A Linguistically Oriented Approach to Speech Synthesis by Rule

Mark Tatham

Paper read to the Annual Meeting of the Linguistic Society of America, Washington D.C., December 1970. Reproduced from *Occasional Papers* No. 12, Language Centre, University of Essex (1971, 1-9).

Speech synthesis by rule has been with us now for a decade, and significant developments have been made not only in he quality or naturalness of speech produced, but also and more importantly in the theory behind the formulation of the rules themselves. The quality of the speech is interesting only insofar as it reflects an improvement in control strategy — it is not interesting, from the point of view of linguistics, if it reflects an improvement in synthesiser design whether this be a hardware or computer simulated device). It has now been made clear that present synthesiser approaches with the usual terminal analog or the less usual vocal tract analog are capable of producing an output convincing enough in its quality as far as the casual listener is concerned. However, it should be noted that the output, particularly from a terminal analog synthesiser, cannot, by the very nature of its minimal parametric structure, simulate the complete waveform of human speech. This fact is quite unimportant when synthesis is being used for normal man/machine interface applications such as communication between computer and operator, but becomes much more important when the synthesised speech is being used as a stimulus in perceptual experiments, and becomes crucial if synthetic speech is being used as a test of a speech production model.

Quite obviously, any model of speech production, if it is at all complete, will make reference to an input — somehow derived from a phonology — and attempt to explain transformation of the input into the neural control of the muscles associated with speech; it will explain how particular articulations result from particular muscle contractions and how particular sound waveforms are derived from these articulations. In doing so, the model is constrained to generalise as much as possible concerning those properties of speech production which are universal — often ‘innate’ in some sense — and those which are language, dialect or idiolect specific. The model will be partly mentalistic, partly neurological, partly physiological, partly mechanical, partly acoustic, and so on. More importantly, partly linguistic and partly not.

As a piece of hardware, then, the synthesiser itself can be seen to be an implementation of a comparatively small part of any speech production model — that part connected, as far as a terminal analog is concerned, with the soundwave and its typical components, and, as far as a vocal tract analog is concerned, that part connected with the varying acoustic transfer function of the mobile vocal tract. There has been little attempt to produce a hardware device simulating muscle contraction.

Clearly, then, it is the control of a synthesiser that is of most interest to us as linguists — not the synthesiser itself. Indeed, I wish to assert that the synthesiser constitutes an ‘extra’: the implementation of a speech production model might well stop short of actually attempting to drive a synthesiser. Why, then, are we concerned with speech synthesis at all The answer is simple: speech synthesis is not making a synthesiser work, it is making a device capable of modelling the speech process in a productive way — synthesising speech production, not synthesising the result of speech production.

As such then, speech synthesis is concerned mainly with modelling the control of the human vocal tract: interest lies in explaining just how and why particular muscles contract to produce particular articulations at particular times.

It could be argued that existing strategies in synthesis by rule constitute motels of speech production, Coker at Bell Telephone Labs and Haggard at Cambridge University have

synthesis by rule systems which proceed in terms of articulation rather than in terms of the acoustic properties of speech — a dormant approach until comparatively recently. I want to examine how such systems could be construed as models of speech production — though I must add that neither of these researchers has claimed this, and indeed Haggard (1970) has denied that he has a model of the production process.

Typically an articulatory approach proceeds as follows. Initially, tables of values are stored in the control computer which consist of feature matrices. These features relate to certain articulatory parameters — such as tongue height, lip opening, state of glottis, and so on — and are arranged so as to combine to give the appropriate configuration for a particular phone entry in the table. The feature representation is hierarchical: certain features dominate certain others with respect to necessity of their individual realisation in the synthesiser. Thus, for example, lip rounding during, say, [y] in French is crucial to the articulatory realisation of that segment, but the same feature is not at all crucial (and in fact may be realised as + or -) for the English consonantal segment [k]. The feature specification table is almost invariably context free — that is, segments are few in number, there is no separate representation for that particular [k] which is realised preceding [u] distinct from that particular [k] realised preceding [i] (where the former exhibits lip rounding ant the latter lip spreading).

This is one reason for the hierarchical structure of the feature representation, of course. Certain features may be realised differently depending on context and these features are changed according to context sensitive rules supplied in another table. Rules in this second table take the form of stating that certain features fount in the feature representation of a particular segment may have their values changed according to the segmental environment. Typically features are seen to be realised in a running matrix as a rule governed interaction between context ant intra-segmental necessity.

It is interesting, though, that the segmental representation is only context free with respect to phonetic context: contextual phonological variants are represented normally as separate entries — palatal and velar [1], for example, are usually entered separately since they are not phonetic, but phonological variants (at least in English) of a single phoneme.

