FORMAL AND STATISTICAL MODELS OF SPEECH TIMING: PAST, PRESENT, AND FUTURE

George D. Allen, Dental Research Center, University of North Carolina, Chapel Hill, N.C. 27514, USA

Let us begin this paper, whose goal is to review the kinds of models that have been developed in studies of timing in speech production and to suggest some possible directions for further research, by addressing briefly the general nature of models. Although the usual sense of the word "model" is that of an analogy, there is much room for differences in usage. On the one hand, we can have a "descriptive model", which models a set of observations, or data; a full-blown theory, on the other hand, models a complex and usually interacting set of constructs. Intermediate between these two extremes lies the single hypothesis, which is a projection of a theory onto a subspace of smaller dimensionality (often a single dimension) and which is "tested" by comparing it with a set of data. There are no theoretical boundaries between these three types of models, and most studies which model some aspect of speech timing contain elements of two (but seldom all three) of these categories.

Besides differing in the complexity of the structures which they reflect, models also differ in the intended accuracy with which they reflect those structures. Some models, for example, are intended primarily as conceptual guides, with only a loose fit between them and any existing data. Such models motivate the design of further studies and the analysis of data gathered by them, with the usually explicit goal of validation and refinement of the original model. The "chain" and "comb" models suggested by Bernstein (1967) were of this sort and have served as the basis for many recent studies of speech timing control. Other models are tailored closely to data or some other real world phenomena, their intent being more to parameterize the data (for example to permit comparisons of these parameters between different groups of speakers) than to explain the process whereby the data are generated. Klatt's (1975) study is an example of this kind of data-matching model. As with the descriptive vs theoretical distinction mentioned above, these extremes also allow much room for differences among models: about the only commonality among models in their "goodness-of-fit" to the data is that no model fits as well as its proponent would like.

A third difference among models is what I have chosen to term "formal" vs "statistical", though here again there is no true boundary between them. This contrast is exemplified by the difference between "regression" and "correlation" in statistics, the first being used to describe the form of the relationship between two measures (e.g., if A is 10 cm taller than B, then A may be expected to weigh about 5 kg more), the second an estimate of the strength of that relationship (e.g., A will weigh 5 ± 1.4 kg more, 95 percent of the time). Lindblom and Rapp (1973) have thus in this sense developed a formal model of segment duration, whereas Kozhevnikov and Chistovich (1965) carried out the first of many statistical studies seeking significant negative correlations between the durations of successive segments as evidence of temporal compensation within production units.

Let us now review some past and present models of speech timing and its control in terms of these different general features. This review unfortunately cannot begin to cover the wealth of studies that now exist in this area. It would be useful, for example, to try to relate models of production to models of perception. Instead I shall restrict attention here to just a representative sample of models of timing in speech production and hope that perhaps the symposium itself will bring about the more complete discussion this topic deserves.

One major class of models has concerned the durations of segments, the earliest studies dealing with vowel duration in English (House and Fairbanks, 1953; Peterson and Lehiste, 1960; Kim, 1966). Although all of these were primarily descriptive in their origin, Kim's was the most theoretical in its intent. By explicitly labeling the branches on his tree with fixed durational values to be attributed to plus- vs minus-tense vocalic nuclei or plus- vs minus-voicing of the arresting consonant, he cut some of the ties his model had to the data which generated it and aligned it as well as he could with the constructs of distinctive feature theory.

More recent models of segment duration are those of Lindblom and Rapp (1973) and Klatt (1975), mentioned earlier. Both of these are descriptive models, though the data they describe, and thus their derivative models, are different. Lindblom and Rapp used phonologically restricted nonsense material and described variations in a segment's duration as a function of the number of segments, syllables, and words following it in the phrase. Klatt, on the other hand, used a meaningful paragraph, sacrificing control over word- and phrase-length comparisons while retaining contrasts in local segmental and prosodic context and adding syntactic contrasts. Interestingly, both of these studies describe segment variation as a contextually conditioned reduction in duration from a longest "base" form; several other related studies (e.g., Nooteboom, 1972; Umeda and Coker, 1975) have done the same, and Keating and Kubaska (1978) have suggested a role for this process in speech development.

Although these carefully constructed models are in substantial agreement as to the major dimensions required for describing the durations of segments in the phonologically restricted speech samples from which they were derived, other investigators have suggested that they do not model "real" speech. Umeda and Coker (1975), for example, present an alternative model, based again on measured segment durations but from corpora that are less constrained by laboratory conditions than, say, Lindblom and Rapp's (1973) or Nooteboom's (1972) data. Their data, and therefore their model, show the same local contextual effects as the others' (e.g., neighboring segment and syllable types, degree of stress, syntactic word classes), but the longer term effects (number of syllables remaining in the word, and words remaining in the phrase) are absent. This difference shows clearly one of the principle hazards associated with models derived from data: an apparently important component or dimension of the model may turn out to be an artifact of the observational situation. In this particular case the issue remains open.

