Cad.Est.Ling., Campinas, (43):9-36, Jul./Dez. 2002

TEMPORAL ORGANIZATION OF SPEECH UTTERANCE: A C/D MODEL PERSPECTIVE

OSAMU FUJIMURA
(The Ohio State University, Dept. Speech & Hearing Sc.)

RESUMO *Este artigo discute o modelo C/D na qualidade de quadro teórico lingüístico visando a descrever as características temporais de enunciados com referência à organização prosódica geral. Um trem de pulsos sílaba/fronteira, com magnitudes controladas, representa o esqueleto da função de base de um enunciado, determinando completamente sua organização métrica. Contornos vocálicos, tonais e demais contornos fonéticos representam a melodia da função de base. O padrão temporal de sílabas individuais é calculado pela distribuição de suas magnitudes, levando-se em consideração a intervenção de fronteiras com magnitudes controladas. A magnitude da sílaba é realizada como duração juntamente com outras propriedades fonéticas tais como um componente de abertura da mandíbula e forças incrementais de gestos vocálicos e consonantais. Além disso, a constituição de um padrão prosódico envolve um controle tonal independente. A fonologia lexical pode especificar características acentuais (magnitude da sílaba), tonais ou ambas, dependendo da língua, mas toda língua usa as duas características ao nível da frase. A natureza inerentemente dinâmica da fala é fundamental nesta nova abordagem.*

1. THE C/D MODEL

In this paper, we first summarize some basic characteristics of the C/D model as a descriptive framework of utterance representation. We then discuss how this model represents the metrical organization, *i.e.*, rhythmic structure of an utterance, as a stress modulation pattern. Independently from stress control, tonal (voice pitch) control, which is physiologically implemented as laryngeal adjustment, manifests both lexical and phrasal phonological feature specifications. In languages like English, default voice pitch changes often reflect the stress pattern without specific tonal control. According to the C/D model, stress control, unlike voice pitch control, automatically and directly accompanies changes in syllable duration. Jaw opening also reflects the syllable magnitude, but it reflects other phonologically controlled properties of the syllable as well. Magnitudes of boundaries that intervene in the syllable string also contribute to the phonetic metrical pattern of the utterance.

Within each syllable, the temporal organization of articulatory and phonatory gestures is quantitatively computed, given the system parameters, set to reflect the utterance situation. Thus, at the output of the Converter, the skeleton of the utterance is represented by the syllable-boundary pulse train. This skeleton is associated with the melodic specification of phonological features.

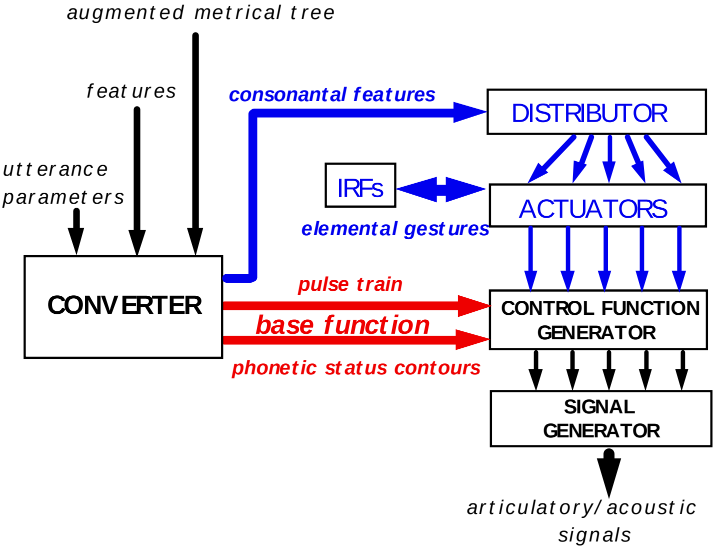

Fig. 1: Block diagram: the C/D model.

There is no role for phonemic segments in this theory. Phonological features pertaining to syllables in the lexical representation, along with features characterizing phonological phrase structure, are interpreted by the Converter (Fujimura, 1992; 1994) (see Fig. 1). Gesture implementation proceeds syllable by syllable, according to the magnitude of each syllable. Phonetic phrases are organized to incorporate the gestural effects of each boundary, according to its magnitude, into the string of syllables, to produce control time functions in individual physiological dimensions. The control functions reflect discourse-conditioned utterance characteristics as well as phonological properties of phrases. Once the string of syllables and boundaries is implemented as the control functions, incorporating mutual temporal relations among syllable gestures, the syllable boundaries cease to be identified (Leben, 1999). Acoustic signals often exhibit discontinuities, due to the inherent nonlinearity of mapping from articulatory control variables to acoustic signal parameters (Fujimura, 1990). These acoustic or spectrographic discontinuities have been interpreted as the boundaries of traditional phonemic segments, but timing of such manifestations varies greatly depending on the prosodic context. There is no synchronization of articulatory and phonatory gestures at such acoustic discontinuities except for incidental physical interactions, *e.g.*, between the supraglottal pressure change and the state of the vocal folds.

1.1. Base function

The base function of an utterance consists of a skeleton, reflecting the syntagmatic organization of the linguistic form used, and a melody, reflecting its paradigmatic phonological information. Both are continuously variable phonetic control functions that reflect prosodic conditions for the particular utterance and discourse. The syllable-boundary pulse train that represents the skeleton is converted to a temporal sequence of syllable triangles with inserted boundary (half) triangles.

Each syllable is associated with syllabic (such as vocalic and tonal) gestures, forming a syllabic segment of the set of phonetic status contours of the base function at an abstract level of phonetic representation. Phonetic status contours hold assigned static (target) values within each syllabic segment, switching from one value to another simultaneously at an abstract syllable boundary. There may be syllables that are not assigned any target value for a given gesture dimension due to the lack of a pertinent phonological feature specification. In this case of underspecification, the current version of the C/D model assumes two possible treatments: (1) assigning a neutral value to the syllabic segment, which may be interpreted as a resting state of the muscles involved for the unspecified control variable, or (2) interpolating the variable between the ending edge of the preceding syllable and the beginning edge of the succeeding syllable. Except for the interpolation treatment of underspecified syllables for some of the status contours, the abstract melodic contours of the base function are step functions of time.

Phrasal adjustments modify the local and global properties of this string of syllables, producing a more concrete form of utterance representation. As one of the phrasal adjustments of control functions, each pseudo-step function is transformed into a smooth continuous time function through an application of a filter with a prescribed step response function. Each dimension of the control function has its inherent step response function, which may be different for phrase onset and phrase offset at a particular phonological phrasal level. This smoothing process corresponds to the traditional concept of coarticulation (Lindblom, 1963). However, the process in general is more complex than coarticulation and can be described quantitatively by step response functions of more than one type, involving the mathematical process of convolution integral in time (Fujisaki & Hirose, 1982). Note, however, that smoothing takes place at the level of physiological control functions and the smoothing characteristics (prescribed as a step response function) vary among different control dimensions dealing with articulatory phenomena. Therefore, the resultant control functions do not reveal simultaneous changes corresponding to the underlying syllable boundaries. The smooth control variables often produce more or less abrupt changes in mechanical movement and acoustic signals, due to the inherent nonlinearity of signal generation processes, as seen in articulatory implosion and explosion, or voice onset and offset.

Vocalic and tonal feature specifications for syllables determine the main aspects of the melody of the base function. Mandibular movement manifests metrical syllabic control mixed with its inherent gestures reflecting phonetic effects of syllable features, vocalic and tonal. Both articulatory and phonatory gestures along with special temporal manipulations may also pertain to boundaries, as observed in Japanese sokuon (obstruent gemination, roughly), which is phonologically specified in the lexicon by means of a special syllable concatenator (Fujimura & Williams, 1999). In some special cases, given a particular language, the phonetic implementation process may produce an epenthetic syllable, as suggested by Williams (*in press*) for a Spanish complex obstruent (spirantized) manner feature in word-initial position. Morpheme, word, and phrase boundaries also affect laryngeal and supralaryngeal control over extended time domains beyond syllable boundaries, and their manifestations may be observed beyond the bounds of the pertinent unit. In addition, phrasal units generally manifest global phonetic characteristics such as tonal and articulatory declination.

The step response to the intersyllabic switching of a phonetic status contour may well implement a movement in the middle of either the preceding or the succeeding syllable. For example, a dynamic pitch that cannot be described by consonantal elemental gestures may be linked to the edge of the syllable. It may also produce temporal non-monotonic change of the signal property, deviating from the traditional concept of coarticulation. Thus, for example, the slight pitch rise before the characteristic pitch fall in Tokyo Japanese when the lexical pitch accent is implemented (Poser 1984) can be described by a non-monotonic step response function. Such a movement behavior around the syllable boundary may be described as an inherent property of the syllable boundary (Hayata, 1997) that follows the syllable to which accent (kernel) is traditionally assigned (Hattori, 1961). This account amounts to assuming an accent-specific syllable concatenator (Fujimura & Williams, 1999). The step response for articulatory vocalic gestures may also show a tendency of a temporal return toward the rest position around the syllable boundary, discussed as a “trough” effect by Lindblom *et al.* (2003). This effect can also be treated as an implementation of a default property of the syllable boundary, a default characteristic of syllable concatenators.[1]

To summarize, the output information of the Converter is divided into three types. One is the phonetic skeleton represented by the magnitude-controlled syllable-boundary pulse train. The second represents the melody, the information that may be interpreted to represent a generalized concept of prosodic information. This output form deals not only with the traditional suprasegmental information (Lehiste, 1970), but also with vocalic and mandibular (possibly also velic, depending on the language) aspects of articulation. This type of information about the utterance is represented by the multidimensional phonetic status contours of the base function. The third type of information pertains to consonantal perturbation gestures that occur locally around syllable margins as discussed in the next subsection.