The rules of the second table, whose function is to modify the original featural specification of a segment according to its phonetic context, generally reflect mechanical properties of the vocal tract itself and the way in which these are time governed. However, it is unfortunately often the case also, that, perhaps for the sake of economy, individual rules are a collapse of a larger number of rules which might sometimes conceal different levels in a speech production model.

Let me give a simple example considering the lip rounding/spreading feature only for the moment — and assuming only three values for this feature: round, neutral, spread. This feature could be marked ‘round’ for [u], ‘neutral’ for [k] and ‘spread’ for [i] in the featural representation table. Rules in the segmental context table will give us no change in the value of this feature for [u] or [i] — the feature is high in the hierarchy for these segments; but they will give us a change for [k]. [k] before or after [u] has lip rounding and before or after [i] has lip spreading at normal utterance rates. Thus [uk] with lip rounding for the [k] and [ki] with lip spreading for the [k].

But what happens when I put these together ant say [ukí] or [úki]. Simplifying a little, we can say that during the [k] there is a rapid shift from lip rounding to lip spreading: according to our hierarchy, lip rounding must be there for [u] and spreading for [i] — the change from one to the other occurs during the [k] which has this feature low valued in the hierarchy. The rules are arranged to ‘track’ the change from round to spread during the intervening neutral consonantal segment.

What is important, though, and this is the point I am making, is that no speech production model which attempted any completeness, would explain the change in this simple way. True, the articulation, or properly the spatial configuration of the lips has changed in a given time from what may be regarded as ono extreme to the other, ant indeed this may have been some

high- level mental ‘intention’ in some sense — but we are additionally constrained to explain how this intention is accomplished.

The configuration of the lips is not a single parameter in terms of muscular control: a number of muscles sometimes contracting simultaneously combine to give us the resultant configuration. Thus, if we divide these grossly into two groups — lip spreading and lip rounding muscles — we can see that investigation is required to see just what is going on when [uk] is followed by [i] or [u] is followed by [ki].

I recently conducted a very simple experiment along these lines, because I hypothesised that [uk] has lip rounding muscle contraction, that [ki] has lip spreading muscle contraction, and that [uki] has both lip spreading *and* lip rounding muscle contraction during the [k]. The hypothesis was confirmed using elementary electromyography techniques. In [uki] a muscle (*orbicularis oris*) associated with lip rounding contracted for as much and for as long as in [uk] alone, and a muscle (the *zygomaticus major* — among others) associated with lip spreading contracted as much and for as long as in [ki] alone: indeed what was happening was that rounding contraction and spreading contraction co-occurred for a period, the one decreasing and the other increasing, giving rise to a change in the spatial configuration of the lips from rounded to spread. Just what might have been expected.

Thus in this case a feature table in the synthesis control strategy should indeed represent this particular feature, but there should be an intervening level of rules indicating just how the change from rounding to spreading is accomplished in terms of controlling individual muscles or muscle-groups to perform the operation.

The experiment and the inference might be considered rather trivial were it not for another more major contribution. I mentioned earlier that commonly synthesis strategies employ a lookup table which lists segments out of context and that contextual variations are supplied later. Incorporating this notion into a speech production model has important theoretical consequences which form the basis of current discussion in the literature concerning whether or not these segments should be regarded as completely devoid of context marking or not (Wickelgren 1969, MacNeilage 1970, MacKay 1970, Whitaker 1971). Wickelgren suggests context marking and the others oppose this view.

It is easy to misinterpret Wickelgren’s point of view and indeed his opponents have exaggerated his claim, but as I see it he would suggest that in the sequence [uki] the [k] is not a [k] which is simply undergoing mechanically constrained featural change by rule, but is in fact a particular [k] associated with the [u- ... -i] context to start with — in other words, that there are as many [k]s as there are possible contexts for [k] to be specified initially in the table. The experiment sought to show that this is not the case.

If the contextualist position is correct then the electromyographic activity associated with lip rounding through the [k] in [uk], particularly with respect to its duration and off-set rate would be different from the lip rounding contraction associated with [k] in [uki]. Experimentally in fact I compared [uki] with [uka], since [a] does not demand a change in the lip feature specification of [k] — certainly not to spreading. In all possible combinations of [u], [i] and [a] with [k] intervocalicly and varying stress, there was seen to be no context sensitive change in the lip rounding or lip spreading contractions: there were obvious spatial configuration changes, however, as discussed earlier — explained simply by mechanical rules. The experiment is not at all conclusive for a number of reasons, but points towards the solution already generally adopted: phonetically context free specification for individual segments.

