponding to simultaneously overlapped or blended features are determined from those of individual component features according to either empirical rules or automatic learning. The nonlinear relation between the task variables and the model-articulators is approximated by trainable Multi-Layer Perceptron (MLP) neural nets, which serve as a generic device for data interpolation in a multi-dimensional space. Another trainable MLP is used to approximate the nonlinear articulator-to-acoustics mapping.

An architecture for speech recognition is presented in Fig. 1 based on the approach described above using the overlapping articulatory feature model interfaced to task-dynamic model of speech production. Language-universal components include the feature primitives in the phonological model and in most subcomponents in the task-dynamic (phonetic) model. The decoding strategy (recognition search) is similar to that in the current HMM-based technology, but the way the acoustic likelihood is computed is drastically different. The structured, probabilistic phonological and phonetic models used here have a high degree of parameterization, which allows parsimonious yet accurate characterization of the recognizer capable of language independent, speaker independent, speaking-style independent, and unlimited-vocabulary speech recognition.

One most crucial component of the task-dynamic model is the nonlinear relation between task variables and model-articulators and that between model-articulators and acoustics. In the speech recognition architecture of Fig.1, a trainable neural net-

[2] Many-to-one nature of the relation gives rise to compensatory articulation (or motor equivalence) and to coordinative structure in this model.

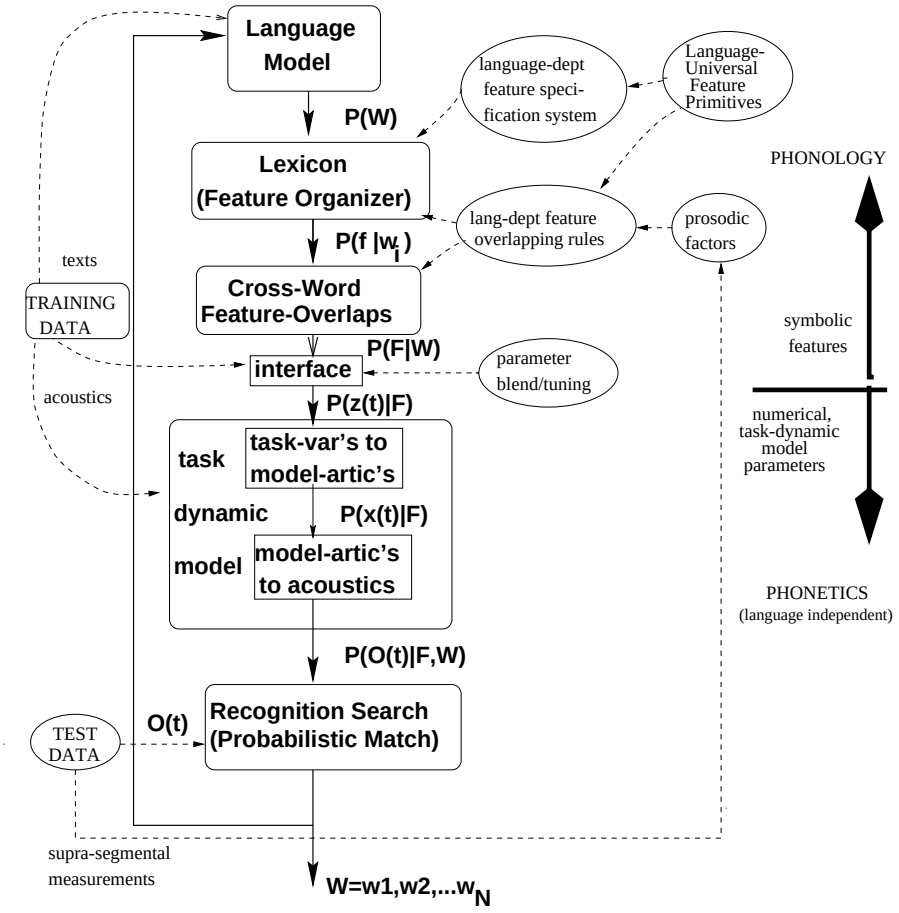

FIGURE 1. An architecture of speech recognition using feature-based phonological model interfaced to statistical task-dynamic model of speech production

work is used for approximating these relations. The general topology of the network is shown in Fig.2, with the network unit connections strongly constrained by speech production mechanisms .