[1] Note that Lindblom *et al*. (2003) call a vowel sequence in more than one syllable, if it does not contain any onset consonant, a diphthong.

1.2. Margin gestures

The third type of information provided by the Converter at its output is consonantal perturbation, as proposed by Öhman (1967). Consonantal features specify margin properties of individual syllables in the lexicon, and they are implemented as local perturbation functions superimposed onto the control functions computed for the base function in the pertinent physiological dimensions. The local functions representing consonantal gestures are an assembly of elemental gestures. Prototype time functions are stored in a table as impulse response functions (IRFs) that represent inherent characteristics of individual elemental gestures, such as the apical stop, using the tongue tip/blade, and the labiodental fricative, using the lower lip. An IRF varies in its function shape, including its inherent peak amplitude and peak timing relative to the excitation pulse.

When the Converter identifies a set of consonantal features within each syllable component, it determines which elemental gestures are to be implemented by which articulators. The Distributor accordingly assigns pertinent specifications to one or more of the Actuators, each representing an elemental gesture, for specific articulatory implementation. The implementation of each elemental gesture is performed by selecting the pertinent replica (pocs pulse) of the syllable pulse for the syllable component, onset, coda, or each syllable affix (p-set or s-fix). This pulse is temporally displaced outward from the syllable pulse by an interval directly proportional to the syllable magnitude. It excites the selected impulse response function (IRF), so that the consonantal gesture is created with an appropriate amplification, reflecting the syllable magnitude, and at the inherent time, relative to the occurrence of the excitation pulse, in the utterance.

For example, an utterance in isolation of the word ‘kit’ /kIt/ in English reflects a phonological representation of the lexical item. It contains the onset specification {dorsal^{O}, stop^{O}}, where the superscript O indicates that the feature specification pertains to the onset component of the syllable[2]. The Distributor, based on the identification of this onset, transmits this elemental gesture specification, along with the syllable number in the utterance, to the pertinent Actuator that handles an onset dorsal stop.

Note that our phonetic implementation system is language-specific; for English, if the phonological specification is {labial^{O}, stop^{O}}, then a bilabial stop will be implemented in onset; if the feature specification is {labial^{O}, fricative^{O}}, then a labiodental fricative is implemented automatically.[3] Such phonetic implementation detail is part of the property of each elemental gesture, as stored in the IRF table, as

[2] The voiceless feature is unmarked for obstruent manners in the current version of the C/D analysis.

[3] In our feature system, we assume that English has a place specification {coronal} for palatals along with {labial, apical, dorsal} for obstruents. The {coronal, stop} specification, for both onset and coda in English, is implemented as the affricate [tS] as in ‘church’ and, if there is a concomitant specification {voice}, as [dZ] as in ‘judge’.

the proper local control function. The following simple example illustrates the computational procedure, according to the current version of the C/D model.

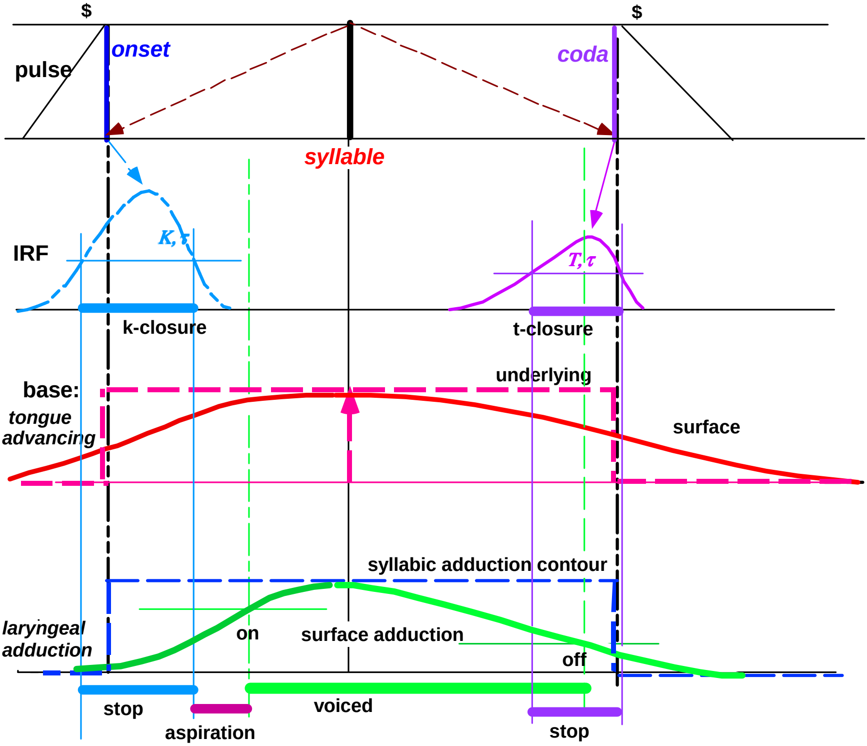

Fig. 2*:* ‘kit’ (CD diagram)

First (see the top panel), a syllable triangle is constructed around the syllable pulse for the core of the syllable. The shadow angle[4] between the left or right downward dashed arrow and the vertically constructed syllable pulse at the center is fixed for all the syllables within a certain discourse domain of the utterance. This triangle defines a “core duration” for the syllable, by determining the left and right edges, *i.e.*, onset and coda edges, respectively. Second, an onset pulse is erected at the onset edge, and a coda pulse at the coda edge, both with their magnitude copied from the syllable pulse. This process is replicated for each syllable pulse in the utterance. Third, the onset and coda pulses, respectively, excite the pertinent IRFs to implement

[4] For the syllable type of our example, the triangle is symmetric.

abstract onset and coda elemental gestures (curves labeled K, τ and T, τ, for dorsal and apical stops, respectively), as shown in the second panel.

The curves depict the local control function for the movement of the crucial articulator, *i. e.*, the tongue dorsum (vertical position, roughly) and tongue tip/blade, respectively. The gesture for each stop consonant, one in syllable onset and the other in syllable coda, is a rising and returning ballistic movement. These local time functions depict abstract control functions as though they represented the position of, say, the center of gravity of the articulatory organ, for explanatory purposes. As the articulator moves up, the curve crosses a certain threshold value depicted by a horizontal bar, which is an indication of the time of contact of the surface of the articulator with the roof of the mouth. The same horizontal bar is used for the stop release in the descending movement returning to the base position, which is where the articulator should be without consonantal perturbation, according to the nucleus to nucleus movement of the base function. The two threshold crossing points of each curve thus, figuratively, indicate the moment of stop closure and stop release, respectively. The time interval, marked by a thick horizontal bar labeled k-closure and t-closure, respectively, can be interpreted as the stop closure duration of the onset and coda consonantal gestures.

The articulator, let us say the center of gravity of the dynamically effective part of the tongue in each consonant, continues to move up after the tongue blade surface completes the stop closure of the vocal tract. Similarly, when it returns after attaining the peak position, it keeps moving down before the closure release. The peak position is not directly observed and it varies, depending on the force of articulation (Malecót, 1955): it reflects the syllable magnitude. Note that the gesture curves in the CD diagrams are meant to show control functions or their underlying motor commands, which may be expected to pattern most closely as muscle activity such as state of contraction. Electromyographic recordings of muscle activity show that the force generated within the muscle is itself a smooth function of time. The position of a flesh point of the articulator to be observed, for example as the pellet position in the microbeam data, is not directly represented in the diagram, but the threshold position bar suggests a saturation of such an observed surface position as it collides with the roof of the mouth. An exact statement of the movement representation must be based on a quantitative simulation of the signal generation process, such as proposed by Wilhelms-Tricarico (1995).

We synthesize control functions as physiological time functions, superposing consonantal elemental gestures syllable by syllable and articulator by articulator, *i.e.*, each independently coordinated set of muscles for a specific phonetic purpose. It should be noted, however, that the mapping between feature specification and gesture implementation is complex and is certainly not one-to-one. This is true even for the vocalic *vs*. consonantal categorization (see Sproat & Fujimura, 1993 for discussion).[5] After this process of control function generation, the phonetic implementation process is a highly nonlinear computation of speech signals (see Fig. 1). This last stage of the C/D model can be interpreted as a computational simulation of natural signal generation by the biological and mechanical speech apparatus (see Fujimura, 1998 for some relevant discussion).

