A Speech Production Model for Synthesis-by-rule*

Marcel A. A. Tatham

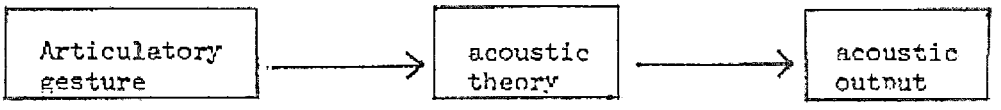
*Sponsored in part by the National Science Foundation through Grant GN-534.1 from the Office of Science Information Service to the Computer and Information Sciences Research Center, The Ohio State University.

A Speech Production Model for Synthesis-by-rule

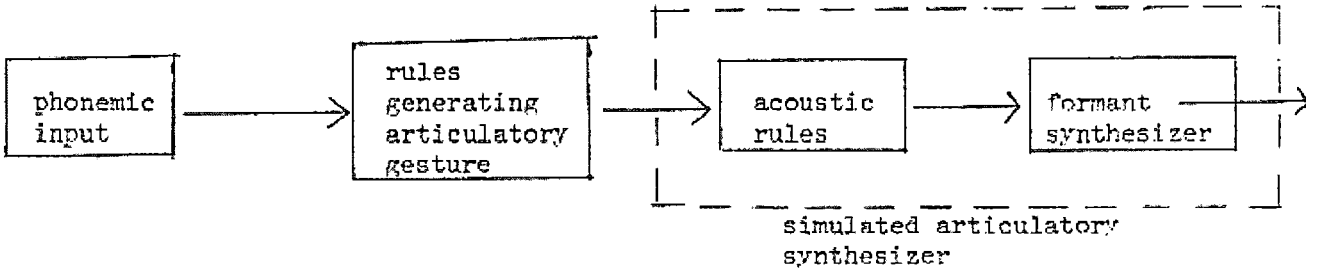
Marcel A. A. Tatham

1. Introduction

Problems in the design and operation of vocal-tract analog synthesizers suggest an interim handling of articulatory synthesis-by-rule. Relationships have been established between the acoustic output of speech and the articulatory gestures required to cause that output (Fant 1960, 1965; Flanagan 1965). These relationships are usually diagrammed:



As a preliminary then to articulatory synthesis the actual speech output from the device can be generated from a formant synthesizer which is immediately preceded by the rules of the acoustic theory, thus:



In the system described by Werner and Haggard (1969) a 'phonemic' input is turned into a number of time-varying articulatory parameters which correspond to spatial measurements in a stylized vocal-tract. A computer then applies a set of acoustic rules which are entirely extra-linguistic to generate from these articulatory parameters control signals which will operate a standard formant (or terminal analog) speech synthesizer.

Several very basic and influential assumptions underlie this approach, some of which reflect a non-linguistic viewpoint. Unlike

a true vocal-tract analog, where in the ideal situation the sounds produced could not sound or be other than absolutely natural, this system relies on a terminal analog synthesizer. I have discussed elsewhere (Tatham 1970) some of the difficulties of synthesis in general and indeed it is noted in Werner and Haggard that the conversion tables are not free from 'tricks' (p. 5) which are designed to make the speech sound more natural. But a more striking criticism can be levelled at the use of terminal analog synthesizers.

Historically the parameters of terminal analog synthesizers and their control have been derived as a result of two criteria: (a) visual and (b) perceptual. (a) It is quite clear from the literature since 1950 that the visual inspection of spectrograms has dominated these choices (Lawrence 1953; Liberman et al. 1954). It was observed that in spectrograms vowel-like sounds exhibit two or three major formant areas in the frequency/amplitude domain: often a (possibly correct, but this is an 'after-the-fact' discovery) assumption was made: such obviously audible and therefore acoustically major parameters were clearly going to be the most perceptually relevant and ought therefore to be synthesized faithfully. (b) Later, relevance of individual parameters was established using perceptual experiments (Liberman et al., 1954). There is no doubt that perceptual criteria dominate approaches in terminal analog synthesis today. That a variety of stimuli can often produce similar perceptual responses cannot be denied and the absurd limit might be reached where for the sake of economy (of computer time, output interface, programming, bandwidth restriction in telephony, etc.) the output of the terminal analog will reduce even more its identity with real speech--yet sound similar.

These criteria of economy are never justified on linguistic grounds and particularly hard to defend in terms of speech production. That a 'nasal' can be perceived by juggling with formant amplitudes only clouds accurate modelling of the speech production system and may even bias perceptual experiments--the fact that the perceiver can be fooled is comparatively trivial.

There is every reason to suppose that synthetic speech is a tool of considerable value in providing stimuli for perceptual experiments, but up till recently it has been a tool that may have been handled with more confidence than has been justified. It remains to be seen whether subjecting the listener to distorted artificial speech (rather than distorted real speech), as in the classical experiment by Broadbent and Ladefoged (1960), enables us to infer anything about the perception of real speech.

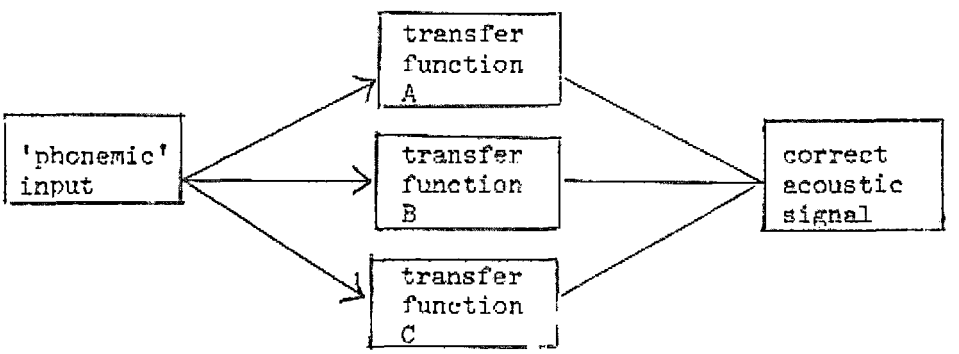
These of course are the reasons for preferring articulatory synthesis. As mentioned earlier vocal-tract analog synthesizers are still not entirely satisfactory--but there seems no reason to wait. The approach offered by the Cambridge group seems admirable: generate as an interim output articulatory parameters. Under ideal conditions these would be converted to parameter control signals for a vocal-tract analog, but pending this they can be converted to terminal analog control signals using the rules of the standard acoustic theory of speech production. If the speech is to be used

for perceptual experiments then absolutely nothing is gained, however: properly the system should stop at the articulatory level.