There are many other studies of segment duration that deserve recognition here, and much more that might be said concerning those studies which have been mentioned. Because of space limitations, however, let us move on to a second major class of speech timing models, those which have dealt with the control of the articulatory time program. Aside from the oversimplified but heuristically useful "isochronic" model of English stress (cf., e.g., Pike, 1945), the first model of speech timing control appeared in Kozhevnikov and Chistovich (1965). As noted earlier theirs tried, via statistical techniques, to identify temporal compensation within production units, their underlying goal being to validate either the "chain" or "comb" model proposed by Bernstein (1967). Because of procedural artifacts inherent to their method, however, they recognized that they could not decide the issues from their data, and so they abandoned the temporal domain in favor of the articulatory. Some later investigators (e.g. Lehiste, 1972; Wright, 1974) were not so cautious and claimed evidence for temporal compensation in spite of warning by Kozhevnikov and Chistovich (1965) and Ohala (1970) that variations in speech rate and measurement error could mask any true effects. Allen (1973, 1974), on the other hand, tried to circumvent the methodological problem by proposing a statistical model which used a statistic that was insensitive to rate variations and by including an explicit estimate of measurement error. In agreement with Ohala (1970, 1975) he found no evidence for temporal compensation within the freely spoken phrase, thus supporting the "comb" model (though only weakly, since a statistically negative result can never be strong evidence for any hypothesis).

In addition to examining the relative validity of the "chain" and "comb" models, Allen's model had the additional advantage of yielding a measure of the speaker's timing control accuracy. In one study (Cooper and Allen, 1977) this model was partially validated using speakers whose timing control was known to be poor, and in another (Tingley and Allen, 1975) the developing ability of children's speech timing control was charted. As a result of these limited successes, Allen (1978) suggested that the methodological limitations inherent in earlier statistical approaches to the study of speech timing control may yet be overcome.

Although Kozhevnikov and Chistovich (1965) and most other investigators were seeking to discover units of speech production, Allen (1973) was equally interested in determining the nature of the mechanism for speech timing control. Following Huggins (1972), Allen distinguished two possible models for such a mechanism ("capacitor discharge" vs "neural counter") and discussed evidence for and possible consequences of each. For example, although Creelman (1962) writes that his data are incompatible with any periodic clock for temporal discrimination, thus arguing against a cyclically activated neural generator, both Michon (1967) and Kristofferson (1976) present data with distinctly periodic components. No direct comparison of the various models suggested so far for controlling speech timing has been performed, however, and the issue remains open.

This brief sampling of models of speech timing may be summarized as follows. (1) Most studies modeling timing in speech production either have described the temporal properties of known production units, such as segments, or have sought evidence of unknown units or the mechanisms whereby they are produced. (2) Although there have been some methodological differences among studies, their results have been in substantial agreement, at least within major classes of models. (3) Many important issues raised by these studies are apparently testable, but great care will be required to avoid methodological pitfalls.

What is the shape of things to come in this area of study? Will tomorrow's models be refined variants of today's, or will new concepts force a radical restructuring of our thought? The answer, I believe, is "both". For some purposes, such as practical speech synthesis, refinements and straightforward extensions of present descriptive models will be adequate for some time. Here the output must be acceptable as fluent speech, but the process by which it is generated need not model human (neuro-) physiology.

There is already under way, however, at least one radical restructuring, which will affect profoundly the form of models of speech production and perception. Turvey (1975) and several of his colleagues have argued persuasively for what they call an "action theory" of speech production, in which the motor system's normal reflexes are organized into ever higher levels of coordination, the highest level of all being sensibly describable only in terms of the overall goal, or plan, of the action. Such mainstays of traditional speech production research as "segment", "coarticulation", and "motor unit" become, in this view, projections of the plan onto subspaces of greatly reduced dimensionality, so simplified in most cases as to obscure the "true" process of production. Fowler (1977) has examined the implications of this kind of theory for models of speech timing, giving us a good opportunity to glimpse at least the immediate future.

At a rather deep level of conceptualization, we may see more explicit appeal to the goals of the speech timing model; that is, it will be not only acceptable but even necessary to consider the function of temporal structure in order to understand adequately what we observe. For example, such a statement as "Speech is made to be spoken" (Allen, 1975) will become literal rather than figurative truth.

Models of "intrinsic timing", as Fowler (1977) terms them, may impose far more explicit constraints on the domain of control than do many present-day models. Since in that view the temporal figure is as much a part of the speech act as, say, its neuromuscular features, intrinsic timing is an inherent property of the act, coterminous with it, not something that is imposed on it by an external timing generator that exists before and after as well as during. Hence it would be improper to speak, for example, of "the effect of speaking rate on segment duration", since the effect is really on the whole structured act within which the segment is embedded.

Some models already refer explicitly to domains of temporal constraint. Lindblom and Rapp (1973), for example, use one parameter to describe the effect of the number of syllables following within the same word and a second for the number of words following within the same phrase. Allen (1973) restricts his model of timing control to effects within the breath group. Other local constraints, such as neighboring phonemic context, are commonly imposed. Even so we may soon find the focus changing in our consideration of domains of temporal constraint; since the timing is intrinsic to the act, we would seek either to isolate acts as delimiters of temporal domains or to identify differences in timing control as evidence of action boundaries. We have often done this before, but usually intuitively or even unconsciously, and with segmental phonology and orthography as our guides. Following "action theory" into hierarchical systems of coordinated reflexes may bring us some interesting surprises.

