J. Acoust. Soc. Jpn. (*E*) **13**, 1 (1992)

Phonology and phonetics—A syllable-based model of articulatory organization

Osamu Fujimura

Division of Speech and Hearing Science, The Ohio State University, Columbus, OH 43210-1002, USA

A new model for describing articulatory organization is proposed, demonstrating the possibility to basically deviate from the traditional segment-concatenation coarticulation theory. This Converter/Distributor (C/D) Model is based on syllables as the concatenative units, and combines feature-based specifications of demisyllables with a novel concept of syllable and boundary strength and timing represented by abstract impulses. Consonantal gestures, superimposed upon a vocalic base function, are implemented as a combination of stored impulse response functions for individual elemental gestures, in respective articulatory dimensions associating actions with articulators. The C/D model is intended for description and explanation of natural speech, and emphasis is placed on its ability to handle articulatory variability due to a wide variety of extraphonological as well as prosodic effects according to simple principles.

Keywords: C/D model, Demisyllable, Nonlinear phonology, Phonetic implementation

PACS number: 43. 70. Bk

1. INTRODUCTION

Recent developments in liguistic theory, in particular nonlinear phonology,[1] advocate an inherehtly multidimensional organization of speech materials, often referring to more or less independent controls of the articulatory organs such as the tongue body, tongue tip, the lips, the velum and the larynx, over a relatively large organizational domain consisting of many phonemic segments (see, *e.g.* McCarthy (1988)). Each dimension (in the original autosegmental theory a segmental plane as opposed to a tonal plane, considered "melodies") has its autonomy,[2] but features that serve as elemental units to form melodies are interdependent. This interdependence is mediated by the organizational structure called "skeleton," which uses a phonemic segmental node (root) for representing functional relations among features in the form of "feature geometry."[3] Syllables are often crucial units in phonological descriptions, but most thoretical treatments introduce syllabic units in the derivational process of "prosodic organization" of segments.[4] In such theories, in contrast to the traditional (linear) phonology[5] based on the classical theory of distinctive features (Jakobson, Fant and Halle (1963)), the binary values of phonological features for each (phonemic) segment are not fully specified. Therefore, the identity of each phonemic segment is not established as its inherent value underlying the phonetically realized physical event; rather, the phonetic realizations of individual consonantal and vocalic events must depend on the context. The contextual effects are much more complex and significant than in the classical view of segment concatenation and coarticulation (see Lindblom (1968, 1990)).

Nevertheless, the coarticulation theory based on what might be identified as segmentalism, has provided the theoretical basis for virtually all discussions of speech organization in phonetics. Notably, the speech synthesis by rule using a text input, as exemplified by Klatt (1987), has demonstrated, with explicit formulation, both the adequacy and the remaining inadequacy of the segmental model.

In the current phonological literature, on the other hand, phonetic phenomena that can not be accounted for by the concatenation-coarticulation model tend to be treated by phonological rewrite rules called allophonic rules. For example, Halle and Mohanan [1985] proposed a rule for changing the "light" /l/ to "dark" /l/ in English under certain syntagmatic conditions. Sproat and Fujimura (1989, forthcoming) criticize this move, based on articulatory movement and electromyographic data under a variety of boundary conditions, and propose that such allophonic variation should be accounted for by numerical phonetic implementation rather than symbolic rewrite rules. In this connection, they indicate a need for a new theory which deviates basically from the segmental theory. Observed phonetic (particularly articulatory) facts deviating from the prediction of the segmental concatenation-coarticulation theory are found abundantly in the research literature (Keating (1988), Browman and Goldstein (1985, 1986, 1988, 1989), Fujimura and Lovins (1978), Krakow (1989), also see Kohler (1990)). Browman and Goldstein have proposed that a gestural score, a sparsely specified abstract representation, is implemented by virtue of the biological system of speech motor control as a temporally organized multidimensional pattern of events.

In short, there is no comprehensive phonetic theory that responds to this need.[6] The present article outlines an emerging theory of phonetic implementation that attempts to replace the segmental accounts.[7] Instead of phonemes, or the simultaneous bundles of distinctive features forming columns of feature matrices in phonological representations, in both lexical and postlexical descriptions, the new theory suggests a different way of representing syllables with minimal specifications of feature values, leaving many aspects of physical characteristics in speech production undetermined, until articulatory gestures are physically implemented in a specific utterance situation. Consonantal gestures are treated separately from vocalic gestures that form syllable nuclei. The flow of vowel to vowel movement forms a base function, somewhat akin to "prosodic" aspects of speech, onto which consonantal gestures are superimposed as "perturbations," as proposed earlier by Sven Öhman (1967).

The theoretical framework assumes the form of generative description, taking as its input abstract phonological and other specifications pertaining to an utterance of the linguistic material in a given situation, and deriving articulatory (and acoustic) signals as the output. A realistic 3-dimensional dynamic model of the tongue and other articulatory organs is a critical ingredient of this theory. It is assumed that the apparent violations of the superpositional principle in mapping from phonological representation to speech signals is due to the nonlinearity of the peripheral physical system, and therefore factors involved in the core parts of the theory can be treated separately in parallel. A new finite-element model of the tongue is being produced by Wilhelms and his colleagues (Wilhelms, personal communication, also see Wilhelms and Wu (1991)) as the central component of the nonlinear, large-deformation signal generating component.