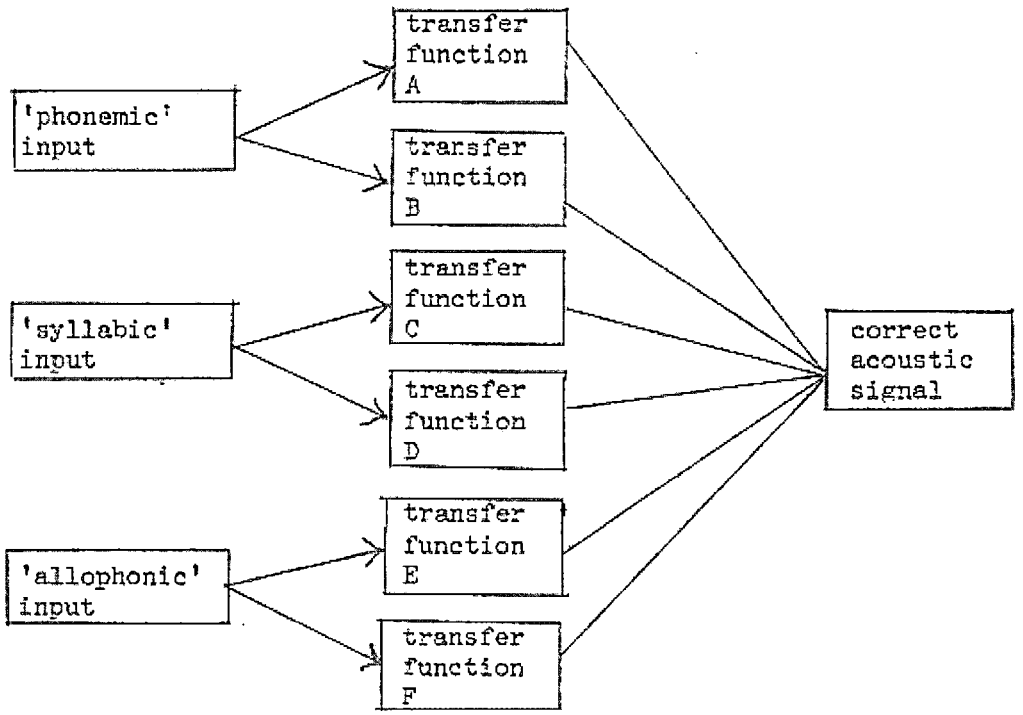
2. The Production Model

The present paper outlines a proposal for a model of speech production which is linguistically dominated and which might serve as the basis for an articulatory speech synthesis program. Linguistics has a lot to say about speech production which cannot be inferred from mere reproduction by lookup table of observed articulatory configuration, however. There is much more implied: it is not enough to generate correct vocal-tract shapes in an economical way for exactly the same reasons that it is not enough to generate 'correct' sounds. I have overstated my case deliberately; the incorporation as standard procedure these days of 'targets' and computed transitions based on contextual information is more than simply elegant--it does rest on linguistic theory: namely on the model which assumes that the phonological elements are phonemic (or quasi-phonemic) in nature and that allophones (or most of them) are the result of neuro-mechanical inertia at some low level in the system (Öhman, 1964). There is evidence that this view is inadequate however: the work of MacNeilage and Declerk (MacNeilage and Declerk 1968) has shown that there are segmental overlap phenomena not directly attributable to neural or muscular inertia (see also Tatham 1969a).

The synthesis strategy should be based on a coherent theory of speech production and the system described here will take into account one particular theory. But since there are competing theories a further function of the entire synthesis system will be to test the production theory. Notice, however, that the obtaining of a correct output is no test, just as the obtaining of correct perceptual response is no test of accurate acoustic signal. Consider, for example, the following possibilities:



In such a system there is no internal basis for evaluation of the model and the economy criterion mentioned above will not do as a measure of correctness. Consider the even worse situation:



Not only now is there no way of evaluating the various contending transfer functions but there is no internal way of evaluating the contending input types--but both these are candidates for evaluation and a measure of 'correctness' is crucial to the theory. Accordingly the transfer function must not simply be a mathematical formula which happens to provide the correct output from a particular input: there must be exterior constraints on this function.

Thus for example it is now established that muscle movement is not a continuously programmed system but a ballistic system controlled by temporally-spaced and non-continuous command situations. So at time T1 a muscle will be instructed to GOTO target t1 (this GOTO might be based on a given spatial movement, or on a given muscular tension), at time T2 an instruction will arrive to GOTO target t2 (a number of updating and correction signals may have arrived between T1 and T2)--but it is <u>not</u> the case that at T1 there is a 'start-moving' command and at T1+1 a 'move-a-little-more' command and at T1+2 a 'continue-moving' command and at Tt1 a 'right, -hold-it' command and at Tt1+1 a 'relax-a-little' command, etc.

Now what would these two different models of muscle command mean to the design of an articulatory speech synthesis system? The

'continuous command' theory would require the computation and supply of a moment-by-moment command signal at some level in the synthesis system corresponding to neural signals arriving at the muscles. The GOTO, ballistic theory would require a single computation (of the target value command) and the supply of just this single command (if necessary repeated for updating purposes) at the same level in the system.

The choice of the GOTO command in the synthesis strategy in the articulatory model presupposes a formula for the actual contraction of the muscle upon receipt of the full command--i.e., a factor expressing the inertia of the muscle in question. This factor is clearly derived at a level different from the level at which the instruction was computed (since it is a property of the muscle itself). But in addition it will also be argued that the computation could not have been accomplished without this piece of data.

For example, consider just one parameter of the muscle contraction: rate. In order to contract X amount, time Tx is necessary. Now, Tx will not alter the value (in a simple model) of command Cx, but it will alter its timing of delivery relative to the desired timing of achievement of contraction CNx. Thus a command signal Cx is computed based on at least two input channels: (a) the need for contraction of the particular muscle [command from a higher level] and (b) data about the rate of inertia of this muscle [data from a lower level].

Any synthesis strategy which began by observing in EMG or other data collection system that contraction for X began Tx before the achievement of X and simply arranged for this to be simulated would not even include the power of descriptive adequacy, let alone explanatory adequacy. But a strategy which includes the generalized data that this muscle is always (in the simple model) inert by a certain factor and computed a temporal change of the delivery of the instruction would provide explanatory adequacy--i.e., it would be able to predict in a transparent fashion the exact timing of the start of contraction of the muscle and would further provide the correct slope of the rate of achievement of that contraction.

In its simplest form then the model of speech production will have an input level to be equated with the input to the motor control system of human speech. It is not necessarily the case that this level is to be further equated with the level of systematic phonetics output from the phonological component of a transformational grammar--although this could be made to be the case.

