The British Telecom Research Laboratories Text-to-Speech Synthesis System - 1984-1986

[PART II]

Katherine Morton

Reproduced from 'Speech Production and Synthesis' - unpublished PhD thesis, University of Essex 1987, pp. 161-172. Based on reports to British Telecom Research Laboratories during the period 1984-1986 (part of a series of research contracts over a nine year period). The reports themselves cannot be placed in the public domain for reasons of confidentiality and intellectual property rights.

A - THE PHONETICS AND PHONOLOGICAL UNDERPINNING

The study of phonetics over the past 20 years can be divided into several different areas:

- articulatory phonetics, which is concerned with the neurophysiology and motor control of speech, and the anatomy and geometry of articulations;
- acoustic phonetics, which is concerned with describing the properties of the speech waveform and the aerodynamics underlying its production;
- auditory phonetics, which is concerned with the anatomical and neurophysiological mechanisms of hearing and with cognitive processes involved in the perception of speech.

In the SbR (synthesis-by-rule) part of JSRU-based synthesizers like the BTRL system facts about only the acoustic subcomponent are normally considered useful for speech synthesis. Although, if a particular effect has to be faked in some way it is helpful to understand some aspects of perception. [H]

B - THE CURRENT PROGRAM

The program supplied by JSRU is constructed in levels which have general similarity to a linguistics descriptive model; and it is important to keep the distinction clear between various levels in the program. The type of segment used for representations of utterances changes as the utterance is processed through these levels from input text to output signals for the synthesizer.

The initial input is in ordinary orthographic symbols. From this is derived a representation in phonemic symbols, then a representation in what today would be called extrinsic allophonic symbols, and finally a representation expressed in terms of intrinsic allophonic symbols. In the JSRU software system the underlying linguistic/phonetic description did not

explicitly state the distinctions between these types of segment. Two labels only were used: phoneme and allophone. These were not always used systematically in the way we would use them today.

Rules linking text to phonemic strings are called orthography to phoneme conversion rules The next set of rules, linking phonemic representations to extrinsic allophones, are phonological rules. Those linking extrinsic allophonic representations to strings of intrinsic allophones are phonetic rules. This is the general layout of the segmental aspect of the SbR program, using current linguistics terms to describe it.

The prosodics processing runs in parallel, and consists of a general procedure in which intonation contours are assigned to sentences based on scanning the input sentence for punctuation marks. Stress is assigned according to a small rule set and some lookup tables. Prosodics is the most difficult area to deal with, and this is apparent in the speech output of SbR systems of this type. One way of dealing with some deficiencies is by including a really well-designed dictionary (see **Dictionary** below).

In both the phonetics and phonology of the SbR system, the linguistic rules are similar in type. They consist of context sensitive rewrite rules which take the general form

A alpha B → A beta B

where alpha is a segment or segments rewritten as the segment or segments beta, and where ~ and B are each strings of null, one or more segments. alternative notation more commonly used in phonology is

alpha → beta / A - B

That is, given a context for segment string alpha of left A and right B, where A and B are strings of null, one or more segments, then alpha rewrites as segment string beta.

In order to make it easy to understand and manipulate procedures in the program, sets of such rules should be grouped in blocks at different levels, and operate in an ordered way on segments in the appropriate contexts. As segments move through the components of the system, algorithms which apply to them should be fully documented. A gloss of the algorithm which expresses a linguistic/phonetic rule should be stated in rule format that can be easily understood.

One serious problem with the current program (JSRU) is that it is extremely difficult to understand. The structure is not apparent except in the most general way, and no documentation exists to explain to someone who has not been intimately involved in developing the program exactly what is happening at each stage, what representational symbols are being processed, and at what point one level ends and another begins.

Translation into a more commonly used language with full documentation should enable a more directed approach to one of the main objectives of the Project: suggesting changes to the current system to bring about an improvement in the quality of the speech output.

C - LP.DAT CHANGES

A major area of work this year has been systematically updating the cell values for allophonic segments in the LP.DAT file [the segmental lookup table of parameter values for canonical forms]. There have been three sources of information for this updating:

- the phonetics literature,
- experimental work we have conducted on the acoustics of human speech in comparison with synthetic speech produced by the BTRL synthesizer, and
- listening tests.