1.3. Features and gestures

Fig. 2 also shows, in the lowest panel, how the voicing contour may be depicted. The syllable /kIt/, specified as {dorsal^O, stop^O, apical^C, stop^C, front, high} is voiceless at both onset and coda, since both margin feature specifications are interpreted as obstruent consonants and there is no specification of {voice} (see Fujimura & Williams, 1999). This phonological specification of the syllable, in combination with the assumption of isolated utterance as the context in this figure, results in a delayed adduction of the larynx as the onset step response to the laryngeal adduction control function’s upward step (broken curve labeled “syllabic adduction contour”). The associated smoothly curved line (surface adduction) is the step response for this control dimension, including both the beginning adduction and ending abduction of the vocal folds. Quite different levels of oscillation threshold height, representing adduction and abduction, are drawn to indicate when the vocal fold oscillation should start and when it should cease as functions of glottal approximation, under an assumed transglottal pressure, which is determined by the utterance condition and the discourse context of the phrase. The on and off threshold values of glottal approximation for vibration are considerably different because of the strong hysteresis observed for each vocal fold oscillation cycle. Based on the threshold crossing points in the upward and downward changes of the glottal width curve, voice onset and offset times, and thereby voicing time interval, are suggested in the figure. Note that the articulatory threshold crossings (next to top panel) and voicing threshold crossings (bottom panel) occur independently, resulting in a considerable aspiration period. In particular, the voice onset time, as typically observed in phrase-initial voiceless stops in English, corresponds to the discrepancy between articulatory release and voice onset. A similar discrepancy may be observed between the t-closure and voicing cessation toward the end of this utterance, which would be observed as a voicing continuation into the stop closure. A quantitative computer simulation of the vocal fold vibration process (Titze, 1994) would account for more details, including continuous changes in voice source signal characteristics, rather than the approximation by a discrete on-off switching as shown here.

[5] Even though it is plausible to assume that the extrinsic tongue muscles are used primarily for vowel articulation and intrinsic muscles for consonantal gestures, as Öhman suggested in the 1960’s, the phonology-phonetic mapping does not have to be one-to-one.

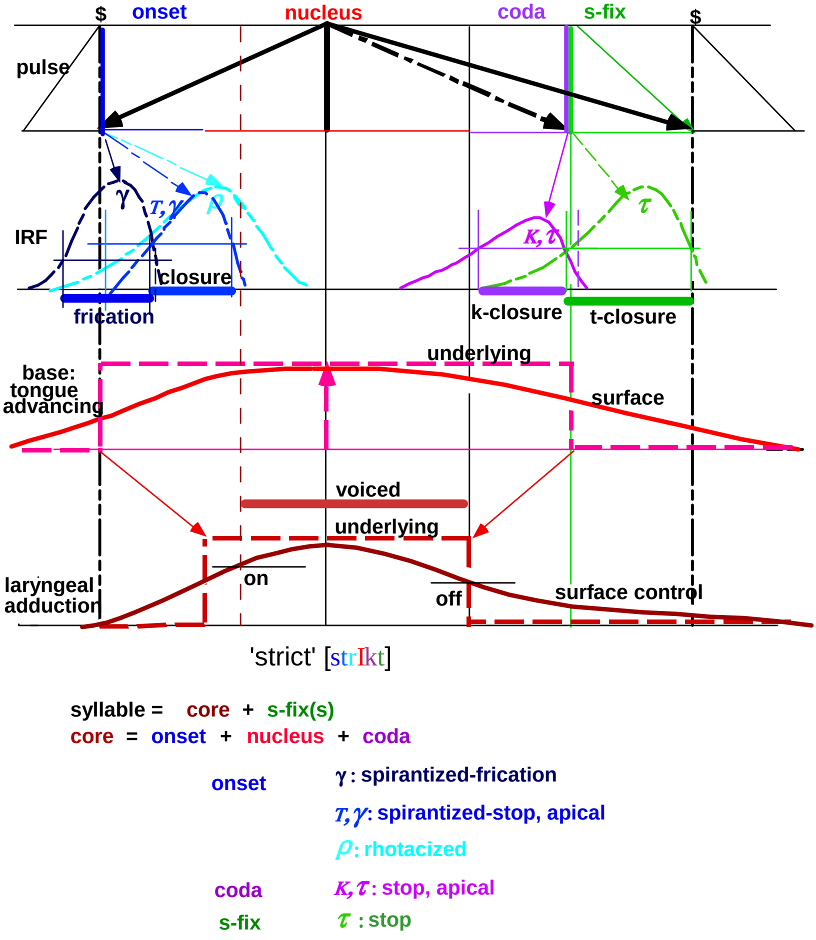

Fig. 3: ‘strict’ (CD diagram)

However simple or complex a gesture cluster may be, the temporal relations among elemental gestures are assumed to be fixed within each syllable component (onset, coda, or each of the syllable affixes, if any), regardless of the syllable magnitude. For example, the onset of the syllable /strɪk.t/, for the word ‘strict’, involves a set of elemental gestures: frication |γ^{O}| and apical stop closure |T^{O}, γ^{O}| for {spirantized ^{O}, apical^{O}}, and rhotacization |ρ| for {rhotacized^{O}} (second panel from top, Fig. 3). These elemental gestures are all evoked by the same onset excitation pulse, according to the unordered set of features {spirantized^{O}, rhotacized^{O}, apical^{O}}.[6] Their peak gestures occur in a certain temporal order, as approximately represented by the phonetic transcription [str], as the result of the inherent properties of their IRFs, rather than by any extrinsic order specification. Both the frication and the apical stop are the manifestations of the obstruent feature pair, manner and place, *viz.* {spirantized, apical}. This situation is similar to the nasalization (velum lowering) and labial stop closure occurring concomitantly but asynchronously (see Krakow, 1999), implementing the phonological feature pair {nasal^{O}, labial^{O}}, and particularly, {nasal^{C}, labial^{C}}, in 'mom'. The frication elemental gesture represented by | γ^{O} |, in 'strict', is different from the apical fricative |T^{O}, σ^{O}|, but the former is assumed to be always the same tongue tip/blade frication gesture, regardless of the concomitant place feature (for example, {spirantized^{O}, labial^{O}} for 'spy' and {spirantized^{O}, dorsal^{O}} for 'sky'). In other words, all the temporal characteristics of this complex set of consonantal gestures are (at least to the current order of approximation) an automatic consequence of designing impulse response functions of individual elemental gestures independently, given the language, dialect, speaker idiosyncrasy, *etc.*

The timing of voice onset, relative to the onset pulse, is determined according to the step response function that is evoked by the phonetic status contour producing the voicing control function. Like articulatory closure onset and offset times, actual voice onset time depends on the syllable magnitude. Given the syllable magnitude value, there is a rather limited difference between the temporal span of the IRF for {stop} and that for the set of gestures as a whole for {spirantized}. The difference is due to the difference in IRFs, not the assigned time interval for the onset part of the syllable. The result is, for the complex gesture frication + stop closure for [sk], the k-closure must occur later for [spirantized} than for {stop] (as a matter of designing the IRFs) to accommodate the preceding frication without being completely obscured acoustically. While the closure period in [sk] is generally shorter than in [k] (as the properties of the IRFs), the control function for voicing is basically the same for the two cases like 'can' and 'scan'.[7] The predicted result is that the time interval between the stop release and voice onset is shorter (less aspiration) for 'scan' than for 'can'. The phonetic status contour is characterized by an off-to-on switching of the step function for laryngeal adduction/abduction control. This switching time, relative to the syllable core edge, is not sensitive to the complexity of the onset feature specification but is determined only by presence or absence of the onset feature {voice}.This property of the voicing contour does not change even when a concomitant feature (like {rhotacized} in the example 'strict' above) is included. Consequently, there is considerable devoicing of the acoustic segment for the liquid in onset when it is accompanied by a voiceless

[6] For a new view of phonological representation of syllables without order specifications, see Haraguchi (1999).

[7] We are assuming that the syllable onset (the left edge of the syllable triangle) is set regardless of the gesture content of the onset. In some cases, particularly concerning the final syllable margin, different syllable types may set the edges of the syllable core differently, manipulating the shadow angle (see below).

obstruent, as in 'tie' *vs*. 'try'. The r-gesture in 'try', which is assumed to be the same as in 'rye', must overlap largely with the stop closure-release gesture in time, and a large part of the r-gesture must be contained within the unvoiced part of the syllable.[8]

The temporal relations between the onset and coda gesture complexes vary depending on the syllable magnitude (Fujimura, 2000b). In terms of gesture sequence, the temporal order of elemental gestures reflects the IRF characteristics of the individual elemental gestures. Generally, an elemental gesture that is considered more sonorant exhibits peak activity closer to the center of the syllable, *i.e.*, further away from the onset or coda pulses inward, compared with less sonorant gestures. Coda sonorant gestures, including nasals,[9] have more widely spread activity inward than corresponding onset gestures, often showing the peak activity well within the spectrographic vowel portion of the syllable. This implies that the temporal ordering of peak activities for a concomitant set of consonantal gestures is inverted between onset and coda, as seen in the contrast between /sleJ/ and /els/ for /l/ and /s/. In terms of segmental phonotactics, the less sonorant obstruent occurs more toward the edge of the syllable than the more sonorant lateral consonant. In terms of the C/D model, the frication elemental gesture that is evoked by the obstruent feature {fricative} exhibits its activity peak closer to the excitation pulse than the sonorant gesture for {lateral}. The spirantized obstruent events are not inverted to conform to this sonority cycle principle (Clements, 1990), as seen in 'skat' [skæt] *vs*. 'task' [tæsk]. Thus we see an opposition between 'ask' [ask] {spirantized^C, dorsal^C, low}, *vs*. 'ax' /ak.s/ {stop^C, dorsal^C, fricative^S, low}.[10] Such minimal opposition with respect to segmental ordering could not be possible if the sonority cycle principle were observed, however the definition of sonority might be given as an inherent phonetic property of the phoneme. The manner feature {spirantized} is commonly seen as a s-fix (only /st/ since s-fixes are always implemented with an apical place feature in English) in words like 'next' /nek.st/ {nasal^O, apical^O, stop^C, dorsal^C, spirantized^S, front}.[11]

[8] As a second approximation detail, there may be some "repulsion" of concomitant gestures, resulting in "spilling over" of some of the onset events into the time interval of what might be considered the "nucleus region". Conversely, an articulatory clash in signal generation may cause an earlier stop release for [t] in the presence of the rhotacization gesture, resulting in a longer aspiration for 'try' than for 'tie'.