The theory is designed to accommodate much utterance-specific characterization of speech phenomena, reflecting not only various prosodic characteristics within contemporary phonological and phonetic theories, such as intonation and accent, but also the expressive nature of speech utterances including speakers' styles and emotional states. A comprehensive descriptive frame work of such extra-phonological characteristics of utterances, however, is outside the scope of the current study. Some preliminary data of realistic conversational speech are being collected and analyzed by Donna Erickson (personal communication, also see Fujimura, Erickson and Wilhelms (1991)).

An empirical and computational approach to test this hypothetical model using extensive articulatory data in the form of the X-ray microbeam pellet movement patterns is being developed in cooperation with John Josephson of the Laboratory for AI Research, OSU, based on his abduction algorithms for inference with best explanations (Josephson 1987, 1990, 1991). This evaluation procedure is conceptually similar to the method of analysis by synthesis (Stevens, 1960) in modeling speech perception.

This exploration into a new descriptive framework of speech signals in relation to abstract representation is strongly motivated by, and has crucial implications for, the theoretical framework of phonological representation in lexical and (postlexical) phrasal phonology. Extensive surveys of relevant linguistic data dealing with a large number of languages, and their interpretations according to the proposed framework, must await large scale investigations. Preliminary attempts are being made in cooperation

with two colleagues with experitise in phonological theory: Cari Spring and Zhiming Bao.

2. C/D MODEL, INPUT REPRESENTATION AND PROSODIC EFFECTS

The proposed theory of phonetic implementation, which we shall call the C/D model, consists of four components: converter, distributor, a parallel set of actuators, (signal) generator. These components are serially ordered as a computational algorithm for deriving speech signals from abstract specifications of an utterance (see Fig. 1). At the output level of each component (and the input level of the next component), there is a representation of the utterance. The series of such representations for the same utterance at consecutive levels exhibits the transformation from the abstract symbolic representation to observable physical numerical variables (time functions).[8]

The input specification given to the C/D model contains a string of syllables, each syllable being identified as a linearly ordered string of syllable prefixes (henceforth p-fixes), a syllable core, and syllable suffixes (s-fixes). The string of syllables is associated by (1) prosodic structural information, *e.g.* a metrical tree,[9] and (2) extraphonological information represented by supplemental symbols, each of which is attached to a boundary symbol identifying the affected unit (syllable, word, phrase of any type, or the entire sentence, paragraph, or discourse).

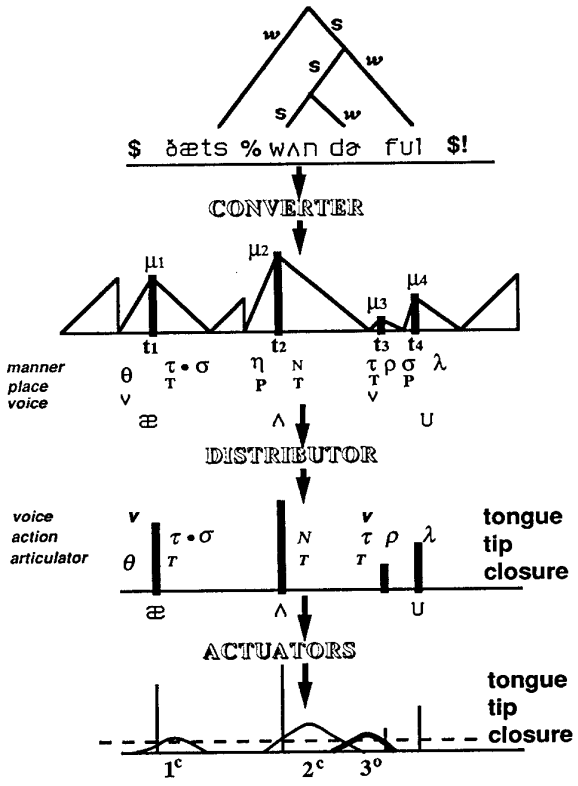

Fig. 1 C/D Model (signal generator not shown).

Syllable features and the s-rep

The syllables are represented by an unordered set of features specified for the onset, another set for the coda, another for the nucleus, and optionally unordered sets of occlusive consonantal features specified for each of the ordered syllable affixes (p-fixes and s-fixes). A common vocabulary of consonantal features are used for both onset and coda specifications, while nucleus specifications are given in terms of a distinct vocabulary of vocalic features.[10] This paper is primarily concerned with (General American) English, and in this language affixes are always a single s-fix (see below) placed to the right of the core. In what follows, the treatment of consonantal features and the concepts of syllable affixes as opposed to syllable cores are equivalent to that of demisyllables (see Fujimura (1979, 1989b), Clements (1989)).[11] Features, surrounded by curly brackets in this paper, are all privative (*i.e.* unary) and underspecified.[12]

Some English examples of the minimal specifications for each demisyllable as a whole may be useful for clarification. In both onset and coda, /g/ is specified as {dorsal, stop, voiced}, /l/ as {lateral}, /sp/ as {labial, spirant}; in onset, /kl/ as {dorsal, lateral}, /sm/ as {labial, spirant, nasal}, and in coda, /mf/ as {labial, fricative, nasal}, /nt/ as {apical, stop, nasal}; /nd/ is analyzed as /n/ {apical, nasal} in coda and +{stop} in s-fix, the latter requiring no specifications of voicing nor place by general requirements imposed on s-fixes. Generally in either initial or final demisyllable, no more than one place[13] specification is given: {spirant} for phonemic /s/ in clusters /sp, st, sk/ (both initial and final positions) and /sm, sn, sl/ (initial position only) is implemented as an accompanying pre-closure frication (which also affects the phonetic value of the phoneme /p/ in /sp/ by eliminating its aspiration, for example). Phonologically, as seen in /sl/ {apical, lateral} in opposition to /fl/ {labial, lateral, fricative}, there is no place specification for {lateral}, and the opposition {apical} *vs.* {labial} exhibits the place distinction of the entire demisyllable.

