TR - H - 184

The C/D Model as a Dynamic, Non-segmental Approach

Osamu Fujimura

1996. 1.17

ATR人間情報通信研究所

〒619-02 京都府相楽郡精華町光台2-2 ☎ 0774-95-1011

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Telephone: +81-774-95-1011
Facsimile: +81-774-95-1008

The C/D Model as a Dynamic, Non-segmental Approach[1]

by

Osamu Fujimura
The Ohio State University, Columbus, OH 43210-1002, U. S. A.

1. The C/D model

The Converter/Distributor (C/D) model is a theory of phonetic implementation for computing the articulatory and acoustic signals as time functions. As its input, the specification of an utterance is given both in terms of phonological representation and parameter values for the utterance/discourse situation. It uses syllables as the "segmental" units, and the content of each syllable is specified by phonological features, each of which pertains to one of the syllable components: onset, coda, or nucleus for the syllable core, or a syllable affix (p-fix or s-fix).

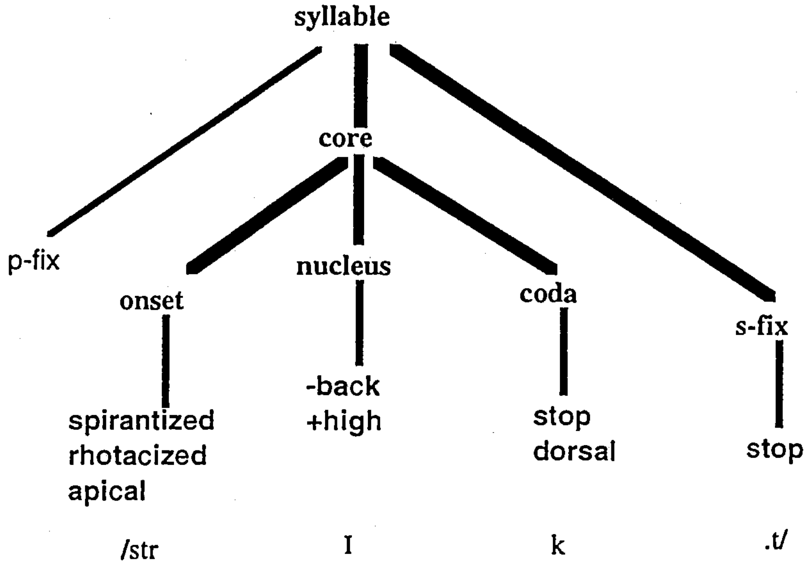

Fig. 1. 'strict': syllable structure and features.

This figure shows, as an example, the syllable structure of the English word 'strict'. There is no p-fix in English. We may note:

The onset /str/ is specified by an unordered set of unary phonological features: {spirantized; rhotacized; apical}. The articulatory closure-release, in addition to the preceding frication noise, is implied by the feature {spirantized}. In a situation where the apical fricative is not followed by a stop consonant, in the phonemic terms, another manner feature, for example {nasal}, is added along with the place; the implication of the oral closure-release still applies. For

[1] A written version of the oral presentation given at the Symposium on Dynamic, Nonsegmental Approaches to Phonetics (John Local: coordinator), The XIIIth International Congress of Phonetic Sciences, Stockholm, August 1995. The main text is included in the Proceedings of the congress, Volume 3, pp. 10-17, as a preprint entitled "Prosodic organization of speech based on syllables: the C/D model". The preprint version did not include any figures, and this paper briefly summarizes and supplements the proceedings paper. This work was conducted in part while the author was visiting ATR/HIP during the summer 1995, and was in part supported by a fund provided by ATR/ITL for the author's research and education activities at Department of Speech and Hearing Science and Center for Cognitive Science, The Ohio State University.

the English word 'smile', for example, the onset specification is an unordered set of features {nasal; spirantized; labial}.

The coda /k/ is likewise specified by unary features: {stop; dorsal}. The voicelessness of an obstruent is unmarked, evoking voice cessation at the syllable margin unless {voiced} is specified, and this implication for coda automatically extends to the s-fix(es) by a universal convention.[2]

The nucleus /ɪ/ is specified by binary features: {-back; +high}.[3]

The s-fix /t/ is specified by only one unary feature: {stop}, which is a manner specification. In English s-fixes, the place is always {apical}, and they are implemented always with the tongue tip/blade. The manner is one of the obstruent features {stop, fricative, spirantized, interdental}.

