Proceedings of The Institute of Acoustics 2A12

SPEECH SYNTHESIS BY RULE FROM A LOW LEVEL FEATURE BASED INPUT

J.E.CLARK

MACQUARIE UNIVERSITY, N.S.W. AUSTRALIA

Acoustic domain rule systems for generating highly intelligible continuous speech from a string of discrete linguistic elements using a terminal analog synthesiser have become an established research technique during the past decade and a half. Most early rule systems operated on the strategy of directly transforming a broad phonetic or phonemic transcription into a continuum of parametric data (1,2). More recently, rule systems (3,4) have been developed which explicitly separate the rules dealing with high level and relatively abstract phonological processes from those dealing with low level phonetic detail. Such strategies are a recognition of the fact that these represent rather different aspects of the speech generation process, and allow appropriate modes of rule expression to be exploited in each area.

The system described in this paper is a set of low level phonetic rules designed to operate from a narrow phonetic transcription having a segment and feature format which is intended to give quite detailed control over the acoustic parametric continuum derived from it. The rules thus have rather less context sensitivity built into them than in more conventional systems, and therefore generate (automatically) a smaller range of intrinsic allophones. This is a quite deliberate strategy, since at the present state of knowledge of the nature and precision of higher level neural encoding of speech motor commands and their relationship to the bio-mechanical performance limitations of the vocal tract it is not easy to make clear cut distinctions between what might, and might not, be low level allophonic intrinsic to the (peripheral) speech production mechanism itself.

All acoustic domain rule systems, including this one, necessarily make assumptions about the degree to which allophonic variation should be automatically (and thus in a sense intrinsically) generated by low level context sensitivity in the rule system without being specified in any segmental form. Most commonly this occurs by deriving the parametric continuum from putatively 'normal' segment target values with duration and degree of undershoot fromthese being determined by the parameter transition rules. The present system attempts to make a greater degree of such variation specifiable at a low yet segmental level of representation, and therefore able to be explicitly specified in linguistic terms at the 'bottom end' of a phonological rule system, rather than being somewhat less accessibly embedded within the purely parametric rule component. This is rather the reverse of conventional systems which seek to free the user from the need to define the acoustic parametric continuum. Such an objective, whilst appropriate for high level rules, is not necessarily so for a low level system or rule component. Furthermore, the system allows considerable control over the nature of such intrinsic allophonic variation as the system does automatically generate through the choice of values used in establishing the user-accessible data base.

SPEECH SYNTHESIS BY RULE FROM A LOW LEVEL FEATURE BASED INPUT

The basic organisation of the rule system is shown in fig.1. The so-called parameter generating model is simply a collection of computational routines and rules which accept and interpret the low level phonetic input string and transform it into the appropriate quasi-continuous parametric data to drive the synthesiser with the aid of information from the resident data base. Prior to input to the parameter generating model, the input string is subject to checks for appropriate syntax, and allows editing for corrections and changes to the string. The user accessible data base consists of range and scaling information for the rules and computation routines of the parameter generating model, which are themselves dimensionless, as well as the parametric data upon which they operate. This has the advantage of allowing a clear-cut distinction between the inherent characteristics of the rules and computing procedures themselves, and those imposed upon them both by their scaling and the data upon which they act. It also allows for considerable 'fine-tuning' of individual aspects of the rules without serious interaction in the system as a whole.

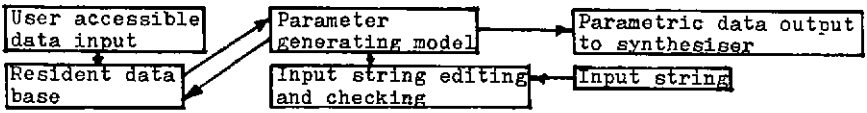

Fig.1

The transformation from any form of segmental representation to an acoustic parametric continuum is a problem central to most forms of synthesis by rule. Being a step from the abstract to the concrete it has inherent difficulties, as the relationships between the segments and the continuum are often non-linear, time overlapping, and complex, as well as context sensitive. In the present system an acoustic syllable based strategy has been employed, in contrast to the more commonly used seg-segmental concatenation strategy, as noted earlier. This simplifies some of the rule structure necessary to account for certain aspects of the above relationships, but at the cost of a rather larger data base. At least two other recent explorations of this general syllable based approach have been reported in the literature.

The syllable remains a contentious unit in speech and language research, and it has been argued that it lacks the psychological reality of segmental units and any satisfactory definition, particularly for purposes of segmentation in polysyllabic words. Despite this, there is evidence both from production and perception studies that it may be an important unit of encoding. The present rule system is based on procedures for the prediction of CV and VC structures, which in turn provide the basis for constructing more complex syllable types. In the case of polysyllabic structures involving VCV sequences, the problem of syllable segmentation in assigning the intervocalic consonant to a particular syllable has been avoided by treating such consonants as shared or common to the structure of both the adjacent syllables.