The corresponding phonetic implementations

specify articulators, such as the same tongue tip for both {lateral} (in applicable situations) and {apical} (in combination with {stop}, {fricative} {spirant}, or {nasal}). The feature subset {voiced, voiceless} constitutes a type which we may designate as "voice features"; nasal, lateral, roticized and glide (labio-velarized and palatalized) involve vocalic gestures and constitute another subset (resonant features).[14] Vocalic gestures can be involved in phonetic implementations of consonantal features, not only sonorant features such as lateral and nasal (see Sproat and Fujimura (forthcoming)), but also obstruent features (*e.g.* Japanese apical stop, fricative and nasal consonants are phonetically palatalized preceding the vowel /i/).

Prosodic structure and boundaries

It has been shown that phonological phrase structures are different from syntactic trees (see Selkirk (1984), Inkelas (1989)). There is also articulatory evidence that such phrasing effects can be observed as a locally uniform speech rate increment/decrement with discontinuous changes at the boundaries (see Fujimura (1986; 1990a, b). In the C/D model, it is assumed, tentatively, that all prominence information is incorporated, at one level of representation (within the converter), into strengths of syllables, *i.e.* syllable pulse magnitudes, and thereby syllable intervals (see below). The type and strength of adjacent boundary may interact with the syllable strength in the computation.

Intonational structure has been intensively studied (see *e.g.* Pierrehumbert and Beckman (1988); for reviews, Fujimura (1989a) and Kubozono (1992)), and is related to the issue of temporal organization (see Edwards and Beckman (1988)).

Utterance characteristics

The syllable-boundary strength pattern represented by the time-amplitude modulated pulse train (see below) is first computed based on phonological information, and then is adjusted for discourse features such as {focus} and discourse organizing principles such as intraparagraphic declination (see Hirschberg and Pierrehumbert (1986), Pierrehumbert and Hirschberg (1990)). Specifications of extraphonological expressive characteristics such as {casual} and {irritated}, as well as speaker and speaking style characteristics of the utterance, are further considered for adjusting the impulse time series. These specifications of utterance characteristics are also encoded into the set of vocalic gesture specifications in specific articulatory dimensions, for example, pertaining to voice quality control as well as mandible height control, in superposition with implementations of vocalic features as base functions.

One unique feature of the present model is that, despite all these adjustments of prosodic control, the inherent consonantal gestures maintain constant and invariable impulse response functions. This idea is motivated by the observation that certain consonantal gestures that are less affected by the complex constraints imposed by the highly nonlinear physical articulatory apparatus are observed to be relatively invariant in selected parts of the movement patterns (icebergs, see Fujimura and Spencer (1983), Fujimura (1986)).[15]

Utterance conditions vary widely. The scope of the current study is too limited to cover all these conditions, and a systematic description of structure and representation of these conditions and characteristics requires large future efforts. Two specific utterance situations are being studied (Westbury and Fujimura (1989, in press), Erickson, personal communication) in connection with the present theoretical work:

(1) Correction of erroneous digits in a dialogue verifying a street address, interpreted as the placement of focus on the corrected word and marked by a symbol "!" placed immediately to right of the left-edge word boundary "#", as in /$ noW % ɪts # faJv #! faJv # naJn # paJn # striJt $/; this symbol placed on a word is interpreted to represent a focus feature causing prominence.

(2) Repeated correction of the same error in a similar situation: /*$ noW % Its # faJv #! faJv # naJn # paJn # striJt $/. The asterisk may be interpreted as representing an emotional feature "irritated."

3. CONVERTER

The first component is the converter. It converts the symbolic representation of the prosodic and extraphonological specifications of an utterance into a temporal sequence of syllables and boundaries, each of which is represented by one impulse in a time series and is assigned a computed set of numerical values of strength (pulse magnitude μ, see Fig. 1) and timing (time of occurrence τ). Symbolic specifications of the syllable identities are (1) passed on to the next stage of derivation (distributor) and (2) interpreted by the converter identifying relevant information concerning syllable types, which deter-

mine, according to a syllable-type table, the multiplicative coefficients α's and β's of the pulse interval formula (see below).

Boundary types, reflecting the hierarchical structure of the phrasing pattern, are tentatively assumed to be (1) partly absorbed by computation into the syllable and boundary strength information, which in turn generates the timing information of each syllable impulse, and (2), as the shadow computation implies, evaluated as an add-on syllable interval (see Fujimura (1987)).[16] Part of the interval due to a phrase boundary is implemented as a pause interval, in interaction with utterance feature specifications, which may in part vary randomly in actual utterances. It may be assumed that articulatory gestures are computed regardless of this pause implementation, but intonation contours may interact with the latter (see *e.g.* Uyeno *et al.* (1981)).

