Articulatory Speech Synthesis by Rule: Implementation of a Theory of Speech Production

Mark Tatham

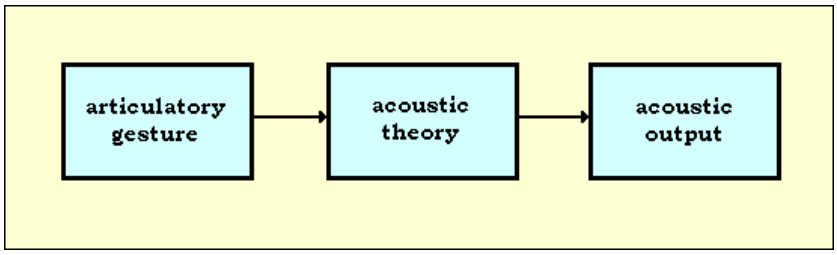
Reproduced from: *Report of the National Science Foundation Grant No. CN-534.1.* Also in *Working Papers,* Computer and Information Science Research Center, The Ohio State University (1970).

INTRODUCTION

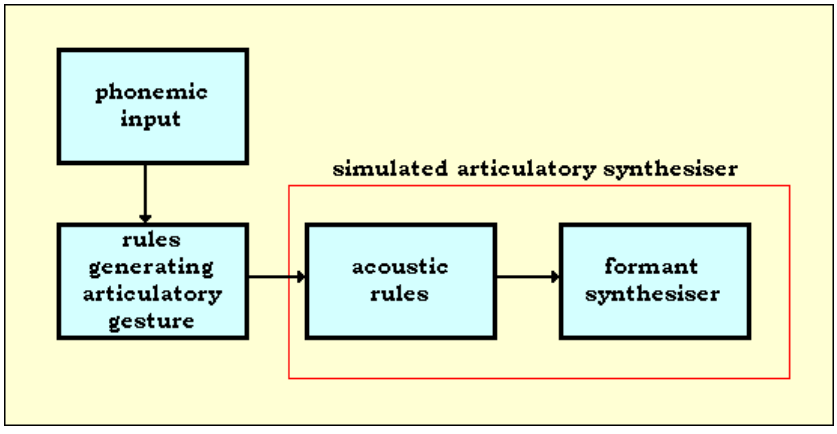
Problems in the design and operation of vocal tract analog synthesisers suggest an interim handling of articulatory synthesis by rule. Relationships have been established between the acoustic output of speech and the articulatory gestures required to cause that output.*

**footnote:* It is true that sometimes an 'identical' acoustic signal can result from two different articulations, but this phenomenon is of no importance here because we are starting from the gesture and ending up with the sound.

These simple relationships are usually diagrammed:



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



In the system described by Werner and Haggard (*Cambridge Report* No. 1, 1969) a 'phonemic' input is turned into a number of time varying articulatory parameters which

correspond in each case to spatial measurements (usually not interrelated) in a stylised vocal tract. The 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 synthesiser.

Several very basic and influential assumptions underlie this approach as handled by the Cambridge Group, 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 synthesiser. Indeed it is noted in Werner and Haggard that the conversion tables are not free from 'tricks' which are designed to make the speech sound more natural. But a more striking criticism can be levelled at the use of terminal analog synthesisers.

Historically the parameters of TA synthesisers have been derived as a result of two criteria: a. visual and b. perceptual.

It is quite clear from the literature since 1950 that the visual inspection of spectrograms has dominated choice of parameters. It was observed that in vowel-like sounds for example there were two or three major formant areas in the frequency spectrum: often a naive (possible correct, butt his 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 synthesised faithfully.

Later, relevance of individual parameters was established using perceptual experiments. There is no doubt that perceptual criteria dominate approaches in TA synthetic speech 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 TA 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 supposed 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 has been unfortunately handled.

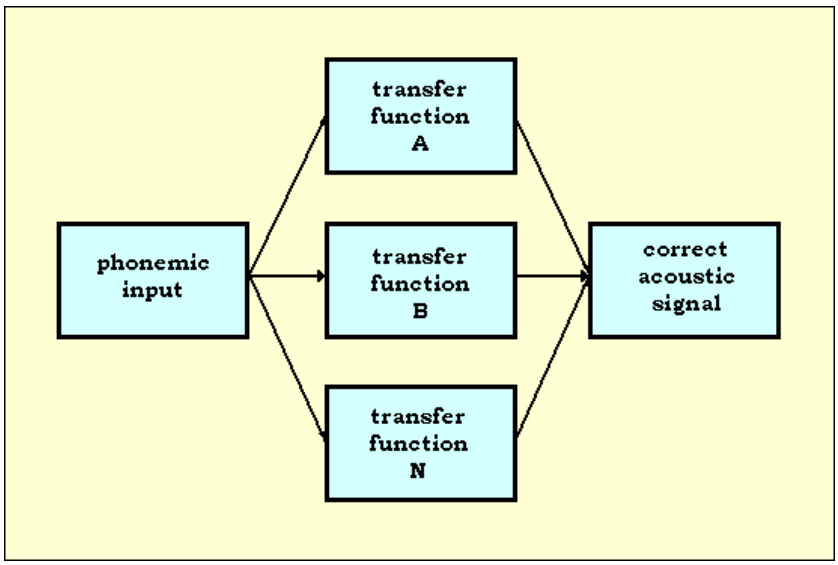
These of course are the reasons for preferring articulatory synthesis. As mentioned earlier VT synthesisers are still not entirely satisfactory — but there seems no reason to wait. The approach offered by the Cambridge Group seems admirable: generate as an interim output articulatory parameters. Under ideal conditions these would be converted to parameter control signals for a VT, but pending this they can be converted to TA 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.