The values now entered in the tables are optimized with respect to each other. It is important to remember that entries in the LP.DAT file are not independent. Within the file they are interdependent in the sense that perception of speech rests on relative judgements on the part of the listener, not absolute judgements. Some of the values are dependent on the hardware they are used to drive. Any further adjustments to the hardware synthesizer BTRL may make will change the final acoustic output of some of the entries in the tables.

D - TRANSITIONS

We have taken the JSRU conjoining algorithms and extended them to produce transition shapes which more nearly resemble those of natural speech. There are three main types of transition, depending on the rank of classes of segments. These types are explained, together with examples illustrating their effectiveness. Critical values for the appropriate parameters in the LP.DAT have been adjusted for input to the algorithms to generate more natural looking transitions.

It should be pointed out that the SbR system supplied by JSRU to BTRL was intended as a research tool. Consequently many values in the tables as supplied were not optimal but were intended as suggestions as a starting point for detailed research which JSRU did not carry out. We have seen our task as making up some of that detailed research to enable the research tool to become a fully-fledged development system for BTRL's specific needs.

In the case of the transition algorithm provided, corrections have been suggested and these are detailed later in the report. (See Comment [F]}

E - AN ALTERNATIVE METHOD FOR HANDLING TRANSITIONS [I]

In earlier reports, we suggested an alternative method of handling transitions in the SbR program which details an algorithm intended to provide much smoother transitions between segments. The method retains fixed parameter values in the tables and preserves the notion of segment. No one would disagree with the idea that accurate replication of speech requires some conjoining algorithm and is a major problem to be solved in any SbR system. This problem results from the decision to base synthesis systems on phoneme segments which are invariant and represent idealized articulatory target values.*

> [**footnote:* The JSRU system in common with other formant synthesis systems includes lookup tables specifying synthesizer driver parameter values for individual segments.]

Some synthesis systems have been developed based on tables containing diphones, demi-syllables, triphones, whole morphemes (with a trade-off in storage and a reduction in size of the rule set). Approaches which have been tried out that do not presuppose single sound segments can sound more natural compared with the usual JSRU approach. But if the notion of target values for individual sound segments is retained, it is necessary to improve the transition algorithm.

Some researchers have claimed that the shape of transitions is relatively unimportant perceptually and that straight-line conjoining as in the JSRU system is adequate. This may be the case for perception of single words in isolation. However, in listening to longer stretches of synthesized speech, part of the listener's feeling that something is not quite natural is removed if the sentences consist of words with transition shapes more like those of natural speech. Listening tests comparing straight-line transitions with hand-crafted transitions derived by direct measurements from spectrograms of human speech confirm this. Furthermore, it can be shown that in human speech

- transition movement is not symmetrical,
- discontinuities occur in only a few cases, and
- transition type is dependent on phonological or phonetic context.

Although to some extent the JSRU algorithm can approximate the general shape of transitions, the curved shape characteristic of natural speech cannot be generated by rule so

easily. It would be possible to specify a large number of allophones and context sensitive rules to access them, but a resulting rule set which could adequately generate natural sounding speech would be very large.

As has been noted elsewhere, three types of basic shape were identified: fast, medium, slow. Algorithms are presented to generate these shapes for conjoined allophonic segments. This approach was generalized to minimize transition discontinuities, and is presented in this program. This was developed further by defining transition types more carefully and introducing a weighting function. This enables the algorithm to characterize the missed target characteristic of natural speech.

The general principle worked quite well as a procedure, giving a close match to human inter-segment transitions. Shapes approximate natural speech more successfully than the original algorithm with its abrupt transitions and discontinuities.

F - FUTURE DEVELOPMENT

We suggest that changes to the system should take into account developments in linguistics over the last decade. Including another subcomponent of phonetics - linguistic phonetics - could be useful in developing a more sophisticated system. In general, linguistic phonetics is concerned with how different areas of phonetics interact with each other and with other components of linguistics. In the JSRU model, interaction of phonetics with phonology is taken into account somewhat , but a more accurate SbR system will include a description of the relation between phonetics, semantics and syntax, as well.