Certainly there will be identity in the temporal respect. Phonology contains only a notional time: that of sequencing of segments. One immediate function of the production model is to transform this notional time into a less abstract time whose segments (whatever these should be decided to be) are organized on more than a sequential basis. Notice the important observation that such a model does not mention 'real-time'--indeed, it would be difficult to know what is meant by a 'real-time' model, in the

correct sense of 'real'. Possibly a real-time model of speech production would deliver an output from an input in exactly the same time as a human being and with that time subdivided in exactly the same way as in the human being. Systems are said to operate in real-time, which means that there is no storage or slow-scanning of the data to make up for deficiencies in handling capacity in the simulator. Real-time notions do not affect the validity of the model or its ability to test the accuracy of its assumptions. It would be easy in advocating a performance model to misuse the term real-time; performance models merely have time other than notional time--they do not need to be real-time models or systems.

The segmental-input type will be assumed. There is enough psychological evidence for us to assume that there is a reality to segments. Strings will be assumed to be segmented in terms of extrinsic allophones--not in phonemes. This is nearly the case with all synthesis-by-rule systems. However our definition of segments will need to be different from that already established by researchers such as Mattingly (1968). It is not the case that the input segments will be phonemes except where the language has an idiosyncratic subdivision of phonemes which cannot be said to be co-articulatory (the classic example is: $L \rightarrow \{ \overset{l}{\underset{r}{}} \}$ in English). We will adopt the theoretical standpoint that rules of this type (properly allophonic rules that form part of the phonology, rather than the phonetics) have been applied to all segments. Thus, besides $L \rightarrow \{ \overset{l}{\underset{r}{}} \}$, we will also have X→x and Y→y, etc.; it is enough to argue this point on the grounds of symmetry alone, but we assume also that any phoneme is subject to a group of allophonizing rules, one function of which (in the model) is to switch levels of abstraction.

Thus the input to the model is characterized as a level expressed in terms of extrinsic allophones (Tatham 1969b). Recent experimental investigations of speech indicate an important and initial factor immediately influencing the ascription of time features to these segments. It seems to be the case that in C1VC2 utterances there is a motor-control link between C1 and V which cannot be explained by any low-level system or co-articulatory effect (MacNeilage and Declerk 1967; Tatham and Morton 1968b). Where this linkage or cohesion is introduced is not clear.

Notice the theoretical standpoint has been adopted that postulates that the cohesion has been introduced at a sub-phonological level. The point still needs to be argued in publication, but we will assume for the moment that although it is possible to construct a phonological component based on syllable segments (or segments of a similar kind) this is a theoretically clumsy and non-productive concept in abstract phonological theory. We shall assume (possibly wrongly--but decisions need be taken in a working model that complete the system; this is the difference between a working and a non-working model) that initial-CV cohesion is established at the motor-level. It is not crucial to the model (since it will

satisfy the data without further speculation), but we might assume that the nature of the motor-control system is such that in speech this cohesion must be imposed.

Evidence from acoustic experiments (Lehiste 1970) supports the CV-cohesion theory, since these two elements remain non-compensatory--that is, complementary--under conditions of rate variation. The variation of rate is a factor to be accounted for crucially later. The data indicates that in cases of temporal strain on the overall word, compensation effects will occur between the V and C2 elements. This indicates temporal elasticity between V and C2 and temporal cohesion between C1 and V: that motor cohesion is observed (preceding paragraphs) is sufficient for us to introduce an actual linkage here which we could express with markers, thus:

≠C1V-C2≠	where ≠ indicates a syllable boundary, juxtaposition (as C1V) motor-cohesion and-temporal compensation.

operating as constraints in much the same way as +, ≠, etc., in the higher linguistic levels. The notation needs further thought because there will have to be rules deleting boundary symbols: these rules may have to be time-constrained. E.g., in cases of low rate speech we might well have

≠C1V-C2≠C3V-C4≠

where C2 and C3 are identical extrinsic allophones (e.g., 'black cat'); in high-rate speech we may want to add the rules:

(i)	xC2≠C3y	xC5y	where C2 = C3 = C5
(ii)	xCV-CVy	xCV≠CVy	where (i) and (ii) are ordered.

Thus so far extrinsic allophones have been linked (for English) in two ways: motor-cohesion and temporal compensation. This composite linkage provides us with a complete syllable unit [≠C1V (-C2)≠] which still retains identity of its internal constituents. This is important because at this point there are two possibilities in the speech-synthesis strategy:

> look-up tables providing (initially) non-temporal (from the segment-sequencing viewpoint) information are consulted. These tables can be organized in one of two ways: (a) syllable-types are listed, (b) segment-types are listed.

(a) Each possible syllable type (we are concerned here only with the C1V part) is listed as a non-analyzable unit exhibiting two temporally-spaced GOTO targets. This will not be chosen because (i) (a theoretical reason) non-analyzability is rejected; (ii) data (often derived from slips-of-the-tongue experiments) indicate that the cohesion is not final. [note, however: slips-of-the-tongue

experiments are confusing because there is often no evidence whether the slip has occurred at the phonological level (supports (a)) or at the phonetic level (supports (b)) (see, for example, Boomer and Laver (1967))].

(b) Segment types are listed together with an external set of rules (i.e. external to the segments) which determine cohesion. If cohesion is similar between all ClV possibilities this simply takes the form of a composite rule indicating in which motor parameters cohesion takes place and to what extent (NB the effect of this high-level cohesion on lower-level co-articulation, etc. will be discussed later). This solution satisfies the theoretical criterion of maximum generalization and compares favourably with the listing system of (a).

So far we have considered the characteristics of the input to the model and an initial stage intended to establish cohesions detected in experimental data. The theoretical model further assumes, as adjunct to the notion of GOTO control, that phenomena such as co-articulation are low-level rule-governed processes. These low-level processes are held to be true universals inasmuch as they reflect tendencies (predominantly inertial) of the neuro-muscular/mechanical system.

At this point it becomes necessary to discuss whether there is any attempt in higher-level programming to overcome such tendencies. So far, unfortunately, there is no definitive instrumental evidence, but it is assumed in the present model that (at least) a ternary system exists in the motor handling of (most of) the inertia-based effects. Inertia effects exist (they must, since all electrical or mechanical systems in the universe possess them)--the question is: are these effects handled in any systematic way; is any higher-level account taken of them? The ternary system in the model at the level postulates that one of three possible modifications exist: (i) counteract the effect, (ii) permit the effect, (iii) enhance the effect (these could be understood as -, 0, +, where 0 indicates the unmarked state). It is not clear from published data on co-articulation (Öhman, 1964, 1966, 1967; MacNeilage and Declerk 1968) (including over- and under-shoot) effects whether all mechanical or other inertia can be modified: presumably further data will be forthcoming; meanwhile the model will account, in the most simple way, for the existing data.

