SPEECH SYNTHESIS BY RULE ; WHY, WHAT AND HOW?

S.G. Nooteboom, I.H. Slis and L.F. Willems

There are more than one reasons why speech researchers may be interested in systems o speech synthesis by rule. Such systems eventually achieve practical importance to voice response units in man-computer communication, as part of a reading machine for the blind, a spoken encyclopaedia or to the automatic generation of taped sets of spoken instructions for certain tasks such as wiring telephone exchange units (e.g. Flanagan, Coker, Rabiner, Schaefer, Umeda, 1970).

One may also be interested in speech synthesis by rule as a research tool in the domain of phonology and speech perception. The importance of synthesis by rule in this respect has been stressed a number of times (by, among others, Liberman, Ingemann, Lisker, Delattre and Cooper, 1959, Lisker, Cooper and Liberman, 1962, Mattingly, 1971, Klatt, 1971, Holmes, 1972). It is this interest which constitutes the main motivation for the work in our Institute on a system of speech synthesis by rule. As this work involves considerable effort it seems worth while to give some thought to such questions as: Why is synthesis by rule a desirable research tool? What properties should the system have? How can these be achieved? In an attempt to answer such questions let us restrict our considerations to terminal analog synthesis. Operationally, we define a system of speech synthesis by rule as a system which accepts a discrete, typed input in terms of phoneme-like symbols and some additional symbols (e.g. stress marks, word, morpheme and phrase boundaries) and automatically converts this input into intelligible speech.

why?

We will put forward three condiserations which, in our view, make it desirable to work on a system of speech synthesis by rule.

a. *Generation of stimuli*

Such a system may provide the means of rapid and easy generation of large sets of stimuli for perceptual experiments. In principle, all stimuli which can be made by a rule system, can also be made by ad hoc specification of the parameters. It is obvious, however, that, if we wish to generate large sets of stimuli, especially if the stimuli consist of whole words, word groups or sentences, this soon becomes very laborious. Once a suitable rule system, or even part of it, has been developed, many experiments can be carried out in much less time than before.

b. *Heuristic strategy*

The second consideration in favour of speech synthesis by rule as a research tool might well be the most important. Working on synthesis by rule is a powerful and inspiring heuristic strategy. In attempting to synthesise intelligible and reasonable sounding speech we are forced to make explicit hypotheses concerning perceptually relevant acoustic properties of speech. In the frequent failures of our attempts we are immediately confronted with many things we do not know about speech. In this way we readily run into research problems we would not have thought of otherwise, or would not have considered seriously, but which, nevertheless, may be of fundamental importance to understanding the processes of speech production and perception.

An example is provided by the growing amount of experimental work on intonation and temporal organisation of speech in a number of speech laboratories. The renewed interest in this domain seems at least partly inspired by the failure of existing synthesis by rule systems to shape the pitch contours and temporal structures of speech in a perceptually satisfactory way. This leads not only to a search for useful prosodic rules but also to the more important question of the perceptual part played by prosodic patterns in speech.

Another, possibly related, research problem which comes into focus in work on speech synthesis by rule is the problem of what we would like to call "perceptual integration". All synthesis systems known to us exhibit in their output signals instances of sound segments which perceptually do not seem to be part of the perceived speech patterns, and which are difficult to locate in the utterance. We may say that they cannot be "perceptually integrated" with the rest of the speech into recognisable patterns. This has a disturbing effect on speech intelligibility. It also immediately raises the question: What perceptual mechanisms are responsible for our ability to hear normal speech as perceptually integrated patterns, and under what conditions do these mechanisms fail?

A third and final example is of a more linguistic nature and concerns the input of the rule system. An unstructured string of input symbols corresponding to phonemes or speech sounds is not suitable for being converted into intelligible and acceptable connected speech. Relevant information concerning morpheme, word and phrase boundaries, lexical stresses and pitch accents would be lacking. This information is needed for the shaping of perceptually satisfactory pitch contours, temporal structures, and for inserting speech pauses in the correct places. The question we are confronted with here is what information should be provided at the input, and what information can be found by automatic analysis of the input string. If the system is supplied with word boundaries, in many cases lexical stresses, morpheme boundaries and phrase boundaries may be found by automatic analysis, but as yet no way has been found to predict the correct positions of pitch accents. It seems that information on both the syntactic and the semantic structure of the phrase or sentence is needed in order to do so. The basic rules involved still escape explicit formulation, however.

Making and testing of complex models

The relation between a linguistic, discrete representation of the sound level of language, in terms of phonemes or speech sounds, and an acoustic specification of speech is a complex one. Speech synthesis by rule provides an excellent means of making explicit models of this relation, and of keeping the models testable. The acoustic regularities of speech are the result of interactions of a great many factors. In accounting for these interactions we need intricate rule systems that can rapidly be revised and tested both quantitatively and perceptually. One of many possible examples is provided by the temporal organisation of speech. Actual segment durations in speech result from interactions between the feature composition of the segments, the internal structure of the syllable (for consonants especially the structure of the consonant cluster), degree of stress, position in word and phrase and other factors. Such interactions can be modelled in a set of quantitative rules, but without implementing these rules in a synthesis system it would be difficult to test them as to their acoustic and perceptual effects.