The prosodic structure is specified as the input to this implementation process in the form of a metrical tree for each sentence, augmented by a numerical mark optionally attached to any node of the tree according to the utterance situation. Other utterance parameter values for the discourse and speaker conditions are also specified, as seen in the upper left corner of Fig. 2.

The model comprises four serially ordered components, as shown in Fig. 2:

1. The converter converts the augmented metrical tree to a pulse train representing linearly concatenated syllables and boundaries as a time series. The pulse height representing the syllable magnitude, shown by the length of each thick vertical bar in the output of the converter, is computed for each syllable based on an algorithm similar to the metrical grid computation suggested in Liberman & Prince (1977), with the numerical augmentation discussed below. According to the syllable magnitude, a shadow computation as depicted by the slant lines descending from the top of each syllable pulse, to the left and to the right, determines the duration assigned to each syllable (the left and right parts may be interpreted as initial and final demisyllables, see Fujimura (1979)). The horizontal dashed arrows indicate the extent of shadows for placing onset and coda pulses, respectively, which are represented by vertical dashed bars labeled 1^o, 1^c, *etc.* for onset of syllable 1, coda of syllable 1, *etc.* (see also Fig. 3). The linear pulse train with controlled magnitude-time values, when associated with phonological feature specifications, constitutes a comprehensive representation of the prosodic organization of the utterance, apart from the parameter values like those at the upper left corner of Fig. 2 which condition the entire utterance. Thus, this pulse train, by definition in this model, is the prosodic organization of the utterance, while the phonological features, such as specifications of glide elements (equivalent to the tense-lax opposition of the vowel), can affect the relevant shadow slopes. The utterance parameters determine, by a fixed algorithm, the set of slope values for syllable components (p-fix, onset, coda, s-fix, or nucleus) and boundaries (for left and right shadows separately) within the entire utterance.
2. The distributor assigns phonological features for each syllable component to pertinent articulatory dimensions (many to many mapping). An articulatory dimension corresponds to an elemental gesture (one-to-one). An example is given in Fig. 2 for the feature combination of {apical} τ and {stop} T assigned to an articulatory dimension "tongue tip/blade closure-release".
3. A parallel set of actuators evoke pertinent elemental gestures, given the phonological features specified, to be implemented by the signal generator in their inherent articulatory organs (or a set of designated muscles using a particular

[2] There is a small set of exceptional words in English which involves the voiceless nominal morpheme 'th' as in 'warmth', 'width', *etc.* This may suggest that the voicing agreement of the syllable affixes with the syllable core margin (*i.e.* onset or coda) as a convention regarding the marking of {voiced} for obstruents should be treated by a morphologically conditioned constraint within the lexicon.

[3] The representation scheme for the syllable nucleus is only tentative and awaits further investigation.

weighted pattern). The tongue tip/blade closure-release as an elemental gesture (implemented as a ballistic motion) is exemplified figuratively in three instances (corresponding to /t/, /n/, and /d/) by the evoked impulse response functions, pertaining to the same articulatory dimension.

4. The signal generator computes the signals, articulatory and acoustic, based on a simulation model of the human articulatory system. This last component of the model is not shown in Fig. 2. The dashed horizontal line in the output of the actuator for tongue tip/blade closure-release suggests an effect of the signal generator saturating the gestural time functions above a threshold value, reflecting the inherent nonlinearity of this component. The computational principles of this nonlinear dynamic 3-d simulation of the speech apparatus are discussed in Wilhelms-Tricarico (1995).

basic utterance rate = 5.5 syllables/sec.
excitement = 4.5
formality = 0.3
dialect: Columbus OH 1
speaker: male.young 4