The present proposed system for articulatory speech synthesis strategy is based on a model of speech production which is linguistically dominated. Linguistics has a lot to say about speech production which cannot be inferred from mere reproduction by lookup table of observed articulatory configurations (the Cambridge approach). There is much more implied: it is not enough to generate correct VT 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. Evidence now abounds that this early view is inadequate. The system described here commits itself to one of several

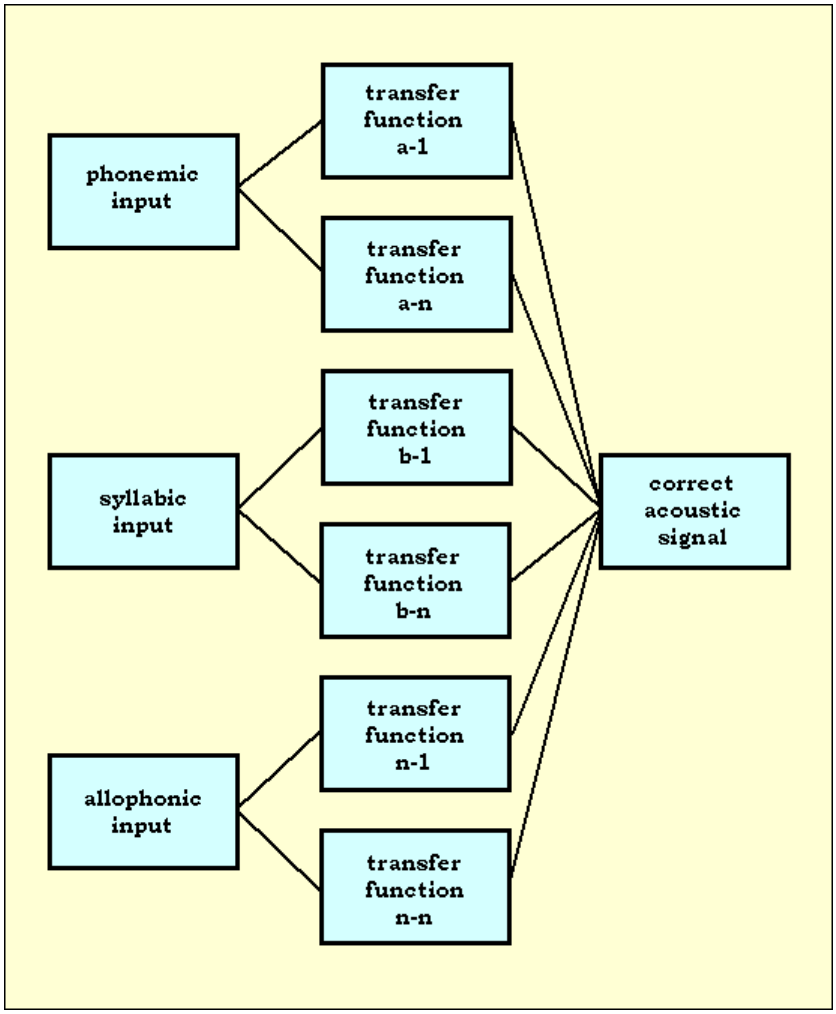
contenting theories of speech production and where possible alternative (but less preferred) strategies will be suggested.

THE THEORY

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, bust as the obtaining of correct perceptual response is to 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 criteria mentioned above will not do as a measure of correctness. Consider the even worse position:



Not only now is there no way of evaluating the various contending transfer functions but there is not internal way of evaluating the contending input types — yet 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 tensing), 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 *not* 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 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.

My 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 argue that the computation cold 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 is 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* 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 who includes the generalised 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 (which may not, certainly will not, be linear).

DETAILS OF THE MODEL

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 identify 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 organised 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. Real time is a misused term among performance model advocates. 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 if not to features as well. Segments and features, though will each need to possess a different kind of reality. Strings will be assumed to be segmented in terms of extrinsic allophones — not in phonemes. This is nearly the case will 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, PhD thesis). 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 coarticulatory (the classic example is: L —» {dark l, clear l} 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 —» {dark l, clear l}, 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 allophonising rules, one function of which (*in the model*) is to switch levels of abstraction.

Thus the input to the model is characterised as a level expressed in terms of extrinsic allophones (Tatham 1969 — Classifying Allophones). 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 coarticulatory effect (McNeilage and Declerk 1967, etc.; Tatham and Morton 1968, etc.). *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 decision need be taken in a working model that complete the system; this is the difference between a working model 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 — Acoustical Society of America paper) 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=

operating as constraints in much the same way as +, #, etc. in the higher linguistic levels. The notation needs further explanation, though, because there will have to be rules deleting boundary symbols: these rule 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:

xC2=C3y	xC5y	where C2 = C3 = C5,
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:

Lookup tables providing (initially) non-temporal (from the segment sequencing viewpoint) information are consulted. These tables can be organised in one of two ways: a. syllables types are listed, b. segment types are listed.

