Rhythmic Organization and Signal Characteristics of Speech

Osamu Fujimura
Department of Speech and Hearing Science, The Ohio State University

ABSTRACT

The Converter-Distributor (C/D) Model, a generative theory of phonetic implementation, describes an utterance as a linear string of syllables with intervening boundaries. Its basic component includes phonetic status contours for voicing, tonal, and vocalic gestures. Consonantal elemental gestures, as stored impulse responses, are excited by the syllable pulse and superimposed onto the base function. A magnitude-modulated syllable-boundary pulse train constitutes a skeletal representation of the rhythmic organization of the utterance. All the temporal characteristics of the speech signal are computed based on the input specifications for each syllable by phonological features and the metrical structure, numerically augmented by prominence enhancement specified for the discourse situation, along with system parameter settings for the particular speaker in each discourse. Segmental durations in the acoustic signal vary according to syllable magnitude, not uniformly among consonants and vowels. The C/D model predicts complex patterns of such prosodic effects on segmental duration as a function of fixed threshold values for relating abstract gestures to observable durations of acoustic signals.

(Supported in part by NSF and ATR)

1. INTRODUCTION

Traditionally in phonetic and phonologic sciences as well as speech technology, phonemic segments were the basic units of speech sounds. According to this view, consonants and vowels are concatenated to form a string of segments. The basic principle is that each of such phonemes, assumed to belong to the sound inventory of a given language, can be selected one by one and placed from left to right continguously, forming the phonological representation of any word, phrase, or sentence that can be uttered. There are some junctures which indicate boundaries between consecutive words and phrases in such a linear string representation of an utterance of linguistic forms.

Current phonological theory, in particular nonlinear phonology in various versions, basically deviates from this principle of describing speech. One basic deviation is the multidimensionality of the representation with respect to concatenative organization. The materials are organized in different tiers independently in parallel for phonological words, etc. The specification of the so-called segmental information is not necessarily given segment by segment. For example, if the language is subject to vowel harmony, a particular vocalic feature is specified for the entire word. In this sense, the traditional conceptual division between segmental vs. prosodic (i.e., suprasegmental) characteristics is no longer clear. The concept of distinctive features, as the minimal units of phonological opposition for linguistic forms, is maintained, but the specification domain can vary, and in general, features are underspecified without specifying their values for each segment.

Thus, the concept of minimal opposition requires an explicit discussion of the relevant domains of context.

Phonetics, based on the phonemic segmental theory, assumed target values for each segment, whether in terms of acoustic properties or articulatory position, both for consonants and vowels. A straightforward version of such a theory might state that such target values represent the control variables for each phonemic segment, to which a duration is assigned simultaneously for all variables, resulting in a set of step functions switching from one segment to the next synchronously. The process of coarticulation would smooth out each variable as a time function to obtain the observed variable in each articulatory dimension or the acoustic manifestation. Obviously this classic view in phonetics does not take advantage of progress in phonological theory.

Some radical proposals have been put forth in phonetic theory, deviating fundamentally from the classic segmental view sketched above. One theory called Articulatory Phonology (AP) proposed by C. P. Brown and L. Goldstein [1992], eliminates the distinction between phonology and phonetics and represents utterance characteristics inherent in each word by articulatory gestures that are characteristic of a specific pair of phonemes (typically CV and VC). By doing so, it is claimed, the great variation of speech characteristics for the same word in different utterance conditions can be characterized succinctly with reference to the relative temporal organization of individual articulatory gestures, as executed by different articulators, relatively independently from each other regardless of segmental boundaries. The phenomena to be described in this way include segmental deletion (elision) and insertion (epenthesis) as observed in different languages, which are usually handled by context-dependent rules in phonology. Many phenomena commonly observed as segmental reduction, which tend to vary continuously in the degree of weakening according to the change in prosodic conditions, can be described successfully by this articulation-based model. Apparent shifts in phonetic quality of the same consonant or vowel according to the context, which seem to be ad hoc for languages or dialects, often cannot be captured by the general principle of coarticulation. They can also be described basically as a matter of temporal organization of gestures that can be controlled independently among different articulators.