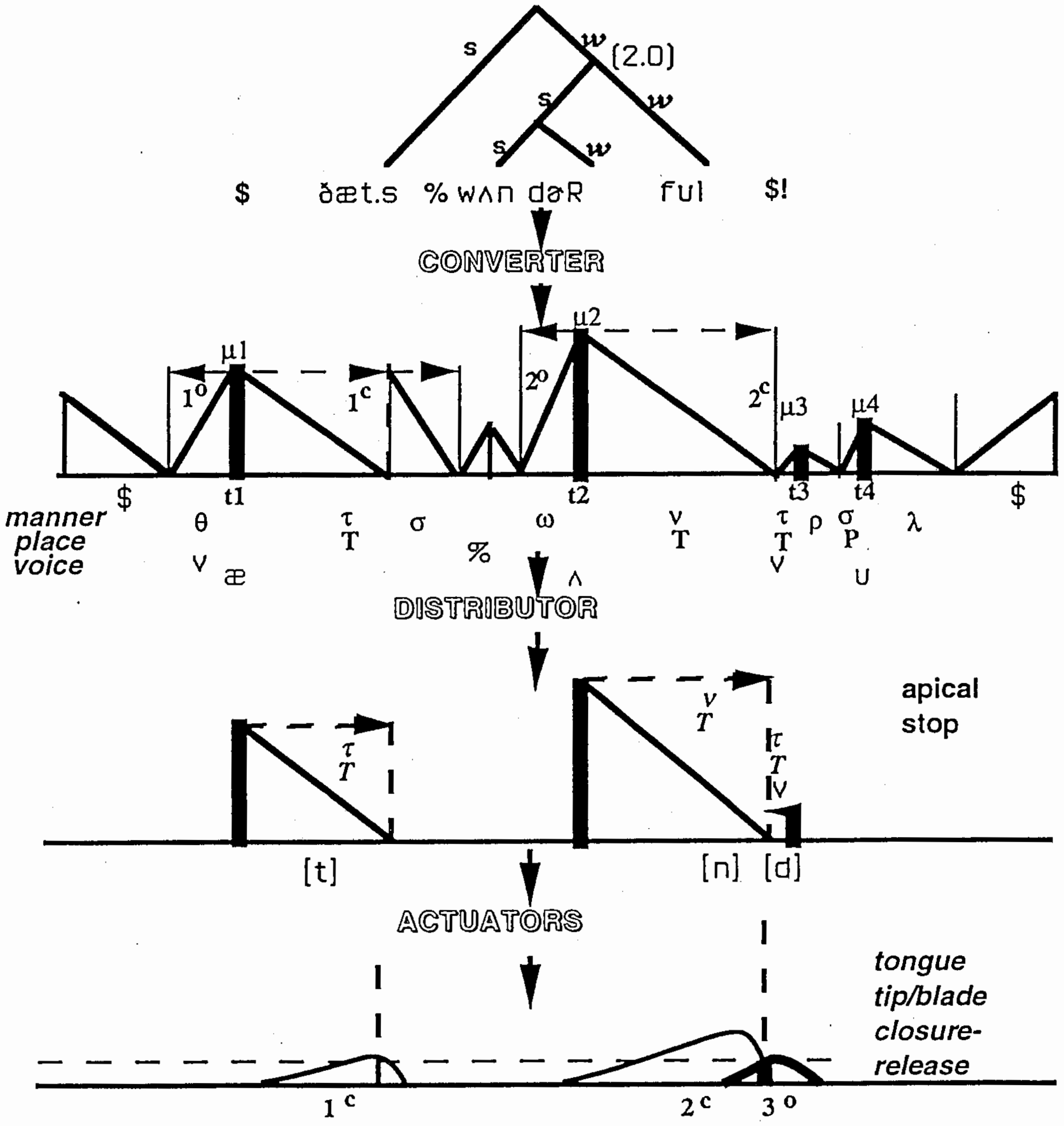

Fig. 2: C/D model components (for an utterance of 'That's wonderful').