Each possible syllable type (we are concerned here only with the C1V part) is listed as a non-analysable unit exhibiting two temporally spaced GOTO targets. This will not be chosen because i. (a theoretical reason) non-analysability 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 often there is no evidence whether the slip has occurred at the phonological level (supports a.) or at the phonetic level (supports 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 C1V possibilities this simply takes the form of a composite rule indicating in which motor parameters cohesion takes place and to what extent.* [**footnote:* The effect of this high level cohesion on lower level coarticulation, etc. will be discussed later.] This solution satisfies the theoretical criterion of maximum generalisation 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 coarticulation 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 (the *must* exist 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 coarticulation (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, in another language where there are four such palatal consonants

- variation will again take place in exactly similar circumstances, and
- 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 the 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 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. 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 low level inertia such that it becomes language idiosyncratic to establish a relationship between the systems resulting in what has been termed an 'articulatory setting'. 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 velarisation, 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 tot he 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 established *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 directly programmed targets which are the result of a relationship established between mechanical tendencies and linguistic demands further resulting in an agreement for a particular and new target for each allophone.* [**footnote:* EMG records do not show that there is any contextual variation of this kind — but interpretation of such data is as yet only scantily formalised.]

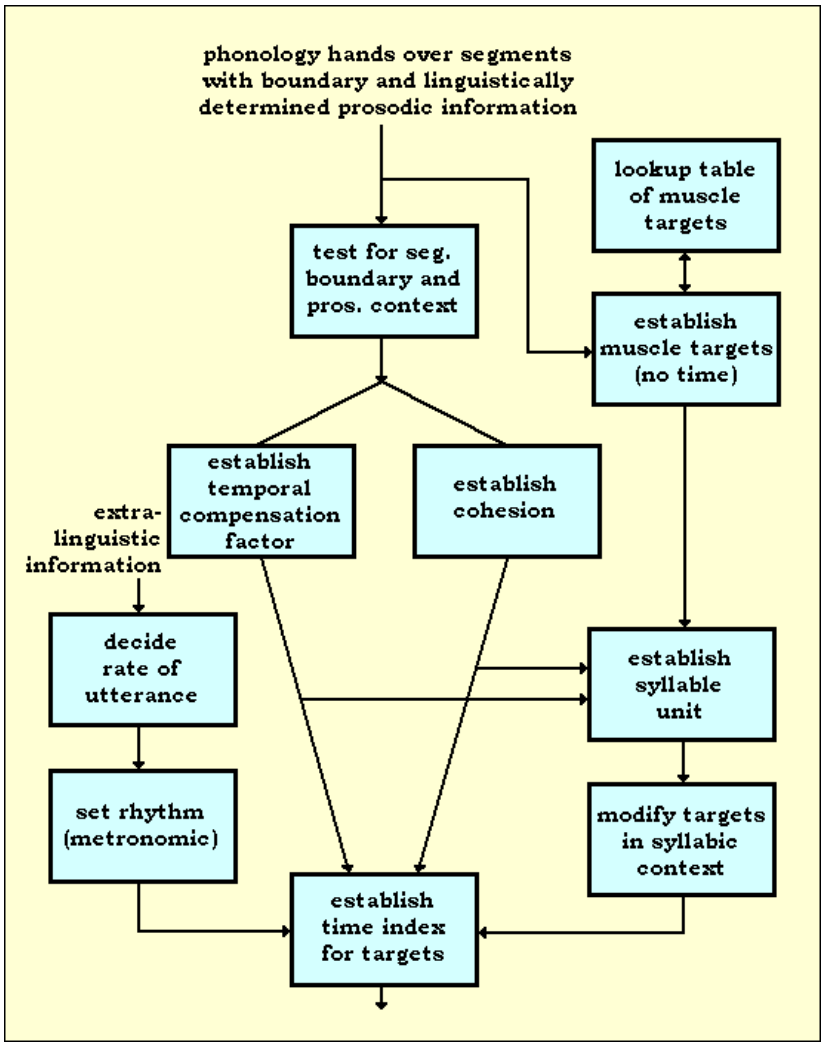
Thus a particular articulatory gesture is the result of a. the linguistically motivated desire to articulate a particular extrinsic allophonic segment, b. the operation of a motor procedural mechanism establishing the cohesion of this segment with syllabic context, c. the insertion of the composite syllabic sized unit (of which this segment now constitutes a part) into the chosen rhythm or rate for this utterance, d., 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 coarticulation limiting factors which limit the freedom of range of articulatory variables but which do not change the established target program.

