MODELS, THEORY AND DATA IN SPEECH PRODUCTION

Joseph S. Perkell

Massachusetts Institute of Technology,
Cambridge, Massachusetts, U.S.A.

ABSTRACT

This paper discusses modeling of the transformation from a linguistic-like input to a sound output in speech production. Such modeling can serve: 1) to formalize the expression of theoretical overviews and 2) as organizing frameworks for focussed programs of experimentation. As an example, a "task dynamic" production model is cited. The model incorporates underlying phonological primitives that consist of "abstract articulatory gestures", and it has been used in an initial attempt to interpret the relation between phonologically-based hypotheses and experimental data. Several issues that arise in such work are discussed, and suggestions are made for an alternative modeling approach.

1. INTRODUCTION

The fundamental motivation of production modeling is to account for the transformation that takes place from an underlying discrete linguistic representation through articulatory movements to a sound output. Global models attempt to account for most of or all of the transformation [cf. 6,30], and more detailed models attempt to account for specific parts of the transformation, such as sound production [cf. 37,8,32,15] or articulatory-to-acoustic relations (via area functions) [7,10,22]. Some other modeling work can be considered somehow to span these two categories [cf. 31,33,35,21].

In this paper I will focus on the contribution of global modeling to a basic understanding of the overall speech production process. In this kind of model, discrete linguistic representations of utterances serve as inputs to a controller which operates on a peripheral apparatus (in control theory terminology, a "plant") which produces sound. If global modeling is to inform us eventually about the nature of the actual input and control mechanisms for speech production, presumably it must incorporate an accurate model of the plant. Thus, in the long run, global models should include specific information about relevant aspects of production such as anatomy, biomechanics, aerodynamics, sound generation and articulatory-to-acoustic relations, all of which exert constraints on the role of the controller. One of the points of this paper will be that global production modeling should also consider *interactions between production and perception* (and lexical access), because perceptual mechanisms also have an influence on the control of speech production (and on sound patterns of languages).

At this point, however, not enough is known about the properties of peripheral production mechanisms and perceptual constraints to account for them comprehensively in a global production model. As a result, the peripheral components of such a model have to be represented mostly by abstractions that cannot be related directly to important constraints on speech production, and its input and controller cannot realistically represent hypotheses about the actual form of the underlying input and control mechanisms.

Given this situation, current global models have two potentially important contributions to make. One is as a means of forcing discipline on the formulation of theories of speech production. To the

extent that such theory is incorporated into an implemented model, it has to be stated in precise terms. Some of the examples I will discuss below have already made this kind of contribution. Another equally-important contribution of a global model is as a link between theory and data, in effect, as an organizing framework for a focussed program of experimentation on strategies of speech production. In this arena, much less has been done up to now. The main reasons for this shortcoming are 1) the enormous amount of work involved, and 2) until recently, the lack of adequate tools, not only for efficient model development, but also for gathering and analyzing the right kinds of data in the most useful manner. Work along these lines is just beginning, and I will discuss an example to illustrate an important kind of contribution that global modeling may make in the near future. In the long run, we can hope that an iterative cycle of model development and related experimentation may inform us about the interesting but currently untestable principles that are incorporated into global models. In addition, as work advances on specific peripheral mechanisms, global models will become increasingly realistic and we will gain a much better understanding of relations among peripheral constraints, control strategies and sound patterns.

2. BACKGROUND

Twenty years ago at the VIIth International Congress of Phonetic Sciences, Bjorn Lindblom presented a paper entitled "Numerical Models in the Study of Speech Production and Speech Perception: Some Phonological Implications" [20]. In the following year, Ken Stevens published "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data" [33]. Both of these seminal papers described the use of production modeling, including articulatory-to-acoustic relations, in combination with hypotheses about speech perception, to predict canonical articulatory targets. The targets were related to ideas about the phonological structure of language. Consistent with those ideas, the targets were essentially static, discrete and invariant in nature. Those particular production models and the related phonology had little to say about variability, timing and articulatory movement. The gap between such essentially static and discrete models on one hand and experimental observations of continuous and variable articulatory movement on the other led to suggestions that it might be more fruitful to study articulatory movements and basic physiological mechanisms without being constrained by the limitations inherent in "static" linguistic models [cf. 23].