Another theory of phonetics that radically departs from the classic view of segment concatenation and coarticulation is the Converter-Distributor (C/D) model [Fujimura 1994, 2000; Fujimura & Williams 1999]. This theory, which follows the theory of generative phonology distinguishing phonetics from phonology [Chomsky and Halle 1968], is a model of phonetic implementation, taking its input from the output representation of phonology and generating its output as speech signals. The C/D model dispenses with the concept of the phonemic segment, however, and it uses syllables as the basic concatenative
segmental units. Syllables are represented as a set of abstract feature specifications, which serve in place of individual consonants or vowels as the domain of functional opposition. In contrast, standard nonlinear phonology defines the syllable as a prosodic organization of phonemic segments. In the C/D model, unlike the articulatory phonology, feature specifications of a syllable are interpreted by computation to produce a temporally organized set of gestures.

The C/D model provides a comprehensive framework for describing utterance characteristics under varied utterance conditions, as seen in daily conversation. It is done so by an explicit and quantitative computational process, with a clear distinction between the inherent sound properties of lexical items and other linguistic forms and the speech characteristics of particular utterances. However, we assume that phonetics varies greatly from language to language; the characteristics of a particular language or a dialect set the parameters for this phonetic implementation system. Also, system parameters are set for individual speakers, utterance style, etc. Describing a specific utterance as a whole, given the system characteristics, the description of the language, and its use by an individual speaker in a particular utterance situation, is the function of the C/D model.

## 2. PROSODIC ORGANIZATION

### 2.1. Base Function

The C/D model describes speech organization by separating consonantal gestures (as in Öhman’s [1967] consonantal perturbation theory) from what is called the base function. The base function represents the prosodic organization of an utterance, using a generalized concept of prosody. Specifically, it has its skeleton represented by a syllable-boundary pulse train. This pulse train, a one-dimensional time function, is assumed to carry, at least as a first approximation, all information about the rhythmic structure, i.e., the stress pattern, of the utterance. Along with the skeleton, there are several aspects of melody in the base function, represented as step functions of time with occasional ramps interpolating switching from one target value to another during phonologically unspecified time intervals. They include the control functions for vocalic gestures representing the relatively slow movement from one syllable nucleus to the next, voice pitch control, other voice quality control functions and laryngeal adduction-abduction control which results, by interaction with air stream control, in the voiced-unvoiced switching function. Each melodic control function is represented at an abstract phonetic level by a phonetic status contour,. Thus, Fujisaki’s model of intonation function as the F0 contour, is generalized for different aspects of “prosodic” functions including vowel articulation (but not consonantal articulation).

Fig. 1 depicts some aspects of the base function for an utterance of the sentence ‘That’s the most important’ with emphasis on ‘most’. The top panel shows consonantal elemental gestures as independent but temporally overlapping events in parallel in different articulators (see Section 3.2.). The next panel below represents the skeletal structure of this utterance by symmetric triangles around syllable pulses, which vary in size but not in shape according to the prominence (magnitude) of each syllable.

In this figure, the internal duration allotment for onset and coda subcomponents is not shown except for the s-fix, which is assigned a separate (abstract) time interval (shown by the half-triangle with a solid slant arrow on right). There are two phonetic phrase boundaries, after the phonetic phrase ‘that’s’ (the leftmost triangle) and the emphasized word ‘most’ (the largest triangle) which is followed by a phonetic phrase boundary because of emphasis [Fujimura, Pardo & Erickson 1998], [Mitchell et al, to appear], shown by a dashed half triangle. The dashed arrow at the end of the utterance suggests the phrase-final elongation effect, which may be interpreted as an expansion of the time scale toward the end of the major phrase.

Figure 1: Phonetic status contours: ‘That’s the most important’

The remaining panels of the figure depict three melodic aspects of the base function as abstract control functions: tongue body gesture (advancing-retracting), mandible lowering, and voicing on-off. These contours represent phonetic status changes. The tongue contour is shown in two versions: the dot-dash line is assumed to manifest the inherent phonologic function of the vocalic feature {+back} by positive vs. negative direction of deviation (single arrows) from the neutral value (thin horizontal dot-dash line). This feature-based deviation is phonetically modified according to the syllable magnitude, depending on the positive or negative excess of the latter in reference to a standard syllable magnitude (shown by a horizontal dot-dash line through the syllable triangles in the second-to-top panel). This phonetic readjustment is suggested by the parallel arrows expanding or reducing the inherent articulatory gesture of tongue advancement-retraction. Note that during the time interval assigned to the reduced vowel (the rightmost small syllable triangle) as well as
boundaries, the tongue body gesture is linearly interpolated in this figure.