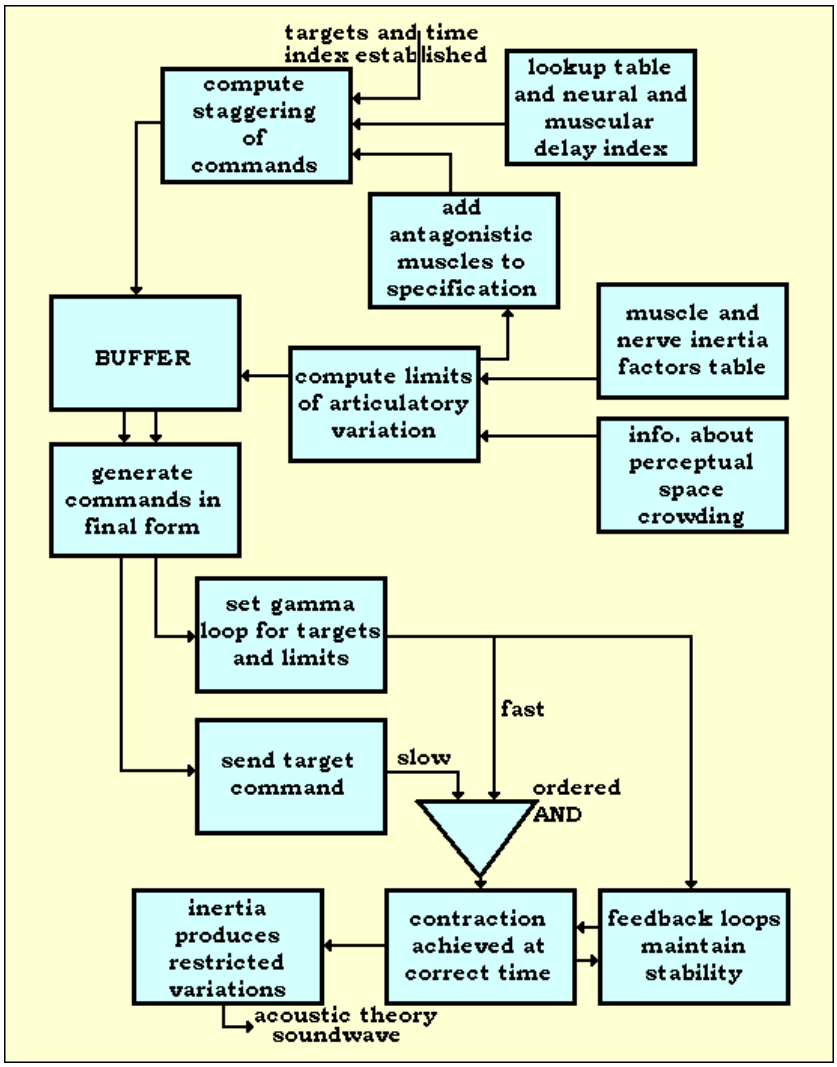
The next diagram represents a simplified version of the present model. Some boxes are tentative (such as the setting of the gamma loop system) but their *function* must occur somewhere to complete the system; they may just be in the wrong place or attributed to the wrong external mechanism — what is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

OPERATIONS INCLUDED IN THE BLOCK DIAGRAM

- Phonology decides which segment (segment = extrinsic allophone in sequential context with morpheme and word boundary symbols, stress pattern, etc.)
- Test for segmental context: this is a hierarchy:
- utterance (or pause group) initial, medial final?
- sub-utterance group: initial, medial, final?
- word initial, medial, final?
- [morpheme initial, medial, final?]
- syllable initial, medial, final?
- Establish whether motor cohesion (lack of temporal compensation) or temporal compensation to operate (depends on hierarchy established under 2.)
- (from 1); look up muscle targets (units not clear): [further unclear: targets for *all* muscles specified? or, marked/unmarked classification — i.e. indication of settings only for relevant muscles for this segment: combination of both?: i.e. segmentally determined feature hierarchy?]
- Establish relationship between targets and cohesion and compensation — i.e. syllable units established
- Decide overall rate of utterance (extra-linguistic?).
- Set rhythm generator according to 6: i.e. metronomic determination.
- Establish how each segment (incorporating 5) will behave temporally in rhythm established by 7.
- Hand over information about segment to motor-command generator (this must be buffered to allow command initiation overlap).
- Establish information about neural line delay.
- Establish muscle response delay.
- Construct muscle command ordering dependent on 10 and 11. (At this point commands for individual muscles are no longer temporally synchronised).
- Consult table about muscle (and other) inertial factors.
- Bring in linguistic information about limits to articulatory variation.
- Recruit addition muscles for limits of articulation maintenance (13, 14, 15 and others may be maturationally wired).
- Set gamma loop for targets and limits.
- Send target command at appropriate time (no low level sequential trigger).
- Muscle contracts within limits, beginning at correct time (within limits). [notice EMG data has temporal limits much finer than amplitude limits]
- Contraction of muscles and articulator movement achieved in correct sequence and at correct time.
- Apply acoustic theory.
- Soundwave output.

TENTATIVE BLOCK DIAGRAM OF THE PROPOSED SPEECH PRODUCTION MODEL





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 (1970), 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 synthesising 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

1. The input shall consist of individual segments which shall be extrinsic allophones (as defined elsewhere) bearing only notional time marking in the form of sequencing.
2. The input shall be indexed with boundary symbols, such as: utterance, group (words, morpheme, syllable): [bracketed classes may not be necessary].
3. The input shall be indexed with certain prosodic features, such as: stress (lexical, group, sentence), intonation (possibly only if marked, but suspect all).

4. Also input will be (extra-linguistic)information derived from decisions about the overall rate of utterance.
5. Speech production is ultimately reducible to articulatory targets (though whether these are stored as representations of shapes, sounds, muscle commands, etc., is not known).
6. These targets are *constant* for the language (irrespective of final output rate, coarticulation, segment position within the syllable, etc.).
7. The hypothesis is temporally adhered to that motor control (not linguistic) dominates the syllabification of segment sequences (— this has not yet been adequately demonstrated).
8. Rate of utterance does not dominate the programming of targets but merely provides a factor which will enhance the effects of (limited, but not directly controlled) system inertia.
9. A function of the motor control system is to stagger (negatively or positively) individual muscle commands to achieve desired articulator 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.
10. 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 coarticulation and variation (over- and under-shoot) which may be permitted.
11. A buffer is required at the point of staggering. The buffer serves two functions: a. permits successive passes of data for rearrangement for staggering, and b. permits 'hold' for output to final peripheral mechanisms of the motor command signals.
12. This model (despite Wickelgren) holds that gamma loop and any similar mechanisms are used for two functions: a. to 'set' limits, and b. to hold them.
13. It further holds that 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.
14. It further holds that fast conducting neurons will permit command signals to arrive at a muscle *during* the previous 'segment', thus generating the right-to-left effects observable in EMG and other data.