Additional objections to such models [9] argued that timing in speech production should be a consequence of intrinsic properties of underlying units, instead of being specified extrinsically, as a separate component of the production process [cf. 25]. Such ideas about intrinsic timing have been a source of inspiration for work on a production model at Haskins Laboratories. The model is an influential component of a long-term, ongoing attempt to account for articulatory timing, kinematics and systematically-conditioned variation in speech production [cf. 29,30] in a way which is synergistic with an evolving phonological theory [3] and a theory of speech perception [19]. For this reason, and also because it is arguably the most well-developed effort of its kind, I will examine the Haskins Model (HM) as a way of illustrating some of the points raised above. I will also refer to other modeling efforts, but those references cannot do justice to the large amount of work that usually goes into production models.

My approach will be the following. I will briefly describe the HM and some initial experimental work which has been guided by the model. Then I will mention several issues raised by this work. Finally I will propose an alternative modeling approach that may have promise for the future.

3. PRODUCTION MODELING AND EXPERIMENTATION AT HASKINS LABORATORIES

3.1 The production model

Saltzman and Munhall [30] describe a "dynamical approach to gestural patterning in speech production" with which they "attempt to reconcile the linguistic hypothesis that speech involves an underlying sequencing of abstract,

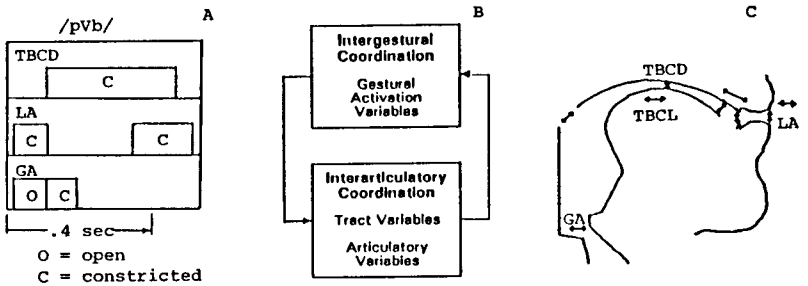

Figure 1. Components of the Haskins Laboratories production model. A: An input gestural score showing gestural activation intervals for the tract variables "tongue body constriction degree" (TBCD), "lip aperture" (LA) and "glottal aperture" (GA). B: The two-level dynamical model. C: The vocal-tract outline of the articulatory synthesizer. The arrows indicate tract variable coordinates; some are labeled as in part A, with the addition of "tongue body constriction location" (TBCL).

discrete, context-independent units, with the empirical observation of continuous, context-dependent interleaving of articulatory movements."

The fundamental, invariant unit in this approach is the abstract gesture.[1] Combinations of abstract gestures underlie phonetic segments, so in a rough sense, abstract gestures can be thought of as having a role similar to that of phonetic features. They characterize what can be done with the vocal tract, in combination, to produce speech sounds, but unlike traditional features, they are characterized by intrinsic dynamics.

Figure 1 is a schematic diagram of the major components of the HM. A "task dynamic" model (part B of Fig. 1 - [29]) is the controller for an articulatory synthesizer, which serves as the plant (part C - [27]) and produces an acoustic output. In the current stage of model development, an utterance-specific "gestural score" (part A - [3]) provides the input to the task-dynamic model in the form of a time-varying invocation of abstract gestures (each represented by a shaded rectangular pulse in one of the rows in part A of the figure). The gestural score is generated according to rules of Browman and Goldstein's Articulatory Phonology [2]. The functional sophistication and mathematical complexity of the task-dynamic model preclude a concise explanation that is also comprehensive, so the following description is necessarily oversimplified (see [12]).

In the task dynamic model (Fig. 1, Part B), there are two interacting levels. At the higher, "intergestural" level, abstract gesture combinations are specified from information in the gestural score, so they will appropriately influence vocal tract movements during the utterance. The lower, "interarticulator" level contains two sets of coordinates. The formation and release of linguistically-significant *vocal-tract constrictions*, such as lip aperture, tongue dorsum and blade constrictions (as well as constriction locations) are specified in a *tract variable coordinate system.* Articulatory movement is generated by modeling the influence of each discrete abstract gesture in the tract variable coordinate system as a time-invariant second order system (characterized as a *point attractor),* with a characteristic stiffness, damping and equilibrium point. Thus, gestural activations determine the relative timing characteristics and evolving parameter values of a dynamical system expressed in terms of tract variable coordinates.