The second-to-bottom panel illustrates the abstract control function for jaw opening (mandible lowering relative to the maxilla). In this depiction, it is assumed that the amount of jaw opening in its prosodically controlled component, which is shown in this picture, directly represents the syllable magnitude (see for empirical testing of this hypothesis, Fujimura et al. [1998], Erickson et al. [1999], Fujimura [2000], Mitchell et al. [to appear]). The lengths of the downward arrows therefore duplicate the syllable pulse height values.

The bottom panel of the figure depicts the voicing contour of this utterance. This phonetic status contour is abstract in the sense that it represents the underlying control of laryngeal adduction (voice on) vs. abduction (voice off), rather than actually observed acoustic voicing change, since they are related to each other through a complex nonlinear mapping process, as suggested in Fig. 3.

2.2. Syllable-Boundary Pulse Train

Boundaries of various magnitudes can be evaluated for the phonetic phrasal organization of a particular utterance, as a numerically augmented metrical tree. As seen in Fig. 2 (a), phonetic augmentation, due to emphasis, is represented by the number in parentheses (2.0), which is attached to a node for the word ‘wonderful’. This results in an expansion, by the factor 2.0, of the head syllable /wøn/ of the emphasized word (Fig. 2 (b), see also Fujimura & Erickson [1994]). It is also assumed in this figure that this emphasis results in phonetic rephrasing, introducing a phrase boundary before the emphasized word, indicated by the percentage mark (compare panels (c) and (d)).

The syllable-boundary pulse train thus generated as a time function represents, by definition according to the C/D model, the rhythmic structure of an utterance. In other words, the model computes various signal characteristics, articulatory or acoustic, of observable physical events, such as stop implosion/explosion, voice onset/offset, based on the putative pulse train and associated phonetic gestures implementing phonological feature specifications.

2.3. Stress Pattern in English

In English, the stress pattern pertains to phonological opposition of words. For example, the verb ‘import’ is specified by a stress pattern different from the noun ‘import’, along with what is considered identical segmental information. When this metrical structure of a word, even in isolation, is implemented as an utterance, inevitably there are effects of phrasing and phrase boundaries observed. The manifestation of the inherent phonological property of the word is modified, particularly toward the end of any major phrase. The manifestation of the stress pattern in English involves many physical variables: temporal span, F0 peak value, voice quality in physical dimensions such as the spectral envelope of the voice source signal and the intensity (power), spectrum, including amplitude, of turbulent noise.

It should be noted that linguistic prosodic control in English as observed in an utterance is not limited to stress pattern or rhythmic organization and the F0 contour based thereon. For example, a yes-no question typically is uttered with an intonation pattern that shows a distinct rise in voice pitch toward the end of the sentence. This pitch rise is basically separate from the stress pattern, which inherently controls the temporal organization of the utterance as well as pitch. Therefore, it can be readily seen that English, along with stress or rhythmic control has independent voice pitch control. The point is that in English, unlike Chinese, Japanese, or many African languages, this is strictly a phrasal phenomenon. Within the lexicon, tonal specification that is independent from the metrical pattern representation, is not applicable in English. The metrical pattern for individual words, as well as phrases, can be represented by a metrical tree [Liberman & Prince 1977], a metrical grid [Prince 1983], [Hayes 1994], or according to more recent work a parenthesis system [Halle & Idsardi 1995]. There are no additional tone features along with the lexical accent from which this metrical structure is derived in the English lexicon.
In contrast, in Tokyo Japanese, there is no metrical pattern specified in the lexicon. Tonal features specify lexical accent patterns to phonologically distinguish different words. This is most effectively represented by the specification of which syllable in the word (or accent phrase) is marked for accent [Hattori 1961] in the case of nouns, and whether the word is accented or not in the case of verbs and adjectives. The lexical accent is manifest primarily as a pitch fall around the end of the nucleus of the accented syllable. There is no other phonological specification in the case of the Tokyo dialet within the lexicon (see Haraguchi [1979], McCawley [1968], Poser [1982], Pierrehumbert & Beckman [1988], Purnell [1999] among many other sources for descriptions of Japanese accent and intonation patterns). At the phrasal and discourse level, however, phonetic stress pattern plays an important role for verbal communications in Japanese.