The converter first evaluates the input representation and computes pulse magnitudes for syllables and boundaries. The ordered string of magnitude-specified pulses is then interpreted, and time values are assigned to syllable impulses by the following linear algorithim:

$$\tau_i - \tau_{i-1} = \alpha \cdot \mu_{i-1} + \beta \cdot \mu_i,$$

for the i-th and $(i-1)$-th syllable. The multiplicative constants α and β may be considered figuratively to determine "shadow angles" of each syllable pulse on the right and left sides, respectively (see Fig. 1).[17] A similar shadow is defined only on the left side of each boundary pulse. The basic idea is simple: if there is a strong (*i.e.*, prominent) component, it is implemented phonetically with a larger duration (see Westbury and Fujimura (1989; in press)), probably as well as a higher voice pitch and enforced voice quality and intensity. The rest of the speech material is "pushed away" in time. The timing computation based on the strength pattern is later affected by several factors through modifying the constants α and β: in particular, all the α's and β's are globally adjusted for the specific speaker and speaking style characteristics.

In Fig. 1, in the output of the converter, syllable pulses are represented by thick vertical lines, and boundary pulses by thin vertical lines. Syllable pulse strengths are denoted by μ_1, μ_2, μ_3 and μ_4, for the four syllables in 'That's wonderful,' and each of these syllable magnitudes is shown by the pulse height. Greek (with an exception of N) symbols under the shadowed pulse train represent "manner" (τ for stop, σ for fricative, N for nasal, λ for lateral, ρ for roticized; θ for interdental). The next lower row shows the pertinent articulator (in small capitals, P for lips, T for tongue tip, C for tongue blade, and K for tongue body), specified only for occlusives. The third row shows voicing of obstruents. Note that voiceless obstruents and voiced sonorants are unmarked.[18] These feature symbols are placed on the left side of the pulse for the initial demisyllable, and on the right side for the final demisyllable.

S-fixes are denoted by a manner symbol (in this case σ) preceded by a middle dot. The existence of an s-fix constitutes a distinct syllable type attribute, and warrants a more gradual righthand shadow slope. Also, the impulse response functions for s-fixes are given different parameters in comparison with those for coda.[19] In English, s-fixes are necessarily apical obstruents, and are voiced when and only when the tautosyllabic coda is voiced; therefore, the only feature specification for each s-fix is that of manner, *viz.* stop, fricative, spirant (/st/), or concomitantly stop or spirant and fricative (/ts/ or /sts/). According to this formulation, English syllables have only single s-fixes. The behavior of the morphemic 'th' as in 'sixths,' 'strengths' [strɛŋkθs] (/stre{N, K}·θ·s/ derived lexically from /stro{N, K}+θ+s/), 'warmth,' *etc.* is exceptional, and violates the voicing agreement of the s-fix also.[20]

At the bottom line of the converter output representation, vowel symbols, as an abbreviation of vocalic feature specifications, determine the inherent characteristics of the syllable nucleus. Note, in particular, that the reduced syllable (the third syllable) is indicated by a short syllable impulse, its time (τ_3) and magnitude (μ_3) being determined by the prosodic information, but is not provided with any vowel identity. A schwa in a reduced syllable is thus given a slot as a syllable nucleus without its content (see Chomsky and Halle (1968); Browman and Goldstein (1990)). The consonantal perturbation often modifies the vowel quality even at the time the peak activity of the vowel gesture is observed. Particularly, final resonant features such as diphthongization, nasalization, lateralization, roticization, accompany strong vocalic gestures[21] such as tongue body retraction and velum lowering, which, by virtue of the temporal characteristics of the impulse response functions assigned for these gestures,[22] manifest their peak effects in the central portion of the syllable nucleus in the time domain.

4. DISTRIBUTOR

The syllable feature specifications are interpreted by the distributor, and are distributed using a feature table, for relevant actuators that are responsible for the corresponding articulatory gestures. The distributor first evaluates for each demisyllable or affix the mapping from the set of specified phonological features into a corresponding set of articulatory gestures. Each gesture may be interpreted as a combination of articulator and action. The distributor passes the syllable impulse specifications unchanged on to actuators.

Articulatory gestures are constructed as follows. For each specified (if any) place feature, P, T, C, or K, a feature table provides a crucial articulator, *i.e.* the lips, the tongue tip, the tongue blade, or the tongue body, in the case of English. A concomitant (*i.e.* tautodemisyllabic) manner feature is looked for.[23] If a {stop} feature τ is identified, an articulatory dimension is established for the stop elemental gesture in the identified articulator. A {fricative} feature σ in combination with an identified articulator similarly establishes the frication elemental gesture in another articulatory dimension. A combined specification of τ and σ results in an integral affricate elemental gesture, the temporal sequencing of the articulatory actions being prescribed by the inherent parameter values of impulse response functions, generally according to the vowel affinity (see Fujimura (1975)) or sonority cycle (see Clements (1989)) principle for consonantal gestures. A {nasal} feature specification has its projection onto the stop elemental gesture in the articulator which corresponds to the concomitantly specified place feature. It also has a projection in the articulatory dimension velum lowering, which has a stored impulse response that has temporal characteristics entirely different from the response function in the oral closure dimension, even though both respond to the same syllable impulse. Similarly, a specification of the feature {lateral} has a projection in the tongue tip as the designated articulator, at least if it is in the onset, but this articulator designation is entirely achieved by the feature table, regardless of the concomitant place feature specification (consider an example such as 'plot' where the concomitant place is labial, or 'belt' where the place is apical). It also has phonetic projections in the tongue body retraction and in the tongue blade narrowing as inherent elemental gestures. Note that the impulse response function to be evoked by any elemental gesture depends crucially on the onset-coda distinction, and possibly, to a limited extent, on the concomitantly specified features. The parameter values of the impulse response functions in the table is also sensitive to the particular dialect as well as the language, and the idiosyncrasy of the speaker.