The tract-variable specification is transformed into motions of *model articulators* such as the lips, jaw and tongue body in the (midsagittal-plane) space of

1. Confusion can arise from use of the term "gesture" to denote an abstraction. To avoid such confusion, I will use the term "abstract gesture" to denote the abstraction, and "movement" to denote physical or simulated articulatory movement.

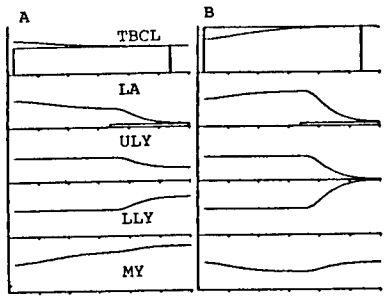

Figure 2. Example gestural activation levels and tract-variable and model-articulator trajectories for the utterances /Xib/ (A) and /Xaeb/ (B) (X = schwa).

the articulatory synthesizer, i.e., the *model articulator coordinate system.* The transformation (mapping) from tract variable to model articulator coordinates is strictly kinematic; it accounts in part for two facts: the motions of more than one articulator can be influenced by a single tract variable, and some articulators can be influenced by more than one tract variable [cf. 25].

All dynamical properties of this system reside in the controller, and biomechanical properties of the vocal tract are not represented directly. (The term "stiffness" refers to a characteristic of the abstract point attractor, and not to the actual biomechanical properties of the musculoskeletal apparatus.) In the formative stages of model development, the importance of considering biomechanics and aerodynamics is acknowledged, but bypassed for practical reasons.

In order to begin to understand some functional characteristics of the model, consider the very simple examples shown in Fig. 2. It illustrates gestural activation levels and tract-variable and model-articulator trajectories for the utterances /Xib/ (A) and /Xaeb/ (B) (X = schwa). In each panel, the row labeled TBCL contains an activation pulse specifying (target) level and duration along with a parameter value trajectory for the tract variable (abstract gesture) "tongue body constriction location"; the row labeled LA contains the same kind of information for the tract variable "lip aperture"; and rows labeled ULY, LLY and MY contain articulatory trajectories for the vertical upper lip, lower lip and mandible positions, respectively. The horizontal axis represents time. The TBCL and LA activation pulses overlap in time and have the same durations for both utterances. For the two utterances, the LA pulse has the same magnitude, indicating bilabial closure; but the different vowels (/i/ and /ae/) invoke different levels of TBCL activation. The TBCL trajectories evolve according to the activation levels and second order dynamic responses of their corresponding abstract gestures. The LA tract variable is defined with respect to the positions of the lips and jaw. Even though LA is not activated during the vowels, it has a changing trajectory which differs during the vowel portions of the two utterances because of the active influence of the vowels on the jaw (via the jaw-tongue synergy). The LA trajectories move toward closure when that abstract gesture is activated, but the rate and magnitude of movement differs, depending on the vowel-specific starting point at the onset of LA activation. For the two utterances, the pairs of ULY and LLY trajectories have the same overall shape, but different vowel-dependent rates of change and end points: since the vowel /ae/ is more open, greater lip movement is required to reach closure. The JY trajectories differ in response to the overlapping influences of the two different pairs of tract variable trajectories, since the mandible positioning is affected by both lip and tongue body positioning.

This example illustrates some (but not all) of several important characteristics of the task-dynamic model. The model accounts for: coarticulation (as "coproduction" of sequences of (partly) overlapping abstract gesture complexes), overlapping influences of multiple abstract gestures and tract variables on movements of individual articulators (as a result of "blending" of abstract gestures), and movement of articulators when they are not under active control (as governed by articulator-specific "neutral attractors" - see also, section 3.2 below).

Relative timing of the activation of the set of abstract gestures for each speech

segment and the timing of sequencing of segments is currently specified extrinsically by the input gestural score, so the goal of accounting for intrinsic timing has not yet been reached. Future development of the model will incorporate additional layers of intrinsic dynamics, implemented in the form of neural networks [30,16] for inter-gestural timing within segments and timing among successive segments in an utterance.

