Model Building in Phonetic Theory

Mark Tatham

Reproduced from *Language Sciences*, February 1971. Copyright © 1971 Mark Tatham.

This is a revised version of a Forum Lecture given to the July 1970 meeting of the Linguistic Society of America.

This work was supported in part by the National Science Foundation through Grant GN–534.1 from the Office of Science Information Service to the Computer and Information Sciences Research Center, The Ohio State University, and in part by Grant B/SR/6733 from the Science Research Council to the Language Centre, University of Essex.

This paper sketches a tentative proposal for a speech production model that is linguistically dominated and that might also serve as the basis for speech synthesis by rule strategies based on articulation rather than on the more common acoustic basis. There is no doubt that speech synthesis has reached a level of sophistication where it can be used as an experimental tool in linguistics and it now makes sense to think of bringing together work in model building and speech synthesis. I do not intend to discuss the actual implementation of the model, but merely to underline that if it is to be tested using speech synthesis then it must be as complete and explicit as possible.

The model postulates that any one articulatory gesture is the result of

- the linguistically motivated desire to articulate a particular extrinsic allophonic segment,
- the operation of a motor procedural mechanism establishing the cohesion of this segment with syllabic context,
- the insertion of the composite syllabic–sized unit (of which the particular segment now constitutes a part) into the chosen rhythm and rate for this utterance,
- (in English — since most of our data is from this language. and I hesitate to generalise on this point — see the reference below to Kozhevnikov and Chistovich) the modification of segmental duration depending on the stressed/unstressed pattern within the rhythm (I do not intend to discuss this point in the present paper),
- the generating of a target program associated upwards (i.e., linguistically) with this segment and horizontally (i.e. motor-wise) with the syllable unit,
- the appendage of coarticulation limiting factors which determine the freedom of range of articulatory variables but which do not change the established target program.

In its simplest form the present model starts at a level representing the input to the motor control system of human speech. There is no reason to suppose that we could equate this level with that of systematic phonetics output from present day phonological models — I do not wish to involve myself here in questions of incompatibility between the competence oriented phonology and the performance oriented phonetic component (see Mansell, *forthcoming*).

One aspect of the input, however, will be the same for systematic phonetics and the input to the production model: that of lack of any absolute time indexing. Systematic phonetics has only notional time indicated simply by the linear sequencing of segments; a similar sequencing of segments will be the only temporal indexing required of the units input to the speech production model. One immediate function of the model will be the addition of durational markings to the segments.

A segmental–type input is thus assumed. The units of the segmentation will be extrinsic allophones, not phonemes — where by extrinsic allophone is meant an abstraction derived from a morphophonemic level *via* a phonology and embodying properties known by the speaker to be those of the sound pattern of language/his language, but not embodying properties known by the speaker to belong to speech production. Knowledge of speech production is properly a fact of the internal working of the phonetic component and not of its input. I am distinguishing between those voluntarily generated articulatory properties of a given segment and those involuntarily generated properties of the same segment (see Tatham 1969a).

The definition of segment at the input is different from that assumed by many researchers in synthetic speech (typical of whom is Mattingly 1968). It will not be the case, as is usually taken, that the input segments are phonemes plus those cases where the language has an idiosyncratic subdivision of phonemes that cannot be said to be coarticulatory in origin (the classical example in English being: L → {palatalized l, velarized l} — phoneme /L/ becomes palatal or velar [1] dependent on context. Besides such commonly recognized idiosyncrasies, the present model will require for every phoneme /X/ a similar transformation (say, to [x]), which will be environmentally constrained according to linguistic (i.e., phonological) demands. All phonemic segments will have undergone phonologically constrained transformations in the form of allophonizing rules before the start of the present model — where once again those allophonizing rules derive only the extrinsic and not the intrinsic allophones (thus excluding non–systematic and non–phonologically determined assimilation, for example). That this is the case is anyway a given assumption of this paper; but if the point need .be argued, then previous researchers have neglected the existence of a phonology whose function is to capture generalities of the sound pattern of a language and thus have an output characterized as extrinsic allophones. Inelegancies of phoneme input models can be seen as further evidence for the reasonableness of the position adopted here. I am not concerned with a device that reads the printed word (see Tatham 1970a).