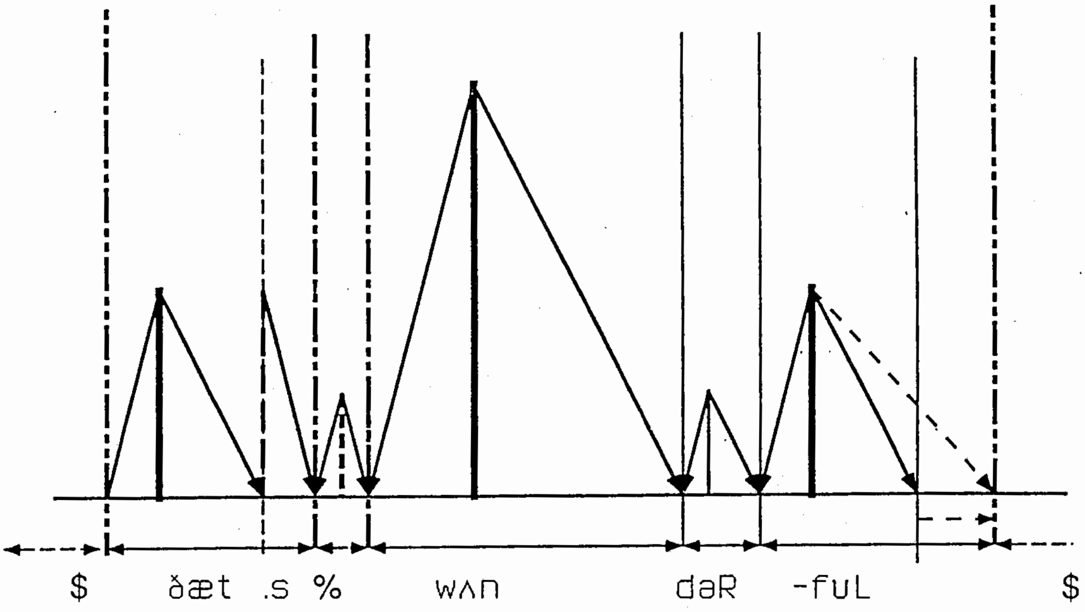

Fig. 3 pulse train and shadow computation: 'That's wonderful'