3.2 Use of the Haskins Model to interpret experimental data

Browman and Goldstein [3] have proposed an articulatory phonology in which (abstract phonological) articulatory gestures are the "atoms out of which phonological structures are formed"; phonological structures are hierarchical "constellations" of gestures; and phonological regularities can be captured by representing constellations of gestures in gestural scores, which can be generated by rule. Some aspects of abstract gestural representations (i.e., those specifying different articulator sets) are "categorically distinct", that is, each set defines a separate phonological category. Other aspects (e.g., location and degree of constriction, stiffness) are not distinct in this way; they are hypothetically determined by relations among production, acoustics and perception as suggested by some of the above-mentioned modeling work [33,21]. It is claimed that information in the gestural score identifies particular lexical entries; phenomena such as non-canonical pronunciations in fluent contexts, segment deletions, insertions, assimilations, etc. can be characterized by orderly modifications of the gestural score. The abstract gestures of articulatory phonology are the same as those of the task-dynamic model, so control of the model with gestural scores can be used to test "phonologically-based" hypotheses.

Browman and Goldstein [4] have used the task-dynamic model to help interpret data from an experiment on the production of the vowel schwa, motivated by the observation that schwa assumes the quality of neighboring vowels. They wanted to investigate two alternative hypotheses: 1) schwa has a specific target which is coproduced with a neighboring stressed vowel, or 2) schwa is completely unspecified for tongue position. Movements of points on the lower lip, jaw, and the blade, mid and rear of the tongue dorsum were measured for one subject using the x-ray microbeam. Utterances were of the form /pV_1pX'pV_2pX/ (where X = schwa). Analysis of tongue point displacement data suggested that the V_1-V_2 trajectory was influenced by an independent schwa target, especially as evidenced by a decrease in tongue height during the schwa when V_1 and V_2 were both the vowel /i/.

The experiment was replicated in simulations with the task-dynamic model, using several different control strategies, observing the simulated articulations and performing listening tests of the acoustic output from the simulations. The control strategy that produced the most convincing result was one in which an active gesture for the medial schwa completely overlapped the gesture for V_2 and control regimes for V_1 and V_2 didn't overlap, as proposed previously [2]. The failures of alternative schemes were instructive, particularly one in which there was no active schwa gesture, but instead a gap with no vowel gesture specified between the end of V_1 and beginning of V_2. During that interval tongue motion was due to relaxation of the tongue body to its neutral position, as well as jaw motion, called into play by the bilabial gestures for /p/. The result was a decrease in tongue height during the schwa when V_1 was the same as V_2. The decrease agreed with the x-ray data for when both vowels were /i/, but disagreed when the vowels were /a/: with /a/, the simulated tongue height decreased during the schwa, but it increased (toward a neutral configuration) in the x-ray data.

It would be possible to offer alternative interpretations of the data; an apparently "successful" simulation cannot "prove" the hypotheses of Browman and Goldstein or "validate" the modeling approach. The main point of this example is rather that it illustrates how production modeling can serve as a means of focusing experimentation. I suggest that if such efforts with the HM can progress in a productive and appropriate fashion, a large, coherent body of experimental

data will result and we will have learned more about speech production than from an equivalent amount of less-well-guided research. However, trying to do this kind of modeling work raises a number of issues; some of those issues may be specific to the Haskins approach and some of them are inherent to any similar modeling effort.

4. MODELING ISSUES AND CHALLENGES

Before considering issues raised by the Haskins work, it should be noted again that their approach is unique in its comprehensiveness and the extent to which it has been implemented and tested.

4.1 The underlying theory: how valid are its basic assumptions?

As Browman and Goldstein progress, presumably they will be conducting more listening tests to investigate alternative control schemes in the production modeling. What will they be asking the listeners to do in these tests? I suggest that Browman and Goldstein will have to deal increasingly with the issue of what are appropriate acoustic and perceptual criteria, perhaps along the lines suggested by work on lexical access. It seems to me, largely on intuitive grounds, that not enough emphasis is given to acoustic and perceptual mechanisms in an approach which places such strong emphasis on gestures. As will be discussed further below, this issue is particularly important for features with prominent acoustic correlates that result from abrupt transitions at moments of vocal tract closure and release. Such features may not be efficiently characterized in terms of abstract gestures.