[9] Nasal stops are both sonorant and obstruent (see Sproat & Fujimura, 1993).

[10] Note that these feature specifications are complete for each syllable using an underspecification scheme. Note also that we are here talking about American English, therefore, the vowel for 'ask' and 'ax' are the same ([ae] in the contentional phonetic transcription, /a/ in the phonemoidal transcription [Fujimura & Erickson, 1997]). In this paper we use the phonemoidal transcription in slashes.

[11] English does not have a voice specification for s-fixes. The only feature specification for an s-fix is one of the obstruent manners: stop, fricative, or spirantized. The phonetic status contour switches automatically to |unvoiced| if the coda feature specification contains an obstruent manner and there is no {voice} specified. The voicing phonetic status value extends its final value (for coda) into the s-fix portion. In languages where {voice} can be specified for syllable affixes, voicing status can switch according to phonological specification within a sequence of affixes, as long as the voicing contour principle is observed.

Syllable affixes are implemented outside the coda or onset, their elemental gesture amplitudes being determined according to the same syllable magnitude. For each s-fix, for example /t/ or /s/ in ‘acts’ /ak.t.s/ {stop^{C}, dorsal^{C}, stop^{S1}, fricative^{S2}, low}, the IRF as its proper elemental gesture is excited by its time-shifted syllable pulse: the s-fix2 pulse is erected outside the s-fix1 pulse (same as the coda pulse) sequentially to make the half triangles contiguous to each other.[12] In English, more than one s-fix can occur, when the syllable involves morphemic suffixes, as in ‘acts’ and ‘lends’ /len.d.z/ {lateral^{O}, nasal^{C}, apical^{C}, stop^{S1}, fricative ^{S2}, front}). There is no extrinsic order specified for a string of consonantal segments within each syllable component.[13] The temporal pattern of gesture overlapping (see Browman & Goldstein, 1992 for some relevant discussion) also is determined quantitatively by the impulse response functions of individual elemental gestures, according to the current version of the C/D model.

This margin gesture implementation scheme assumes some innovation of the phonological feature system, deviating from the traditional Jakobsonian concept and subsequent feature geometric conventions (Sagey, 1986) (see also McCarthy, 1988). Non-obstruent features are not associated with any place specification. It is assumed, universally, that voicing occurs in the phonetic status contour without any internal break within any syllable (voicing contour principle, Fujimura & Williams, 1999. See also Haraguchi, 1999).

1.4. Phonology and phonetics

The same feature specification may evoke different phonetic events (muscle activity patterns) depending on the phonetic context, particularly the syllable component in which it is implemented. For example, English lateral gesture in coda may not employ the tongue tip gesture at all, which robustly characterizes the lateral in onset. As observed in many studies (see Krakow, 1999 for nasals), sonorant gesture implementations are generally variable in many ways. Despite variability of phonetic events, phonological and phonetic resyllabification processes (Borowsky, 1986) seem to support the identity of the feature that moves without changing its phonological identity between heterosyllabic, but temporally adjacent, coda and onset.

Phonetic gestures are implemented by specific articulatory mechanisms based on human anatomy and physiology. While speech characteristics in many ways pertain to the perceptual properties of the speech signals as acoustic events, it is not to be denied that signals must be produced to be heard. We must consider and understand how speech production works, in order to be able to represent speech phenomena with effective generalization from a phonetic point of view. It should also be emphasized

[12] Multiple syllable affixes are numbered inside out as s-fix1, s-fix2, *etc*.

[13] At most one place feature is assumed to be specified within each syllable component in English and many other languages. Apparent exceptions in some languages are being investigated.

that speech production is performed based on pre-established linguistic and other social communicative conventions. Such conventions obviously reflect phonetic constraints, pertaining to both production and perception. It may well be the case that the design of such phonetic systems of language reflect inherently biological motivations, including respiratory principles and rhythmic patterns of the human body. However, describing the phonetic principle and process as they are, a synchronic linguistic description as the C/D model attempts, is a separate issue from explaining how speech or language evolved in anthropological history.

Physical and physiological constraints must be excluded as much as possible from the phonological description. In order to approach this goal, we need to incorporate the physical process with its inherent properties as the medium for describing speech phenomena, separate from the functional description of how such mechanisms are selected and controlled for characterizing phonological distinctions. The information about phonetic characteristics of speech, of course, is based in part on the phonological function of phonetic units. Phonology deals with patterns of oppositions among different linguistic forms. If issues pertain to non-distinctive differences of signals, describing those phenomena, in principle, should not belong to a phonological representation, either lexical or postlexical. What are called allophonic rules should not be handled by phonology if phonetics can handle them. A syllable-based phonetic representation can handle them more effectively than representations based on phoneme-size segments because the context specification for allophonic variation is largely contained within each syllable. By designing the feature system for minimal contrasts in the domain of the syllable, rather than phonemic segments as autonomous units, a much less redundant representation can be obtained without imposing *ad hoc* constraints. We could not do this before because we tacitly had to assume that phonetics must be universal and must represent phonemic segments as autonomous units in any intrasyllabic, as well as intersyllabic, context.

The representation of utterances by the C/D model in a generative descriptive format is, conceptually, a logical continuation of generative phonology, as Chomsky and Halle (1968) discussed in their *Sound Pattern of English*. Given a more powerful phonetic implementation model, however, generative phonology can transfer much of the description to phonetics. The concept of systematic phonetics is controversial (Fujimura, 1970). Whether the method of description is generative or constraint-based, the representation scheme is the basic issue (Fujimura, 1996). The C/D model maintains a strict distinction between phonology and phonetics. However, we must acknowledge that phonetics must be different for different languages. Once we accept this language specificity[14], then there are a number of issues that do not pertain to phonology as a pattern of distinctive opposition, but can be handled in phonetics coherently and more exactly.

[14] Language specificity of phonetics can be considered a numerical parameter setting of the system.

2. METRICAL STRUCTURE OF SPEECH

As a different approach for understanding the temporal organization of speech, the concept of isochronic or quasi-isochronic organization of some phonetic units has been most recently discussed in various forms of biological oscillator models (see *inter alia*, Barbosa (2002) for such a system involving more than one unit). The C/D model does not assume an organization of a string of syllables according to a constant pace of rhythm, particularly if the language uses stress control. Instead, we assume a temporally progressive formation of an array of contiguous syllable triangles and boundary gaps with variable sizes. The base length of each syllable triangle represents an abstract syllable duration, which, in our current model, is simply proportional to the syllable magnitude, given the syllable type (see below). At this level of description, both syllables and boundaries have numeric magnitudes, and the hierarchical categories of phrases are considered only as phrasal features for implementing their paradigmatic effects rather than categorical distinctions in a syntagmatic organization.

The boundary gaps are represented by triangles (half triangles) of a fixed angle, directly relating a boundary strength (magnitude represented by a boundary pulse height) to the temporal gap length between the consecutive syllables. This gap may be interpreted as an abstract pause duration, even though, concretely, there may be some acoustic signals.

Phrase-final elongation is a well-recognized phenomenon pertaining to boundaries (Lehiste, 1980), independent from syllable stress patterning. In addition to the boundary magnitude, phonological features of phrases (often associated with their edges) may have to be considered in phonetic implementation of gestures (see Sproat & Fujimura, 1993). The abstract pause, representing each boundary half triangle, may appear in speech signals in different forms: an interval of complete silence, a weak phonation or articulation “spilled over” from the preceding syllable by a prolongation of some of its gestures by parametrically affecting impulse response functions, or an expansion of the local time scale of the entire set of control time functions due to phrase-final lengthening[15].

The principle that the metrical organization of an utterance is represented as a linear string of syllable-boundary pulses would imply that, apart from the insertion of occasional boundaries, all the temporal characteristics of speech signals simply reflect the control of syllable magnitudes. Given that syllable triangles are similar, *i.e.*, the shadow angles are constant throughout a pertinent discourse domain, the temporal patterning of an utterance would be the same if the stress pattern is the same, regardless of what syllables are used. However, we assume[16] that different syllable types are associated with different shadow angles (in effect) of the triangles that

[15] See Byrd & Saltzman (2003) for a recent study based on task dynamics concepts. Fujimura (1987) discusses a phrase-hierarchical elastic model of duration equilibrium, where boundary units can be inserted as additional units.

[16] This is one of the refinements to the original version of the C/D model (Fujimura, 1992; 1994a).

represent the whole syllables (Fujimura, 1994b; *in press*). For example, extra syllable weight, as determined by the set of coda features in the phonological specification of the syllable, can add some additional duration assignment for the syllable, either by manipulating the right-hand angle of the syllable triangle, or by adding a supplemental half triangle as is done for s-fixes.[17]

In any case, the point is that the distribution of syllable (and boundary) magnitudes throughout an utterance is not assumed to be governed by any principle of regular repetition of metrical units directly in our model. However, the computation of syllable magnitudes can reflect some general principle in a manner compatible with the C/D model, with respect to its input specification scheme. At the same time, this consideration based on the C/D model also implies that, if there is no stress control, and if only one syllable type is used, the temporal organization of speech should be quite regular and isochronic with respect to syllables. In Japanese, for example, there is no stress control within the lexicon. Apart from the use of stress and phrasal boundaries in postlexical phrasal implementation, therefore, speech of this language exhibits unmodulated regular rhythm. The traditional recognition of morae as the temporal units (Hattori, 1961) reflects the fact that this language uses two syllable types, shorter and longer, depending on whether the syllable coda is used or not. The coda is traditionally considered a moraic consonant.[18] The time interval assigned to the coda component, however, is typically considerably shorter than the short syllable.