In JSRU-based SbR systems, two linguistic levels are implemented: phonology and phonetics. These systems see phonetics as a continuation of phonology and not as a different type of modeling, accounting for different kinds of data. Over the last five years a new component , cognitive phonetics (proposed by members of ASTL-Essex), is being recognized as useful in simulations. This component characterizes the phonetic processing which produces speech, but which in addition is controlled by cognitive processing. There are as yet no commercial systems incorporating true simulation of the human language processing system. [J]

G - THE DICTIONARY

The current JSRU dictionary is very small, containing a few exceptions to the initial orthography to phoneme rule set. The idea of substituting a more comprehensive dictionary is a major departure from the view of the 6~s model which believed that the more generated by rule, and the less looked up in tables, the better.

There are however different kinds of linguistic-based dictionaries. A dictionary which simply gives some kind of phonological or phonetic rendering of an orthographic word is not very useful, particularly since in most dictionaries the two levels of phonology and phonetics are not clearly separated. There is not much point to a simple lookup table of words beyond the comparatively few exceptions already noted to the orthography to phoneme conversion rules. Such a strategy might lead to a limited vocabulary synthesizer, thus losing the important feature which JSRU managed so well in comparison with competing systems in the 6~s -generality, enabling the synthesizer to produce any sentence of English.

The need for a more sophisticated dictionary and decisions about the best format stem from defects in the a-syntactic, a-semantic segment-oriented model on which the JSRU system is based. Linguists have known since the middle G()s that it is not useful to think of the components of a grammar (semantics, syntax, phonology, phonetics) as theoretical abstract constructs which are completely separate from each other, Dictionary entries would ideally contain information that would allow interaction between the different components of the linguistic descriptive system.

The concept of a phonology/phonetics interdependent with semantics and syntax is the goal: this has been misinterpreted by non-linguists applying linguistic descriptions to practical problems such as speech synthesis and automatic speech recognition. This misunderstanding

unfortunately carried over into the JSRU SbR system and into systems which developed from it. The rules of linguistics are not a set of algorithms; they are not 'programmable' into a working system such as SbR in an attempt to simulate the human speech production process. [K]

Descriptive linguistics is not about simulation, whereas SbR software is about simulation. It is possible that some of the difficulties in applying linguistics/phonetics to speech synthesis systems may be due to the differences between description and simulation. However, the JSRU synthesizer itself is very good, perhaps because it is a true simulation of the acoustic system of human speech. If the SbR program which drives it were a true simulation also then the speech output might be very much better.

Many of the questions concerning a usable dictionary system for synthesis are about the interdependence of the components in linguistics and about the problems of description vs. simulation. For example, to what extent does the present SbR system fail because it is a-semantic or a-syntactic? Certainly in rendering prosodics which is not adequate. In addition segmental rendering does not capture variability, which is a major distinguishing characteristic of human speech. [L] Variability itself results in part from cognitive aspects of the semantics and syntax of language processing.

Various dictionary-based systems have been demonstrated over the last decade, and all of them succeed better than the 6~s idea of a generative system but which, because of a misinterpretation of ideas in linguistics, is not a true simulation. Early and effective systems which are now being revised include the Telesensory Prose-2000, DECtalk and, it seems, Infovox. Researchers are discovering how to establish an efficient and practical dictionary format.

Since the synthesis system's output is sensitive to dictionary format, the dictionary under design at Essex University [subsequently incorporated into SPRUCE] (not part of the BTRL sponsored SbR project) is conceived as multi-dimensional (an orthography-to-phoneme dictionary is one-dimensional) and dimensionally open-ended. It incorporates a semantic/syntactic parser which

off-loads some of the information normally just listed in the dictionary,

partly bridges the gap towards a proper cognitively-oriented system, and

enables a more generative approach than is otherwise the case with a simple dictionary.

A dictionary-based system for synthesis-by-rule has advantages, but it needs careful design. At worst the system would output speech of a similar or lower quality than now, at best of improved quality - though still not up to the naturalness of human speech. It must be emphasized that the dictionary format in respect of the relationship between levels should be carefully thought through. It is important to make certain that the decision for any particular degree of complexity is carefully taken in a well principled fashion.

H - THE FEMALE VOICE

As a preliminary to extending the scope of the JSRU SbR system, we were asked to look at what might be involved in adding a female voice to the system. To that end, we have examined some of the characteristics of female voices as opposed to male voices. It is clear that simply raising the fundamental frequency is inadequate for simulating the female voice.