It remains to be determined how Articulatory Phonology does as a phonology. As one alternative, Halle and Stevens [11] discuss a system (derived from [5,28,24]) which is also hierarchical, but has as its primitive elements more traditionally-defined features. Those features play a role in speakers' "knowledge of the language", in that they capture phonological distinctions and transformations. The features are described as falling into two categories, articulator-bound (i.e. executed by particular articulators - "labial" by the lips, "coronal" by the tongue blade) and articulator-free ("continuant", "sonorant" and "syllabic" which have robust acoustic correlates and are not tied to the action of any one articulator). This division is somewhat similar to Browman and Goldstein's division of abstract gestural primitives into those that are and are not "categorically distinct". The hierarchy of primitives in both points of view is anatomically-based. However, in contrast to abstract gestural primitives, the more traditional features discussed by Halle and Stevens are based as strongly in acoustics and perception as they are in production. One of the challenges to Articulatory Phonology will be to determine how well it can account for phonological regularities and processes such as assimilation, in comparison to other systems. A major challenge to proponents of any system is to account for the relationship between underlying representations and kinematic articulatory behavior; by its nature, Articulatory Phonology claims to incorporate that relationship.

4.2 Toward more realistic models: consonant production, aerodynamics and biomechanics

For justifiable practical reasons, the HM does not yet try to account for aerodynamical and biomechanical properties of speech production. However, a modeling approach which is based on concentrating all dynamical behavior in the controller may be difficult to adapt in the future to incorporate biomechanics and aerodynamical factors, which are critical for the simulation of consonant (and voice) production. Accounting for consonant production in the task-dynamic model and in Articulatory Phonology are likely to present some of the most formidable challenges for this work.

Regardless of the choice of modeling approach, incorporating aerodynamics and especially biomechanics is very difficult, because so little is understood about those mechanisms. Scully's [31] synthesis work, which includes aerodynamic factors, implies a large increase in number of details that have to be specified and major problems with specifying parameters at moments of consonant closure and release. The aerodynamic and resulting acoustic events which occur at those critical moments may be especially important as cues for perception and

lexical access, particularly in signaling temporal landmarks and as correlates of certain distinctive features [cf. 34]. When future models do incorporate biomechanical and aerodynamical constraints, the role of the controller in such models probably will be more representative of the type of control actually performed by the central nervous system. Presumably, that control will reflect: 1) the strategies speakers use in producing perceptually-important temporal landmarks, and 2) whether timing is extrinsic, intrinsic or due to a combination of factors which takes listeners' needs into account while conforming to internal constraints (also see Section 4.4).

4.3 Implications of incorporating intrinsic timing and introducing neural networks into production models

Saltzman and Munhall [30] observe that connectionist models can embody the knowledge constraining the performance of serial activity, including coarticulatory patterning. Jordan's recurrent network model (see [16]) can be used to define a time-invariant dynamical system with an intrinsic time scale that spans a sequence, and it has been used to simulate sequencing and coarticulation with a feature-like input. As mentioned above, future work with the HM will attempt to use connectionist models to account for inter-gestural temporal coordination within segments (for properties such as voicing onset time) and for timing of sequences of segments. This endeavor will also be extremely challenging, especially when it attempts to account for the complex timing relationships that have been observed in acoustic measurements as well as other perceptually-salient acoustic characteristics of the resulting signal. The future incorporation of intrinsic timing also raises questions about the ultimate role of currently-hypothesized representations of phonological regularity in the form of gestural scores.

4.4 Using models to evaluate data and vice versa

In the results of Browman and Goldstein [4], there are large differences between the "most successful" model output and the articulatory data. Some of these differences are unavoidable, because the model does not attempt to take into account a number of kinds of intra- and

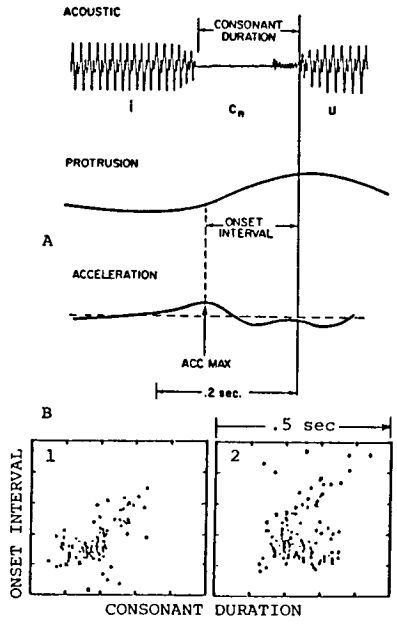

Figure 3. A: Schematic illustration of acoustic, lip protrusion and acceleration signals vs. time, illustrating two measurements, "consonant duration" and "onset interval". B: Plots of the illustrated measures for two subjects.