Thus, consider a language with only two palatal consonants of any one manner-type. Assuming a dominance of maximal differentiation (a psychological constraint) these will take the target forms of back and front (velar and alveolar, say), but this detail is comparatively unimportant. What is important is that the present model will predict a very wide variation in the point of contact of each consonant (but with little, if any, overlap) directly correlatable with segmental context. Thus preceding a front vowel, the consonant will exhibit a front allophone, etc. The model will further predict that this is the 0 or unmarked case--i.e., that there is no voluntary effort made to make the tongue less subject to context effect.

The model will predict, however, that, in another language where there are four such palatal consonants, (a) variation will again take place in exactly similar circumstances and (b) such variation will be very much more limited than in the case of the two-consonant language. The present model prefers to express this marked situation in precisely that two-level (or aspect) way maintaining the original inertia-derived rule and limiting it with a second, linguistically-determined rule. Thus the marking rule does not collapse two quite distinct and opposing tendencies--one quite a-linguistic and the other quite linguistic and concerned with maintaining perceptual clarity. Exactly the same phenomenon will be predicted for languages having a small number of distinctive vowel phonemes: the range of over-shoot and under-shoot variation will be considerable compared with a language with a larger number of vowels where the risk of perceptual confusion is that much greater if some kind of control is not exercised.

Notice that if control is to be exercised, the knowledge of the inertia effect must be possessed in advance by the control mechanism. This has got to be the case in this model; simple non-adjustable feedback systems cannot be relied on solely for one very simple reason [but there is an allowable alternative solution]: Language L1 with 3 palatal consonants and Language L2 with 5 palatal consonants both share a target value for one of their consonants--yet the range for L1 will be greater than the range for L2. But the model postulates a GOTO signal which will be identical in each case. Feedback cannot control the range of variation unless that feedback has been 'set' with respect to its limits: that such a possibility exists is well attested in the neuro-physiological literature (see, for example, Matthews 1964). But the feedback cannot be set unless there is prior knowledge of the inertia that will occur and the steps that must be taken to contain the variation within the linguistically determined limits.

It could be argued that a relationship exists between the linguistics system and the bi-level inertia system such that it becomes language idiosyncratic to establish a relationship between the systems resulting in what has been termed an 'articulatory setting' (Drachman 1970). That there is a tonic state of the musculature (called 'basic-speech-posture') is undeniable and similarly that certain languages exhibit a predisposition for certain prevalent (usually secondary) phonetic characteristics (like velarization, retroflexion, predominance of lip-rounding, etc.). But we have only to discover one language with a small number of vowel phonemes with wide articulatory variation and at the same time with a large number of palatal consonants with a small degree of variation for this hypothesis to become suspect.

A second argument against this hypothesis is that it lends too much status to the low-level systems and gets them unsystematically involved in high-level phonological processes by postulating that phonological processes 'carry-along' with them arbitrary handling of the muscular and articulatory system.

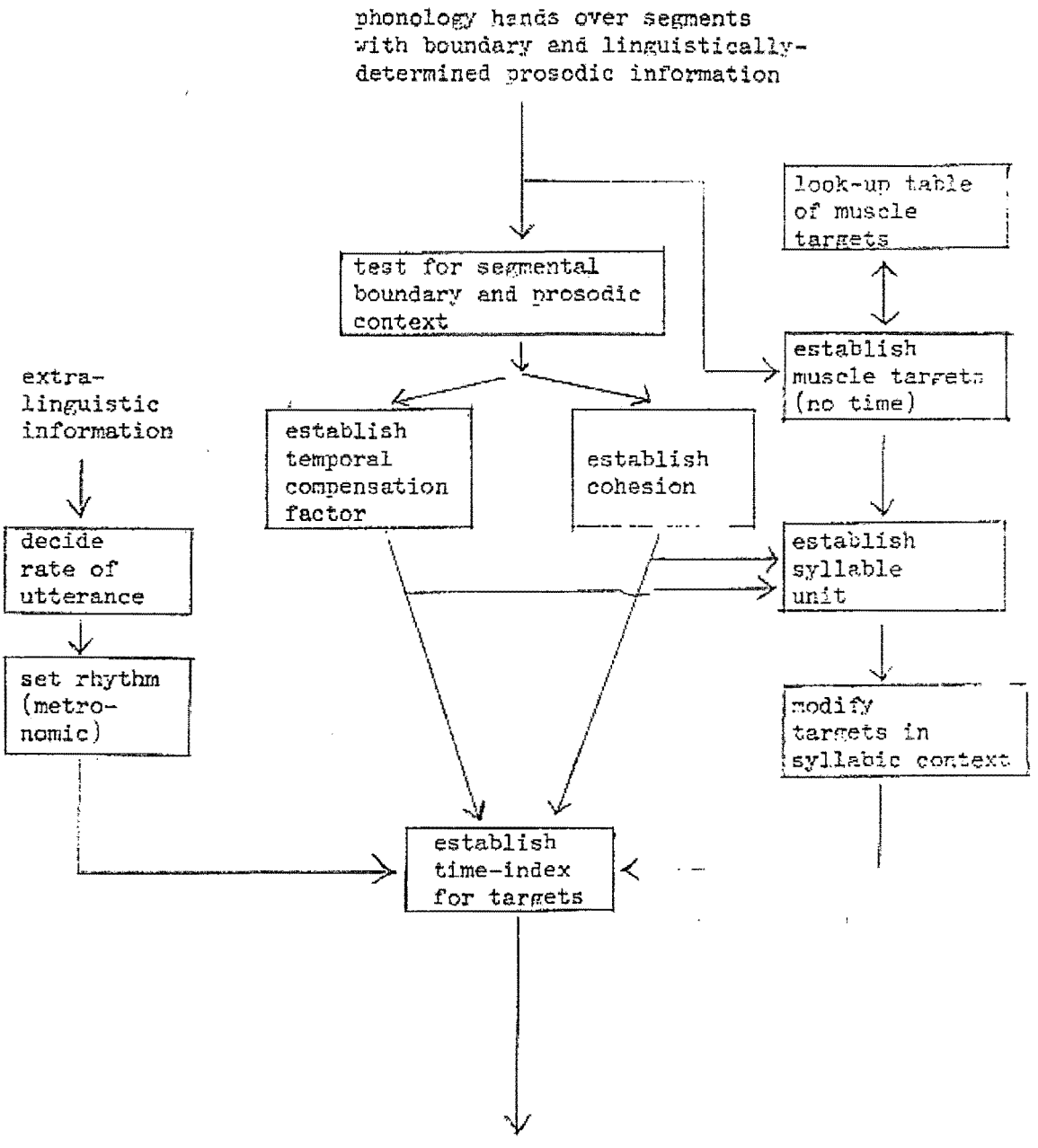
A third argument against this hypothesis is that it does not adequately account for the range of variation exhibited by a segment