An initial procedure for time indexing these segments has been suggested by recent experimental investigations in the temporal organization of speech. It seems to be the case that in C1VC2 utterances there is a temporal motor control link between the C1 and the V that we are unable to explain by any low level part of the motor system or coarticulatory effect (MacNeilage and Declerk 1968; Tatham and Morton 1968). It has not been shown experimentally where this cohesion might be introduced.

I am adopting the theoretical standpoint that the C1V cohesion occurs at a sub-phonological level [note 2]. Further, I suggest that the phenomenon can be regarded as an innate property of the motor cortex organizing the control of speech: I can find no evidence that this is not the case [note 3].

It is assumed, then, that C1V motor control cohesion is introduced during the process of motor encoding of higher level units. There is evidence from recent acoustic experiments (Slis 1968; Lehiste 1970) supporting the cohesion theory by showing that at the acoustic level there is less variation at different rates of utterance between the relative timings of Cl and V, than between V and C2 in the C1VC2 utterance. Furthermore, the duration changes observed in rate variation in the V and C2 elements are systematic — in fact, compensatory. Lehiste infers that it seems to be the case that an effort is made to maintain an overall CVC duration and that if the V or C2 depart from their 'normal' (originally intended) values a compensatory effect takes place in the other of the two segments [note 4].

Thus, it is suggested that (in the case of a CVC utterance) temporal motor cohesion between C1V and temporal compensation between V and C2 can form the basis of a stage in the model at the level of motor encoding of the input units that will have the function of adding a detailed time dimension that was not present at the input. This time is not absolute, since it will depend largely on the much higher level decision as to overall rate of utterance. Once that is decided, then given a few rules about what happens to particular environments under different rates, it should be possible to compute actual timings for each segment.

We could set up a simple notation illustrating the working together of cohesion and compensation:

1. +ClV–C2+

stands for a C1VC2 utterance such as the English word ‘cat’, where ‘+’ is a syllable boundary, and where juxtaposition (as C1V) is motor cohesion, and where ‘–’ is temporal compensation .

Using this notation, we might write the following:

2. +C1V–C2+C3V–C4+

where C2 and C3 are identical extrinsic allophones of the input string (e.g. in the utterance ‘black cat’).

This form of notation might be used to write rules of the following kind (again ‘black cat’ as the example):

3. (i) xC2 +C3y —» XC5y

where C2 = C3 = C5

3. (ii) xCV–CVy —» xCV+CVy

where (i) and (ii) are ordered.

Application of rule (i) to the input string +C1V–C2+C3V–C4+ will produce: .

4.1 +C1V–C5V–C4+

and application of rule (ii) will produce:

4.2 +C1V+C5V–C4+

thus accounting for the unreleased [k] of ‘black’ in this utterance at normal and fast rates.

The composite linkage notion proposed for temporally indexing C1VC2 monosyllables establishes the complete syllabic unit, *viz*. +CV(–C2)+, where C2 may be optional (Tatham and Morton 1968), but retaining identity of the internal constituents. Electromyographic data (MacNeilage and Declerk 1968, etc.) may indicate motor cohesion between certain segments, but it does not indicate loss of identity of the segments within the syllables. It is important to stress that we do not want loss of this identity at any point in the model.

Consider, as an illustration of the consequences of the question syllable listing *vs*. segment listing, two possibilities in the model, each requiring lookup tables providing non–temporal information:

Each possible syllable type (noting only, for the moment, the ClV part is listed as a non–analyzable unit exhibiting two temporally spaced targets. this possibility is not chosen for the model because (i) from a theoretical standpoint non–analyzability of a composite unit is rejected, and (ii) slips–of–the–tongue experiments may indicate that cohesion is not final (although we often cannot say whether it is at the phonological or phonetic level that some slips occur — see, for example, Boomer and Laver 1967).