3. SYLLABLE STRUCTURE AND SEGMENTAL DURATION PATTERNS

The rhythmic organization of a speech utterance is represented by the base function, in particular the syllable-boundary pulse train. In order to relate this abstract temporal organization to observable physical events, in particular acoustic signal characteristics, we need to discuss the internal structure of each syllable and its signal manifestations. Effects of altering the syllable magnitude on acoustic phonetic quantities such as segmental durations are of central concern in this sense. We shall speculate on what our theory predicts to see how we can evaluate the validity of this theory by empirical, particularly acoustic, data. It also turns out that some hidden system parameters of the C/D model may eventually be inferred to some extent from the pattern of variation of segmental durations as the function of syllable prominence.

3.1. Syllable Components

In the input representation of each syllable, the C/D model assumes phonological specification for each component of the syllable core, i.e., nucleus, onset, and coda. As additional syllable components, syllable affixes are optionally specified if the language allows them. They can be a p-fix or an order-specified sequence of p-fixes preceding the syllable core in time, or an s-fix or an order-specified sequence of s-fixes succeeding the syllable core. Phonological features are designated for their applicable syllable component and are under-specified, considering the syllable as the domain, to represent phonological contrast in the given language (see Fujimura [1996a, 1997] and Fujimura & Williams [1999]). English allows one or more s-fixes (if the second s-fix represents a morpheme) but no p-fixes, and s-fixes are allowed only in word-final position. A s-fix in English must fulfill the following conditions (see [Fujimura 1976, 1979] for the basic idea, although the current formulation below differs in detail to conform to the C/D model framework).

1. It must be an obstruent produced with the tongue tip (phonologically place-specified as [apical]).
2. Voicing is not specified. The voicing status of the s-fix is an extension of that of the tautosyllabic coda.

(3) A manner feature {stop}, {fricative}, or {spirantized} must be specified (the only feature specification for s-fixes in English).

Not all s-fixes are morphemic, as seen in ‘tax’ (/ts/), ‘lend’ (/ld/), ‘act’ (/t/), ‘text’ (/st/), a single s-fix specified by a single feature (spirantized)). The manner feature [spirantized] is phonetically implemented with unvoiced apical frication preceding an oral stop closure or a nasal stop closure. The phonological place specification pertains to the choice of the crucial articulator of the stop closure, whether oral or nasal, while the articulator of the frication is always the tongue tip. If [spirantized] is specified with the place feature and not [nasal], concomitantly, the accompanying oral closure is implied. Voicing is never specified (no voicing contrast in the context of [spirantized]). In the case of nasal stops, as seen in ‘smell’ and ‘snow’, which occur only in onset, the oral closure is also implied, but the nasalization is specified by an additional feature specification [nasal] (see Fujimura & Williams [1999]).

The temporal organization of the English word ‘splatter’, uttered in isolation, is exemplified in Fig. 3. (All figures in this paper are speculative, given for the purpose of explaining how the model works.) The top panel shows syllable triangles for the first syllable with main stress followed by the reduced syllable. In this figure, the large symmetric triangle with downward arrows on left and right corresponds to the nucleus portion of the syllable with main stress. To its left, there is a half triangle with a downward left arrow; this corresponds to the onset component of the same syllable. Similarly, a small symmetric triangle and the small half triangle to its right for coda represent the second syllable. Note that, in this example, the half triangles for the coda of the first syllable and the onset of the second syllable are missing, and the symmetric triangles, large and small, corresponding to the two nuclei are next to each other. The ambisyllabic “flap” does not have any allotted duration [Fujimura & Williams 1999].

The dotted half triangles to the left and to the right of the entire utterance (word uttered in isolation) represents the time allotment for the phrase boundaries identified by dollar signs. The thick vertical bar at the center of each symmetric triangle is the syllable pulse, its height representing the syllable magnitude. The angles formed by this bar and the slant arrows on both sides are the same and constant throughout all triangles representing syllables in the utterance. The syllable pulse is duplicated with the same height to the left of the edge of the symmetrical triangle (end point of the arrow) as the onset pulse.