Summarising, we may state that it is desirable to work on speech synthesis by rule (a) in order to obtain a rapid and easy way of generating acoustic stimuli for experiments on speech perception, (b) as a heuristic strategy in speech research, (c) for modelling intricate interactions underlying acoustic regularities of speech, and testing these models both acoustically and perceptually.

what?

Below we formulate some requirements and properties desirable for a system set up for the purposes outlined above.

a. Ease of operation

A synthesis system as we conceive it is used in the laboratory by more people than only those actually involved in its development. All researchers of the laboratory staff, guest workers and students who wish to use the system must be in a position to do so. This implies that operation should be easy and easily acquired.

b. Visual, acoustic and numerical feedback

For all purposes mentioned it is important not only to have the facility to listen to the output and tape it, but also to have an immediate check on the parameter values generated, both visually on an oscilloscope, and numerically in a printed output. One should be able to select for feedback the parameter values one is interested in.

c. Special input for parameter values

It is desirable to have facilities for controlling the value of one or more parameters on the spot without changing the rule system. This may concern an acoustic parameter for one particular segment in the speech chain, the rest being controlled by the rule system, or a rule coefficient of the system, affecting all segments to which the rule applies. The desired parameter values could be supplied to the system either by a typewritten input or by some other manual control.

d. Exchangeability of rules and subsets of rules

It is desirable to be able to suppress each rule or subset of rules temporarily and replace it by a new one. This holds for all levels of organisation, which implies that the system should have a clear organisation in smaller and larger subsets of rules, all with their own names, and as independent of the other subsets of rules as possible. This would enable us to compare and test alternative rules or rule sets.
We may wish, for example, to replace the rules for vowel durations in stressed syllables, or the whole set of rules governing segment durations, or all prosodic rules.

e. Retraceability of acoustic effects

The system should be so organised that it is, in principle, always possible to retrace acoustic phenomena in the output to the rules or rule interactions responsible for them. This requirement is extremely important and extremely difficult to meet. It is important for a rather self-evident practical reason. If it is not met, it will often be very difficult to correct defects in the system because one simply does not know which rules to change. It is also important for a more theoretical reason, because if it is not met, we cannot consider the rule system to be a useful, insight giving model of the interactions underlying acoustic regularities in speech. We cannot use the rule system to explain how these interactions work. The rule system itself becomes yet another complex phenomenon to be explained.

This danger is real. A more or less complete set of rules for the generation of speech soon becomes so complex that even the people actually building the system may easily lose track of the chains of causes and effects in the system. This should be avoided as best as possible by a rigid and clear organisation. The inclusion of ad hoc rules, the interaction of which with other rules is not known, should be considered bad practice. In this respect it is important to keep up an accurate and easily readable description of the system and its workings.

f. Compatibility with speech production models

Although we are limiting our considerations to terminal analog synthesis, it is good to keep in mind that many of the regularities the system has to simulate in real speech stem from properties of the human speech production system. The explanatory power of our rule system will be considerably greater if it is so organised that specific rules or subsets of rules can be related to existing models of speech production.

For example, in writing rules for formant movements, it seems advisable to incorporate constraints which follow from dynamic models of speech production. It might even be worth while to generate separately the effects of the dynamics of articulatory movement and those of the acoustic theory of speech production in order to enhance the generalising power of the rules.

Similarly, it seems worth while to write separate rules for prosodically conditioned durational variation and durational variation caused by the short term dynamic behaviour of the articulatory organs, and make lower level interaction rules specify the actual durations.

In writing rules for consonant clusters we may introduce units corresponding to articulatory segments, which, owing to the considerable overlap in time of articulatory gestures, are not directly reflected in the acoustic signal, but which help in writing rules with some generalising power.

Keeping our rules as much as possible compatible with speech production models may guide us in creating a rule system which has a clear internal organisation and which is an explicit formulation of non-trivial aspects of what we know about speech.

how?

In this final section of our paper we describe briefly how our system for speech synthesis by rule is organised and how we have implemented some of the desired properties described above. This system replaces an older one which was based on a rather extensive hardware machine, the IPOVOX II, to which a software rule system was added. This older system has been described earlier (Slis and Muller, 1971, Slis 1971). In the present system the hardware has been confined to the actual signal generator, a digital hardware synthesiser (Rockland), which is driven from a computer (Philips P 9202, 16 bits, 16 K). A disk memory is used for additional storage capacity. The Rockland synthesiser needs fresh information about all parameter values at each period of the fundamental frequency ("glottal period"). Thus in the rule system the input in terms of discrete phoneme-like symbols has to be transformed into a sequence of segments corresponding to glottal periods, each with appropriate parameter values. Our rule system is at present confined to Dutch.

a. Segmental organisation of the programme

A main feature of the programme is that the segmental organisation of the input is sustained as long as possible in the rule system. Thus most of the synthesis rules operate on segments of phoneme size, each such segment in principle corresponding to an input symbol. For practical reasons this correspondence is not quite kept up in the case of vowels. All vowels consist of two symbols in the input. The rule system, however, interprets a phonologically short vowel as one phoneme-like segment, and a phonologically long vowel or diphthong as two segments.

b. Changing parameter values

For each phoneme of Dutch, standard values for a set of 20 parameters are stored in a table. These values do not necessarily correspond to any realisation of the phoneme but are rather so chosen that the rule system is optimised. The parameters are: Vowel formant frequencies F_1, F_2, F_3, F_4, F_5 and their band widths B_1, B_2, B_3, B_4, B_5, a nasal formant F_{N1} and a noise formant F_{ns} with their respective band widths B_{N1} and B_{ns}, amplitude of voice, noise and hiss (after formant filtering) A_V, A_{Ns}, A_{Asp}, pitch level F_0 and segment duration DUR. The twentieth parameter is not used at present.