This computation of the temporal organization as described above may be interpreted to be “opportunistic” if it is viewed as a speech production model of the progression of one syllable after another. There, the timing of the next syllable is determined by the abstract syllable duration (which is dependent on the syllable magnitude) of the preceding syllable. Certainly it is not automatically produced with a fixed pace of isochronic units such as feet.[19] But the syllable-boundary magnitude distribution pattern must be given as the prescription of the utterance one way or another, as part of the input to the C/D model. This (phonological or phonetic) input information itself may be constrained somehow to conform to such a temporal pattern of the resultant control signals. Barbosa’s data (see Barbosa, 2002; Barbosa & Arantes; *in press*) seem to suggest such a prescriptive constraint. In our current model, the average syllabic rate over a discourse unit is controlled as one of the system parameters that sets an overall or average speed of utterance. Different types of

[17] Browman and Goldstein (1988) provide evidence that the temporal organization principle is different between syllable initial and syllable final consonantal clusters. Provided that the syllable final structure is adjusted for syllable types, this finding may well be compatible with the assumption of the C/D model.

[18] There are cases of trimoraic but monosyllabic onomatopoeic expressions and loan words, typically a combination of a vowel elongation or palatal glide and nasal in coda, *e. g.*, /koHN/ (corn) and /kaHN/ (the sound of a large bell). Loan words with the palatal glide often form disyllabic words in Japanese, *e. g.*, /ko-iN/ (coin).

[19] The effects of syllable magnitude or durational adjustments due to syllable types are not absorbed within each unit, to which a fixed duration is assigned; if they were, a isochronic rhythm would result (see Fujimura, 1987; also Nooteboom, 1997 for relevant discussion).

discourse-specific modifications of the syllable prominence pattern, such as magnitude enhancement and boundary manipulation due to focus, are also implemented within the phonetic process according to the input specification, in addition to the overall system parameter setting.

Our theory of phonetic implementation is not necessarily a model of real-time speech production: rather, it is a description of how a given utterance is organized in its phonetic description. How utterances in general are cognitively planned given a discourse situation is a different issue. Sternberg *et al*. (1988) suggested, based on their psychological experiments, that the motor program was produced prior to starting the execution of speech utterance for chunks roughly the size of stress groups. Speech utterances are not executed syllable by syllable (much less phoneme by phoneme) on the fly. The use of impulse response functions that are not physically realizable, allowing the impulse response to extend into the future, is justified on this ground, if the model is used to simulate the actual speech production process.

One more notable point is that tonal processes are independent from the stress pattern computation. Even though, statistically, pitch is strongly correlated with stress in English (Bolinger, 1958) as well as many other languages, stress control is independent from tonal control, as discussed elsewhere (see Fujimura, 2001; *in press*). Therefore, the description above about stress should not be interpreted by itself as an account of what is generally called the intonational patterns of speech.

By default, syllables are placed next to each other in a temporal sequence, unless there is a boundary between them. According to the underlying juxtaposition of syllabic units, the control function generation process produces smoothed time functions for physiological control of anatomical structures in the base function. The local control functions (IRFs) for consonantal gestures are smooth functions to begin with. Furthermore, the inertia involved in the physical movement of each articulator adds, in signal generation, mechanical smoothness to the time functions representing the positions of effective centers of gravity of anatomical substructures.

The operation of syllable concatenation may deviate from the direct juxtaposition of adjacent syllables. This is observed either when there is a phonologically specified special syllable concatenator or when a special concatenator is introduced within phonetics according to the given phonological and phonetic environment (Fujimura & Williams, 1999). Of course, gesture modification is not limited to a part of a syllable or a syllable as a whole. According to phonological and para-phonological properties of phrasal units, phonetic gestures can modify an extended, as well as local, temporal domain of the phrasal units[20], producing specific gestural characteristics associated with phonological phrase boundaries. Such speech characteristics constitute prosodic effects, but their implementations can be characterized by articulatory or phonatory gestures as well as temporal modulation.

The phrase-final pitch rise, as in yes-no questions, is an example of tonal (as opposed to stress) phrasal gestures in English, whereas English has no tonal

[20] Note that an utterance of a word in isolation always has accompanying phrasal characteristics.

specification within the lexicon. It should be emphasized, however, that in many utterance situations, a voice pitch change in so-called stress languages like English often reflects the stress pattern (Ladefoged, 2001). In many utterance situations, as observed in phonetic laboratory experiments, lexical stress patterns are implemented as respiratory control accompanied by mandibular maneuvers. This control pattern automatically manifests itself, for physiological and physical reasons, as default modulations of voice fundamental frequency. In usual room acoustic situations, the F0 variation and durational modulation may be the most robust physical variables reflecting stress control. Variation of voice quality, such as source signal spectral envelope (Pierrehumbert, 1989; Fujimura, Cimino & Sawada, 1995) and formant frequencies also change considerably and meaningfully (Erickson, 2002; Menezes, 2003) conveying syllable prominence control (see also Laver, 1980).

3. SYLLABLE MAGNITUDE EFFECTS ON ACOUSTIC SEGMENTAL DURATION

When we control prosodic patterns by altering degrees of prominence of different words within an utterance, *e.g*., by attaching contrastive emphasis to a selected word, the durational patterns of acoustically defined segmental durations change drastically. Generally, an emphasized word, or typically its head syllable with main stress, expands its durational appearance in the spectrogram, but the expansion is not uniform over time. Roughly, the temporal expansion concentrates its effect on what is acoustically observed as the vowel duration, but consonantal durations, whatever the definition may be, also are affected. There are cases, well known in various languages, where the phonetic appearance of the consonantal segment, rather than the vowel segment, is totally deleted, when the syllable is weakened in the phrase. In some cases, consonantal or vowel segments, spectrographically, are created where there are no phonological specifications (epenthesis).

These phenomena are sensitive to prosodic conditions. The variation, when quantitatively observed, is continuous, manifesting different numerical degrees of syllable magnitude. If we assume a certain nonlinear process for signal generation, succeeding the linear process of gesture implementation in the abstract phonetic process up to the point where we generate control functions (see Fig. 1), we can predict quite intriguing patterns of what should be observed acoustically in the form of traditional segmental duration patterns as the syllable magnitude is manipulated (Fujimura, 2000b). Also, this line of thought seems to be useful in describing vowel reduction and deletion, observed widely in various languages including English. Leben and Fujimura (2001) (also see Leben, 1999) discuss some observed facts and phonetic interpretation pertaining to what are called extra-short vowels in Kwa languages.

As an exercise to appreciate possible effects of syllable magnitude on acoustic segments to be observed in spectrograms, we can make some simple assumptions about the signal generation process and try some quantitative predictions (Fujimura, 2000b). The jaw opening is increased when the syllable magnitude is large due to prominence of the syllable, given a particular vowel (Erickson, 2002). The tongue surface tends to be lowered, due to this mandibular effect of prosodic control, regardless of the inherent phonetic manifestation of phonological vowel height. However, increasing syllable magnitude also results in an exaggeration of the inherent phonetic properties of the vowel quality, in the form of deviation from the neutral vocal tract configuration. Consequently, when syllable prominence is increased due to stress, the following effects would be observed. If the vowel is phonologically a low vowel, the tongue surface is lowered, partly due to larger jaw opening and partly due to the enhancement of the inherent low tongue surface. If the vowel is phonologically high, the two effects of increased syllable prominence tend to cancel each other, since the jaw opening is made larger and the tongue surface elevation is exaggerated. By assuming a high vowel, therefore, we can minimize this syllable magnitude effect on vowel articulation and concentrate on temporal properties of the syllable constituent gestures (see Macchi, 1988 for discussion of a similar case with respect to lip opening).

Signal generation by speech organs exhibits a three-dimensional deformation of the tongue shape due to its contact with the roof of the mouth and other walls of the oral cavity. The approximate constancy of the flesh volume of the tongue results in a rather complex effect on the tongue shape. As a crude approximation[21], let us assume a simple saturation effect of the articulator's displacement due to hard wall contact (an opposing soft articulator, as in the case of bilabial stops, has basically the same property due to approximate symmetry). In other words, we assume that a flesh point representing the tongue tip (or blade) surface position moves upwards in the first part of the ballistic motion for an apical stop elemental gesture. Then it is completely stopped by collision at some position, *viz*. the articulatory threshold for stop closure (see Fig. 2 and Fig. 3). The threshold position can vary from speaker to speaker, depending on where the flesh point under consideration is, *etc*. In any case, the movement of a flesh point (pellet) used in the experiment may be interpreted as a reflection of the impulse response function as excited by the onset or coda pulse. We assume a CVC syllable being uttered in an appropriate context and we do not consider any phrase boundary effects explicitly here, assuming that they are minimized by selecting an appropriate phonetic context.

[21] See Wilhelms-Tricarico (1995) for a mathematical method (Legendre's undetermined multiplication coefficients in tensor computation) exactly considering the volume constancy constraints.