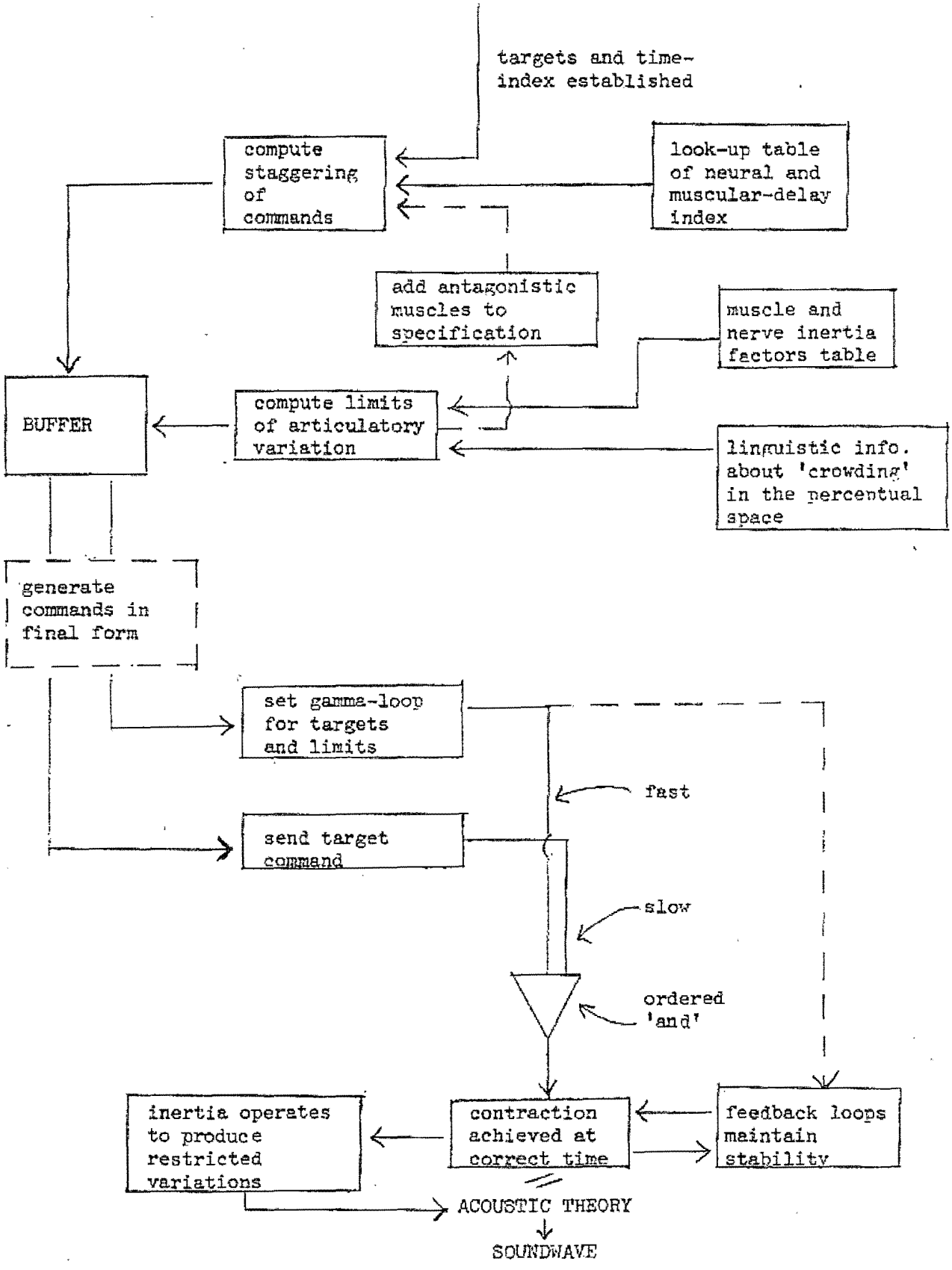
in a constant environment. If there were that much correspondence between mechanical inertia and phonology the articulation would be much more precise. The present model does predict a range of variation in the same segmental context because it only establishes limits for the variation, not new and different targets. I.e., the favoured model postulates a target, establishes the inertia formula, and establishes the limits to be imposed on that formula; the unfavoured model postulates a variety of a relationship established between mechanical tendencies and linguistic demands further resulting in an agreement for a particular and new target for each allophone [EMG records do not show that there is any contextual variation of this kind--but interpretation of such data is as yet only scantily formalized (Cooper 1965)].

Thus a particular articulatory gesture is the result of (a) the linguistically-motivated desire to articulate a particular extrinsic-allophonic segment, (b) the operation of a motor procedural mechanism establishing the cohesion of this segment with syllabic context, (c) the insertion of the composite syllabic-sized unit (of which this segment now constitutes a part) into the chosen rhythm or rate for this utterance, (c), in English, the modification of segmental duration depending on the stressed/unstressed pattern within the rhythm, (e) the generating of a target program associated upwards (i.e., linguistically) with this segment and horizontally (i.e., motor-wise) with the syllable unit, (f) the appendage of co-articulation limiting factors which limit the freedom of range of articulatory variables but which do not change the established target program.

Fig. 1 represents a simplified version of the present model. Some boxes are tentative (such as the setting of the gamma-loop system) but their function must occur somewhere to complete the system: they may just be in the wrong place or attributed to the wrong external mechanisms--what is correct about them is that, if included, then this model satisfies in a true explanatory way, the observables.