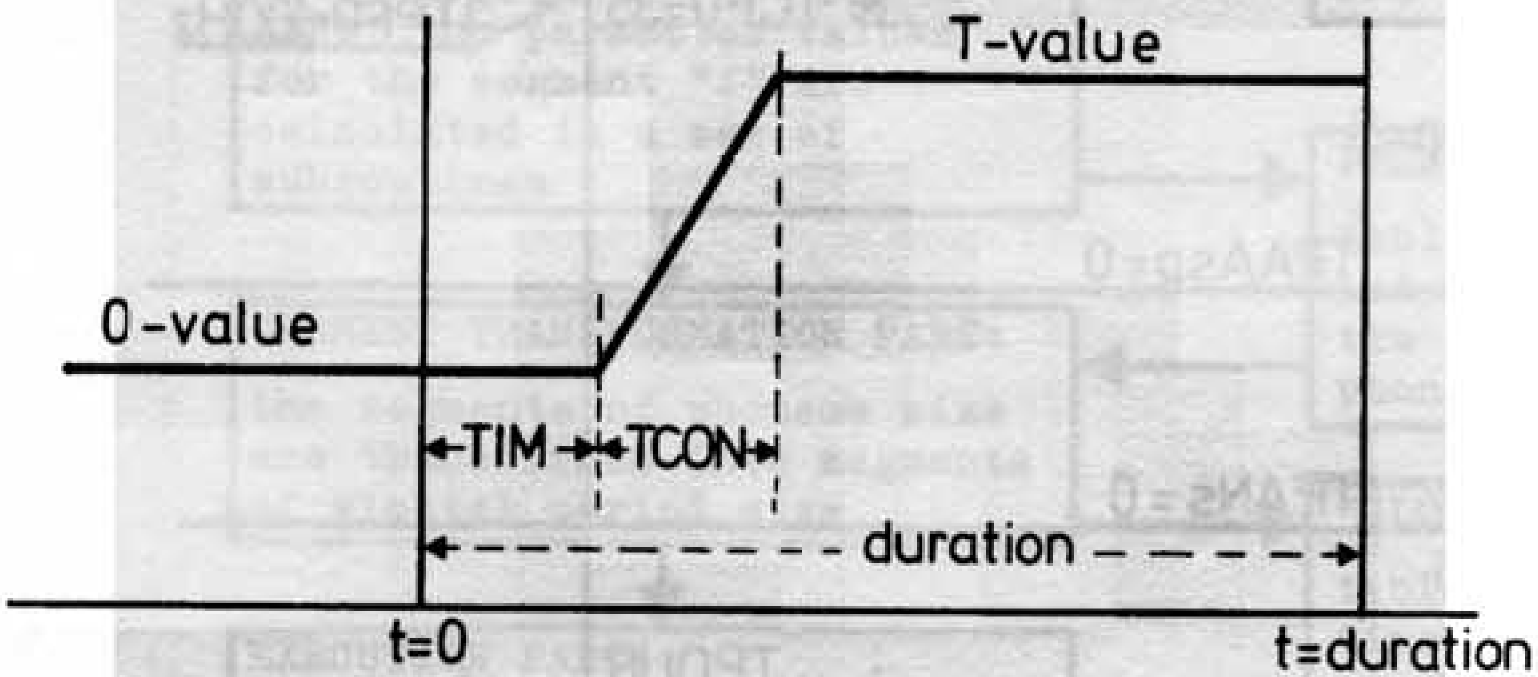
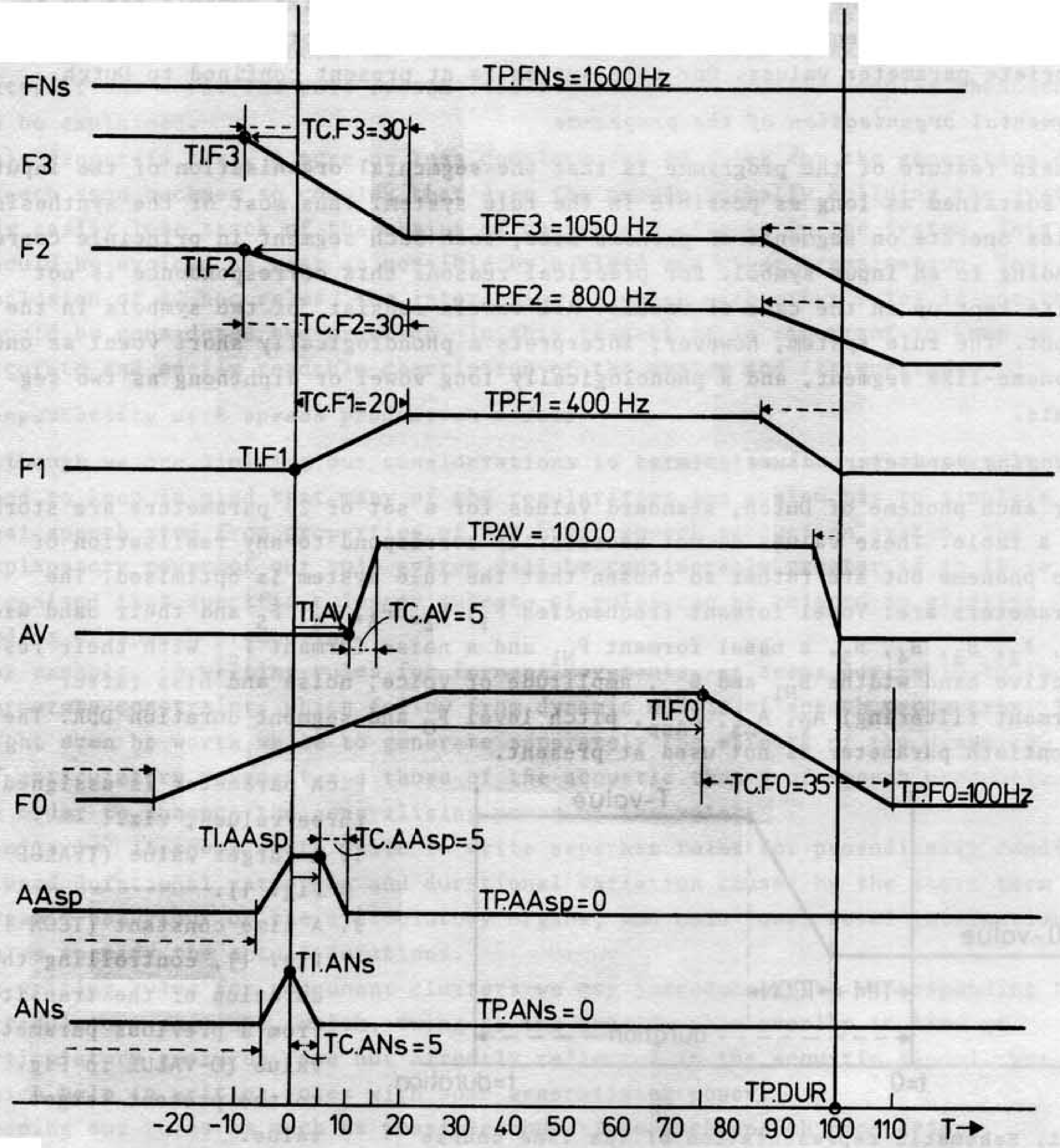