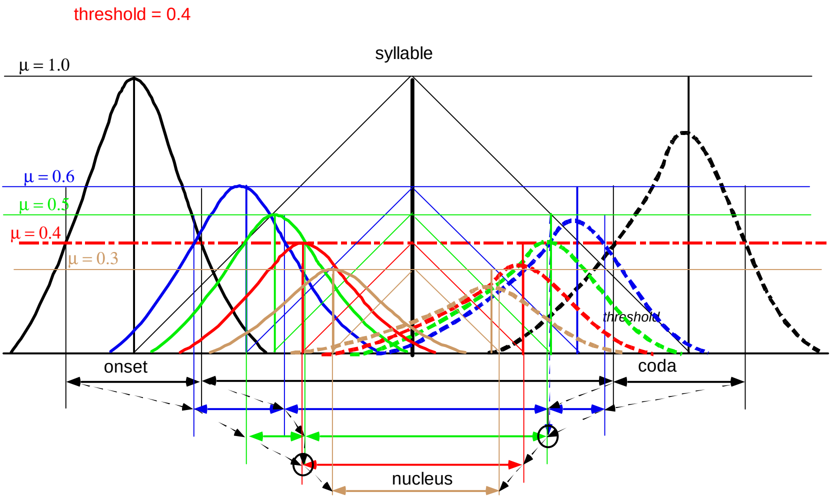

Fig. 4 (a): Syllable magnitude effects on acoustic segment durations (higher threshold)

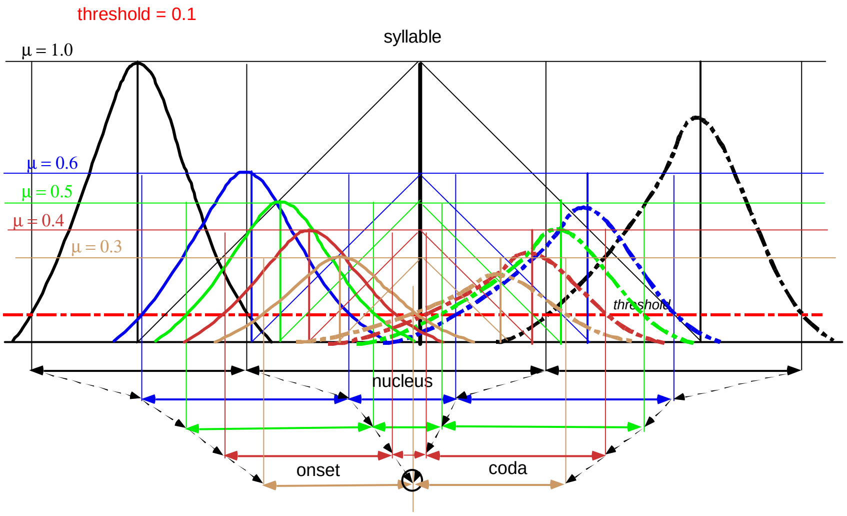

Fig. 4 (b): Syllable magnitude effects on acoustic segment durations (lower threshold)

Fig. 4 (Fujimura, 2000b; c) shows the effect of changing the relative value of the articulatory threshold height for the crucial articulator of a stop consonant in onset and coda positions. The threshold value is one of the system parameters, which are assumed to be constant for the utterance paradigm of the putative experiment. The impulse response functions for the stop consonants in onset and coda are fixed, say, both for the apical (voiceless) stop /t/.[22] It can be seen that the changing pattern of stop closure intervals at the beginning and ending part of the syllable duration shows the crucial role played by the threshold position relative to the flesh point excursion during the utterance of the same syllable associated with different syllable magnitude μ. When the threshold is set relatively high (0.4 in arbitrary scale in Fig. 4 (a)), a progressive reduction of the syllable magnitude (in the same scale as μ) from 1.0 to 0.3 results in a gradual shortening of the stop closure periods. First the coda consonantal duration disappears at $\mu = 0.5$ (see the circle showing the collapsing double-headed arrow to the right), and then at $\mu = 0.4$, the onset consonant disappears acoustically. The vowel duration becomes shorter but it is not as severely affected by the weakening of the syllable as much as the consonants. Fig. 4 (b), setting the threshold position at a lower value (0.1), keeping everything else the same, changes the durational behavior drastically. Now the vowel disappears and the consonants survive. The vowel duration is the remainder of the syllable duration, as directly determined by the syllable magnitude according to the C/D model, after subtracting the stop closure periods as shown in these figures.

4. BOUNDARY MAGNITUDE MANIPULATION IN CONTRASTIVE EMPHASIS

In Fig. 5, two utterances are compared: the reference utterance without correction (thick triangles) and an utterance correcting the first word ‘nine’, in the same phrase ‘nine five nine’. The experimental task of the speaker was correcting an error of the key word (digit in the street address) repeatedly in simulated dialogues between the subject and the experimenter (from Blue Pine data, see Erickson, 1998; Erickson *et al.*, 1998; Mitchell *et al.*, 2000; Menezes *et al.*, 2002; Menezes, 2003). The digit sequence in this example was taken from the sentence: 'I live at nine five nine Pine Street.' The thin-lined triangles represent another utterance of the same sentence in the same dialogue set; in response to the experimenter's erroneous shadowing, a correction was made on the first digit. Time on the abscissa is in decisecond (100 msec). The ordinate is jaw opening interpreted as the syllable magnitude,[23] in an arbitrary scale, the zero corresponding to closed jaw. The horizontal double-headed arrow, under the row of syllable triangles, shows a gap of about 20 msec between the first and second digits (thin triangles), which was introduced by the correction of the first digit. The syllable

[22] There is some complication about interpreting the articulatory variable here as representing the vertical coordinate value of a flesh point: the roof of the mouth is curved and the action against the hard palate of the tip of the tongue is not necessarily vertical (nor normal to the palatal surface). Such details can be discussed effectively only when we actually compute physical and acoustic consequences of physiological control using a realistically complex three-dimensional simulation (Wilhelms-Tricarico, 1995).

[23] In order to exclude the effect of the vowel on jaw opening, all key words, ‘five’, ‘nine’, and ‘Pine’, in this experiment contained the same low-vowel nucleus /aJ/.

magnitudes were estimated according to the jaw opening maxima. The symmetric syllable triangles were drawn with the same shadow angle from the top of each syllable pulse, with the angle chosen to be the largest without causing any overlapping of the triangles anywhere in the dialogue. The resultant gaps between consecutive triangles are interpreted to represent the magnitude of boundaries.

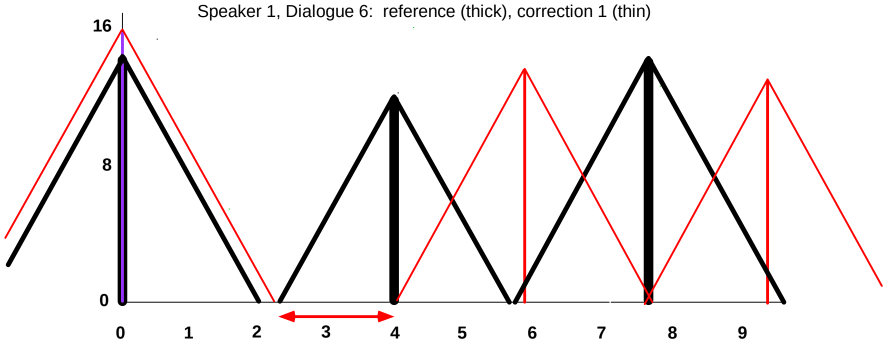

Fig. 5: An example of boundary insertion due to contrastive emphasis

The strategy for implementing contrastive emphasis varies from utterance to utterance and from speaker to speaker. Sometimes syllable magnitudes are controlled primarily, and sometimes boundary magnitudes are manipulated also (as in this example). The boundary inserted has a continuously variable magnitude, and it can be inserted before or after the emphasized syllable (Menezes, 2003). Whether such strategic variation can be completely captured by the skeletal variable represented by the syllable-boundary pulse train remains to be examined quantitatively over extensive conversational data. Voice pitch and voice quality control may play a crucial role in conveying communicative meaning in many conversational situations, particularly when the utterance involves some emotional expressiveness. Menezes (*ibid.*) shows that jaw movement and temporal modulation, both governed by syllable magnitude control according to the C/D model, are more robustly related to the intended and perceived correction than F0 modulation patterns in the Blue Pine data.

5. SYLLABLE MAGNITUDE EFFECTS ON VOWEL QUALITY

In Fig. 6, a hypothetical utterance of ‘It’s an echo.’ is depicted as a CD diagram, showing the effects of syllable magnitude control on vowel articulation. It is found, by observation of articulatory movement patterns, that contrastive emphasis, placed on a word in a sentence, affects vowel quality for different vowels (Erickson, 2002). This finding is consistent with the assumption of an enhanced jaw opening for more prominent syllables, and it also shows that the tongue (and lip rounding) articulation that implements the inherent gestures for the vowels are significantly different when the syllable is emphasized. A plausible interpretation based on the C/D model is that, while the jaw is made more open when the syllable magnitude is increased, the inherent vowel articulation, relative to the given mandible position as a deviation from the neutral vowel gesture, is also enhanced, not just because of emphasis as such, but generally as a function of syllable magnitude.