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 give period of time);
c. limiting factors to govern succumbing to inertia phenomena.

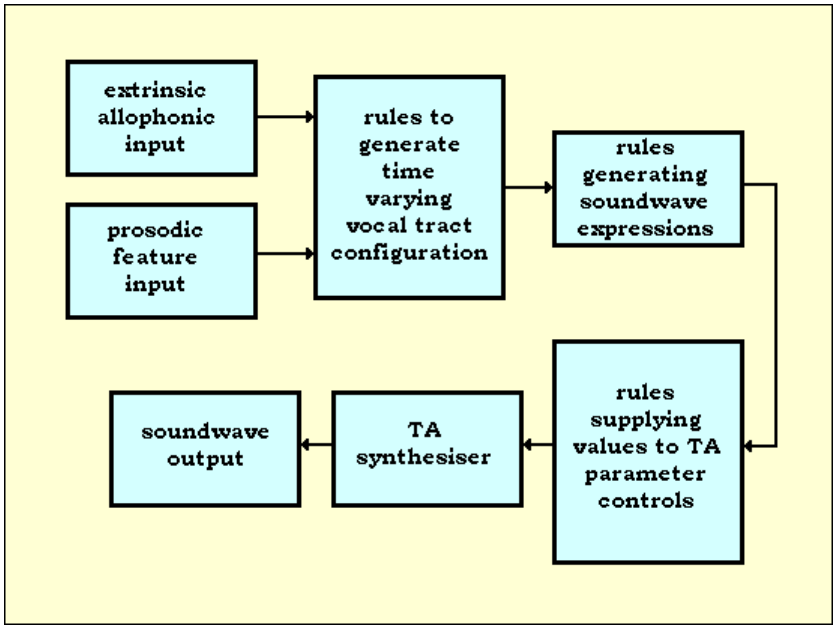
It is therefore held as a theoretical tenet that at the final stages of articulation, *universal* (that is: always operating within this speaker under normal conditions and comparable with similar effects in other speakers) constraints apply to the transformation of the input signal to the muscles (i.e.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

Such a postulate of the role of inertia would predict 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).

The model permits variation in successive repetitions of the same utterance in a way that existing speech synthesis systems do not (we are concerned with only speech synthesis by

rule). Existing programs (Mattingly 1968; etc.) store targets and generate allophones in such a way that neither temporal not target nor transitional output can vary unless the stores are changed.

SYNTHESIS



This paper is not concerned with: The terminal analog synthesiser and its control, for with the tricks which are still needed to obtain a perceptually natural speech output (see Tatham 1970: Int. J. Man-Machine Studies). It is important however to recognise that 'the synthesiser' means the hardware unit *plus* an explicit statement of those tricks (or rather, hardware limitation correction strategies); these two together in principle constitute the idea l synthesiser. It is to this extent that from the point of view of the *theoretical design* of the speech production model simulation, the peripheral output device (the synthesiser) is trivial – just as the printing device of a computer, or the display of TV is trivial. *On no account* may any of the limitations of the synthesiser retract on preceding stages of the simulation unless they are

a. totally explicitly expressed, and
b. utterly removable no matter what the ultimate degrading effect on the output soundwave.

Nor is this paper concerned with the conversion of vocal tract configurations to soundwaves: the conversion formulae are well set out in the literature (see especially Fant).

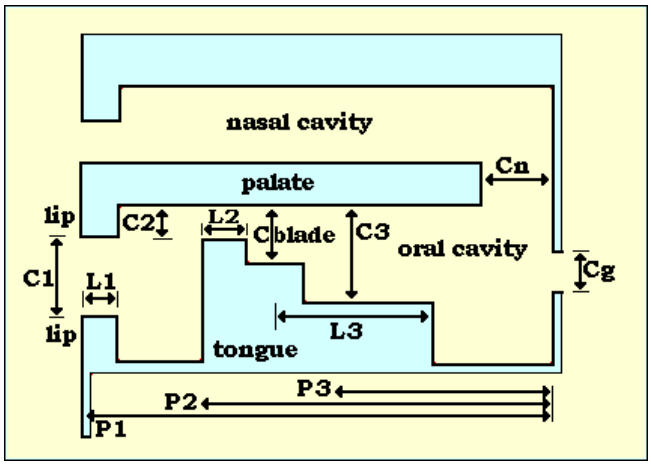
This paper *is* concerned with the explicit implementation of the foregoing model in an attempt to produce some kind of speech output, where 'speech' may simply mean the 'correct' VT configuration at any given time in the particular utterance being generated.

**footnote:* The acknowledged basis for the following description of the proposed system is Werner and Haggard's work and the general principles of synthesis by rule strategy now widely employed and stemming from Kelly and Gerstman (1961), Holmes, Mattingly and Shearme (1964), Mattingly (1968) and several other researchers.

It is proposed to begin with an account and critique of the Cambridge Group's work as published in several papers in their *Speech Synthesis and Perception Report*s No. 1 (July 1969) and No. 2 (February 1970). It is always easy to criticise pioneer work, and my suggestions should not be taken as any failure whatsoever to acknowledge work which will undoubtedly prove a milestone on this seemingly long road to a good and reliable synthesis system. There is no doubt that much of what I have to say will have been noticed and rejected for this or that reason by the Group: this is more likely to demonstrate my inadequacies than theirs.