Fig. 1. Schematic representation of the time course of one of the synthesis parameters.
0-value = the previous target value
T-value = the target value of the present segment
TIM = the onset time of the change towards the new target
TCON = the duration of the change towards the new target

Each parameter is assigned three values, viz.:

1. A target value (TVALUE in Fig. 1).
2. A time constant (TCON in Fig. 1), controlling the duration of the transition from a previous parameter value (0-VALUE in Fig. 1) to the present target value.
3. A value (TIM in Fig. 1), controlling the moment of onset of the transition from one parameter value to the next, with respect to an abstract phoneme boundary (t = 0 in Fig. 1). If the transition starts before t = 0, TIM is assigned a negative value. In Fig. 2 an example is provided of some parameter values and their transitions in a plosive-vowel combination. This may give an impression of the possibilities of the system. The transitions may also follow more complex functions than provided by straight lines.



Parameters	TP=Target (TVALUE)	TC=Dur. of change (TCON)	TI=Start of change (TIM)
FNs =noise formant	TP.FNs	TC.FNs	TI.FNs
F3 =third formant	TP.F3	TC.F3	TI.F3
F2 =second formant	TP.F2	TC.F2	TI.F2
F1 =first formant	TP.F1	TC.F1	TI.F1
AV =amplitude voice	TP.AV	TC.AV	TI.AV
FO =fundamental freq.	TP.FO	TC.FO	TI.FO
AAsp=amplitude aspiration	TP.AAsp	TC.AAsp	TI.AAsp
ANs =amplitude noise	TP.ANs	TC.ANs	TI.ANs

Fig. 2. Example of the-time courses of 8 of the synthesis parameters in a plosive vowel combination.

c. Block diagram of the system

Fig. 3 presents a block diagram of the system. The system is so organised that there are four separate programmes and three data storage tables on the disc. The division into separate programmes is as follows:

1. INPUT part. The input string is stored and the conditions, calculated from the input string, are added to each phoneme.
2. RULE part. Standard parameter values are supplied from storage table. Input conditions of f-1, f, f+1, f+2 are made available. Synthesis rules operate on standard parameter values of segment f.
3. SEGMENT TRANSFORMATION part. The data for each phoneme, calculated in the RULE part and stored, are reorganised in terms of data for glottal period segments.
4. EXECUTION part. The glottal period data, calculated by the SEGMENT TRANSFORMATION part and stored, are transmitted to the speech synthesiser.