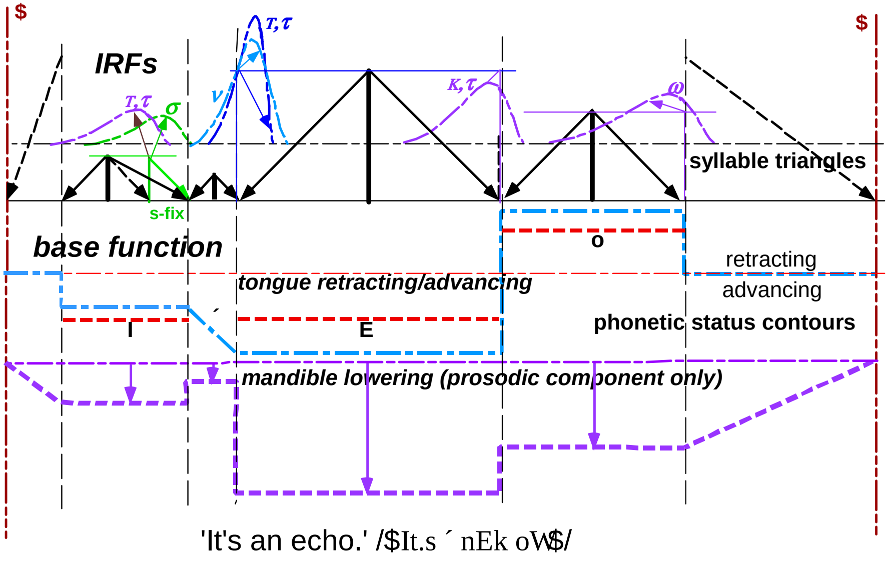
$
$
IRFs
T,τ
σ
ν
T,τ
K,τ
ω
syllable triangles
s-fix
base function
o
retracting
advancing
tongue retracting/advancing
I
E
phonetic status contours
mandible lowering (prosodic component only)
'It's an echo.' /$It.s ´ nEk oW$/

Fig. 6*:* Change of vocalic gestures due to syllable stress

Fig. 6 shows an utterance with normal intonation and focus placed on the word ‘echo’ in ‘It’s an echo.’ The top panel illustrates the elemental gestures. The coda nasal for the indefinite article ‘an’ is assumed to be resyllabified, phonologically, into the onset of the first syllable of ‘echo’. The row of syllable triangles depicts the stress pattern of this utterance. The thick dot-dash line in the middle panel shows the nucleus-to-nucleus flow of vowel gestures as a phonetic status contour of tongue retracting/advancing in the form of a quasi-step function for the succession of syllabic gestures. Underlying this pseudo-step function contour, as shown by the horizontal thick broken line segments, there are target vowel gestures that implement the phonological feature (back/front) specifications of syllable nuclei. We interpret these inherent vocalic gestures as vowel-specific deviations from the resting position (the thin horizontal dot-dash line in the middle panel). These underlying target gestures, as proper deviations from the resting position, are enhanced or reduced as seen in the phonetic status contour (thick dot-dashed line), according to the prosodic condition in the utterance. Specifically, we assume that the excess or shortage of the syllable magnitude relative to a reference magnitude level (thin dot-dashed horizontal line across the triangles) determines how much the vowel-proper deviation from the neutral reference condition should be. In our figure, the distance between the two thick lines (dotted and dot-dashed) for each vowel is made proportional to the incremental magnitudes, which are shown in the syllable triangle panel by the tips of triangles appearing above or below the reference level (thin dot-dashed horizontal line). Reduced syllables are assumed to have no specification for the inherent vocalic feature, and therefore phonetic status is not assigned; the phonetic status contour linearly interpolates the tongue retracting/advancing gesture value for the unspecified syllabic segment.

The “non-segmental” component (in the traditional sense) of mandibular gestures for syllables, the vowel-inherent component set aside, is indicated by downward arrows in the bottom panel (see Macchi, 1988 for a relevant discussion, dealing with a similar situation of labial articulation). This downward arrows in the bottom panel represent the effect of syllable magnitude to jaw opening, the arrow lengths replicating the syllable magnitudes in the triangles (pulse height). In addition, when the vocalic feature {low} or {high} is implemented by proper extrinsic muscle contractions, mandibular height adjustments occur along with the tongue height gesture according to the feature specification. The phonetic status contour for mandibular position reflects these two contributions, but this figure does not show this additional vowel specific component.

The phonetically modified gesture control functions are then subjected to coarticulatory smoothing (not shown here), producing the articulatory control functions in the different physiological dimensions. The smoothing characteristics, represented by the step response functions, vary from dimension to dimension.

Exact evaluations of speech behaviors under a variety of utterance circumstances must await extensive data interpretation with the use of an advanced and sufficiently accurate computational simulation of the signal generation process (Wilhelms-Tricarico, 1995). When such computation is achieved, the predicted patterns can be compared with observed data, articulatory or acoustic, to test the validity of the theory.

6. CONCLUDING REMARKS

The C/D model, departing from the basic assumptions of the traditional segmental model of speech organization, assumes syllables as the minimal concatenative units. Within each syllable, phonological features specify the identity of that syllable that contrasts it with others in a given language’s linguistic forms. Phonological features are specified for the syllable as a whole or for each of the margin components, onset or coda of the syllable core, or each of the syllable affixes, p-fix or s-fix, as may apply.

An utterance is organized by its base function and superimposed consonantal gestures. The base function of an utterance comprises the skeletal structure, exhibiting the metrical pattern, and melodic gestures, associated with each syllable. Vocalic and tonal gestures along with mandibular movements exhibit the melody, which is represented by phonetic status contours in individual control functions for signal generation by a coordinated set of speech production mechanisms. Each phonetic status contour is a step function in time except for occasional interpolating ramp segments for syllables for which the target value of the particular control dimension is not phonologically specified. This abstract phonetic contour, switching from syllable to syllable in its value in each of multiple dimensions simultaneously at each syllable boundary, is like the concatenated stream of segments in classical phonology and phonetics, except that the units are syllabic, not phonemic, segments.[24] The pseudo-step functions elicit dynamic step function responses, which vary in their temporal characteristics from dimension to dimension, resulting in asynchronous movements from one syllable to another. Phonetic phrases are formed, implementing the temporal sequences of syllable-to-syllable changes of target values, modifying the melody sequences for global or edge-characteristic phrasal patterns, including various tonal and articulatory declination as well as other discourse-governed utterance characteristics.

Local consonantal gestures are superimposed onto smoothly implemented control functions in individual control dimensions. They comprise elemental gestures retrieved from a stored inventory of impulse response functions, each representing, basically, a ballistic movement pattern as a temporal deviation from the base position of the crucial consonantal articulator. The temporal characteristics, including the location of the peak activity relative to the excitation pulse and spread in time, either before or after the peak activity, vary greatly from gesture to gesture. There are qualitative as well as quantitative (including the choice of the articulator) differences between syllable onset and coda. This dynamic picture of consonantal gestures basically distinguishes the C/D model from the traditional segmental representation of speech, elaborating Öhman's concept of consonantal perturbation. The phonemic model (including all phonological theories except articulatory phonology) depicts speech as a quasi-static phenomenon, switching from one target state to another, regardless of whether the segment is a vowel or a consonant. The prevailing acoustic theory of speech production and perception also is largely based on the quasi-static segmental assumption (Stevens, *in press*). Human perception, however, is more sensitive to temporal changes than to static characteristics of sound. The C/D model deviates from the basic concept of such traditional descriptions of speech phenomena, providing a powerful phonetic description with a dynamic view.

ACKNOWLEDGMENT: This research was supported in part by collaborative research agreements with Dr. Kiyoshi Honda at ATR International, Kyoto, Japan and with Dr. Masaaki Honda at NTT Basic Research Laboratories, Atsugi, Japan. The author benefited from discussions with many research associates and students at The Ohio State University. He would also like to acknowledge the participation of J. C. Williams in this study and related research work and the lasting support of Yoh'ichi Tohkura and Aki Fujimura for the development of the C/D model.

[24] Eleonora Albano (*in press*), in her novel approach, addresses the issue of nucleus-to-nucleus interaction by examining lexical statistics of Brazilian Portuguese.

REFERENCES

ALBANO, E.; FRANCOZO, E.; COELHO, O.; ARANTES, P.; BASSO, R. & ROCES, L. *in press*. The dynamics of V-to-V phonotactics: Lexical statistics and connectionist simulation. *Proc. LP2002*.

BARBOSA, P.A. (2002). Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production, *Proc. Speech Prosody 2002*, 163-6.

BARBOSA, P.A. & ARANTES, P. *in press*. Investigation of non-pitch-accented phrases in Brazilian Portuguese: No evidence favoring stress shift. *Proc. ICPhS 2003*.

BECKMAN, M.E. (1986). *Stress and Non-Stress Accent*. Dordrecht: Foris.

BOLINGER, D.L. (1958). A theory of pitch accent in English. *Word* 14, 109-49.

BOROWSKY, T. (1986). *Topics in the lexical phonology of English*. Doctoral dissertation, U. Mass, Amherst.

BROWMAN, C.P. & GOLDSTEIN, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica* 45, 140-55.

BROWMAN, C.P. & GOLDSTEIN, L.M. (1992). Articulatory phonology: An overview. *Phonetica*, 49, 155-180.

BYRD, D. & SALTZMAN, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *J. Phonetics* 31, 149-80.

CHOMSKY, N. & HALLE, M. (1968). *The Sound Pattern of English*. New York: Harper & Row.

CLEMENTS, G.N. (1990). The role of sonority cycle in core syllabification. In J. Kingston and M. E. Beckman (*eds*.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*. Cambridge: Cambridge University Press, pp. 283-333.

CRYSTAL, T. & HOUSE, A.S. (1997). A note on the durations of American English consonants. In S. Kiritani, H. Hirose & H. Fujisaki (*eds*.), *Speech Production and Language: In Honor of Osamu Fujimura*. Berlin: Mouton de Gruyter, pp. 195-213.