Finally, we should still find as much need in models of intrinsic timing as in our present models for the notions of "temporal compensation" and "timing control mechanism" ("clock"). The assumption that motor action plans are organized hierarchically implies that temporal compensation will appear at all levels below the very highest; otherwise the temporal figure could not be intrinsic to the plan. Moreover, as long as neuromuscular events within the plan do not follow rapidly one upon the other, as fast as the associated lowest level reflexes allow, a controlling mechanism must be assumed to decide when to move on to the next. It could be proposed that the neural structures and pathways responsible for coordinating the action of muscles in space are simultaneously responsible for their temporal patterning as well, i.e., the plan is its own clock. The dissociation of temporal from spatial control in such dysrhythmic conditions as cerebellar ataxia, however, suggests strongly that a separate mechanism will continue to be needed in adequate models of timing in motor action plans.

In conclusion, we may expect that descriptive models of speech timing will continue to be elaborated, with fairly clear lines of historical development from the very earliest descriptions of segment durations. Theoretical models, on the other hand, may be about to undergo substantial modification, as we revise our conceptualization of the speech production process and of the relationship of timing to that process.[1]

References

Allen, G.D. (1973): "Segmental timing control in speech production", JPh 1, 219-237.

Allen, G.D. (1974): "Measurement error in speech timing studies", JASA 55, Suppl. 1, S42 (abstract).

Allen, G.D. (1975): "Speech rhythm: its relation to performance universals and articulatory timing", JPh 3, 75-86.

Allen, G.D. (1978): "Vowel duration measurement: A reliability study", JASA 63, 1176-1185.

Bernstein, N.A. (1967): The Coordination and Regulation of Movements, Oxford: Pergamon Press.

Cooper, M.H. and G.D. Allen (1977): "Speech timing control in normal speakers and stutterers", JSHR 20, 55-71.

Creelman, C.D. (1962): "Human discrimination of auditory duration", JASA 34, 582-593.

Fowler, C.A. (1977): Timing Control in Speech Production, Indiana: University Linguistics Club.

House, A.S. and G. Fairbanks (1953): "Influence of consonant environment upon the secondary acoustic characteristics of vowels", JASA 25, 105-113.

Huggins, A.W.F. (1972): "Just noticeable differences in segment duration in speech", JASA 51, 1270-1278.

Keating, P. and C. Kubaska (1978): "Variation in the duration of words", JASA 63, Suppl. 1, S56 (abstract).

Kim, C.-W. (1966): "The linguistic specification of speech", UCLA: Working Papers in Phonetics, 5.

Klatt, D.H. (1975): "Vowel lengthening is syntactically determined in a connected discourse", JPh 3, 129-140.

(1) The support of NSF grant number BNS 7614345 and NIH grant number RR 05333-16 is gratefully acknowledged.

Kozhevnikov, V.A. and L.A. Chistovich (1965): Speech: Articulation and perception, Joint Publications Research Service, Washington, D.C., 30,543.

Kristofferson, A.B. (1976): "Low-variance stimulus-response latencies: Deterministic internal delays?", Perc.Psych. 20, 89-100.

Lehiste, I. (1972): "Timing of utterances and linguistic boundaries", JASA 51, 2018-2024.

Lindblom, B. and K. Rapp (1973): "Some temporal regularities of spoken Swedish", Papers from the Institute of Linguistics, University of Stockholm.

Michon, J.A. (1967): Timing in Temporal Tracking, Soesterberg, the Netherlands: Institute for Perception.

Nooteboom, S.G. (1972): Production and Perception of Vowel Duration, Doctoral dissertation, University of Utrecht.

Ohala, J.J. (1970): "Aspects of the control and production of speech", UCLA Working Papers in Phonetics 15.

Ohala, J.J. (1975): "The temporal regulation of speech", in Auditory Analysis and Perception of Speech, G. Fant and M.A.A. Tatham (eds.), New York: Academic Press, 431-452.

Peterson, G.E. and I. Lehiste (1960): "Duration of syllable nuclei in English", JASA 32, 693-703.

Pike, K.L. (1945): The Intonation of American English, Ann Arbor: The University of Michigan Press.

Tingley, B.M. and G.D. Allen (1975): "Development of speech timing control in children", Child Devel. 46, 186-194.

Turvey, M.T. (1975): "Preliminaries to a theory of action with reference to vision", in Perceiving, Acting and Knowing: Toward an Ecological Psychology, R. Shaw and J. Bransford (eds.), Hillside, N.J.: Lawrence Erlbaum Associates.

Umeda, N. and C.H. Coker (1975): "Subphonemic details in American English", in Auditory Analysis and Perception of Speech, G. Fant and M.A.A. Tatham (eds.), 539-564, New York: Academic Press.

Wright, T.W. (1974): "Temporal interactions within a phrase and sentence context", JASA 56, 1258-1265.