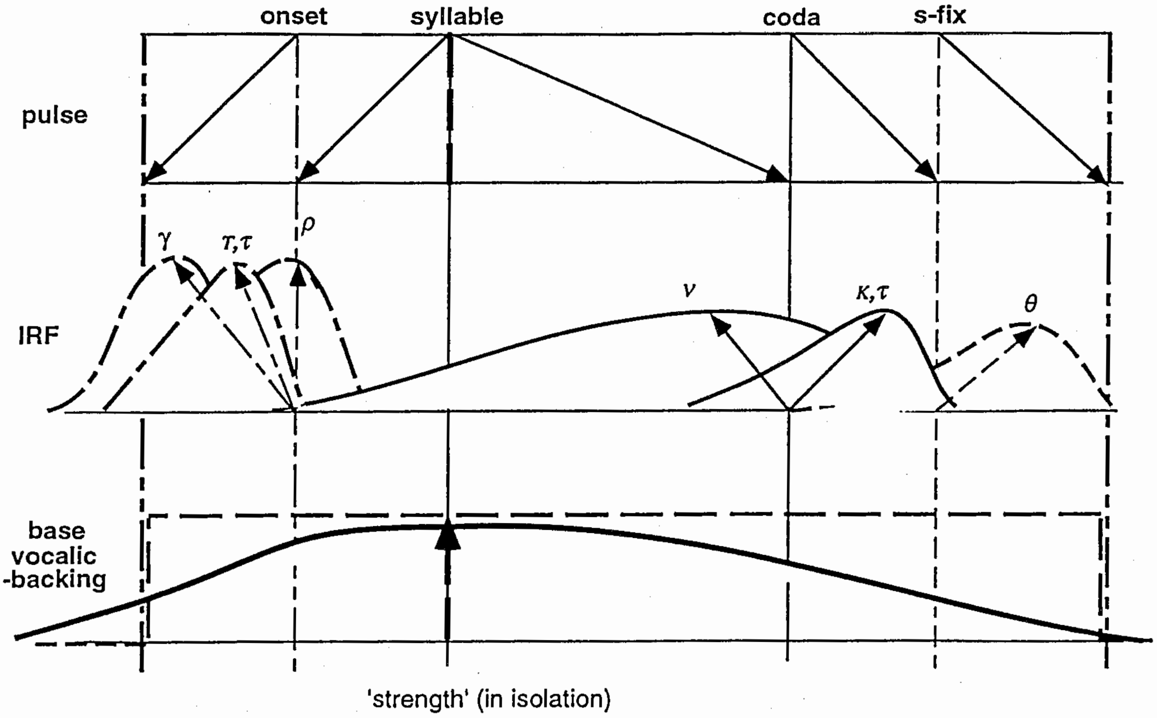

Fig. 4: Pocs pulses and evoked IRFs for 'strength' /strɛNk.θ/ spoken in isolation

The shadow computation algorithm was recently revised to erect subordinate pocs (p-fix, onset, coda, and s-fix) pulses, which inherit the magnitude of their pertinent syllable pulses (see Figs. 3, 4).[4] The assigned syllable durations (double-headed arrows, Fig. 3) are still proportional to the abstract syllable magnitudes, as

[4] In the figures showing pulse shadows, except Fig. 4, the onset and coda pulses are omitted, and the sub-durations due to the subordinate pulses are absorbed into the main (core) shadows.

proposed in the original version of the C/D model (Fujimura (1992)). There is a special slope adjustment for phrase-final lengthening, as seen in Fig. 3. Such prosodic modulations are generally accompanied by specific intonation (F_0 and voice quality) contours (see, *e.g.*, Anderson, *et al.*, 1984) as aspects of the base function along with the vocalic (nucleus-to-nucleus) contour.

Fig. 4 illustrates the shadow computation for temporal placement of pocs pulses and samples of the impulse responses evoked (in different articulatory dimensions, speculatively for the word 'strength' spoken in isolation). Greek symbols (Italic font) in combination with a small capital for stops and fricatives stand for elemental gestures (*γ:* spirantized frication; *T,τ:* apical stop; *ρ:* rhotacization; *ν:* nasalization; *K,τ:* dorsal stop; *θ:* interdental frication).

Fig. 5 shows the effects of augmentation of the word 'wonderful', as in Fig. 3, resulting in a strong temporal as well as abstract spatial expansion of the stressed syllable.

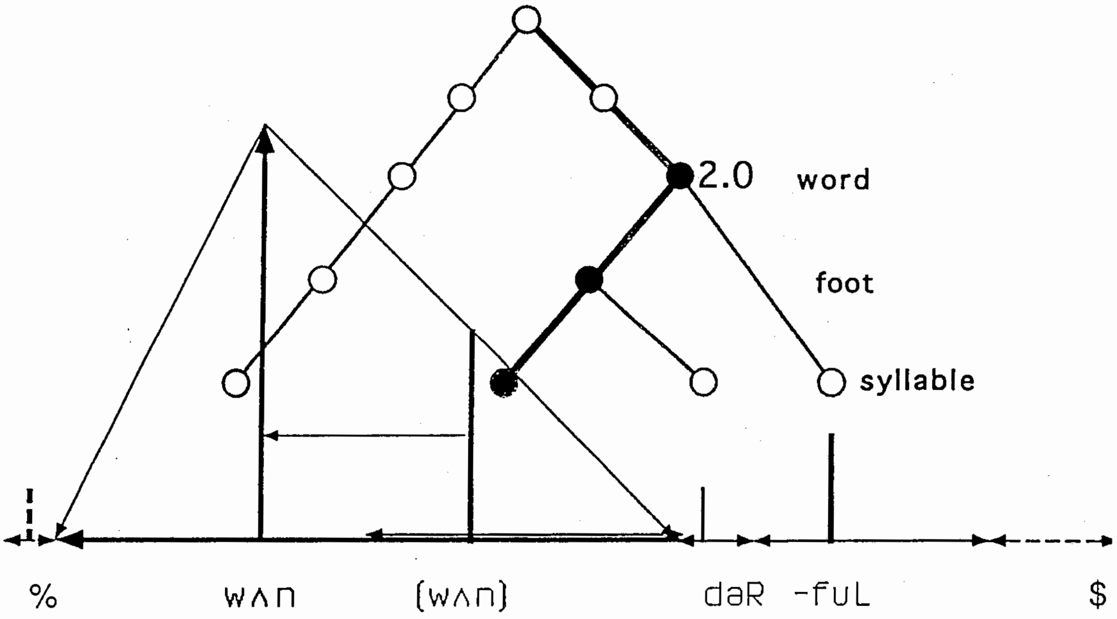

Fig. 5: Numerically augmented metrical tree and phonological units.