Assume an articulatory event (G, μ, τ), where G is an articulatory gesture, which represents the consonantal articulatory implementation of the entire demisyllable or the s-fix. It comprises a set of elemental gestures. For example, the gesture for the onset of 'split' comprises elemental gestures that include the following physical phenomena: frication in the tongue tip for {spirant}, the stop closure in the lips for {labial} and implicationally [voiceless], and the partial contact in the tongue tip, the narrowing in the blade, and the retraction in the body for {lateral}. In the example 'That's wonderful!' given in Fig. 1, the distributor interprets phonological feature symbols into similar but distinct symbols, which abbreviate individual specifications of articulators and actions. The distributor then delivers the pulse train of a phrasal unit (see Fujimura (1990b) for discussion of phonetic phrases) to their responsible actuators, to evoke response functions as time series in multiple dimensions.

For example, the interdental fricative [θ] is pronounced by the tongue tip (an articulator) with a protrusion (action), forming an articulatory gesture |tip protrusion|. The feature symbol θ for {interdental} stands for this articulation which is commonly interpreted as the implementation of the phonemes /θ/ and /ð/. Note that there is no specification for fricative (+continuant) for interdentals. In fact, whether interdental consonants in English are fricatives, stops, or affricates, or perhaps laterals, is debatable from an articulatory point of view. The voicing distinction may or may not affect the articulatory gesture (impulse response function) in general, and in the example of Fig. 1, this information is retained for the actuator implementing tongue tip closure. The feature table contains an entry for {θ} (in Greek) and the distributor passes the elemental gesture designation |θ| (in Phonetic) to the actuator representing [tongue tip protrusion]. In the case of /p/ *vs.* /t/, for example, the distributor combines the specifications for the manner τ and the place P or T to designate dimension [lip closure] or [tip closure].

All relevant actuators are activated simultaneously by the distributor. Variance in temporal occurrences of events in different dimensions for different gesture specifications is due mostly to the difference in response function parameters, to be observed in phenomena such as the time of peak activity relative to the same syllable impulse.

5. ACTUATORS, ARTICULATORS AND ACTIONS, SIGNAL GENERATION

Articulatory gestures to implement feature specifications are constructed for the signal generation, as complexes of concurrent elemental gestures. The gestures are linearly superposed upon each other in individual articulatory dimensions, with inherently specified weighting coefficients dimension by dimension for each feature specification, as well as (multiplicatively) an extrinsic weighting factor μ of each syllable pulse. Articulators are physico-physiological apparatus, such as the tongue tip, the lips, the mandible, the velum, and a few independent laryngeal controls, identified in the signal generator system, and each of them is related to a set of articulatory dimensions representing their anatomical associations. An elemental gesture is a combination of an articulator and an action, and is implemented by an actuator. Actuators correspond one to one to articulatory dimensions, and are computational processes, which implement elemental gestures as proper patterns of activity in proper articulators. The output time function of an actuator represents a time course of physiological (and consequently physical) activity for the given utterance.

Elemental gestures are executed by specifying time functions which represent physiological activity such as patterns of muscle contractions. The impulse time functions constitute a family of mathematical function forms, which, for consonantal gestures, generally (1) starts from the current value of the vocalic base function in the particular articulatory dimension, (2) deviates from it with a certain time constant in the manner prescribed by the specified functional and parameters, and (3) similarly returns to the base function (Fujimura and Wilhelms (1991)). In addition, one of the inherent parameters for the impulse response function of each actuator represents a temporal property roughly to be identified as (positive or negative) delay of the peak event, relative to the triggering pulse. For each articulatory dimension, the base function and the consonantal perturbation function are linearly superimposed before the nonlinearity is introduced by the articulatory system simulation.

The signal generator as a physical model computes the movement of the tongue and other parts of the articulatory system. As an example of articulatory signals, it would generate pellet movements as observed in the computer-controlled X-ray microbeam (Fujimura, Kiritani and Ishida, 1973; Kiritani, Itoh and Fujimura, 1975; Nadler, Abbs and Fujimura, 1987). Pellet movements and resultant acoustic signals would be the output signals of the model. Interaction between different articulatory dimensions takes place in the physical system, and is, in many cases, conceptually straightforward (*e.g.* between lower lip position and mandible position). An effective implementation of such principles, however, requires a fairly accurate 3-dimensional simulation model of the tongue and associated anatomical structure in combination with realistic physiological control (see Kiritani *et al.* (1976), Fujimura and Kakita (1979), Kakita, Fujimura and Honda (1985)). It is now clear that such a model also must represent dynamic properties of the system, considering inertia, as well as many factors of nonlinearity. Movement saturation causes nonlinear processes such as soft clipping[24] of the position time functions, exhibiting a flat plateau of different durations due to different syllable impulse magnitudes. The treatment of interactions between adjacent impulses may involve other types of nonlinearity. For example, simple repulsion or attraction principles between closely adjacent impulses can be implemented: the converter senses, in a form of feedback from a particular subsystem (articulator) of the signal generator, events like clashes, and makes timing adjustments of the consecutive syllable impulses to keep two activities separate in time, without affecting other inputs for different actuators (see Fujimura (1986); Browman and Goldstein (1989)). In contrast, most phenomena of articulatory merging as well as hard coarticulation (see Fujimura and Lovins (1978)) can be accounted for by the saturation in articulator position (*i.e.* soft clipping).

NOTES

[1] There are two other meanings of nonlinearity: one is the nonlinear application of ordered rules (see Halle and Vergnaud (1987)), and the other is nonlinearity of a system in the sense of not obeying the superposition principle, as extensively discussed in physics and en-

gineering literature.

