PROSODIC EFFECTS ON ARTICULATORY GESTURES -- A MODEL OF TEMPORAL ORGANIZATION

Osamu Fujimura, Donna Erickson and Reiner Wilhelms

The Ohio State University
Division of Speech and Hearing Science
Columbus, OH 43210-1002, U. S. A.

ABSTRACT

A theory of phonetic implementation based on articulatory gestures and their temporal organization is proposed. It is compatible with Öhman's early insight (consonantal perturbation), which in effect assumes a separate tier for vowel to vowel movement as the base, and consonantal gestures superimposed on this base. The segmental constituent units are syllables, each of which is specified by demisyllabic feature values. A generative description is given as a series of computational modules: a converter, a distributor, a concurrent set of actuators, and a signal generator. Implications regarding various conditions of prosodic control are suggested.

1. CONVERTER

A computational procedure is shown in Fig. 1. The converter, as the first component of our model of phonetic implementation, "converts" an abstract phonological representation to an annotated linear pulse train. The converter maps phonological feature specifications for each demisyllable into a set of articulatory gestures. In Fig. 1, τ represents a stop gesture, σ fricative, θ interdental, η glide, N nasal, λ lateral, ρ retroflex; T specifies apical articulation, P bilabial. The letter v stands for consonantal voicing. Vocalic gestures are separately treated, and are represented here by phonemic symbols (none for reduced vowels). Syllable affixes (see Fujimura [4]) are separated by a dot from the final demisyllable.

Time and magnitude are assigned to each pulse (represented by vertical lines). The pulses belong to one of two types: syllables (shown with thick vertical lines), or boundaries (thin vertical lines). Each pulse is associated with minimal phonological feature or boundary type specifications. The phonological tree and the metrical grid, or equivalent abstract representations, constitute the primary basis of computation of the magnitude values of all pulses. Prominence specifications (see an exclamation mark), reflecting factors such as contrastive emphasis, the degree of excitement, *etc.*, are also absorbed into the numerical specifications of time and magnitude. Utterance-related specifications (speaking style, speaker characteristics, *etc.)* are retained as annotations attached to utterance phrases. An utterance phrase constitutes the domain of motor programming as an integral unit of utterance (see Fujimura [6,7]), and affects both the impulse response functions and the parameters of the signal generator.

The timing characteristics of individual gestures are determined by the converter.[1] In Fig. 1, the i-th syllable pulse is located at t_i, and its height represents the magnitude μ_i assigned to each syllable.

Let us assume the interval between two contiguous syllables to be related to the pulse magnitudes by

$$t_i - t_{i-1} = \alpha\mu_{i-1} + \beta\mu_i,$$

where α and β are multiplicative constants which determine "shadows" of each syllable pulse on the right and left sides, respectively. A similar shadow is also defined on the left side of a boundary pulse. This results in a leftward shift of the last syllable in the phrase before the boundary, making the decaying effect of the syllable pulse response function part of the syllable duration, rather than part of the overlapping next syllable. This accounts for the preboundary elongation.[2]

[1] These time values are presumably readjusted *via* feedback signals from the signal generation process. For example, the articulatory repulsion, as discussed by Fujimura [5], apparently pertains to temporally adjacent gestures within the same articulator.

2. DISTRIBUTOR

The distributor distributes the codes produced by the converter to a concurrent set of actuators, each of which represents an articulatory dimension. An articulatory organ may be involved in defining multiple articulatory dimensions. An articulatory dimension may involve more than one organ. The distributor interprets the feature specifications for each demisyllable in terms of articulatory gestures, and distributes relevant syllable pulse information to individual actuators. In the figure, Greek letters in Italic represent the elementary gestures in the distributor output, and Roman capitals represent the specified articulators (see below for further explanation).A family of mathematical functions prescribes the elementary articulatory gestures as time functions represented in terms of a physical measure of the state in each articulatory dimension. A set of muscular units forms a configuration of physical means for implementing cortical control of a specific dynamic event. This integral configuration of each gesture constitutes an articulatory dimension, such as production of a laminar /s/. Separate articulatory dimensions are defined for different manners of articulation, such as stops *vs.* fricatives. The output of the distributor is a replica of the input for each actuator to the extent the information is relevant. Thus the code (N, T) standing for /n/ in the final demisyllable of the second syllable /wʌn/ is interpreted as $\{\tau,T\}$ for the tongue tip (T) closure (τ) dimension and operates in parallel with $\{N\}$ for velum lowering. The impulse response function for the $\{N\}$ gesture (in final demisyllable) is implemented with peak event at about t_2, whereas the $\{\tau,T\}$ gesture has its peak later (see Sproat and Fujimura [10] for similar situations of English laterals). This depiction may appear similar to Browman and Goldstein's gesture score [1,2].