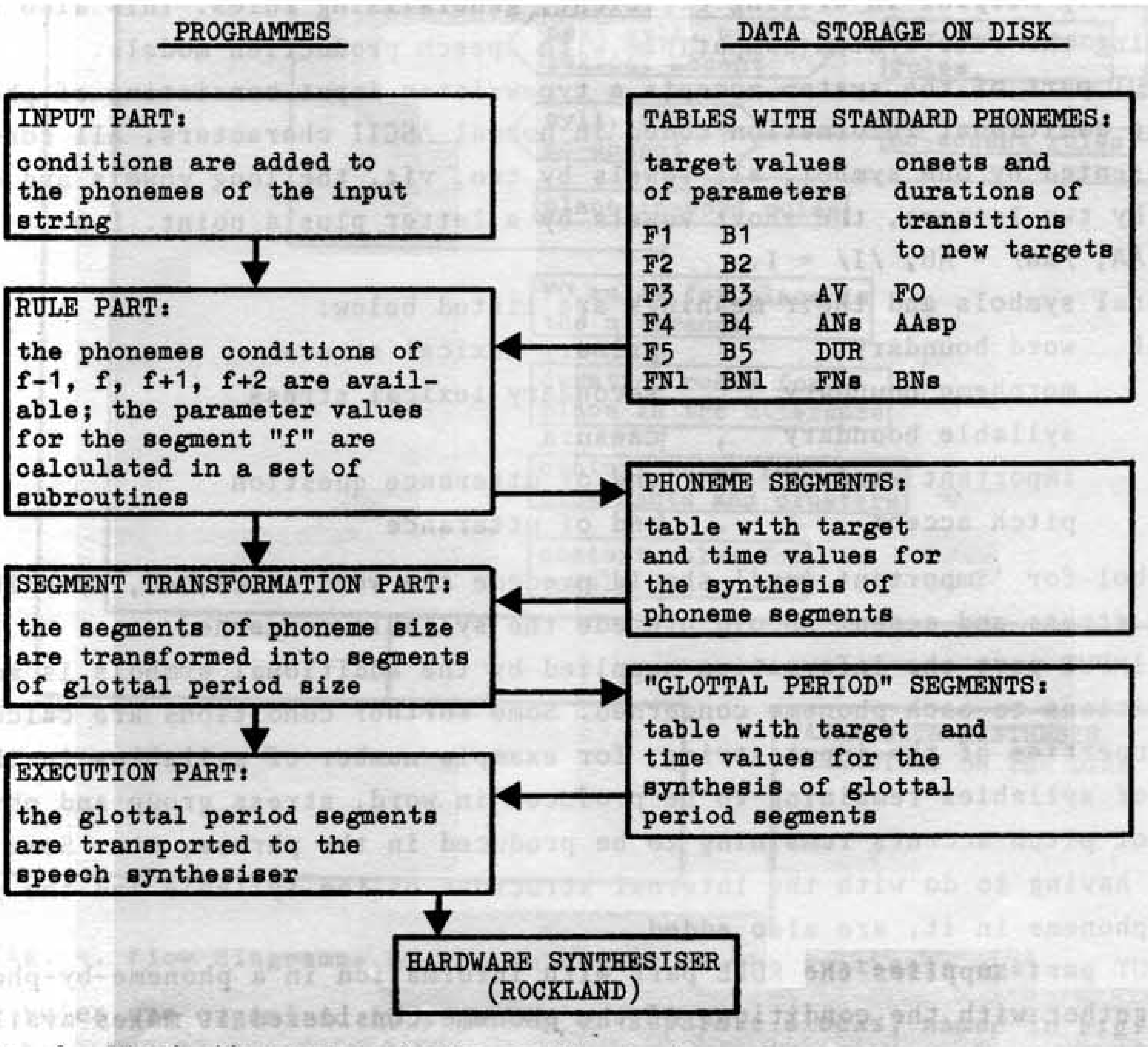

PROGRAMMES
DATA STORAGE ON DISK
INPUT PART: conditions are added to the phonemes of the input string
TABLES WITH STANDARD PHONEMES:
target values of parameters
onsets and durations of transitions to new targets
F1 B1
F2 B2
F3 B3 AV FO
F4 B4 ANs AAsp
F5 B5 DUR
FN1 BN1 FNs BNs
RULE PART: the phonemes conditions of f-1, f, f+1, f+2 are available; the parameter values for the segment "f" are calculated in a set of subroutines
PHONEME SEGMENTS: table with target and time values for the synthesis of phoneme segments
SEGMENT TRANSFORMATION PART: the segments of phoneme size are transformed into segments of glottal period size
"GLOTTAL PERIOD" SEGMENTS: table with target and time values for the synthesis of glottal period segments
EXECUTION PART: the glottal period segments are transported to the speech synthesiser
HARDWARE SYNTHESISER (ROCKLAND)

Fig. 3. Block diagramme of the synthesis by rule system.

Of these four separate parts of the system the EXECUTION part is trivial. Below we consider in somewhat more detail the inputs to the RULE part, the internal organisation of the RULE part, the function of the SEGMENT TRANSFORMATION part, and some additional features that have to do with the flexibility and ease of operation of the system.

d. The inputs to the RULE part

The RULE part has two different inputs, one being supplied by the table with standard phoneme parameters, the other by the INPUT part of the system.