inter-speaker variation that are evident in actual production data. Given other important priorities, it is probably not appropriate to attempt to account for such variation in the modeling at this point. However, it will be challenging to interpret variable data with respect to a model that does not incorporate variability.

Scully's [31] synthesis of a short utterance as produced by two different speakers reveals a number of inter-speaker differences in articulatory, aerodynamic and acoustic parameters. Fig. 3 shows another kind of inter-speaker difference from an experiment we conducted on timing of upper lip protrusion movements for the vowel /u/ [26]. Part A is a schematic illustration of acoustic, lip protrusion and acceleration signals vs. time for an utterance /iC_nu/, illustrating two measurements, "consonant duration" and "onset interval". The experiment used utterances such as "lee coot" and "leaked coot" to vary consonant duration. Part B shows plots of onset interval vs. consonant duration for two subjects. For Subject 1, there was a predominantly linear relation of the two measures. For Subject 2, there was much more scatter in the data. A positive linear relation in the data can be interpreted as reflecting a constraint for the movement to begin around the time of the acoustic offset of the preceding /i/, as suggested by a look-ahead model [cf. 13]. The data containing more scatter can be interpreted as evidence for the fluctuating influence of the preceding constraint, in competition with a constraint for the lip protrusion movement to have constant kinematics, as hypothesized in the task-dynamic framework. Thus, different subjects can exhibit movement patterns that support rather different interpretations.

Such examples point out additional major challenges to production modeling: developing appropriate and objective performance metrics and model optimization procedures, as well as efficient methods for gathering, analyzing and interpreting the most useful kinds of production data. If global models are to be used effectively to evaluate experimental data in the long run, there has to be a dramatic increase in the sophistication and amount of work that compares model output with production data.

5. POSSIBLE FUTURE DIRECTIONS

5.1 Parallel-distributed-processing models (neural networks)

Jordan's work with neural networks appears to be particularly promising in its capability to account for some of the most important characteristics of speech production, while at the same time offering intuitively sound approaches to some of the difficult problems mentioned above (see [16]; also [1]). This work, consisting of two lines of research, seeks to integrate solutions to the problems of "excess degrees of freedom", serial order and learning.

As already mentioned, one component of this modeling has been used to simulate *sequencing* and *context sensitivity* (i.e., coarticulation) of speech movements with a feature-like input. This simulation is accomplished by representing actions as points in a target space which correspond to regions in an articulatory space. A trajectory is found which passes through the regions in articulatory space so that values of articulatory degrees of freedom change minimally over time. This constraint represents an interaction between the serial nature of the task and the existence of excess degrees of freedom.

The problem of using acoustic information to constrain articulatory trajectories is addressed through the mechanism of a *forward model*, which represents the second component of Jordan's work. A forward model is a learned internal model of the transformation from articulatory space to (acoustic) target space. Once the forward model has been learned, it can be used to convert acoustic errors backward into articulatory errors. Thus the system can learn to perform articulatory trajectories on the basis of iterative attempts to achieve specified sequences of acoustic targets [17].

5.2 Toward an alternative global model

The preceding material leads me to suggest that there is a need for an alternative, comprehensive production model. Such a model and the primitive elements it uses for specifying utterances should, in a balanced fashion, take into account as many aspects of the speech communication process as we think are important for a further understanding of speech production. Those aspects include: the nature of phonological regularities, control of the production mechanism, peripheral constraints on articulation and sound generation, articulatory-to-acoustic relations, relations between acoustics and perception, and mechanisms of lexical access. Clearly, such an effort has to have a long range perspective, but I believe enough of the elements exist to begin approaching the problem.