The onset pulse of the first syllable excites IRFs for three elemental gestures, labeled <p>, <l>, and the <P, p> as shown in the second panel of the figure, for the onset cluster /spl/ of the first syllable. The upward arrows suggest that these local time functions are responses to the excitation by the pocs (p-fix, onset, coda, or s-fix, in this case onset) pulse located in time at the origin of the arrows. The utterance-final coda gesture, identified by >>, is evoked by the small excitation pulse, i.e. the coda pulse of the reduced syllable at the end of the word.
3.2. IRFs and Consonantal Threshold

Fig. 4 illustrates a description of the temporal organization of an utterance of the word ‘conclude’ in isolation with respect to consonantal gestures and their acoustic segmental durations. The top panel is similar to that of Fig. 3, except for the addition of dashed diagonal lines from the top of syllable pulses (thick vertical bars), which show overall time spans for each of the two contiguous syllables. Also, the downward arrows originating from the top of each syllable pulse in Fig. 3 are now replaced by simple line segments, to simplify the figure. In some previously published figures (e.g., Fujimura, Pardo, & Erickson [1998]), this overall syllable triangle is used without showing the core-margin time allotment separately. The downward dashed arrow in this figure toward the right edge of the panel suggests the phrase (utterance) final elongation.

In the lower panels, different articulatory dimensions are shown in separate panels for elemental gestures: onset dorsal stop (tongue body position), coda nasal (velum lowering), and onset lateral gesture (lateral tongue body position), other elemental gestures, such as tongue body retraction, are not shown here. It should be noted that elemental gestures do not occur in synchrony for the same syllable component such as the dorsal stop gesture and lateral apical gesture for the onset of the second syllable. The asynchronism which staggers the two concomitant onset gestures results in the sequenced consonant cluster in the traditional phonemic sense. But this sequence cannot be reversed as a phonological contrast. In the C/D model, this fact is represented by inherent temporal characteristics of the stored IRFs themselves, relative to the excitation pulse (in this case onset pulse). The apparent difference in the beginning time and duration of the [l] segment between ‘late’ and ‘slate’, for example, is explained, according to the C/D model, as the physical interaction between the tongue tip frication generation and the observable sonorant sound of [l].

3.3. Segmental Duration in the Acoustic Signal

A consonantal gesture in general is assumed to take place basically as a ballistic motion, returning to the base position automatically under one command [Fujimura 1994]. Correspondingly, the elemental gesture curve monotonically rises...
and falls and this total action is represented by an integral impulse response function that is inherently fixed for the elemental gesture. The movement time function is the response to the excitation signal, i.e., a pulse. When this time function as a control function crosses a saturation threshold position, the time value of the function can be interpreted approximately as the onset or offset of some signal characteristics, such as stop closure and frication (a more detailed model may use different threshold values for onset and offset, respectively). Thus, the time interval between the rising and falling threshold crossings represents, roughly, the duration of the acoustic segment as usually measured in acoustic phonetics [Lehiste 1973].

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5}
\caption{CVC duration patterns for two threshold values}
\end{figure}

Fig. 5 depicts the interaction between the syllable magnitude and the consonantal saturation threshold for any CVC syllable, according to what the C/D model predicts. The same set of control functions, according to different syllable magnitude values, are used for two threshold values. The different choices of the threshold value, as a system parameter of the model, result in two qualitatively different patterns of segmental duration. When the threshold is set high at 0.4 (arbitrary scale), relative to consonantal gesture peaks, the consonantal gestures disappear and the vowel segment survives, as the syllable prominence, shown by different magnitude $\mu$, is reduced from 1.0 to 0.3 (see the bottom of Fig. 5 a). Since the IRFs are generally lower in peak for coda than for onset, the onset obstructant gestures are more susceptible than onset gestures of the same type in their durational reduction and segment deletion (see the circles). When the threshold value is lowered to 0.1 without changing anything else, the consonantal durations remain relatively unaffected while the vowel segment becomes deleted as the syllable magnitude is reduced down to 0.3 (see the circle).

Vowel deletion and consonant (particularly coda) deletion are often observed in different languages, and in many cases this process is phonologized or lexicalized. The C/D model may explain why these happen and why one of the possible processes, rather than the other, happens. Also, since the system parameters of the model are not observable directly, these phonetic behaviors of segmental duration as observed quantitatively under different prosodic conditions may be useful for inferring system parameters, in this case the consonantal saturation threshold.

4. ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation of the U. S. A. (BCS-9977018) and Advanced Telecommunications Research.

5. REFERENCES

Fujimura, O., Pardo, B. & Erickson, D. Effect of emphasis and irritation on jaw opening. *Proceedings of European Speech Community Association, Conference on Sound Patterns of*