When a numeric augmentation is specified to a node of the metrical tree, the head among the daughter nodes receives the augmentation. This augmentation value is transmitted through the sub-tree dominated by the designated daughter node recursively until the augmentation is implemented on the syllable with the main stress of the word that is selected as the phonetic head of the augmented phrase. This augmentation may be implemented by multiplying the originally computed (*i.e.* grid-height based) magnitude by the numerical factor assigned. Fig. 5 illustrates this process, in the case of an emotive emphasis placed on the word 'wonderful' in the sentence 'That's wonderful', as we see also in Fig. 2 above. The first syllable with the main stress /wʌn/ of this word is magnified by the numerical factor 2.0 in its pulse height, and the timing of the syllable pulse, relative to others, is shifted leftward from the original position, according to the modified pulse height and its associated shadows. Note that the slopes of the shadows (on left and right of the pulse) remain unchanged, and therefore, in Fig. 5, the right-hand shadow of the shifted enlarged pulse passes the tip of the original pulse.

The magnitudes of boundary pulses are computed according to the levels of the phonological boundaries and the augmentation of any element within the phrasal unit. The base function is computed from a series of syllable/boundary pulses. The nucleus-to-nucleus flow of vowel gestures is represented by this base function, along with intonational variables and mandible movement which reflects syntagmatic modulation of the flow of the stress-dependent vocalic gestures for syllable margins. In other words, the consonantal gestures are superimposed onto this base function reflected through the interaction among physical articulatory organs in their time courses of movement, as implemented by the signal generator. An inherent impulse response function (IRF) of each elemental gesture is evoked by the pertinent pocs pulse with their amplitude and timing according to the magnitude and time values of the pertinent syllable pulse, as discussed above.

The C/D model, in contrast to Articulatory Phonology, recognizes the distinction between phonology *vs.* phonetics. But the phonetic implementation process mixes numerical variables (pulse values) and symbolic variables (features). The phonetic system is language-dependent, as well as speaker idiosyncratic, through the choices of system parameters including the table entries of the IRFs.

Fig. 6 summarizes the discussion above by an overall picture of the prosodic organization of the utterance 'That's wonderful!' The thick solid vertical lines represent full syllables, the thin solid line the reduced syllable, and the thick dotted vertical line the phrase boundary (/%/). The dot-dash line running in the middle of the figure horizontally depicts the abstract vocalic base function representing the tongue body backing gesture for the {+-back} vocalic feature. See also the dashed line in Fig. 4 which similarly shows an abstract step function for tongue body fronting; the smooth solid curve superimposed shows a concrete version of this time function as a control function generated by a physically non-realizable filter (see Fujimura, in press) as an inherent property of the signal generator. The base function in jaw movement is also shown in Fig. 6. The mandibular movement is controlled to return to the rest position for reduced syllables as well as phrase boundaries, whereas the vocalic base function shows a linear interpolation ignoring these non-substantive elements. Note that the difference in mandibular position between the open vowels /æ/ and close vowel /ʊ/ is not reflected in the control function since this difference is explained by the interaction in the signal generator between the lingual gestures aiming at different target positions and mandibular opening, in the case of vowel gestures. In Fig. 6, some of the consonantal gestures are also exemplified as impulse response functions evoked in different articulatory dimensions (rounded, retracted, and rhotacized).