[2] In addition, the parameters of the impulse response functions may be sensitive to the magnitude of the following boundary pulse.

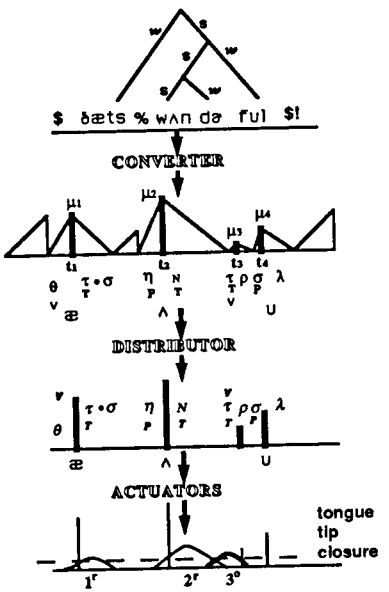

Fig. 1 A computational model of consonant implementation (signal generator omitted)

3. ACTUATORS

Each consonantal actuator receives the time-magnitude pulse information with respect to consonantal gestures for (1) an initial demisyllable, (2) a final demisyllable, (3) syllable affixes (in sequence) as applicable [4]. Different gestures are assigned to separate articulatory dimensions. Multidimensional processings take place concurrently in different actuators triggered by syllable pulses. Each actuator evokes the demisyllabic response, gesture by gesture. Elementary gestures do not require sequencing information within each demisyllable. The impulse response function contains a parameter that shifts the peak activity relative to the time value of the syllable

pulse. The parameter values are prepared in conformity with the inherent vowel affinity [3], [4]).[3] For consonantal gestures, each syllable impulse evokes an impulse response function in each of the relevant articulatory dimensions for each demisyllable. For example, the i-th syllable pulse in the phrasal domain under consideration has an assigned time t_i and a magnitude μ_i, and it triggers a feature event function $\mu_i\, g_\tau^o(t - t_i)$ in the dimension tongue tip raising, where $g_\tau^o(t)$ is the impulse response function for the gesture τ, stop closure, for the initial demisyllable (indicated by the superscript o). The time value t_i is designated for the i-th syllable. For the feature lateral, two articulatory gestures are relevant: (1) tongue tip raising for partial contact and (2) retraction of the back of the tongue body due partly to tongue blade narrowing [10]. The symbol λ in the distributor output in Fig. 1 stands for these two articulatory dimensions in an abbreviated form. Similarly, *N*, as in {*N,T*}, stands for the redundant *N* and τ of the nasal apical stop. The closure gesture τ is actually specified at the pertinent actuator, but is not shown in the figure. If the final consonant of the syllable is /m/, then the syllable pulse evokes the feature event function $\mu_i\, g_P^T\, (t - t_i)$ of the final demisyllable in the bilabial closure dimension *P*, and also $\mu_i\, g_N^T(t - t_i)$ in the velum lowering dimension N. This function for the final nasality shows a maximum velum lowering at about the peak occurrence of the syllable nucleus, *i. e.*, $t = t_i$. Each actuator, within the pertinent articulatory dimension, compiles the feature event functions evoked by the syllable pulses within the time domain of utterance phrase. The event time functions are added according to the linear superposition principle, and the resultant time functions for different articulatory dimensions, together with the vocalic base functions, are passed on to the signal generator. The same family of functions is used for prescription of all the consonantal gestures, with parameter values selected for individual articulatory dimensions. Each elementary event starts with the base position moving in the direction of the prescribed vocal tract constriction (in the three-dimensional sense), and then automatically returns to the base position. Different time constants are specified (in the gesture table for each actuator) for starting and ending trips. The lowest panel of Fig. 1 intends to suggest such occurrences of response events, for the final demisyllables of the first and the second syllables, and the initial demisyllable of the third, in the dimension tongue tip closure. The situation of the apparent target reaching short of the roof of the mouth, as in the case of /s/, is presumably a mechanical or physiological consequence of saturation in an inherently three-dimensional system.