As said above, the table with standard phoneme parameters contains maximally 20 parameters for each phoneme, each parameter being assigned 3 values, viz. target value, duration of transition, and onset of transition. These values are abstract with respect to phoneme realisations, and are chosen so as to optimise the rule system. Perhaps they most closely resemble phoneme realisations in optimally pronounced monosyllables, but they are not identical.

The memory space for one of the parameter values is actually used for a different purpose. A 16-bit memory word, belonging to the table with standard parameters, is filled with information on the feature composition of the phoneme. The features used are quite straightforward distinctive features, such as place of articulation, voiced/voiceless, degree of opening, etc. The classification obtained in this way is extremely helpful in writing efficient, generalising rules. This also helps us in keeping the rule system compatible with speech production models.

The INPUT part of the system accepts a typewritten input consisting of phonemes and some additional information coded in normal ASCII characters. All consonants are presented by one symbol, all vowels by two, viz. the long vowels and diphthongs by two letters, the short vowels by a letter plus a point. E.g.
/a:/ = AA, /au/ = AU, /I/ = I.

Additional symbols and their meanings are listed below:

blank	word boundary	'	primary lexical stress
=	morpheme boundary	"	secondary lexical stress
-	syllable boundary	,	caesura
+	important word	?	end of utterance question
↑	pitch accent	.	end of utterance

The symbol for 'important word' should precede the word concerned, symbols concerning stress and accent should precede the syllable concerned.

In the INPUT part the information supplied by the additional symbols is added as conditions to each phoneme concerned. Some further conditions are calculated from properties of the input string, for example number of syllables in the word, number of syllables remaining to be produced in word, stress group and phrase, number of pitch accents remaining to be produced in the phrase, etc. Some conditions having to do with the internal structure of the syllable and the position of the phoneme in it, are also added.

The INPUT part supplies the RULE part with information in a phoneme-by-phoneme way. Together with the conditions of the phoneme considered it makes available the conditions of the preceding phoneme and the two following phonemes.

e. The internal organisation of the RULE part

In Fig. 4 we have attempted to give a schematic account of the internal organisation of the RULE part of the system. The leftmost block is in itself a trivial part of the system, calling on the necessary information and the subroutine rules. The whole block called 'subroutine rules' is the interesting part of the system. It can be replaced by any appropriately formulated alternative set of subroutine rules. As it is, the subroutine rules are hierarchically organised. An incomplete impression of this hierarchical organisation is presented in Fig. 5. The sub-