9. Discussions: Machine speech recognition

The above sections described a feature-based phonological model interfaced with a task-dynamic model of speech production. While many details need to be specified, this model can be regarded as the a simple version of the "key" used by speaker to encode phonological messages and simultaneously as a functionally compatible "internal" model used by human listener in decoding speech (perception). In machine

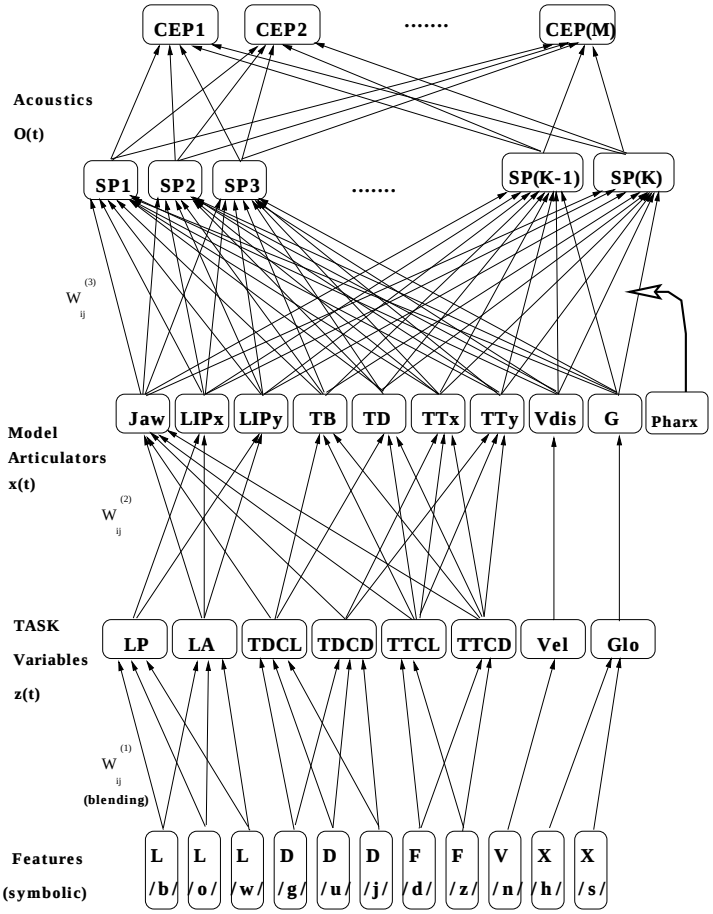

FIGURE 2. Neural-net implementation of static nonlinear mappings in the task-dynamic model

recognition of speech, this production-oriented model is used by the recognizer to accurately and succinctly characterize the dynamic pattern of the observed speech signal, thereby providing an accurate term in class probability $P(\mathbf{O}|W)$ of Eqn.(1). According to Bayesian decision which we adopt as an engineering strategy for the analogous human listener's cognitive interpretation of the auditorily received speech information, optimal recognition performance would be achieved given accurate estimate of $P(\mathbf{O}|W)$.

The proposed production-oriented approach to speech recognition is based on statistical characterization, via functional approximation, of the signals at various levels of the human speech "chain" — phonological, motor-task, articulatory, acoustic, and auditory levels. In particular, statistical relations among the signals at these levels are functionally approximated. This, therefore, contrasts sharply with mo-

tor theory, direct-realist theory, and acoustic-auditory theory of speech perception (section 3) which all insist on existence of phonetic invariance either at particular level(s) of the speech production process, or at the output of such a process. It also contrasts sharply with the analysis-by-synthesis model of speech perception (section 3) in that the speech recognition decision is made in an integrated manner by evaluation and comparison of a posteriori probability $P(W|O)$ given possible candidate hypotheses[3], rather than performing step-by-step inversions from the auditorily received signal to the final perceptual object of linguistic messages.