Fig. 1: TENTATIVE BLOCK DIAGRAM OF PROPOSED SPEECH PRODUCTION MODEL




targets and time-index established
compute staggering of commands
look-up table of neural and muscular-delay index
add antagonistic muscles to specification
muscle and nerve inertia factors table
BUFFER
compute limits of articulatory variation
linguistic info. about 'crowding' in the perceptual space
generate commands in final form
set gamma-loop for targets and limits
fast
send target command
slow
ordered 'and'
inertia operates to produce restricted variations
contraction achieved at correct time
feedback loops maintain stability
ACOUSTIC THEORY
SOUNDWAVE

Operations in Fig. 1

1. Input from the phonology decides which segment (=extrinsic allophone in sequential context with morpheme and word boundary symbols, stress pattern, etc.) is required at a particular point in the utterance to be generated.
2. Test for segmental context:
   (a) utterance (or inter-pause group): initial, medial, final?
   (b) (any) sub-utterance group: initial, medial, final?
   (c) word initial, medial, final?
   (d) morpheme initial, medial, final?
   (e) syllable initial, medial, final?
3. Establish whether motor-cohesion (lack of temporal compensation) or temporal compensation is to operate (this depends on the answer to 2).
4. (from 1) Look up muscle targets. If targets are expressed in terms of degrees of muscle contraction we have to ask the theoretical question: should targets for all muscles be specified irrespective of whether or not some are not involved in this particular segment, i.e., should a marked/unmarked system of classification be introduced? A combination of full entries with marked system would produce a segmentally determined feature hierarchy which is an issue of theoretical importance.
5. Establish relationship between targets and cohesion and compensation--i.e., establish syllable unit.
6. Decide overall rate of utterance.
7. Set rhythm generator according to 6: (i.e., provide metronomic determination).
8. Establish how each segment (incorporating 5) will behave temporally in the rhythm established by 7.
9. Hand over the information about the segment to the motor-command generator (this must be buffered to allow command-initiation overlap).
10. Establish information about any neural line-delay in the system.
11. Establish the muscle-response delay.
12. Construct a muscle command ordering dependent on 10 and 11 (at this point commands for individual muscles involved in the production of a particular segment are no longer temporally synchronized).
13. Consult table about muscle (and other) inertia factors.
14. Bring linguistically-determined information about any limits to be imposed on any articulatory variation which is likely to occur.
15. Recruit any additional (antagonistic included) muscles which may be necessary to maintain the limits of articulation variation decided under 14.
16. Set gamma-loop for targets and limits.
17. Send command at appropriate time (notice that this model does not assume the possibility of any low-level sequential triggering of commands).
18. The muscles contract within the specified limits, beginning at the correct time to achieve synchrony of articulator movement associated with a particular segment. Notice that EMG data shows that contraction seems to have finer temporal than amplitude

limits (Tatham and Morton 1968a).
20. Apply the acoustic theory.
21. Output soundwave.

Implementing the above model is well-nigh impossible, for several reasons, principal among which is that there is just not enough data for most of the boxes (even if the boxes themselves are correct). Take, as an obvious example, the temporal compensation and motor-cohesion boxes: that these two phenomena exist seems likely at the present time, as we have shown, but even a simple descriptive statement of their details does not exist yet. For the moment this does not matter. What does matter is attempting to use the model's implications for synthesizing speech even if we have to guess at individual values for any item. Guessing reduces reliability of using the working model for perception research, but it is a way of getting at the details for production research.

Before beginning a description of the synthesis strategy let us recapitulate the most fundamental assumptions of the present model--which (however grossly) would need their respective representations somewhere in the synthesis system.

Fundamental Assumptions of the Speech Production Model

A The input shall consist of individual segments which shall be extrinsic allophones bearing only notional time marking in the form of simple sequencing.
B The input shall be indexed with boundary symbols, such as: utterance, group, word, morpheme, syllable.
C The input shall be indexed with certain prosodic features, such as: stress (lexical, group, sentence), intonation (possibly only if marked, but suspect all).
D Also input will be (extra-linguistic?) information derived from decisions about the overall rate of utterance.
E Speech production is ultimately reducible to articulatory targets (though whether these are stored as representations of shapes, sounds, muscle commands, etc., is unknown).
F These targets are constant within individual for a particular language (irrespective of final output rate, co-articulation, segment position within the syllable, etc.).
G The hypothesis is adhered to for the moment that a-linguistic motor control dominates the syllabification of segment sequences at the periphery.
H Rate of utterance does not dominate the programming of targets but merely provides a factor which will enhance the effects of system inertia.
I A function of the motor control system is to stagger (negatively or positively) individual muscle commands to achieve desired articulatory movement at the correct time--the theoretical standpoint is taken that it is not until this late time that staggering occurs. Staggering is computed according to lookup table.
J A lookup table containing inertia factors reacts with command staggering and antagonistic systems, together with psychological/

perceptual information about the crowding status of the perceptual space, to compute limits of co-articulation and variation (over- and under-shoot) which may be permitted.

K Some kind of buffer is required at the point of staggering, serving two functions: (a) it permits successive passes of data for re-arrangement for staggering and (b) permits 'hold' facilities for output to final peripheral mechanisms of the motor-command signals.

L This model (despite Wickelgren 1969) holds that gamma-loops and any similar mechanisms are used for two functions: (a) to 'set' limits and (b) to hold them: it further holds that

M gamma-loop systems are used to provide information about the prior state of the muscles which results in a left-to-right effect observed in the final output; and further

N that fast conducting neurons will permit command signals to arrive at a muscle during the previous segment, thus generating the right-to-left effects observed in EMG and other data (MacNeilage and Declerk 1968).

O A composite signal arrives at the muscle which is a temporally-governed transform of the original extrinsic allophone lookup table target values. This signal embodies: (a) the target value (re-computed); (b) a temporal element (which may just be a re-issuing of the same command for a given period of time); (c) limiting factors to govern phenomena associated with any succumbing to inertia.

P It is therefore held as a theoretical tenet that at the final stages of articulation, universal (that is: always operating within a particular speaker under normal conditions and comparable with similar effects in other speakers) constraints apply to the transformation of the input signal to the muscles to produce the output configuration. This constraint system is rule-governed (i.e., quite predictable) and 'known' to the system which has used this information to compute the limitations or counteraction measures to be applied.