Segment types are listed in the lookup tables together with an external and generalized set of rules that determine the cohesion. If cohesion is similar between all C1V possibilities this simply takes the form of a composite rule indicating in which motor parameters cohesion takes place and to what extent. This solution satisfies the theoretical criterion of maximal generalization and compares favorably with the listing system of (a). But even more importantly it enables the separation of definite language specific units — the phonologically determined input segments — from a complex non–language specific by–product of motor encoding — the syllabification.

I stated earlier that I wish to regard syllabification as an innate property of the motor encoding process; it may be necessary later to establish whether or not this process can be language specifically modified in a way exactly analogous to my proposed handling of the apparently language specific intrinsic allophones (Tatham 1969a).

So far in this paper I have been concerned with a discussion of the form of the input to the production model and outlining a procedure for organizing the input segments into time constrained syllabic type units. As mentioned above, a target programming notion of motor control is incorporated, and associated with this is the notion that phenomena such as coarticulation, overshoot, and undershoot are low level processes predictable by rule and operating right at the periphery. It is held that such low level processes are involuntary and true universals inasmuch as they reflect tendencies of the neuro–muscular/mechanical system (Whitaker 1970).

I have suggested (Tatham 1969a) that we can regard coarticulation, etc., as being universal while admitting the linguistically governed extrinsic control over such effects. What is meant by this is simply that such tendencies are 'known' and can sometimes be overridden for certain language specific purposes — or at least if not overridden then modified to the extent that definite limits can be set on any coarticulatory effect. Such a handling of the intrinsic allophone problem enables an explanation of the observed phenomenon that certain languages permit a particular segment [x] to be coarticulated to different extents. The fact of coarticulation itself is held to be the universal and must be accounted for as such; the fact that coarticulation is limited is seen as a separate parameter. Thus, instead of saying we have in language L, a coarticulation Y of a segment, but in language L2 a coarticulation Z for the same segment, we say that there is a general coarticulation K, which occurs in L1 as Ky and in L2 as Kz where y and z are language specific generalizable phonological demands. This approach renders transparent the relationship between Ky and Kz that is obscured in the non–associable descriptions Y and Z (see also Tatham 1970b).

Notice that if linguistically governed control is to be exercised over a system constrained coarticulatory effect, then there must be knowledge of that effect before the control can be exercised (a phonetic consequence of the theory) and knowledge that this control can be used for linguistic purposes (a phonological consequence when we want to consider where phonological units come from). L1 with three palatal consonants of the same manner type and L2 with five palatal consonants both share, say, a target value for one (say, the central–most) of their consonants — yet the range in Ll for that consonant will be greater than the range in L2 (principle of maximal differentiation — a psychologically/perceptually determined effect). The model postulates a target command that will be identical in its basic form for both L1 and L2. Feedback control will be 'set' for the different range limits: that such a possibility exists is well attested in the literature (see Matthews 1964, Ohman 1967, MacNeilage and Declerk 1968, Tatham 1969b, MacNeilage 1970, Hardcastle 1970, Whitaker 1970). But the feedback cannot be set unless there is prior knowledge stored somewhere of the coarticulation effect likely to occur and of the necessary procedure to be followed to contain that effect within the linguistically determined limits.

To summarize: I have been concerned with setting forth a speech production model that might form the basis of a control program for speech synthesis by rule. The model differs from existing synthesis models in that its input has undergone phonological transformations from the phonemic level. Durational indexing of the segmental input is handled by syllabification and the latter is postulated as a universal (i.e. innate) property, Coarticulation is seen as a true universal, while language specific coarticulatory variants are seen as special linguistically governed modifications of the universal.

NOTES

Of course, the problem is vastly complicated by the probable necessity of extending the model to handle a feature array as input, where it is the segmented individual features that have to be considered, the principle however would be the same.