[2] See *e.g.* Williams (1971) and Goldsmith (1976).

[3] See *e.g.* Clements (1985), and Sagey (1986).

[4] See *e.g.* Steriade (1982), Itô (1986, 1989), Itô and Mester (1986), McCarthy and Prince (1990), Clements and Keyser (1983). See Kahn (1976) and Borowsky (1986) for phonetic discussions arguing for syllables in phonology.

[5] Representatively, Chomsky and Halle (1968).

[6] Current phonological discussions that are based on the root node in feature geometry still assume a phonemic segment as the descriptive unit (the slot on the skeleton). Archangeli and Pulleyblank (personal communication) are attacking this basic issue using a radical approach.

[7] The proposed theory does not specify a representational framework within phonology, while it expects a certain form of its output. It does attempt to prove, however, that segmental description is not necessary at least for a coherent phonetic theory, removing phonetic support for any form of segmentalism in phonology.

[8] In comparison, a similar model based on segmentalism was described by the author (Fujimura (1967, 1972)); see also Nakata (1977).

[9] See Liberman and Prince (1977). The computation of syllable pulse magnitudes (see below) may be similar to the construction of the metrical grid in their theory. Prince (1983) argues for the use of grid without the tree. See also Halle and Vergnaud (1987).

[10] This separation of vocalic features from consonantal features deviates from current phonological literature (but see Steriade (1987)).

[11] The syllable affixes are conceptually similar to appendixes (Halle and Vergnaud (1987)) and extrametrical segments (Hayes (1982)).

[12] *I.e.* many feature values are left unspecified, from the segmental point of view, throughout the process up to the evocation of elemental gestures as impulse responses.

[13] The term "place" for a feature class is used in this paper loosely, not reflecting the distinction between place and articulator in some phonological discussions (see McCarthy (1989)).

[14] Semivowels in onset (/w/ and /y/) and coda (/W/ and /J/) are to be represented by features [labio-velarized] and [palatalized], respectively, and these two features, together with [aspirated] for /h/ and [long] for vowel elongation /H/, form a subclass of "glide features." Note that phonetic implementation, as well as phonological patterning of glide features vary grossly from language to language.

[15] Certain utterance conditions of limited types, such as interruption and emotional disturbance, may cause a temporal distortion that lies outside this description, and still can be accommodated within the current model by an introduction of peripheral modifications of the impulse response parameters, or perhaps a local modification of the time scale itself.

[16] Edwards, Beckman and Fletcher (1991)'s result may be interpreted as an indication that boundary pulses are evaluated to modify parameters of the impulse response functions themselves. It is not clear immediately if this is the case, because, due to the nonlinearity (saturation) of the signal generator, the maximum speed of an articulator's movement, for example, may depend on the syllable pulse magnitude, even when the apparent extent of the excursion is the same.

[17] Pulses are placed in time to make shadows contiguous with each other in this depiction.

[18] The mark "v" specifies that the consonant cluster in the demisyllable is entirely voiced despite the inclusion of an obstruent (stop or friction), which in unmarked cases should be voiceless.

[19] An alternative approach would be to assign a separate impulse to each s-fix, treating it as though it were a separate syllable.

[20] The author is indebted to Briony Williams for this observation.

[21] Note that there is a critical lack of uniform correspondence between vocalic features and vocalic gestures in the present theory.

[22] Such tendencies reflect general (parametrically language specific) principles governing the design of impulse response function tables.

[23] An action may be executed by a particular combination of specific muscles that results in a participation of different parts of an organ, or even more than one organ. The articulator in this theory is a neurophysiological concept rather than physical one.

[24] Softness of clipping stems largely from the three-dimensional nature of the physico-anatomical structure as well as from peripheral feedback.

REFERENCES

Borowsky, T. J. (1986). "Topics in the lexical phonology of English," Doct. Diss. Dept. Linguistics, U. Mass., Amherst.

Browman, C. P. and Goldstein, L. M. (1985). "Dynamic modeling of phonetic structure," in *Phonetic Linguistics —Essays in Honor of Peter Ladefoged*, V. A. Fromkin, Ed. (Academic Press, New York), pp. 35–53.

Browman, C. P. and Goldstein, L. M. (1986). "Towards an articulatory phonology," Phonol. Yearb. **3**, 219–252.

Browman, C. P. and Goldstein, L. M. (1988). "Some notes on syllable structure in articulatory phonology," Phonetica **45**, 140–155.

Browman, C. P. and Goldstein, L. M. (1989). "Between the grammar and physics of speech," in *Papers in Laboratory Phonology* I, J. Kingston and M. E. Beckman, Eds. (Cambridge University Press, Cambridge).

Browman, C. P. and Goldstein, L. M. (1990). "'Targetless" schwa: an articulatory analysis," Haskins Labs. Stat. Rep. **SR-101/102**, 194–219.

Chomsky, N. and Halle, M. (1968). *Sound Pattern of English* (Harper and Row, New York).

Clements, G. N. (1985). "The geometry of phonological features," Phonol. Yearb. **2**, C. J. Ewen and J. M. Anderson, Eds., 225–252.

Clements, G. N. (1989) "The role of sonority cycle in core syllabification," in *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, J. Kingston and M. E. Beckman, Eds. (Cambridge University Press, Cambridge).

Clements, G. N. and Keyser, S. J. (1983). *CV Phonology: A Generative Theory of the Syllable* (MIT Press, Cambridge, MA).