Firstly, I don't really care too much what the speech output sounds like and adopt the theoretical standpoint that a perceptually unconvincing sound is no reason for tinkering with the model. This is based solely on the fact that the hardware devices are intrinsically poor in quality and I would rather assume perfect hardware. Some researchers have judged that above all the speech must sound good; the Cambridge Group do not wholly fall into this category. Indeed they have remarked (Haggard: 'Devoicing of voiced fricatives', No. 1) that a strategy based on lack of adequate hardware, but designed to obtain good perceptual results (like switching between hiss and buzz at the optimum moment to achieve perceptual voiced fricatives where it was impossible to have both hiss and buzz together) has resulted in the question: why does this apparently wrong procedure yield such good perceptual results? In the voiced fricatives example they discovered that the widely accepted notion that these segments are voiced throughout just is not true in a large number of cases for real speech. It is with this reservation that I deny perceptual evaluation of the synthesis output.

The Cambridge model is based on a stylised VT derived from the idea that we can assume a series of interconnected cylinders as a good approximation (Fant 1960: 'Acoustic Theory …'). Adopting this the Group has determined a series of parameters which can be sued to produce stylised configurations which approximate to human articulations. Thus:



The parameters specified are articulatory and eleven in number: they are subdivided into two types: position and closure parameters, *viz.:*

Position:

P1 – lip position (for lip rounding)

P2 – tongue tip position

P3 – tongue body position

L3 – length of closure of tongue body (L3 is a 'free' parameter – see P. 7 of their text (No. 1)

Closure:

C1 – lip closure

C2 – tongue tip closure

C3 – tongue body closure

Cblade – blade closure

Cj – jaw closure

Cn – velar closure

Cg – glottal closure

P1, P2, P3 specify distances from the glottis. L3 is free, but conforms to the rule that 'both closure and degree of fronting or retraction from natural position decrease the length of the tongue body constriction'.

Rules determine from these position and closure parameters a set of 11 constriction parameters which describe the VT shape. The constriction parameters are:

P1, p2, p3, l1, l2, l3, c1, c2, c3, cn, cg

c2 and c3 are derived from C2 and C3 by rule interaction with Cj [c2 is increased slightly with increase in Cj; c3 is increased slightly by a high value of Cj if p3 is in the front of the oral cavity, but decreased if p3 is in the pharynx].

p1 is adjusted by rule according to the value of C1

l1 varies as P1

l2 is a constant

c2 can be replaced by Cblade for /1/ or /ll/

cblade is computed by adding Cblade to c2.

In other parameters:

P2 = p2

P3 = p3

L3 = l3

C1 = c1

Cn = cn

Cg = cg

There is a lookup table of target values for each articulatory parameter. Target values are marked if the parameter in question is critical for the particular phoneme, unmarked (later giving greater freedom for contextual variation) if not. Marking of phonemes is arranged hierarchically or according to assumptions of dominance: all vowel parameters are marked; for consonants only Cg, and Cj and the position of closure of the primary constriction(s) are marked. For nasals Cv is marked and this parameter is also marked for fricatives (velum always closed).

Thus the Cambridge model starts with a lookup table (addressed from the phonemic input) of target values for each phoneme. Each phoneme is featurally classified where each feature is equivalent to 11 articulatory parameters and each parameter has a value. The parameters are further marked or unmarked dictating ultimately the degree of freedom which that parameter holds for that phoneme in any context – [*notice that the fact of marking is not context dependent but an inherent fact of the phoneme and indeed of the parameter itself]. Transitions are calculated on the articulatory parameters, not on the constriction (shape) parameters.

As summarised on p. 12 (*Report* No. 1) the following information is required for the 10 parameters excluding pitch:

"1. For each phoneme, a target value, marked or unmarked for each of 11 articulatory parameters.

2. For each phoneme, its duration, manner class, and rate modifier.

3. Basic rate of change of each of the 11 articulatory parameters.

4. Vocal tract length and neck length of the speaker."

Now let us take a look at some of the basic assumptions which must implicitly or explicitly underlie the Cambridge strategy of far (we have not yet discussed transition computation but only the individual parameters and the way they are set up in relation to each other).

1. For all intents and purposes the VT can be described in terms of simple cylindrical sections.
2. Articulators (cg, tongue, lips) can be described as simple squared-off interrupters of this basic cylindrical configuration.
3. Movement, where movement occurs, can be interpreted in terms of a simple linear function.

There is probably no quarrel with 1 to 3 since these represent known and discussed stylisations that are approximate enough to not affect the results seriously. Perhaps No. 3 could be improved a little, but the computer programming complications probably outweigh the gain.

There is a stored value for every parameter for every phoneme.

4. Some of these parameters dominate each other intrinsically: thus Cg and Cj are marked for all segments.

Some phonemes dominate others in the number of marked parameters: thus vowels dominate consonants (tentatively: fricatives and nasals dominate stops – because Cn is unmarked for these.

MARKING IN THE LOOKUP TABLE IS USED TO DECIDE REDUCTION AND CO-ARTICULATION.