I am side–stepping in the present issues being discussed in Wickelgren 1969, MacNeilage 1970, MacKay 1970, Whitaker 1970; it will be apparent from this paper that I am adopting a position opposed to the context sensitive associative chain hypothesis. The latter has a certain appeal in providing an easy answer to a good many problems, but I find it difficult to imagine a good counter example for the model, and contend that it may be unsuitable because of its power — let alone anything else.

Strictly from the point of view of model building, it is possible to construct a phonology operating on syllabic units, and this would provide an input to the phonetics already exhibiting the cohesion, but I prefer not to assume this form of input. At any rate, if we are to have a working model then the question of the form of the input and the point in the model at which syllabic sized limits have to be introduced must be decided one way or the other, while allowing, of course, the possibility that data will be forthcoming which might completely reverse the decision.

Notice, though, that these data are for English; Kozhevnikov and Chistovich (1965) report temporal compensation between C1 and V in similar situations in Russian.

References

Boomer, D. A. and J. D. M. Laver. 1967. Slips of the Tongue. *Work in Progress* 1. University of Edinburgh, Linguistics Department.

Hardcastle, W. 1970. The Role of Tactile and Proprioceptive Feedback in Speech Production. *Work in Progress* 4. University of Edinburgh, Linguistics Department.

Kozhevnikov V. and L. Chistovich. 1965. *Speech: Articulation and Perception*, Washington, D.C., Joint Publications Research Service 30.543.

Lehiste, Ilse. 1970. Temporal Organisation of Spoken Language. *Working Papers in Linguistics* 4. The Ohio State University, Computer and Information Sciences Research Center.

MacKay, D. G. 1970. Spoonerisms: the Structure of errors in the Serial Order of Speech. *Neuropsychologia* Vol. 8.

MacNeilage, P. F. 1970. Motor Control of Serial Ordering of Speech. *Psychological Review* 77.3.

MacNeilage, P. F., and J. L. Declerk. 1968. On the Motor Control of Coarticulation in CVC Monosyllables. New York Haskins Labs, *SR–12*.

Mansell, P. (*forthcoming*) Linguistic Parameters in Performance Models. *Occasional Papers*, University of Essex, Language Centre.

Matthews, P. B. C. 1964. Muscle Spindles and Their Motor Control. *Physiological Review* 44.

Mattingly, I. C. 1968. Synthesis by Rule of General American English. *Supplement to Status Report on Speech Research*. New York Haskins Labs.

Ohman, S. E. G. 1967. Peripheral Motor Commands in Labial Articulation. *STL–QPSR–4/67*. Stockholm: Royal Institute of Technology .

Slis, I. H. 1968. Experiments on Consonant Duration Related to the Time Structure of Isolated Words. *IPO Annual Progress Report* 3. Eindhoven, Institute for Perception Research.

Tatham, M. A. A. 1969a. Classifying Allophones. *Occasional Papers* 3. University of Essex, Language Centre.

Tatham, M. A. A. 1969b. The Control of Muscles in Speech. *Occasional Papers* 3. University of Essex, Language Centre.

Tatham, M. A. A. 1970a. Speech Synthesis: a Critical Review of the State of the Art. *International Journal of Man/Machine Studies*, Vol. 2.

Tatham, M. A. A. 1970b. A Speech Production Model for Synthesis–by–Rule. *Working Papers in Linguistics* 6. The Ohio State University, Computer and Information Sciences Research Center.

Tatham, M. A. A., and Katherine Morton. 1968. Further Electromyography Data towards a Model of Speech Production. *Occasional Papers* 1. University of Essex, Language Centre.

Whitaker, H. A. 1970. Some Constraints on Speech Production Models. Essex Symposium on Speech Production Models. In *Occasional Papers* 9, University of Essex, Language Centre.

Wickelgren, W. A. 1969. Context–Sensitive Coding, Associative Memory and Serial Order in (Speech) Behavior. *Psychological Review* 79. 1 .