Edwards, J., Beckman, M. E., and Fletcher, J. (1991). "The articulatory kinematics of final lengthening," J. Acoust. Soc. Am. **89**, 369–382.

Edwards, J., and Beckman, M. E. (1988). "Articulatory timing and the prosodic interpretation of syllable duration," Phonetica **45**, 156–174.

Fujimura, O. (1967). "Nihongo no onsei," in *Hoosoo-bunka-kenkyuusho Nizissyuunen Kinen Ronbunshuu*, (NHK Publ. Bureau, Tokyo), pp. 363–404.

Fujimura, O. (1972). "Sooron," in *Onseikagaku*, J. Oizumi and O. Fujimura, Eds. (U. Tokyo Press, Tokyo).

Fujimura, O. (1975). "Syllable as a unit of speech recognition," IEEE Acoust. Speech Signal Process. **ASSP-23**, 82–87.

Fujimura, O. (1979). "An analysis of English syllables as cores and affixes," Z. Phonetik Sprachwiss. Kommunikationsforsch. **32**, 471–476.

Fujimura, O. (1986). "Relative invariance of articulatory movements: An iceberg model," in *Invariance and Variability in Speech Processes*, J. S. Perkell and D. H. Klatt, Eds. (Lawrence Erlbaum, Hillsdale, N. J.), pp. 226–242.

Fujimura, O. (1987). "A linear model of speech timing," in *For Ilse Lehiste*, R. Channon and L. Shockey, Eds. (Foris, Dordrecht, Holland), pp. 109–123.

Fujimura, O. (1989a). "Onsei-on'in kenkyuu no tenboo," in *Nihongo no Onsei-on'in, Nihongo to Nihongo-Kyooiku*, Vol. 2, M. Sugito, Ed. (Meiji-shoin Pub., Tokyo).

Fujimura, O. (1989b). "Demisyllables as sets of features: Comments on Clement's paper," in *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, J. Kingston and M. E. Beckman, Eds. (Cambridge University Press, Cambridge, MA), pp. 334–340.

Fujimura, O. (1990a). "Articulatory perspectives of speech organization," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds. (Kluwer Academic Publishers, Dordrecht).

Fujimura, O. (1990b). "Methods and goals of speech production research," Lang. Speech **33**, 195–258.

Fujimura, O., Kiritani, S., and Ishida, H. (1973). "Computer-controlled radiography for observation of movements of articulatory and other human organs," Comput. Biol. Med. **3**, 371–384.

Fujimura, O. and Lovins, J. (1978). "Syllables as concatenative phonetic units," in *Syllables and Segments*, A. Bell and J. B. Hopper, Eds. (North Holland, Amsterdam), pp. 107–120. (An unabridged version available from Indiana University Linguistic Club).

Fujimura, O. and Kakita, Y. (1979). "Remarks on quantitative description of the lingual articulation," in *Frontiers of Speech Communication Research*, S. Öhman and B. Lindblorn, Eds. (Academic Press, London), pp. 17–24.

Fujimura, O. and Spencer, W. (1983). "Effects of phrasing and word emphasis on transitional movements—location and stability of tongue blade iceberg patterns," J. Acoust. Soc. Am. **76**, S59 (Abstract).

Fujimura, O., Erickson, D., and Wilhelms, R. (1991). "Prosodic effects on articulatory gestures—a model of temporal organization," Proc. XIIth Int. Congr. Phonetic Sciences, Aix en Provence.

Fujimura, O. and Wilhelms, R. (1991). "Time functions for elemental articulatory events," J. Acoust. Soc. Am. **90**, 2310 (Abstract).

Goldsmith, J. (1976). "Autosegmental phonology," Doct. Diss., Dept. Linguistics & Philosophy. MIT, Cambridge, MA.

Halle, M. and Vergnaud, J.-R. (1987). *An Essay on Stress* (MIT Press, Cambridge, MA).

Halle, M. and Mohanan, K. P. (1985). "Segmental phonology of modern English," Linguist. Inq. **16**, 57–116.

Hayes, B. (1982). "Extrametricality and English stress," Linguist. Inq. **13**, 227–276.

Hirschberg, J. and Pierrehumbert, J. (1986). "The intonational structuring of discourse," Proc. 24th Annu. Meet. Assoc. Comp. Linguistics, 136–144.

Inkelas, S. (1989). "Prosodic constituency in the lexicon," Doct. Diss., Dept. Linguistics, Stanford University, Stanford, CA.

Itô, J. (1986). "Syllable theory in prosodic phonology," Doct. Diss., Dept. Linguistics, Univ. Mass. Amherst, MA.

Itô, J. (1989). "A prosodic theory of epenthesis," Nat. Lang. Linguist. Theory **7**, 217–259.

Itô, J. and Mester, A. (1986). "The phonology of voicing in Japanese," Linguist. Inq. **17**, 49–73.

Jakobson, R., Fant, C. G. and Halle, M. (1963). *Preliminaries to Speech Analysis*, 3rd ed. (MIT Press, Cambridge, MA).

Josephson, J. R. (1987). "A framework for situation assessment: Using best-explanation reasoning to infer plans from behavior," Prob. Expert Systems Workshop, pp. 76–85.

Josephson, J. R. (1990). "Spoken language understanding as layered abductive inference," Tech. Rep., OSU Lab. Artificial Intelligence, Columbus, OH.

Josephson, J. R. (1991). "Abduction: Conceptual analysis of a fundamental pattern of inference," Tech. Rep., OSU Lab. Artificial Intelligence, Columbus, OH.