Now, this model does not take account of muscle contraction itself: it implies that certain groups of muscles always operate to produce desired effects (such as a particular tongue configuration) and that consequently it is the *articulator shape* which is important. I cannot tell from the published papers whether this step has been taken for simplification of the control or whether this is deliberately assumed to be the case. If it has been done for control simplification when the underlying theoretical model runs into trouble in the transition tables – which will need to be largely rewritten for each language and possibly with one language will run into a large and perhaps unnecessary number of *ad hoc* environmental constraints.

The most obvious example lies in the observation mentioned earlier that languages differ in their restriction of coarticulation and reduction effects.

TRANSITION COMPUTATION IN THE CAMBRIDGE SYSTEM

In the Cambridge model transitions are computed for all 11 parameters for each successive pair of phonemes pi and pi+1. The transition values are "determined from mid-point to mid-point – that is, starting with the value attained at the mid-point of pi values and calculated for each time sample until the mid-point of pi+1" whereupon, presumably. Pi+1 now becomes a new pi and the following phoneme becomes a new pi+1. There are a number of ‘transition’ types (8), the choice of which is dependent firstly on a notion of dominance (pi dominates pi+1 or pi+1 dominates pi, or neither), which in turn depends on the marked or unmarked status of a particular set of parameter values for the two phonemes; and dependent secondly a choice of rate of transition which is also determined from the marked/unmarked status of the parameters.

In the pair of phonemes one may dominate the other, or *vice versa,* or both may possess equal dominance. A particular transition type is selected according to which of these three possibilities is the case. "There are several reasons why one phone may dominate another; the choice of a transition type does not depend on why a phone is dominant."

Four classes of phoneme are established: pause (PZ), vowel (V), marked consonant: (C), and unmarked consonant (C). [Where "a consonant is 'unmarked' for a certain parameter if the value of that parameter is not critical for the specification of the phoneme. This is signalled to the program by making its target value in the phoneme tables negative".] Obviously, then, consonants will be marked on some parameters and not others and immediately the way is established to handling reduction and coarticulation which occur only in some parameters but not all of particular consonants.

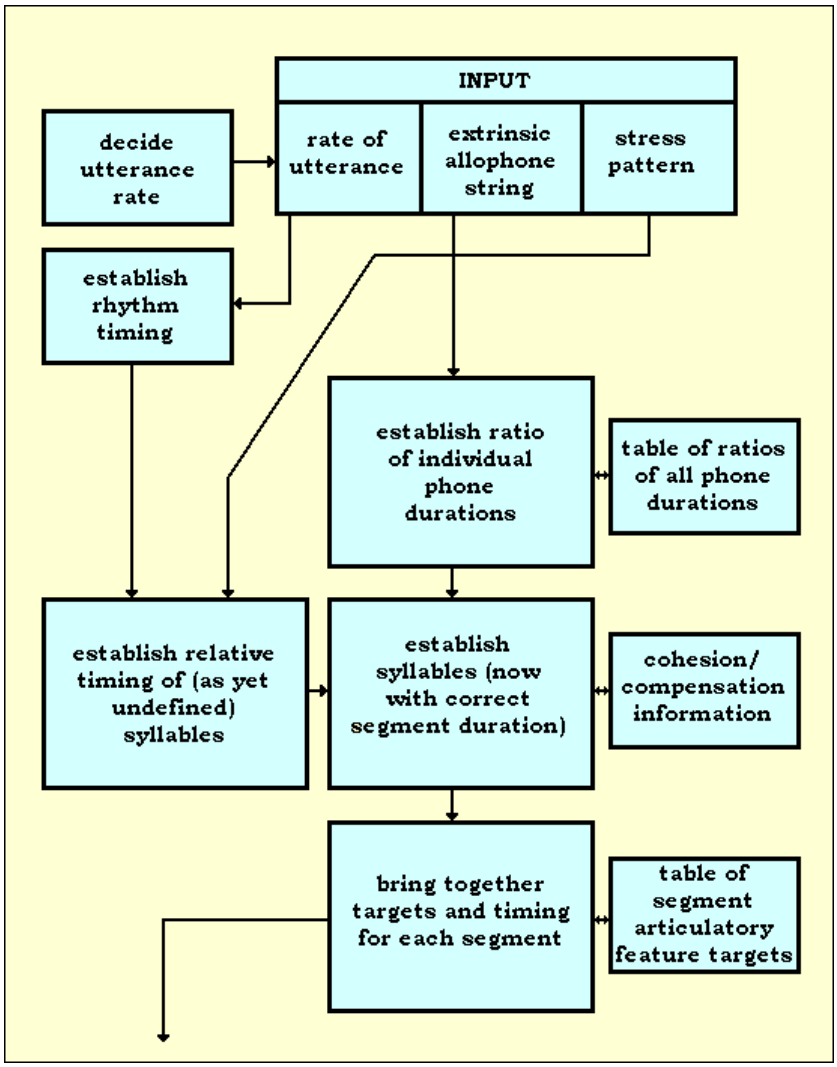
- Consonants are marked on some parameters.
- Vowels are marked on all parameters (except tongue blade and velar closure).
- A semi-vowel is a vowel in C-C context with respect to its closure parameters.
- A semi-vowel is a consonant in contexts not included under 3.
- everything dominates PZ;
- everything but PZ dominates C;
- within closure parameters only (C1, C2, C3, Cj and Cblade), C dominates V."