Similar to the well established practice in modern speech recognition research, the functional form of the speech encoder described in this paper, which comprises the feature-based phonological model interfaced to a task-dynamic model, is fixed a priori, while the parameters of the speech encoder are automatically trained from observable speech data [4]. While the Bayesian-based optimal decoding strategy remains the same, different training methods have different implications in terms of the various phonetic theories of speech perception. Maximum-likelihood training implies phonetic invariance in the parameters of the speech production model characterizing systematic changes of the phonetic variables[5]. For Bayesian-style training (including but not limited to the MAP learning), the implication is that phonetic invariance exists in the probability distribution classes, consistent with the proposal in a recent theoretical framework that speech production goals are specified in terms of regions (distributions) rather than of points [17]. Finally, minimal classification error (discriminative) training will imply non-existence of any kind of underlying phonetic invariance; rather, perceptual contrasts are the primary objective in determining the speech encoder's parameters. If some type of constraints on speech economy are incorporated into the speech model as an encoder[6], then proper balance between the degree of the constraints and the discriminative objective would give a way of implementing the concept of H&H theory (section 3) advocating encoder-decoder or speaker-listener interactions and mutual constraints in the human speech communication process.

10. REFERENCES

[1] Clements N. and Hume E. (1995) "The internal organization of speech sounds," in *The Handbook of Phonological Theory*, J. Goldsmith (ed.), Blackwell, Cambridge, 206-244.

[2] Denes P. and Pinson E. (1973) *The Speech Chain — The Physics and Biology of Spoken Languages*, New York, N.Y., Doubleday Press.

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

[3] This is the same philosophy permeating all modern speech recognition research.

[4] For details, see [6].

[5] To be differentiated from the claim of motor theory that phonetic invariance is in the actual phonetic variables themselves.

[6] See one example in [6] based on the smoothness-prior Bayesian approach developed originally by statisticians.

[4] Deng L. (1992-1993) "A Computational Model of the Phonology-Phonetics Interface for Automatic Speech Recognition," Summary Report of Research in Spoken Language, Laboratory for Computer Science, Massachusetts Institute of Technology.
[5] Deng L. (1997) "Integrated-multilingual speech recognition using universal phonological features in a functional speech production model," *Proc. ICASSP*, Munich, Germany, vol. 2, pp. 1007-1010.
[6] Deng L. (1998) "Computational models for speech production," this book.
[7] Deng L., Ramsay L., and Sun D. (1997) "Production models as a structural basis for automatic speech recognition," *Speech Communication*, August issue.
[8] 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.
[9] Diehl R. and Kluender K. (1989) "On the object of speech perception," *Ecological Psychology*, vol 1, pp. 1-45.
[10] Fowler C. (1986) "An event approach to the study of speech perception from a direct-realist perspective," *J. Phonetics*, vol. 14, pp. 3-28.
[11] Halle M. and Stevens K. (1962) "Speech recognition: A model and a program for research," *IRE Trans. Information Theory*, vol. 7, pp. 155-159.
[12] Klatt D. (1989) "Review of selected models of speech perception," in *Lexical Representation and Process*, W. Marslen-Wilson (ed.), pp. 169-226.
[13] Liberman A., Cooper F., Shankweiler D. and Studdert-Kennedy M. (1967) "Perception of the speech code," *Psychology Review*, vol. 74, pp. 431-461.
[14] Liberman A. and Mattingly I. (1985) "The motor theory of speech perception revised" *Cognition*, vol. 21, pp. 1-36.
[15] Lindblom B. (1990) "Explaining phonetic variation: A sketch of the H&H theory," in *NATO Workshop on Speech Production and Speech Modeling*, W. Hardcastle and A. Marchal (eds.), pp. 403-439.
[16] Lindblom B. (1996) "Role of articulation in speech perception: Clues from production," *JASA*, vol. 99(3), pp. 1683-1692.
[17] 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*, vol.23, pp. 23-35.
[18] 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.
[19] Saltzman E. and Munhall K. (1989) "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, 1, pp. 333-382.
[20] Stevens K. and Blumstein S. (1981) "The search for invariant acoustic correlates of phonetic features," in *Perspectives on the Study of Speech*, P. Eimas and J. Miller (eds.), pp. 1-38.