Kahn, D. (1976). "Syllable-based generalizations in English phonology," Doct. Diss., Dept. Linguistic &

Philosophy, MIT, Cambridge, Mass.
Kakita, Y., Fujimura, O., and Honda, K. (1985). "Computation of mapping from muscular contraction patterns to formant patterns in vowel space," in *Phonetic Linguistics*, V. A. Fromkin, Ed. (Academic Press, New York), pp. 133–144.
Keating, P. A. (1988). "Underspecification in phonetics," Phonology **5**, 275–292.
Kiritani, S., Itoh, K. and Fujimura, O. (1975). "Tongue-pellet tracking by a computer controlled X-ray microbeam system," J. Acoust. Soc. Am. **57**, 1516–1520.
Kiritani, S., Miyawaki, K., Fujimura, O., and Miller, J. E. (1976). "A computational model of the tongue," Annu. Bul. Univ. Tokyo RILP, **10**, 243–251.
Klatt, D. H. (1987). "Review of text-to-speech conversion for English," J. Acoust. Soc. Am. **82**, 737–793.
Kohler, K. J. (1990). "Segmental reduction in connected speech in German: Phonological facts and phonetic explanations," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds. (Kluwer Academic Publishers, Dordrecht).
Krakow, R. A. (1989). "The articulatory organization of syllables: A kinematic analysis of labial and velar gestures," Doct. Diss., Yale University.
Kubozono, H. (1992). "Recent developments of phonetics and phonology," J. Acoust. Soc. Jpn. (J) **48**, 3–8 (in Japanese).
Liberman, M. Y. and Prince, A. (1977). "On stress and linguistic rhythm," Linguist. Inq. **8**, 249–336.
Lindblom, B. (1968). "On the production and recognition of vowels," Doct. Diss., Lund University.
Lindblom, B. (1990). "Adaptions of speech processes: A sketch to H & H theory," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds. (Kluwer Academic Publishers, Dordrecht).
McCarthy, J. J. (1988). "Feature geometry and dependency: A review," Phonetica **45**(2–4), 84–108.
McCarthy, J. J. (1989). "Linear ordering in phonological representation," Linguist. Inq. **20**, 71–99.
McCarthy, J. J. and Prince, A. (1990). "Prosodic morphology and templatic morphonology," in *New Perspectives on Arabic Linguistics* II, M. Eid and J. McCarthy, Eds. (John Benjamins, Amsterdam), pp. 1–54.
Nadler, R. D., Abbs, J. H., Fujimura, O. (1987). "Speech movement research using the new X-ray microbeam system," Proc. 11th Int. Congr. Phonetic Sciences, Tallinn Estonia, 1 (Paper Session 11.4), 221–224.
Nakata, K. (1977). *Onsei* (Corona Publ., Tokyo).
Öhman, S. E. G. (1967). "Numerical model of coarticulation," J. Acoust. Soc. Am. **41**, 310–320.
Pierrehumbert, J. and Beckman, M. (1988). "Japanese tone structure," Linguist. Inq. Monogr. **15** (MIT Press, Cambridge, MA).
Pierrehumbert, J. and Hirschberg, J. (1990). "The meaning of intonation contours in the interpretation of discourse," in *Intentions in Communication*, P. R. Cohen, J. Morgan, and M. E. Pollack, Eds. (MIT Press, Cambridge, MA).
Prince, A. (1983). "Relating to the grid," Linguist. Inq. **14**, 19–100.
Selkirk, E. O. (1984). *Phonology and Syntax* (MIT, Cambridge, MA).
Sproat, R. and Fujimura, O. (1989). "Articulatory evidence for the non-categorization of English /l/-allophones," Linguist. Soc. Am. Annu. Meet., Dec. 27–30, 1989, Washington, D. C.
Sproat, R. and Fujimura, O. (forthcoming). "Allophonic variation in English /l/ and its implications for phonetic implementation."
Steriade, D. (1982). "Greek prosodies and the nature of syllabification," Doct. Diss., MIT, Cambridge, MA.
Steriade, D. (1987). "Locality conditions and feature geometry," NELS Meet. Handb. **17**, 595–617.
Stevens, K. N. (1960). "Towards a model for speech recognition," J. Acoust. Soc. Am. **32**, 47–55.
Uyeno, T., Hayashibe, H., Imai, K., Imagawa, H., and Kiritani, S. (1981). "Syntactic structures and prosody in Japanese: a study on pitch contours and the pauses at phrase boundaries," Annu. Bul. Univ. Tokyo RILP **15**, 91–108.
Westbury, J. and Fujimura, O. (1989). "An articulatory characterization of contrastive emphasis in correcting answers," J. Acoust. Soc. Am. **85**, Suppl. 1, S98(A).
Westbury, J. and Fujimura, O. (in press). "Articulatory correlates of contrastive emphasis in correcting answers in English," in *Speech Perception, Production and Linguistic Structure*, Y. Tohkura, Y. Sagisaka, and E. Vatisiotis-Bateson, Eds. (Ohm Publ., Tokyo).
Wilhelms, R. and Wu, C. M. (1991). "Modeling the tongue with finite elements," J. Acoust. Soc. Am. **90**, 2100 (Abstract).
Williams, E. (1971). "Underlying tone in Margi and Igbo," PhD. Diss. Dept. Linguistics and Philosophy, MIT.
Note: Some of the doctoral dissertations in linguistics are published by Garland Publ., New York.