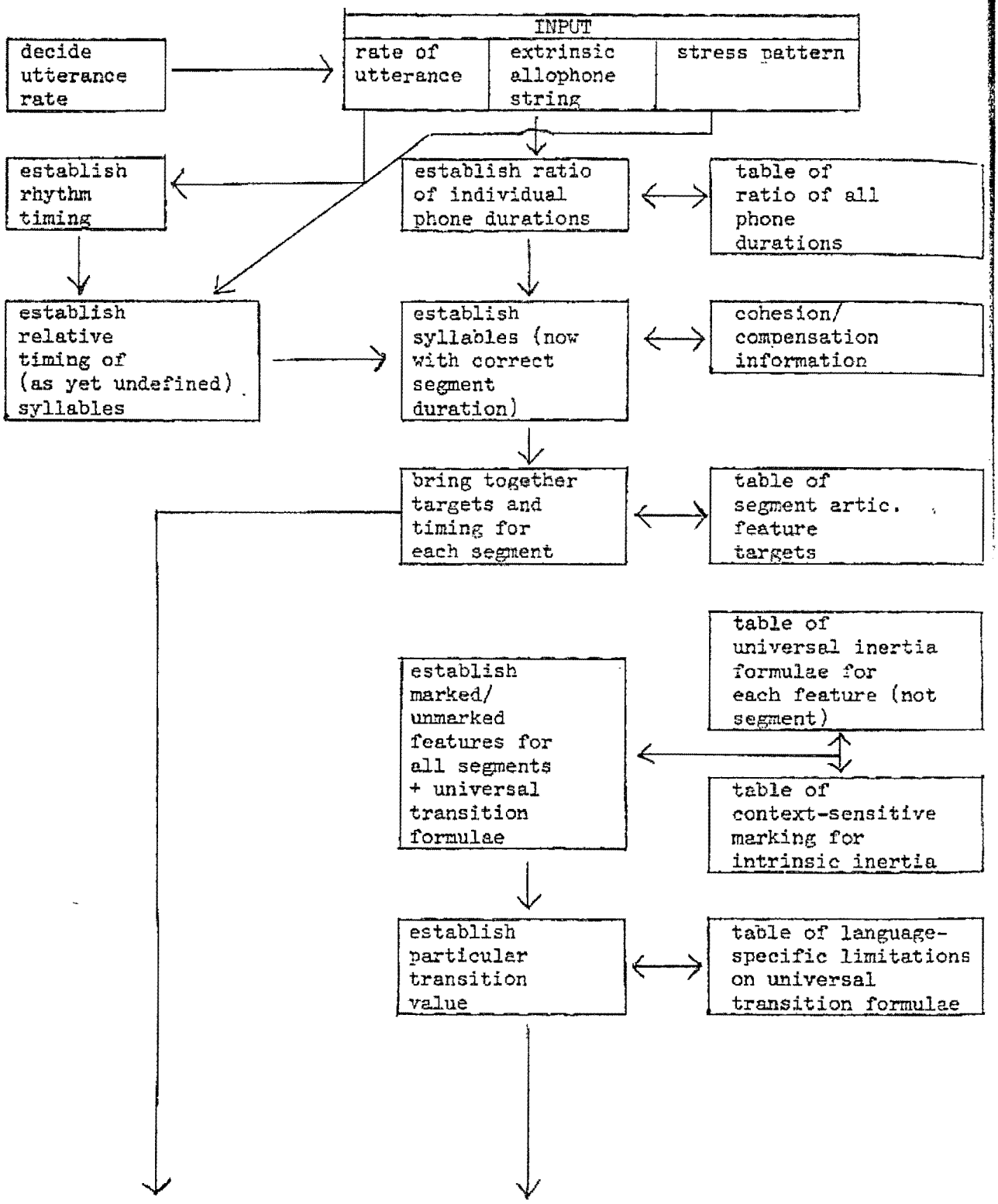
Q Such a postulate of the role of inertia predicts that where constraints were not controllable then no fine differentiation could be required by the linguistic system (a trivial hypothesis, but requiring to be stated).

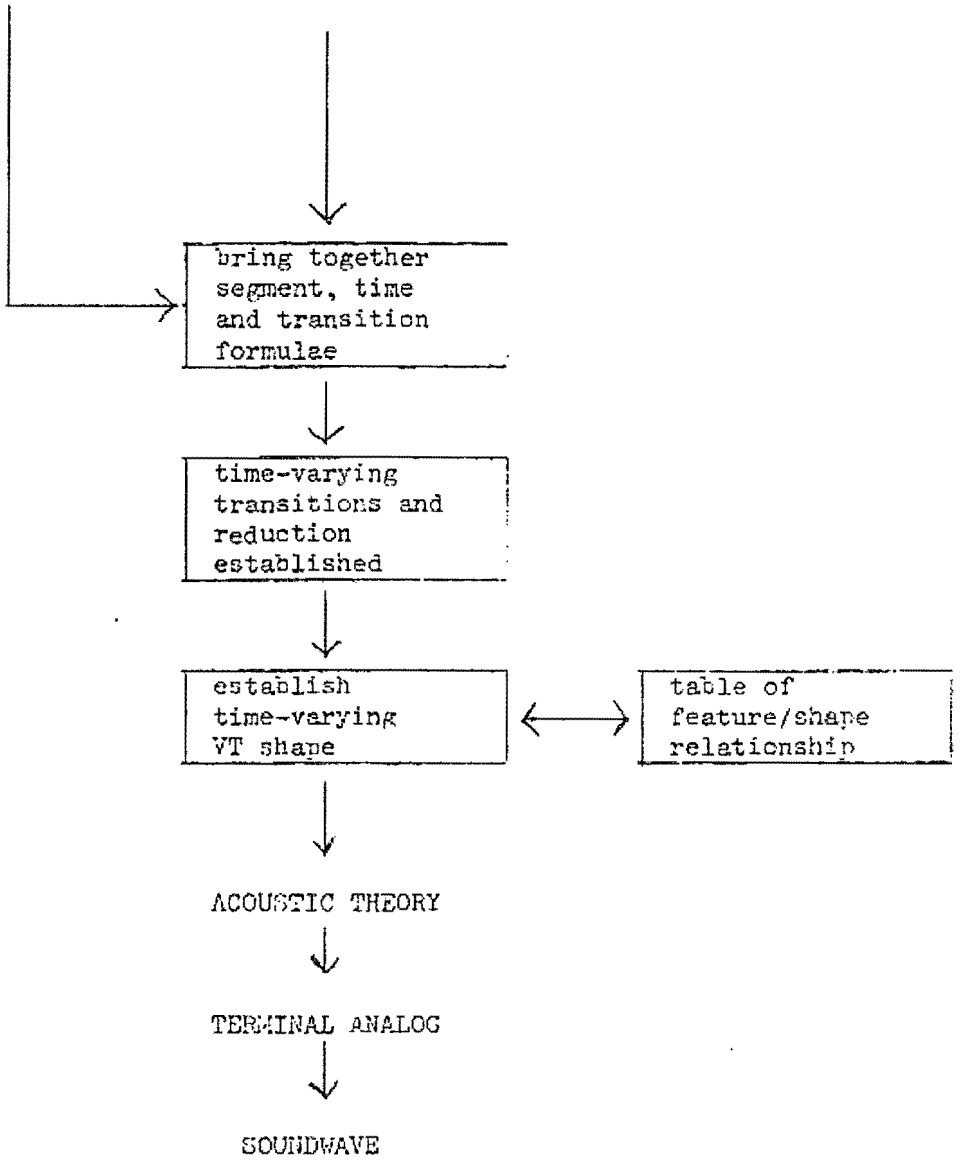
R The model permits variation in successive repetitions of the same utterance in a way that existing speech-synthesis systems do not (we are concerned only with speech synthesis-by-rule). Existing programs (Holmes et al. 1964; Mattingly 1968) store targets and generate allophones in such a way that neither temporal nor target nor transitional output can vary unless the stores are changed.