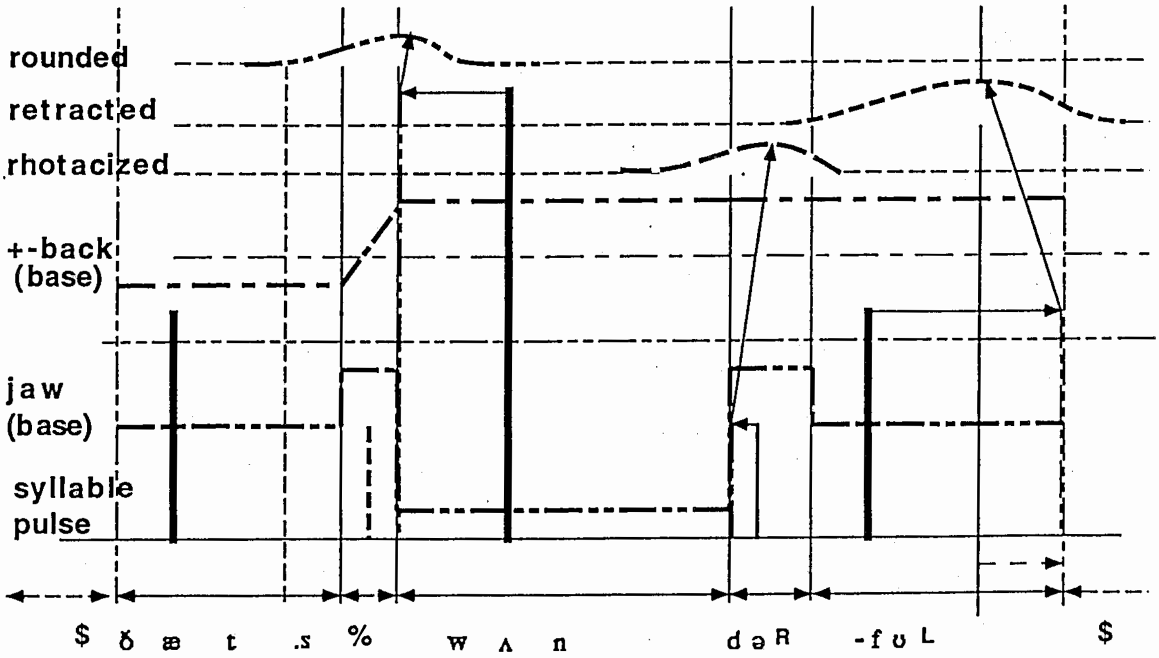

Fig. 6: Overall organization of an utterance 'That's wonderful'.

REFERENCES

Anderson, M., Pierrehumbert, J. B. & Liberman, M. Y. 1984. Synthesis by rule of English intonation patterns. *Proceedings of ICASSP '84.* 1, pp. 2.8.1- 4. Piscataway, N. J: IEEE Service Center.

Fujimura, O. (1979). An analysis of English syllables as cores and affixes. *Zs. f. Phonetik, Sprachwissenschaft u. Kommunikationsforschung,* 32, 471-476.

Fujimura, O., Erickson, D., Wilhelms, R. (1991). Prosodic effects on articulatory gestures: A model of temporal organization. *Proc. XIIth International Congress of Phonetic Sciences, Aix en Provence,* 1, 120-124.

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

Fujimura, O. (1994a). Syllable timing computation in the CD Model. *Proc. International Conference on Spoken Language Processing,* Yokohama, Sept. 94. (pp. 10-11)

Fujimura, O. (1994b). C/D Model: A computational model of phonetic implementation. In E. S. Ristad (ed.), *Language Computations:* (pp. 1-20). Providence: American Mathematical Society.

Fujimura, O. (1995a). Prosodic organization of speech based on syllables: The C/D model. Proceedings ICPhS 95, Vol. 3, pp. 10-17.

Fujimura, O. (1995b). The syllable: Its internal structure and role in prosodic organization. In B. Palek (ed.) *Proceedings of LP'94,* Prague, Czech Republic, Institute of Linguistic and Finno-Ugric Studies, Charles University.

Fujimura, O. & Erickson, D. (in press). Acoustic phonetics. In W. Hardcastle & J. Lever, (*eds.*), *Manual of Phonetics.* London: Blackwell.

Keating, P. A. 1990. Phonetic representations in a generative grammar. *J. Phonetics* 18, 321-34.

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

Sproat, R. W. & Fujimura, O. 1993. Allophonic variation in English /l/ and its implications for phonetic implementation, *J. Phonetics* 21, 291-311.

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