From my point of view, the most rational primitive elements are features which can be convincingly motivated in phonology and, in general, have correlates in production, perception and acoustics. As hypothesized by Stevens [34], collections of such feature specifications could serve as lexical representations. In those representations, manner

features, which have robust acoustic correlates (but are articulator free), specify temporal landmarks. Each landmark is generated by the action of a primary articulator. At the time specified by each of the landmarks, one or more secondary articulations must be coordinated to produce cues corresponding to additional feature specifications (also see [14]). Such representations could serve as the input goals for a controller, as developed by Jordan.

Optimally, the controller would operate on a realistic model of the peripheral production mechanism, which would have correct anatomical, biomechanical and aerodynamic properties. Since we currently cannot build such a model, an articulatory model like the one developed by Maeda [22] could be used in articulatory synthesis. Maeda's model has the advantage of being based on statistical analyses of the articulations of individual speakers. In order to circumvent problems of inter-speaker variation and obtain more realistic synthetic area functions, it would be helpful to obtain enough articulatory data on a few speakers to specify individual versions of such a model and then use the same speakers in subsequent experimental and model-based tests of hypothesized control mechanisms.

Until we know much more about the anatomical, biomechanical and aerodynamical detail necessary for the realistic synthesis of consonants, it may be worth considering an alternative to articulatory modeling in the generation of synthetic utterances for use in perceptual tests: the time-varying articulatory positions generated at the controller output could be used to specify the parameters of a high-quality terminal analog synthesizer [36,18].

If such a modeling effort could be realized, it would be possible to "close the loop" as is being done at Haskins Laboratories, and examine the perceptual consequences of hypothesized production mechanisms in comparison with actual production data. This approach would face many of problems outlined above, but it would be based on a balanced perspective which may be more representative of the speech communication process as a whole.

6. CONCLUSION

Before global modeling of speech production can provide us with real insight about the control of speech production, it will have to come to grips with a number of extremely difficult problems as mentioned above. Undoubtedly, the solutions to many of those problems will receive major contributions from modeling work on a variety of detailed mechanisms. Those mechanisms range from interactions among aerodynamics and biomechanics in the production of transient acoustic cues, to signal processing and feature extraction in the auditory system. In the meantime, work with global models should continue to stimulate our thinking about theoretical issues and serve as organizing frameworks for focussed programs of experimentation.

ACKNOWLEDGEMENTS

I am grateful to Cathe Browman, Nick Clements, Michel Jackson, Michael Jordan, Elliot Saltzman, Ken Stevens and Mario Svirsky for their helpful comments. Caroline Smith, Cathe Browman, Louis Goldstein and Elliot Saltzman kindly generated the plots used in Fig. 2. Preparation of this manuscript was supported by N.I.H. Grant No. CD00075.

REFERENCES

[1] BAILLY, G. (1990). Robotics in speech production: Motor control theory, Proceedings of the Tutorial Day on Speech Synthesis, Autrans, France.

[2] BROWMAN, C.P. & GOLDSTEIN, L. (1986). "Towards an articulatory phonology", *Phonology Handbook, 3,* 219-252.

[3] BROWMAN, C.P. & GOLDSTEIN, L. (1989a). "Articulatory gestures as phonological units", *Phonology, 6,* 201-251.

[4] BROWMAN, C.P. & GOLDSTEIN, L. (1989b). "'Targetless' schwa: an articulatory analysis", presented at the Second Conference on Laboratory Phonology, Edinburgh.

[5] CLEMENTS, G.N. (1985). "The geometry of phonological features", *Phonology Yearbook, 2,* 223-250.

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

[7] FANT, G. (1980). "The relationships between area functions and the acoustic signal", *Phonetica, 37,* 55-86.

[8] FANT, G., LILJENCRANTS, J. & LIN, Q. (1985). "A four-parameter model of glottal flow", STL-QPSR, 4/1985, Stockholm, 1-13.

[9] FOWLER, C.A. (1980). "Coarticulation and theories of extrinsic timing control", *J. Phonetics, 8,* 113-133.

[10] FUJIMURA, O. & KAKITA, Y. (1979). "Remarks on quantitative description of lingual articulation", in B. Lindblom & S. Ohman (eds.), *Frontiers of Speech Communication Research,* Academic Press, London.

[11] HALLE, M. & STEVENS, K.N. (1990). "Knowledge of language and the sounds of speech", presented at the Symposium on Music, Language, Speech and Brain, Stockholm.