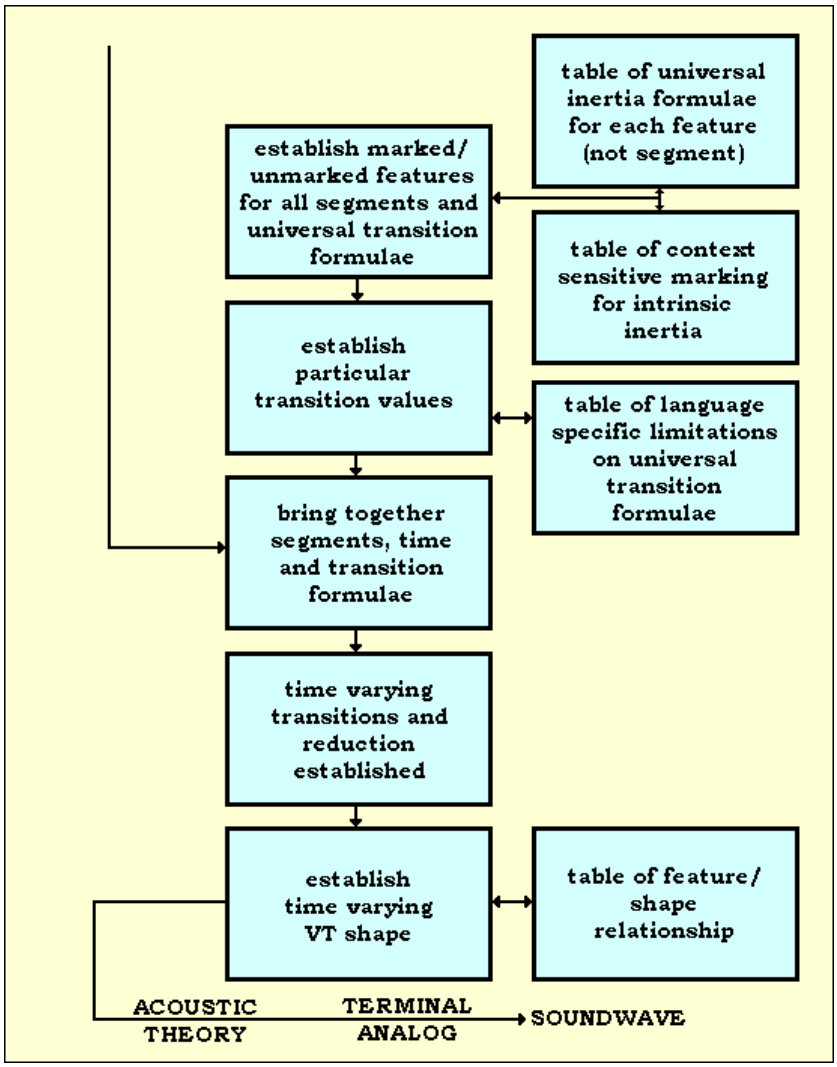
"In a transition between a dominating and a dominated phone (e.g. C-PZ, C-Y on a closure parameter, etc.), whether the dominating phone is first or second, the target values of the dominating phone have more influence on the transition values than the target values of the dominated phoneme. This is because the target value of the dominating phone, if attained (as it usually is) is carried throughout the duration of the dominating phone to its boundary with the dominated phone. Only within the dominated phone do values start to head toward the target value of the dominated phone. Since there is less time to approach this target it is less likely to be reached. Furthermore, the rate at which the values approach the target of the dominated phone within the dominated phone is the rate of change as modified for the dominating phone, where applicable, unless the dominated phone is C; in this case, half the usual rate of this parameter is used. This assures that very slow progress will be made toward the unmarked targets of consonants, and that they will often not be attained. Expressed in another way, the full force of change in an articulatory position is only applied when marked for a phoneme, and this change occurs largely within neighbouring phones, as observed in EMG studies of speech."

For further details of this notion dominance and the types of transition see pp. 1032 of *Report* No. 1. There are basically only two types of transition:

- Type 1, for marked closure parameters of consonants where a vowel dominates a marked consonant or a consonant dominates a vowel. The transition is external to the phoneme and can assume one of three rules:
    i. rapid (2 marked consonants),
    ii. moderate (marked consonant and vowel),
    iii. slow (anything with adjacent unmarked consonant).

  A marked target is not undershot.
- Type 2, for vowel parameters and consonant position parameters. "It is asymmetrical and, depending upon rate, duration and distance, it may undershoot; when it does not undershoot, but reaches target before the midpoint of a phone, some asymmetry may result, the period of target being in general early in the phone. (NB if the following phone is unmarked the reverse may be true."


INPUT
rate of utterance
extrinsic allophone string
stress pattern
decide utterance rate
establish rhythm timing
establish ratio of individual phone durations
table of ratios of all phone durations
establish relative timing of (as yet undefined) syllables
establish syllables (now with correct segment duration)
cohesion/ compensation information
bring together targets and timing for each segment
table of segment articulatory feature targets



DESCRIPTION OF THE 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 (for English) 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 is 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
15. Soundwave.

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

- general ratio of phone durations
- cohesion/compensation information
- articulatory feature targets for all segments
- universal inertia formulae for each feature
- context sensitive inertia information
- language specific limitations of inertia values
- feature/shape relationship.
- *Hypothesis:* a. f. and b (?) are language specific; the rest are universal.
- Utterance specific input information required:
- string of extrinsic allophone segments
- rate of utterance
- stress information
- intonation (?).

SOME OF THE DEFICIENCIES

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.

- 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.
- Quite clearly not all the boxes in the speech production model have been implemented (particularly at the neuro-muscular level).
- 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 for the first time render transparent some of the stages in the speech production process.