Prosodic Assignment in *SPRUCE* Text-to-Speech Synthesis

Mark Tatham
Eric Lewis

Reprinted from Tatham, M. and Lewis, E. (1992) Prosodic assignment in SPRUCE text-to-speech synthesis. In R. Lawrence (ed.) *Proceedings of the Institute of Acoustics* 14, 447-454. St. Albans: Institute of Acoustics

1. THE *SPRUCE* TEXT-TO-SPEECH SYSTEM

The SPRUCE (SPeech Response from UnConstrained English) text-to-speech system is currently under development at Bristol and Essex Universities. It has a number of innovative features which are designed to obtain the most natural sounding possible synthetic output. In its standard version the system generates parameter files suitable for driving the Loughborough Sound Images implementation of the Speech Research Unit parallel formant synthesiser designed by Holmes [Holmes, 1985]. These parameter files specify values for formant frequencies and amplitudes, excitation type and fundamental frequency, each at 10ms intervals.

The parallel formant synthesiser was chosen as the means of generating waveforms for the standard SPRUCE partly because it is widely available, but in particular because this type of parametric approach enables considerable flexibility. Other systems (see O’Shaughnessy, 1987, for a survey of techniques), for example those using representations of speech as linear predictive coding coefficients or some kind of stored waveform synthesis (Le Faucheur *et al.* 1991), do not lend themselves to easy manipulation, particularly in the frequency domain a facility which we have found important during SPRUCE development. The SPRUCE design is however flexible enough to be able to drive such synthesisers.

The overall architecture of SPRUCE has been described in more detail previously [Lewis and Tatham, 1991; Tatham and Lewis, 1992], and consists essentially of three stages. These are

1. Conversion of orthographic text to a sequence of phonological syllables as retrieved from a large word dictionary (the default), or by a separate conversion procedure for words not found in the dictionary.
2. Phonological and prosodic processing dependent on a syntactic parse that has been optimised for these particular tasks.
3. Retrieval and conjoining of representations of acoustic syllables to provide the sequence of frames of parameter values for driving the synthesiser.

Unlike most synthesis systems SPRUCE is primarily syllable based. That is, the unit basis for generating parameter files is the acoustic phonetic syllable, although the system will also generate files based on acoustic units smaller than syllables (allophones, diphones, triphones) or larger (words, phrases, sentences). For certain applications the units may be mixed within a particular synthesis task. Phonological syllables form the basis of higher level processing including prosodic assignment.

In standard mode the system depends on an inventory of around 10,000 normalised, parametrically analysed acoustic syllables segmented from recorded natural speech. The normalisation procedure mostly involves removal of intonational fundamental frequency information (f0), although micro-prosodic effects are preserved. Other normalisation procedures are designed to minimise discontinuities which might arise at syllable boundaries during the conjoining procedure. Generating a sentence sized parameter file consists of assembling the appropriate syllable units from the inventory, making durational adjustments in line with a calculated rhythmic contour and supplying a complete f0 contour for the entire sentence.

The basis for syllable selection is the matching of words in the input text with their phonological syllable structures as retrieved from a large (100,000 entries) word dictionary. The actual acoustic syllable selected later depends on this initial identification of the syllabic structure of particular words, but the outcome is also influenced by subsequent phonological processing. One of the advantages of using a word dictionary is that entries can have markers affixed which provide both syntactic and semantic information (or any other markers found to be necessary). These markers are used in parsing input sentences for subsequent prosodic processing.

The computational tasks involved in SPRUCE's conversion of text to speech consist, therefore, of

1. retrieving syllabic representations of the words from the dictionary to build a sentence of abstract syllable objects,
2. computing the abstract prosodic contour for the input sentence,
3. retrieving the correct acoustic syllable representations from the syllable inventory,
4. conjoining these syllable units,
5. computing the acoustic prosodic parameters,
6. constructing the file of parameter values for forwarding to the synthesiser.

The way these tasks interact is shown in Fig. l as a block diagram of the entire SPRUCE text-to-speech system. An additional set of procedures, not described here, enables pragmatic effects to modify the normal prosodic (rhythm and intonation) contours generated by the basic system [Morton 1992a, 1992b].

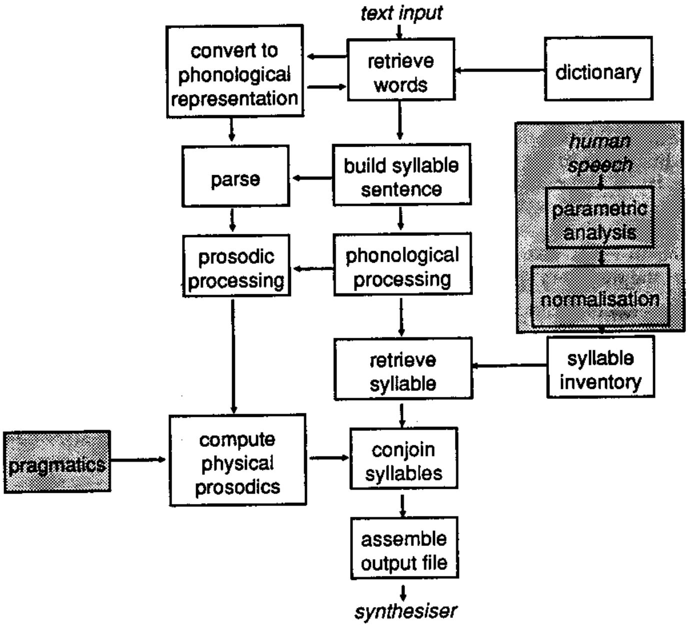

Fig. 1 SPRUCE text-to-speech synthesis.