Initially, we looked at the excitation source and suggest that a source more closely resembling the female glottal pulse would produce a more suitably shaped spectrum. This could contribute to a female sounding voice, but is not the only difference between the two voices.

There is a general view that spectral tilt is a main feature associated with different voices. This is more easily dealt with by hardware changes than by entering values in tables. In particular, in human speech there is considerable variation in the frequencies and amplitudes of formants 4, 6 and 6, which are not accessible to the SbR program, yet which play an important role in determining the individual characteristics of different human voice qualities

including the female voice. The relative amplitudes of the formants can be adjusted in the hardware, if necessary, under software control.

The average female voice tract dimensions are smaller than those of the average male. Not only does this give rise to formant frequencies in vowel and vowel-like sounds which are generally higher for the female voice than for the male voice, but alters the spectrum of aperiodic sounds produced within the oral cavity.

A general conclusion might be that a convincing implementation of a female voice would involve designing into the hardware synthesizer a second periodic excitation source, possibly a second aperiodic excitation source; and in addition in the driving software certainly supplementary tables of specifications for all allophonic segments in respect of formant frequencies and amplitudes. Transition shapes should also be modified.

I - DURATIONS

A pilot study has been conducted on the way in which segment duration in various contexts is handled by the JSRU SbR system. We have detected various anomalies which may contribute to the unnatural quality of synthetic speech.

There are three ways of altering duration:

- the minimum and inherent durations found in the tables can be changed. These are the values which the duration rules work on;
- the duration rules as supplied by JSRU can be modified or new rules can be added;
- in exceptional cases where changing durations and application of rule does not produce the right effect , allophones could be added for specific contexts.

This study has formed the basis of the proposal for more detailed investigation of duration.

J - STOP CONSONANTS

Identification of stops is triggered by acoustic features of stops themselves and by information from formant transitions in the adjacent vowel(s). In fact it has been shown experimentally that vowel transition information alone can produce a high identification score.

In previous reports we have made suggestions for rendering voiced stops by paying careful attention to timing of onset of periodic excitation; this is not handled very well in the SbR tables. In general we have found that the model employed by JSRU as a basis for synthesis of stop consonants has several defects.

The JSRU stop consonant model depends on the general segmental model of speech production adopted in the early 60s which was adequate and forward thinking for the time. But models explaining speech production have developed during the last 10 years (Kent 1976, Nolan 1982). In fact, the notion that speech consists of strings of segments has led to several of the problems associated with synthesizing stops.

Unlike most other sounds stops do not maintain parametric specifications throughout their duration. The whole notion of segments and targets rests on the assumption that specifications are maintained. Stops are regarded as having three phases - the stop phase, during which airflow from the mouth ceases; the burst phase, during which the increased air pressure consequent on the stop phase is released; and the aspiration phase - a period of formant structured aperiodic sound following the burst.

The solution for JSRU was to treat each phase as though it were an independent segment. This met the need for segment specifications to be a-temporal, since in principle no values would change during each phase. Voiceless stops which characteristically have all three phases were rendered as [stop + burst + aspiration] represented as x+xY+xZ (where x stands for the particular stop, Y stands for burst and Z for aspiration), and voiced stops with just the two phases [stop + burst], since these were thought to typically omit the aspiration phase.

This solution avoids several problems, while at the same time giving rise to a few new ones. There are contexts in which voiceless stops occur in which one or more of the phases do not occur (see Ainsworth and Millar 1976). For example, acoustically it is difficult to identify a stop phase in sentence initial position; the burst and aspiration phases of the first of two juxtaposed stops may be omitted, the aspiration phase, and sometimes the burst phase, of a final stop are often omitted.

Any rule which is not context sensitive and rewrites /t/ as T+TY+TZ will fail in such contexts. The word *apt,* for example, is described in human speech as AA+P+T+TY, but in the synthetic version as AA+P+PT+PZ+T+TY+TZ. We have identified some of the anomalies in the system and supplied phonological contexts marking variants from the simple three-phase breakdown of stop segments. The above applies to voiceless stops, but there are some occasions for some speakers where there is a noticeable aspiration phase associated with the voiced stops as well. The aspiration phase is also clearly apparent following voiceless fricatives and affricates - a characteristic feature which should be replicated.