Early Synthesis - an Example of Simulation

Kate Morton

Reproduced from the PhD Thesis Speech Production and Synthesis, June 1987.

Generic text-to-speech synthesis systems involve running a production algorithm which accesses linguistic/phonetic databases. Such a system is illustrated in Fig. 11.

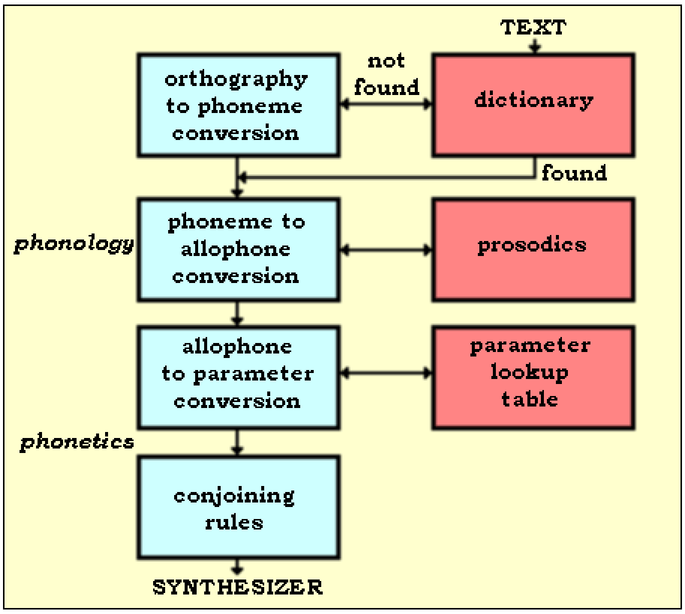

Fig. 11 A generic synthesis by rule system.

The system begins by converting orthographic strings to phonemic strings by applying sets of rules, or by consulting a dictionary giving phonemic representations of words or morphemes. A set of context sensitive rewrite rules is then applied to generate a corresponding allophonic string. These rewrite rules deal with phonological assimilation, vowel reduction, and changes which, in the human being, are under voluntary control and are language dependent. Stress and intonation marks are assigned to this string by interpreting punctuation marks in the text, by dictionary lookup for word stress, or by rule. Semantic and syntactic information are not normally available to the system.

In the next stage, individual segments are parameterized by consulting a table with entries for each segment available to the system. These table entries contain all the information necessary to calculate transitions and provide acoustic parameter values for the synthesizer driver. Segment boundaries are assigned, and transitions are calculated across segment boundaries. The final parameterized string is delivered to the synthesizer driver, and a speech waveform is output (Holmes 1964 *et al.*, Ainsworth 1974). Thus in its usual for a SbR system consists of a set of algorithmic procedures which result in a speech output that is identical each time the same text is input. Human beings, however, produce speech which varies a great deal even on repetition.

Databases used in synthesis systems of this type consist of descriptions derived from linguistics and phonetics as the subjects stood in the 60s and early 70s. Thus the database of phonological rules consists of context sensitive rewrite rules about stress assignment, phonological assimilation, vowel reduction, duration assignment, intonation contours. The allophone database consists of a lookup table specifying intrinsic parameter values for each allophone. In linguistics, these databases comprise generalizations about language and speech, rather than specifications of actual individual events (Chomsky 1957). Effectively, then, the output from a synthesizer is a simulation of idealizations rather than the simulation of real events.

In several ways synthetic speech does not fully replicate human speech, and is sometimes seriously deficient. For example, orthographic to phoneme conversion is not always satisfactory, since there are many exceptions to spelling rules. The prosodics generally is poor. Segment joining can be fairly good, but there are exceptions, so that speech in some cases is unintelligible. The result is that synthetic speech sounds unnatural, machine-like, and is sometimes difficult to understand. It is not of sufficiently high quality for extended listening or good enough for general use in sophisticated systems such as interactive devices.

Variability, which contributes to naturalness, could be added to the system by increasing the size of the databases to include alternative optional rules.

EXAMPLES OF OPTIONAL RULES

The knowledge base contains the following optional rules (for the moment it is not important whether they are phonological or phonetic). Here are two examples of optional rules. In both cases choice depends on non-linguistic factors. Rule xxx could apply, for example, in casual speech, Rule xxy could apply under conditions of contrastive emphasis.

Rule xxx (opt.)

[plosive] → [unreleased] / -- #

'In final position plosives are optionally unreleased.'

Rule xxy (opt.)

V → [lengthened] / [-- , +stress] (C) #

'Stressed vowels at the end of words are optionally lengthened.'

Optional rules constitute part of a linguistic description, but they cannot be included in a synthesis system unless there is same way of choosing which one among them is the right one on any particular occasion. Although included in a knowledge base optional rules cannot be individually retrieved without additional information. Unfortunately the type of linguistics used in current synthesis does not indicate how to select among the alternatives it describes.

One way the synthesis system could produce variability would be by listing a context sensitive rule for each possible occurrence of all items in all environments. However, if such a list could be constructed it would be very large and probably impossible to deal with efficiently.

Therefore, a way is needed to simplify the method of choosing the right optional rule. One method is to enable the system to make a decision based on optimal guessing as to the likelihood of a particular rule suitable for a particular utterance (see SYNth-EX).