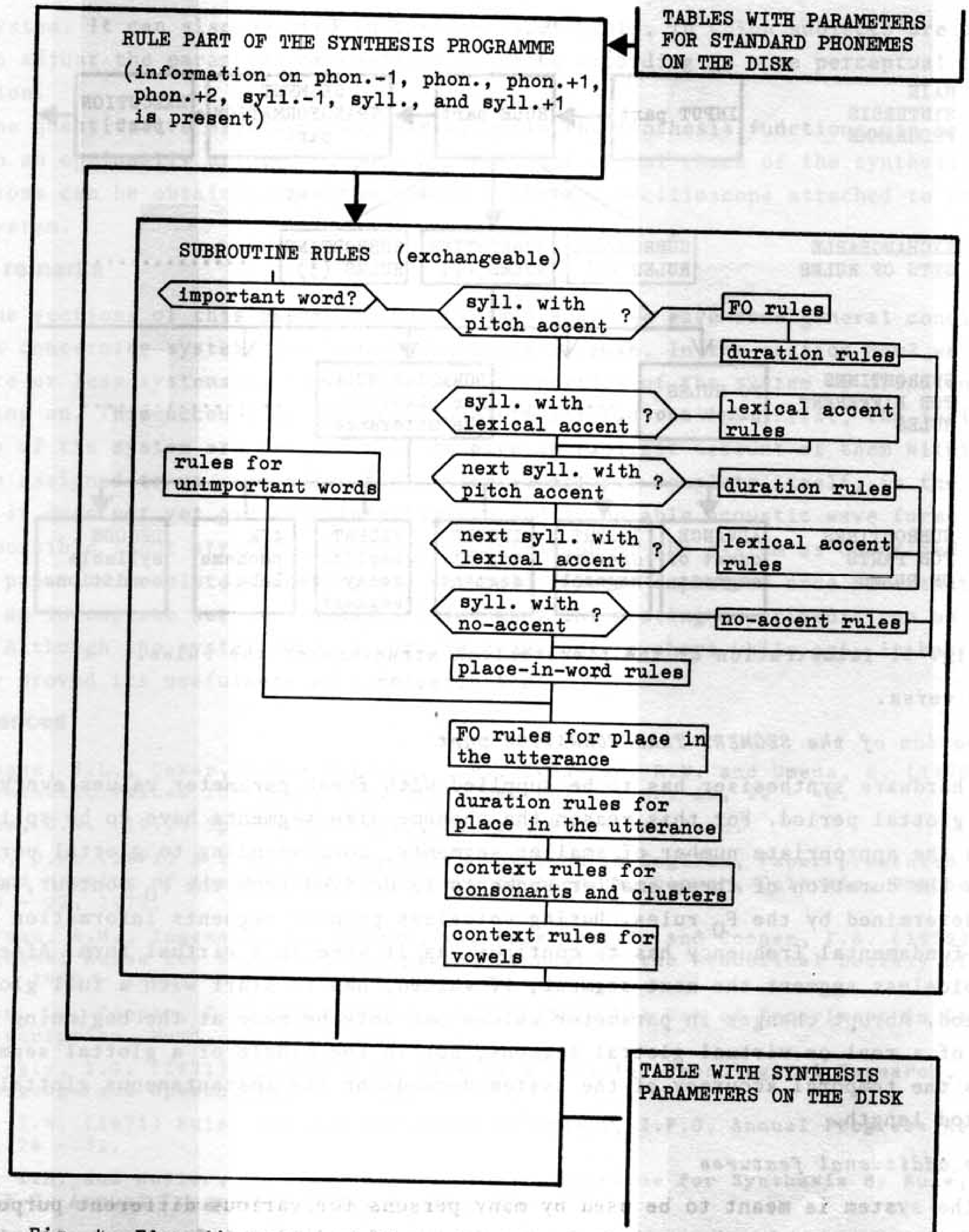

RULE PART OF THE SYNTHESIS PROGRAMME
(information on phon.-1, phon., phon.+1, phon.+2, syll.-1, syll., and syll.+1 is present)
TABLES WITH PARAMETERS FOR STANDARD PHONEMES ON THE DISK
SUBROUTINE RULES (exchangeable)
important word?
syll. with pitch accent ?
FO rules
duration rules
syll. with lexical accent ?
lexical accent rules
rules for unimportant words
next syll. with pitch accent ?
duration rules
next syll. with lexical accent ?
lexical accent rules
syll. with no-accent ?
no-accent rules
place-in-word rules
FO rules for place in the utterance
duration rules for place in the utterance
context rules for consonants and clusters
context rules for vowels
TABLE WITH SYNTHESIS PARAMETERS ON THE DISK

Fig. 4. Flow diagramme of the rule part of the synthesis system.

routine rules are organised in a number of separate blocks, named in Fig. 4, such as F_0 rules, duration rules, lexical accent rules, consonant and consonant cluster rules, etc. Each of these blocks can be replaced by an alternative block. Each block is split up into a number of smaller blocks. This is exemplified in Fig. 5 for the block DURATION RULES for place in the utterance. The hierarchical organisation of the subroutine rule system helps us in keeping track of the internal workings of the system and makes it possible to exchange rather easily subblocks of subroutine rules or individual rules. In some cases, however, things may become somewhat more complex than they appear to be, owing to necessary interactions between blocks of subroutine rules, as, for example between F_0 rules and duration rules. It may be necessary to cause F_0 movement to affect duration, and