For purposes of rule organisation, the syllable is treated in a traditional way as having three components, a central and obligatory peak, and an optional onset and coda. The syllable peak is occupied by the vocalic

nucleus, and provides the primary reference from which parametric characteristics of the remainder of the syllable are calculated. The onset and coda are occupied by consonants whose contextual relationship to the syllable are defined in the input string features. The syllable based nature of the synthesis strategy is reflected in certain of the features attached to segments in the input string and which define syllable rather than segment oriented aspects of the parametric continuum. The phonetic representation of the input string has a format of line by line entries in which each line represents one segmental element in the sequence of speech. The initial symbol of the line defines the segment type, and the associated features define its specific characteristics in that particular segment and syllable context. Symbols defining the start and finish of speech sequences, as well as intervals of silence are also provided. The general form of a CVC sequence would be:

$	start of speech sequence
S $[F_1 \cdots F_n]$ $\langle F_1 \cdots F_n \rangle$	consonant (prosody features<>optional)
V $[F_1 \cdots F_n]$ $\langle F_1 \cdots F_n \rangle$	vowel
S $[F_1 \cdots F_n]$ $\langle F_1 \cdots F_n \rangle$	consonant (prosody features<>optional)
##	end of speech sequence

Vowel targets are specified in cardinal vowel terms which refer to a modified form of the cardinal vowel system auditory space using the traditional 16 primary and secondary cardinal reference points, but with a unique central counterpart for each reference vowel. They are also re-arranged into rounded and unrounded vowel spaces to avoid problems of discontinuity in their acoustic representation associated with changes in lip rounding between adjacent cardinals. The auditory arrangement of the unrounded space is shown in fig.2.

Incremental changes in values of vowel targets relative to a given cardinal value may be made in five steps of height and fronting within the vowel space as shown in fig.3 using vowel target features H±N and F±N. The same may be done using the centralised form of the cardinal value by adding the feature C to the target specification. The duration of the target is specified in N 10mS increments by the feature SLN, and where glides or diphthongs are required, two or more targets may be joined by a glide with a similarly specified duration using the feature GLN. The cardinal reference values are specified in terms of $F_1F_2F_3$ and are stored in the resident data base. Incremental values specified by the input string are calculated by an interpolation and estimation method between these points in the $F_1F_2F_3$ space. The vowel prosody features define the pitch and amplitude for the duration of the entire syllable peak, that is, the vowel target(s) and associated glide(s), together with the consonant-associated

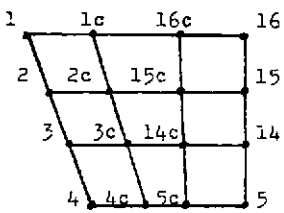

Fig.2

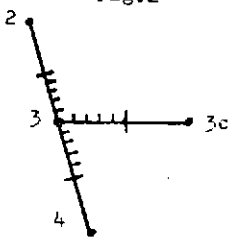

Fig.3

transitions which together comprise the complete vocalic nucleus. There is a choice of four pitch contours defined by the feature PCN operating over a range of 10 equal pitch increments the magnitude of which are are user defined in the resident data base, together with a 11th larger increment for terminal pitch falls. The contour types are; linear pitch change, falling, rising, rise-fall, and fall-rise, with contour shapes constructed in straight line sections. The pitch levels used in a given syllable are defined in the associated vowel prosody feature set at the start (PSN), finish (PFN), (and in some instances) medial (PMN) points of the contour. Rules for preserving essential aspects of pitch change of contours in syllable nuclei of varying lengths are applied automatically in all cases except that of linear pitch change. Provision also exists for varying the intensity contour (ICN) of the vocalic nucleus (although it has not been greatly used in practice) as well as setting the intensity itself (IN) in 2dB increments. The vocalic nucleus may also be made voiced or voiceless by selection of the features V or VL respectively. The form of the vowel and associated feature input is thus:

ε[SLN H N F N C] <PCN PSN (PMN) PFN ICN IN V(L)>

Consonants in the onset and coda of the syllable are considered in two components, the (relatively) steady state part corresponding to a period of radical constriction or occlusion in the vocal tract, and the spectral changes evdenced as formant transitions moving to orfrom the vowel target in the vocalic nucleus. Essential aspects of each are defined by the features associated with consonant segments. In the steady state part these are, duration (SLN), occlusion duration (OLN) (for stops and affricates and intensity (IN), and in the vocalic nucleus formant trajectory offset (LONN) and transition duration TTNN). Onset of voicing may also be specified, but applies variously to the steady state or formant transition components depending on syllable context, segment type and feature value.

Formant transitions are initially computed in the frequency domain only as the trajectory of a point in a three dimensional acoustic space defined by $F_1F_2F_3$ axes. The endpoints of the trajectory are defined by the formant values of the vowel target in the syllable peak and by a set of formant values selested from a 4x4 matrix of consonant 'locus' values each corresponding with the 16 (4x4) cardinal reference points in each vowel space, and which thus functions context sensitively. The purpose of the formant trajectory feature (LONN) is to allow the consonant endpoint of the trajectory to be shifted back towards the vowel or extended beyond its normal value as required. Transition duration defines the time occupied by the trajectory when transformed to the time domain, and there is provision in the data base to make this a linear or variable index power law thereby setting the shape of the transitions. Both features have 'normal' values for stressed syllables which are applied unless N is given a value other than 0.The NN is required for specifying the syllable context of the consonant (CV VC or VCV). In the case of consonant clusters, only one C is considered as shared, and the remainder assigned to one or other syllable with parameter computation proceeding outward from the vowel target in the syllabic nucleus in all cases.

References

1. J.N.HOLMES et al 1964 L&S 7,127-143. Speech synthesis by rule
2. L.R.RABINER 1968 Bell Sys. Tech. J. 47,17-37. Speech synthesis by rule
3. R.CARLSON & B.GRANSTROM 1975 STL QPSR 1,17-26. A text to speech system
4. D.H.KLATT 1976 IEEE ASSP-24 5,391-398. Structure of a phon. rule comp.