[12] HAWKINS, S. (in press). "An introduction to task dynamics", in D.R. Ladd & G.J. Docherty (eds.), Proceedings of the Second Conference on Laboratory Phonology, Cambridge University Press.

[13] HENKE, W.L. (1967). "Preliminaries to speech synthesis based on an articulatory model", Proceedings of the 1967 IEEE Boston Speech Conference, 170-177.

[14] HUFFMAN, M. (manuscript). "Articulatory landmarks: Constraining timing in phonetic implementation".

[15] ISHIZAKA, K. & FLANAGAN, J.L. (1972). "Synthesis of voiced sounds from a two-mass model of the vocal cords", *Bell System Tech. J., 51,* 1233-1268.

[16] JORDAN, M.I. & ROSENBAUM, D.A. (1989). "Action", in D.A. Posner (ed.), *Foundations of Cognitive Science,* M.I.T. Press, Cambridge, MA, 727-767.

[17] JORDAN, M.I. & RUMMELHART, D.E. (1990). "Forward models: Supervised learning with a distal teacher", *Cognitive Science.*

[18] KLATT, D.H. (1980). "Software for a cascade/parallel formant synthesizer", *J. Acoust Soc. Am., 67,* 971-995.

[19] LIBERMAN, A.M. & MATTINGLY, I.G. (1985). "The motor theory of speech perception revised", *Cognition, 21,* 1-36.

[20] LINDBLOM, B.E.F. (1971). "Numerical models in the study of speech production and speech perception: Some phonological implications", Proceedings of the VII International Congress of Phonetic Sciences, 71-73.

[21] LINDBLOM, B.E.F. & SUNDBERG, J.E.F. (1971). "Acoustical consequences of lip, tongue, jaw and larynx movements", *J. Acoust. Soc. Am., 50,* 166-1179

[22] MAEDA, S. (1990). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model", in W.J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modeling,* Kluwer Academic Publishers, Dordrecht, The Netherlands.

[23] MACNEILAGE, P. (1972). "Speech physiology", in J.H. Gilbert (ed.), *Speech and Cortical Functioning,* Academic Press, New York.

[24] MCCARTHY, J.J. (1988). "Feature geometry and dependency: A review", *Phonetica, 45,* 84-108.

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

[26] PERKELL, J.S. & MATTHIES, M.L. (manuscript). "Temporal measures of labial coarticulation for the vowel /u/".

[27] RUBIN, P.E., BAER, T. & MERMELSTEIN, P. (1981). "An articulatory synthesizer for perceptual research", *J. Acoust Soc. Am., 70,* 321-328.

[28] SAGEY, E. (1986). "The representation of features in nonlinear phonology". Ph.D. dissertation, Massachusetts Institute of Technology.

[29] SALTZMAN, E.L. (1986). "Task dynamic coordination of the speech articulators: A preliminary model", *Experimental Brain Res., Ser 15,* 129-144.

[30] SALTZMAN, E.L. & MUNHALL, K.G. (1989). "A dynamical approach to gestural patterning in speech production", *Ecological Psychology, 1,* 333-382.

[31] SCULLY, C. (1990). "Articulatory synthesis", in W.J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modeling,* Kluwer Academic Publishers, Dordrecht, The Netherlands.

[32] SHADLE, C.H. (1986). "Models of turbulent noise sources in the vocal tract", *Proc. Inst. of Acoustics, 18,* 213-220.

[33] STEVENS, K.N. (1972). "On the quantal nature of speech: evidence from articulatory-acoustic data", in P.B. Denes & E.E. David (eds.), *Human Communication, a Unified View,* McGraw-Hill.

[34] STEVENS, K.N. (1988). "Phonetic features and lexical access", presented at the Second Symposium on Advanced Man-Machine Interface Through Spoken Language, Hawaii.

[35] STEVENS, K.N. (1989). "On the quantal nature of speech", *J. Phonetics, 17,* 3-45.

[36] STEVENS, K.N. & BICKLEY, C. (in press). "Constraints among parameters simplify control of Klatt formant synthesizer", *J. Phonetics.*

[37] TITZE, I.R. (1984). "Parameterization of the glottal area, glottal flow and vocal fold contact area", *J. Acoust Soc. Am., 75,* 570-580.