3. The Synthesis Strategy

Fig. 2 is a very tentative suggestion for implementing the above model. Most of the information required for the lookup tables does not exist.

Fig. 2. DIAGRAM OF SYNTHESIS PROCEDURE




bring together segment, time and transition formulae
time-varying transitions and reduction established
establish time-varying VT shape
table of feature/shape relationship
ACOUSTIC THEORY
TERMINAL ANALOG
SOUNDWAVE

Description of the synthesis system

1. Together with the decision to input a certain sequence of extrinsic allophones an overall rate for the utterance is decided; this could take the form 'fast (+), standard (0), slow (-)'.
2. A rhythm is established which might take the form of an actual time for duration between tonic-stressed elements.
3. Information from the input about the stress pattern establishes the placement of stressed and unstressed elements within the established rhythm.
4. By reference to a table of the relative durations of all phones in the language the relative durations of the actual input allophones in sequence are established.
5. (3) and (4) are combined to establish the actual timing of each element (segment) by including information about cohesion and compensation.
6. Reference is made to a table of target values for each feature for each of the segments in the utterance and this information is brought together with the timing information already established.
7. A table of universal inertia formulae associated with each articulatory feature and a table of marking values for these inertia formulae dependent on segmental context are brought together to establish which features of the utterance segments are marked or unmarked for transition and what the transition formulae are for these features on a universal basis.
8. By lookup table of the language-specific limitations to be applied to these transition formulae, particular formulae are substituted for the output of (7).
9. Feature targets, segment timing and specific transition formulae are brought together.
10. Transitions and reduction, etc., are computed.
11. VT shape is established by means of a lookup table related to the output of 10.
12. Acoustic theory is applied.
13. Conversion of the output of the acoustic theory to TA parameters.
14. Operation of TA to output (5) soundwave.

Notice that prior input information for storage purposes (lookup) is required for:

(a) general ratio of phone durations
(b) cohesion/compensation information
(c) articulatory-feature targets for all segments
(d) universal inertia formulae for each feature
(e) context-sensitive inertia information
(f) language-specific limitations of inertia values
(g) feature/shape relationship.

Hypothesis: (a), (f) and (?b) are language specific; the rest are universal.

Utterance-specific input information required:

(a) string of extrinsic allophone segments
(b) rate of utterance

(c) stress information
(?d) intonation

Some of the deficiencies

A The system is tentative for the moment and much of it could not be implemented except on a very ad hoc basis because values for most of the lookup table information are not available.

B In particular no mention has been made of intonation and how this is derived; no mention has been made either of how amplitude control (e.g, for stressed vowels) is derived.

C Quite clearly not all the boxes in the speech production model have been implemented (particularly at the neuro-muscular level).

D Particularly unsatisfactory is the way segment, timing and transition formulae (9) are brought together in one big obscure computation.

It is hoped that this is the beginning of a speech-synthesis-by-rule system which will render transparent some of the stages in the speech production process.

References:

Boomer, D. S. and J. D. M. Laver (1967) "Slips of the Tongue," Work in Progress No. 1, U. of Edinburgh Linguistics Department.

Broadbent, D. E. and P. Ladefoged (1960) "Voice judgments and adaptation level," Proc. Royal Soc. B. Vol. 151.

Cooper, F. S. (1965) "Research techniques and Instrumentation: EMG," Proc. Conference: Communicative Problems in Cleft Palate: ASHA Report No. 1.

Drachman, G. (1970) "Rules in the Speech Tract," Papers from the Sixth Regional Meeting of the Chicago Linguistic Society, University of Chicago.

Fant, C. G. (1960) Acoustic Theory of Speech Production, Mouton: The Hague.

Fant, C. G. (1965) "Formants and Cavities," Proc. 5th International Congress of Phonetic Sciences, Münster, 1964, Karger: Basel/ New York.

Flanagan, J. L. (1965) Speech Analysis, Synthesis and Perception. Springer Verlag: Berlin.

Holmes, J. N., I. G. Mattingly and J. N. Shearme (1964) "Speech Synthesis by Rule," Language and Speech 7.

Lawrence, W (1953) "The Synthesis of Speech from Signals which Have a Low Information Rate," Communication Theory, ed. by W. Jackson: New York/London.

Lehiste, Ilse (1970) "Temporal Organization of Spoken Language," Working Papers in Linguistics No. 4, Computer and Information Sciences Research Center, Ohio State University.

Lieberman, A. M., P. Delattre and F. S. Cooper (1954) "The Role of Consonant-vowel Transitions in the Perception of the Stop and Nasal Consonants," Psych. Monograph 68.

Matthews, P. B. C. (1964) "Muscle Spindles and Their Motor Control," Physiol. Review 44.

Mattingly, I. G. (1968) "Synthesis by Rule of General American English," Supplement to Status Report on Speech Research: Haskins Labs: New York.

MacNeilage, P. F. and J. L. Declerk (1968), "On the Motor Control of Coarticulation in CVC Monosyllables." Haskins Labs, SR-12: New York.

Öhman, S. E. G. (1964) "Numerical Model for Coarticulation Using a Computer-simulated Vocal Tract." JASA 36.

Öhman, S. E. (1966) "Co-articulation in VCV Utterances: Spectrographic Measurements," JASA 39.

Öhman, S. E. G. (1967) "Peripheral Motor Commands in Labial Articulation," STL-QPSR 4, 1967. RIT: Stockholm.

Tatham, M. A. A. (1969a) "The Control of Muscles in Speech," Occasional Papers No. 3, U. of Essex Language Centre.

Tatham, M. A. A. (1969b) "Classifying Allophones," Occasional Papers No. 3, U. of Essex Language Centre; also to appear in Language and Speech 1970.

Tatham, M. A. A. (1970) "Speech Synthesis: A Critical Review of the State of the Art," Int. Journal Man-Machine Studies Vol. 2.

Tatham, M. A. A. and Katherine Morton (1968) "Further Electromyography Data Towards a Model of Speech Production," Occasional Papers No. 1, Univ. of Essex Language Centre.

Werner, Edwenna and M. Haggard (1969) "Articulatory Synthesis by Rule," Speech Synthesis and Perception Progress Report No. 1. Psychological Laboratory, U. of Cambridge.

Wickelgren, W. A. "Context-sensitive Coding, Associative Memory and Serial Order in (Speech) Behavior." Psychological Review Vol. 79.1.