[3] An autosegmental computation is assumed at the level of distributor for spreading redundant feature values. For example, in the vocal fold adduction dimension, one specification of voicedness (for the entire consonant cluster) indicated for each demisyllable will be distributed throughout the time domain of obstruent gestures and affixes.

4. SIGNAL GENERATOR

The signal generator, (not shown in the figure), receives the time functions generated by the total set of actuators, and synthesizes them with the vocalic base functions to materialize articulatory movement in an integrated system. Various types of interaction among articulatory dimensions are automatically treated by the physical model of the total system, both within the same articulatory organ (such as the lips or the tongue) and among different organs (such as the mandible and the lower lip). The system is highly nonlinear. In particular, it contains a strong saturating characteristic (soft clipping) so that a large syllable pulse typically results at the output of the signal generator in a plateau of articulatory position as a function of time. In Fig. 1, at the bottom of the figure, the horizontal dashed line indicates this "soft clipping" that takes place in the signal generating process.

The process of generating the (vocalic) base function differs from that for consonantal gestures. The syllable pulse magnitude is transformed by a nonlinear saturating function into a multiplicative factor that represents the degree of achieving the vocalic gesture target, relative to the neutral vocal tract condition (schwa gesture). The response function parameters are assigned for vocalic gestures in such a way that the peak position occurs with no delay relative to the input impulse. Target positions are specified for the peak.

5. CONCLUDING REMARKS

Our current data, obtained by the Wisconsin microbeam facility [8], concern the pellet positions representing sample flesh points on the surface of the articulatory organs along a particular direction of movement, as a crude approximation for the state variable in each articulatory dimension. More exactly, in our future work, the observed pellet time functions will be compared with predicted output functions using a dynamic three-dimensional computational model of the articulatory system. This computational model is currently being developed and will constitute the principal part of the signal generator.

6. REFERENCES

[1]BROWMAN, C.P. & GOLDSTEIN, L. (1985), "Dynamic Modeling of Phonetic Structure", in V. A. Fromkin (*ed.*), *Phonetic Linguistics -- Essays in Honor of Peter Ladefoged* (pp. 35-53), New York: Academic Press.

[2]BROWMAN, C.P. & GOLDSTEIN, L. (1989), "Tiers in Articulatory Phonology, with some implications for casual speech", in J. Kingston and M.E.Beckman (*eds.*) *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (pp. 341-376). Cambridge: Cambridge University Press.

[3]CLEMENTS, G.N. (1989),"The Role of the 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* (pp. 58-71). Cambridge: Cambridge University Press.

[4]FUJIMURA, O. (1979), "An Analysis of English Syllables as Cores and Affixes", *Zs. für Phonetik, Sprachwissenschaft und Kommunikationsforsch. 32*, 417-476.

[5]FUJIMURA, O. (1986), "Relative Invariance of Articulatory Movements: An Iceberg Model", in J.S. Perkell and D.H. Klatt (*eds.*) *Invariance and Variability in Speech Processes* (pp. 226-242). Hillsdale, N.J.: Lawrence Erlbaum Associates.

[6]FUJIMURA, O. (1990a),"Articulatory Perspectives of Speech Organization", in W.J. Hardcastle and A. Marchal (*eds.*) *Speech Production and Speech Modelling* (pp. 232-342). Dordrecht: Kluwer Academic Publishers.

[7]FUJIMURA, O. (1990b), "Methods and Goals of Speech Production Research", *Language and Speech 33*, 195-258.

[8]NADLER, R.D., ABBS, J.H. & FUJIMURA, O. (1987), "Speech Movement Research Using the New X-Ray Microbeam System", *Proc. 11th ICPhS*, Tallinn, Se 11.4.

[9]ÖHMAN, S.E.G.(1967),"Numerical Model of Coarticulation", *J. Acoust. Soc. Am. 41*, 310-320.

[10]SPROAT, R.W. & FUJIMURA, O. (1989), "Articulatory Evidence for the Non-categorization of English /l/-Allophones", *Linguistic Soc. Am. Annual Meeting, Dec.89,Washington, D.C.*