ERICKSON, D. (1998). Effects of contrastive emphasis on jaw opening. *Phonetica* 55, 147-69.

ERICKSON, D. (2002). Articulation of extreme formant patterns for emphasized vowels. *Phonetica*, 59, 134-49.

ERICKSON, D., FUJIMURA, O. & PARDO, B. (1998). Articulatory correlates of prosodic control: Emotion and emphasis. *Language & Speech* 41, 395-413.

FUJIMURA, O. (1970). Current issues in experimental phonetics. In R. Jakobson & S. Kawamoto, (*eds*.), *Studies in General and Oriental Linguistics*. Tokyo: TEC Co., pp. 109-30.

______. (1976). Syllable as concatenated demisyllables and affixes. *J. Acoust. Soc. Am.* 59, Supplement 1, S55 (abstract).

______. (1979). An analysis of English syllables as cores and affixes. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 32, 471-476.

______. (1987). A linear model of speech timing. In R. Channon & L. Shockey (*eds*.), *For Ilse Lehiste*. Dordrecht: Foris Publ., pp.109-23.

______. (1990). Methods and goals of speech production research. *Language & Speech* 33,195-258.

______. (1992). Phonology and phonetics -- A syllable-based model of articulatory organization. *J. Acoust. Soc. Japan (E)* 13, 39-48.

______. (1994a). C/D model: A computational model of phonetic implementation. In E. Ristad (*ed*.), *Language Computations. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 17,* pp. 1-20. Providence, RI: American Mathematical Society.

______. (1994b). Syllable timing computation in the C/D model. *Proc. ICSLP 94, Yokohama, Vol. 2,* pp. 519-22.

______. (1995). Prosodic organization of speech based on syllables: The C/D model. *Proc. ICPhS 95,* Vol. 3, pp. 10-17.

______. (1996). Syllable structure constraints: A C/D model perspective. In B. Agbayani & N. Harada (*eds*.), *Proc. SWOT-II, UCI Working Papers in Linguistics, Vol. 2,* Irvine, CA, pp. 59-74.

______. (1998). Neuromuscular simulation and linguistic control. *Bulletin de la Communication Parlée* (U. Grenoble) 4, 59-63.

______. (2000a). The C/D model and prosodic control of articulatory behavior. *Phonetica* 57, 128-38.

______. (2000b). C/D model prediction of CVC segmental duration for varied syllable prominence. *The Phonetician* 82, 9-21.

______. (2000c). Rhythmic organization and signal characteristics of speech. *Proc. ICSLP 2000,* Vol. I, 29-35.

______. *in press*. Stress and tone revisited: Skeletal *vs*. melodic and lexical *vs*. phrasal. In S. Kaji (*ed*.), *Proc. International Symposium on Cross-Linguistic Studies of Tonal Phenomena, Historical Development, Phonetics of Tone, and Descriptive Studies.* Tokyo University of Foreign Studies, ILCAA.

FUJIMURA, O.; CIMINO, A. & SAWADA, M. (1995). Voice quality control within a sentence: Expressive effects of source spectral envelope change. In O. Fujimura and M. Hirano (eds.), *Vocal Fold Physiology: Voice Quality Control*, pp. 201-15.

FUJIMURA, O. & ERICKSON, D. 1997. Acoustic phonetics. In W. J. Hardcastle & Laver, J. (eds.), *The Handbook of Phonetic Sciences*. Oxford: Blackwell Pub.

FUJIMURA, O. & WILLIAMS, J.C. (1999). Syllable concatenators in English, Japanese, and Spanish. In O. Fujimura, B. Joseph, & B. Palek (*eds*.), *Proceedings of LP'98*. Prague: Charles University Press, pp. 461-98.

FUJISAKI, H. & HIROSE, K. (1982). Modeling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation. *Proc. 13th International Congress of Linguists,* pp. 57-70.

HALLE, M. (1997). On stress and accent in Indo-European. *Language* 73, 275-313.

HARAGUCHI, S. (1999). A theory of the syllable. In O. Fujimura, B. Joseph, & B. Palek (*eds*.), *Proceedings of LP'98*. Prague: Charles University Press, pp. 691-715.

HATTORI, S. (1961). Prosodeme, syllable structure and laryngeal phonemes. In *Studies in Descriptive and Applied Linguistics*. Tokyo: International Christian University, pp. 1-27.

HAYATA, T. (1997). The bearer of an accent is a boundary rather than a segment (in Japanese). *Gogaku Kyooiku Kenkyuu Ronsoo (Studies in Language Education)* 15, pp. 233-48.

______. (1999). *Onchoo no Taiporojii* (Typology of Tonal Systems) Tokyo: Taishukan Pub.

HAYES, B. (1984). The phonology of rhythm in English. *Linguistic Inquiry* 15, 33-74.

JAKOBSON, R.; FANT, G. & HALLE, M. (1963) (3rd Edition). *Preliminaries to Speech Analysis*. Cambridge, MA: MIT Press.

KRAKOW, R.A. (1999). Physiological organization of syllables: A review. *J. Phonetics* 27, 23-54.

LADEFOGED, P. (2001). *A Course in Phonetics* (4th Edition). Orlando, FL: Harcourt.

LAVER, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

LEBEN, W.R. (1999). Weak vowels and vowel sequences in Kwa: Sounds that phonology can't handle. In O. Fujimura, B. Joseph & B. Palek (eds.), *Proceedings of LP'98*. Prague: Charles University Press, pp. 717-732.

LEBEN, W. & FUJIMURA, O. (2001). Extra-short vowels in West African languages. In B. Palek & O. Fujimura (*eds*.), *Proc. LP2000*. Prague: Charles University Press, pp. 83-94.

LEHISTE, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

LEHISTE, I. (1980). Phonetic manifestation of syntactic structure in English. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics (U. Tokyo*)14, 1-28.

LIBERMAN, M.Y. & PRINCE, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry* 8, 249-336.

LINDBLOM, B. (1963). Spectrographic study of vowel reduction. *J. Acoust. Soc. Am*. 35, 1773-81.

LINDBLOM, B.; SUSSMAN, H.M.; MODARRESI, G. & BURLINGAME, E. (2003). The trough effect: Implications for speech motor programming. *Phonetica* 59, 245-62.

MACCHI, M.J. (1988). Labial articulation patterns associated with segmental features and syllable structure in English. *Phonetica* 45, 109-21.

MALÉCOT, A. (1955). An experimental study of force of articulation. *Studia Linguistica* (Lund U.), 35-44.

MCCARTHY, J. (1988). Feature geometry and dependency, a review. *Phonetica* 43, 84-108.

MENEZES, C. (2003). *Rhythmic Pattern of American English: An Articulatory and Acoustic Study*. Doctoral dissertation, Dept. Speech and Hearing Science, The Ohio State University.

MENEZES, C.; PARDO, B.; ERICKSON, D. & FUJIMURA, O. (2002). Changes in syllable magnitude and timing due to repeated correction. *Speech Communication* 40, 71-85.

MITCHELL, C.J. (2000). *Analysis of Articulatory Movement Patterns according to the Converter/Distributor Model*. Master's thesis, Dept. Speech & Hearing Science, The Ohio State University.

MITCHELL, C.; MENEZES, C.; WILLIAMS, J.C.; PARDO, B.; ERICKSON, D. & FUJIMURA, O. (2000). Changes in syllable and boundary strengths due to irritation. In R. Cowie, E. Douglas-Cowie & M. Schröder (*eds*.), *Proceedings of ISCA Workshop on Speech and Emotion*. Belfast: Textflow, pp. 98-103.

NOOTEBOOM, S. (1997). Prosody of speech: Melody and rhythm. In W. J. Hardcastle & J. Laver (*eds*.), *The Handbook of Phonetic Sciences.* Oxford: Blackwell Pub.

ÖHMAN, S.E.G. (1967). Numerical model of coarticulation. *J. Acoust. Soc. Am*. 41, 310-20.

PIERREHUMBERT, J.B. (1989). A preliminary study of the consequences of intonation for the voice source. *Speech Transmission Laboratory Quarterly Progress and Status Report 4*, 23-36.

POSER, W. (1984). *The Phonetics and Phonology of Tone and Intonation in Japanese*. Doctoral dissertation, MIT.

SAGEY, Elizabeth. (1986). *The Representation of Features and Relations in Nonlinear Phonology*. Doctoral dissertation, MIT. (New York: Garland Press 1991).

SPROAT, R. & FUJIMURA, O. (1993). Allophonic variation in English /l/ and its implications for phonetic implementation. *J. Phonetics* 21, 291-311.

STERNBERG, S.; KNOLL, R. L.; MONSELL, S. & WRIGHT, C.E. (1988). Motor programs and hierarchical organization in the control of rapid speech. *Phonetica* 45, 177-97.

STEVENS, K. N. *in press*. Acoustic and perceptual evidence for universal phonological features. *Proc. ICPhS 2003*.

TITZE, I. 1994. *Principle of Voice Production*. Englewood Cliffs, NJ: Prentice Hall.

UMEDA, N. 1975. Vowel duration in American English. *J. Acoust. Soc. Am.* 58, 434-45.
______. 1977. Consonantal duration in American English. *J. Acoust. Soc. Am.* 61, 846-58.

WILHELMS-TRICARICO, R. 1995. Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *J. Acoust. Soc. Am*. 97, 3085-98.

WILLIAMS, J. C. *in press*. Syllable-based phonology of Ibero-Romance languages. *Proc. ICPhS 2003*.