In all probability we are wrong to treat segments in isolation and this is what Wickelgren is primarily pointing out: but his solution is no real improvement. It seems to me that it is wrong to way there is lip rounding or lip spreading for the [k] — rather there is lip rounding and lip spreading associated primarily with the adjacent segments and quite simply this parameter is irrelevant for [k] — hence the hierarchical marking of features.

It can be argued that properly set up, therefore, the rules for synthesis can reflect a genuine speech production model. But there is yet one more important point to be mentioned.

The input to a synthesis by rule program generally consists of a string of what have been called extrinsic allophones — that is, those variants at the output of the phonology — phonological contextual variants rather than phonetic contextual variants. The rules of synthesis — like the rules of the phonetic component or the speech production model — supply though phonetic variants and, as we have seen, provide for the spill of particular features of one segment into adjacent segments. Notice, however, that in the case of all synthesis programs to date, whether they approximate to a true model or not, with a given set of rules the output from the device will always be the same for the same input — that is, if we input a series of segments, say A,B,C, we always got out X,Y,Z, because the same static tables have been used to discover values for the appropriate features and to supply the phonetic contextual variations. But this is not the case with real speech. If we decide to utter a particular word or group of words, then when we do this a second time the phonological output will be the same — but our speech output will be measurably different: it is well known that two phonologically same utterances from even the same speaker are not identical in waveform. But from a speech synthesiser they are.

We could say that our tables of features and rules are equivalent to the rules of syntax and that of course application of the same rules produces identical results if we make a pass through them — that is, the *idealised* utterance is produced. There is something wrong with this. Any speech production model has failed if it does not account for the fact that there is a good deal of variation in successive 'same' utterances. I don't know how to handle this: but it is worth underlining that a true speech production model cannot be similar in concept to the syntactic and phonological components of a competence grammar — we do not want to say the same kinds of things about speech. Somehow or other the output from a synthesiser should vary just as the output from a human being does.

As a start to handling this variation I want to take a quick look at one of the biggest variables — time. No two phonologically identical utterances from the same speaker have identical timing under normal conditions. Yet in synthesis programs as we have them today, time is just another of the features in the feature table, also subject to rules which invariably produce the same output. Some study has recently been made of timing of segments in repeated phonologically same utterances (Lehiste 1970) and it seems that variation in some aspects of timing even when overall utterance rates are kept as near constant as possible is bound up with phonetic syllable structure. Specifically in, say, a simple C(onsonant)V(owel)-C(onsonant) utterance variations in the initial C and the V are less than between the V and the final C where in some sense they are compensatory. It would be quite wrong, therefore, to simply associate a particular segment with a typical duration derived from a lookup table: such timings can only be derived from an interaction between an abstract or notional duration associated with a segment, an input channel specifying overall utterance rate or change of rate, and further information derived *from within the model* (since it is a property of speech production and not phonology) about phonetic syllable structure and how it relates to any one particular segment in its particular context.

It is partly for this reason that we must deny the single channel input which has been a property of all synthesis by rule programs I am aware of. Synthetic speech which is to be more natural in as much as it genuinely reflects one of the most noticeable surface phenomena of real speech — that of phonetic output variation of repeated phonologically same utterances — will have to abandon the all too attractive acceptance of an input which looks like the output of a standard phonology — however useful this may be to engineers concerned with reading machines or similar man/machine communication devices.

I have presented a brief discussion of why I think speech synthesis forms a suitable testing round for a speech production model and tried to point out some of the reasons why present approaches are inadequate. Most importantly linguists engaged on the development of such a model should be constrained neither by the technology of the device itself, nor by near sighted desires to produce acceptable outputs (where acceptable does not mean simply 'convincing') from the simplest possible input. For a genuine working model much more

thought must be given in synthesis programs to distinction between levels and to multi-channel inputs.

References

Haggard, M. (with K. Haycock) (1970) 'Articulatory synthesis by rule', *Speech Synthesis and Perception* 3, The Psychological Laboratory, University of Cambridge

Lehiste, Ilse (1970) 'Temporal organisation of spoken language', *Working Papers in Linguistics* 4, Computer and Information Sciences Research Center, The Ohio Statue University

MacKay, D. G. (1970) 'Spoonerisms: the structure of errors in the serial order of speech', *Neuropsychologia* Vol. 8

MacNeilage, P. F. (1970) 'Motor control of serial ordering of speech', *Psychological Review* Vol. 77:3

Whitaker, H. A. (1971) 'Some constraints on speech production models', Essex Symposium on Speech Production Models, in *Occasional Papers* 9, University of Essex Language Centre

Wickelgren, W. A. (1969) 'Context-sensitive coding, associative memory and serial order in (speech) behavior', *Psychological Review* Vol. 76:1