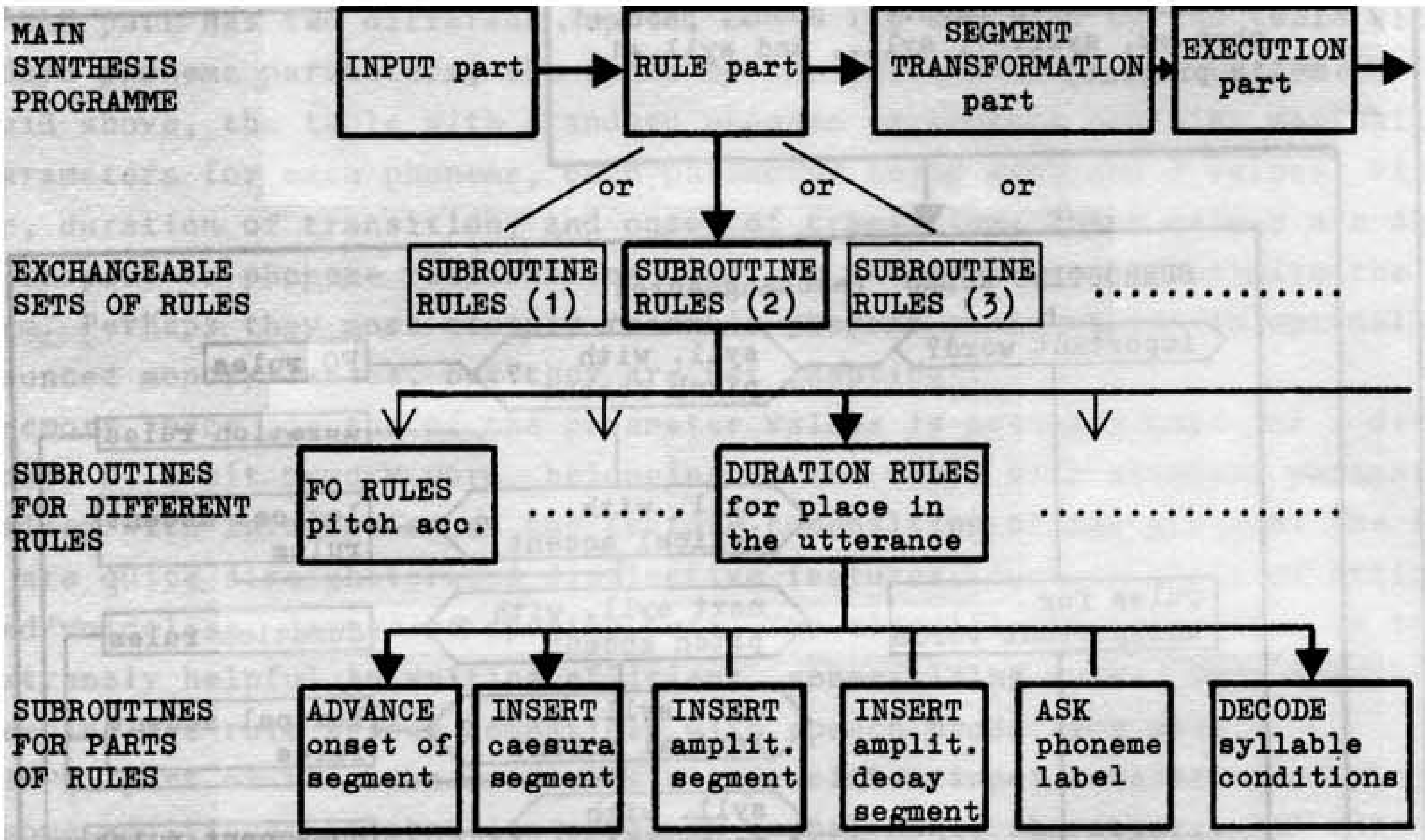


Fig. 5. Illustration of the hierarchical structure of the rules.

vice versa.

f. Operation of the SEGMENT TRANSFORMATION part

Our hardware synthesiser has to be supplied with fresh parameter values every new glottal period. For this reason the phoneme size segments have to be split up into the appropriate number of smaller segments, corresponding to glottal periods. Thus the duration of these smaller segments is derived from the F_0 contour, as it is determined by the F_0 rules. During voiceless phoneme segments information on the fundamental frequency has to continue, as it were in a virtual form. After a voiceless segment the next segment, if voiced, has to start with a full glottal period. Abrupt changes in parameter values can only be made at the beginning or end of a real or virtual glottal segment, not in the middle of a glottal segment. Thus the temporal accuracy of the system depends on the instantaneous glottal period length.

g. Some additional features

As the system is meant to be used by many persons for various different purposes, special attention has been given to the design of additional features enhancing the flexibility and ease of operation.

The operation of any rule or block of rules in the system can be suppressed manually by means of a set of sense switches. This is helpful both in making stimulus tapes and in playing around with the system for heuristic purposes.

Parameter values and coefficients of selected rules can be manually controlled with a set of potentiometers (connected to the computer with an ADC). This can be done at various levels of the programme. For example, one of the parameter values in the table of standard phonemes, or the value of a coefficient on one of the rules can be brought under manual control. In the SEGMENT TRANSFORMATION part of the system a segment parameter in the already assembled utterance can be replaced in this way. This provision gives a flexible and fast way of interacting with the system. It can also be used in on-line experiments, in which subjects are asked to adjust the parameter or coefficient value according to some perceptual criterion.

The quantitative effect of any variation in the synthesis functions can be checked in an optionally printed output. An immediate visual check of the synthesis functions can be obtained with the aid of a storage oscilloscope attached to the system.

final remarks

In the sections of this paper called Why? and What? we gave some general considerations concerning systems for speech synthesis by rule. In the section How? we gave a more or less systematic account of some properties of the system we are actually working on. This account is far from complete for two reasons. First, the individual rules of the system are far too many to give an explicit account of them within the space assigned to us. Secondly, the system is not yet complete itself, in the sense that it does not yet generate intelligible and acceptable acoustic wave forms for all possible input strings. The basic organisation of the system as described in this paper is complete, however, and we would like to emphasise once more that even with an incomplete set of synthesis rules many interesting experiments can be carried out. Although the system came in working order only a short while ago, it has already proved its usefulness as a research tool.

references

Flanagan, J.L., Coker, C.H., Rabiner, L.R., Schaefer, R.W. and Umeda, N. (1970) Synthetic Voices for Computers, IEEE Spectrum, 7, nr. 10, 22 - 45.

Holmes, J.N. (1972) Speech Synthesis, London, Mills and Boon.

Klatt, D.H. (1971) A Theory of Segmental Duration in English, Paper presented at the 82nd Meeting of the Acoustical Society of America, Denver, Colorado, October 19 - 22, 1971.

Liberman, A.M., Ingemann, F., Lisker, L., Delattre, P. and Cooper, F.S. (1959) Minimal Rules for Synthesizing Speech, Journal of the Acoustical Society of America, 31, 1490 - 1499.

Lisker, L., Cooper, F.S. and Liberman, A.M. (1962) The Use of Experiment in Language Description, Word, 18, 82 - 106.

Mattingly, I.G. (1971) Synthesis By Rule as a Tool for Phonological Research, Language and Speech, 14, 47 - 56.

Slis, I.H. (1971) Rules for the Synthesis of Speech, I.P.O. Annual Progress Report, 6, 28 - 31.

Slis, I.H. and Muller, H.F. (1971) A Computer Programme for Synthesis By Rule, I.P.O. Annual Progress Report, 6. 24 - 28.