2. *SPRUCE* PROSODICS PHILOSOPHY

The procedures in the SPRUCE text-to-speech system which handle prosodies are responsible for determining what will be perceived by listeners as speech rhythm and intonation in the synthesised output. As with all other aspects of SPRUCE we are concerned with providing acoustic information that is necessary and sufficient to be reproduced on demand and to cause listeners to believe they are hearing real human speech. Our aim is not necessarily to replicate the sound wave exactly as produced by a human being. The engineering of the system is nonetheless carefully principled and draws on as much consistent supporting theory as possible.

Our model makes the theoretically based assumption that rhythm and intonation in human speech are generated by a two-stage process. The first stage involves cognitive processing, and results in the assignment of an abstract prosodic contour to the sentence to be spoken. The second stage involves a realisation or interpretation of the abstract contour in terms of the physical mechanisms and processes available to a human being for generating a speech sound wave.

We have taken this two-tier approach to both prosodic and segmental processing not simply because linguists and psycholinguists do, but because there is a real advantage to separating the two levels in terms of the building the model. For example it is useful to have a well-defined level where we can describe what it is listeners feel they hear in the sort of terms they might use when questioned about the synthetic signal we are generating. Listeners are not going to tell us that formant two could use a rise in amplitude of 2dB or that the fundamental frequency might well be 3Hz higher on a particular syllable.

Linguists speak of descriptions at a cognitive level – the same idea – and it is because these descriptions are not numerically formulated that they are sometimes dismissed in the engineering environment of speech technology. For all its apparent vagueness, what a listener has to report *must* somehow be related to the acoustic signal; we believe that speech technology is largely about that relationship.

In our model of the human process of reading text out aloud we maintain that the task cannot be accomplished without some understanding of the text on the part of the reader. We can observe that even for the most practised of readers – a newscaster, for example – the task often fails if understanding has not taken place. Language understanding has not yet been satisfactorily modelled, but we have assumed that at least some evaluation of the grammatical patterning of the text being read occurs – that is, the text is syntactically parsed by the reader. This procedure is, we believe, supported by some meaning extraction process, using what we might call a semantic parse to detect logical relationships within and across sentence boundaries, many of which will be significant for correct prosodic assignment.

SPRUCE implements this model, keeping prosodic assignment and acoustic realisation separate. Prosodic assignment is treated as abstract, using a rule system to mark up the input sentences for the most plausible rhythm and intonation contours, using a formal means to express what listeners (or linguists) might have to say. The subsequent acoustic processes translate the marking into adjustments of a default rhythm, increasing and decreasing the acoustic durations of specific syllable sized elements within the sentence, and generate a fundamental frequency vector determined by the intonation marking.

In adapting the underlying theory for the development of SPRUCE synthesis as a system designed to operate in near real time, we felt it necessary to build algorithms as good as we could make them, but which for practical reasons permit errors to occur from time to time. The algorithms are however carefully designed. For example, an error occurring from time to time is the assignment of focal sentence stress on the wrong word: we have tried to make sure that the focal stress is assigned in such cases to a word which *could* bear it. But we have also made sure that we build in the means of response to additional information which, in such an example, *would* give correct assignment if that information were available; and we define what that information would be.

3. *SPRUCE* PROSODICS IMPLEMENTATION

Arguably one of the weakest areas of most text-to-speech systems is their inability to assign error free prosodic contours to the input text. In SPRUCE careful attention has been paid to this area, particularly in. respect of detection and repair of errors. Calculating the prosodies for SPRUCE is a two stage process.

a. Stage one – prosodic assignment

The first stage, which we call prosodic assignment, annotates a phonological syllable string (based on interaction between the input text and the word dictionary) with abstract prosodic markers. There are two types of marker, rhythmic and intonational, and both are assigned in procedures which follow and make use of a parsing process. As retrieved from the dictionary the phonological representation prior to the parse provides information about the syntactic categories of words, and for polysyllabic entries, an indication of their word-stress patterning. There is potential for ambiguity here, as might arise in orthographically identical adjectives and nouns which carry different stress patterns – for example, $[\text{con+tent}]_{\text{adjective}}$ and $[\text{con+tent}]_{\text{noun}}$. Such ambiguities are carried forward for resolution by the parser.

The SPRUCE parser delivers a syntactic parse of the syllable string as well as a semantic parse in which some logical relationships between words in sentences and between sentences are indicated. The parse algorithm has, been optimised to provide the minimum amount of information necessary for subsequent prosodic assignment; for this reason it is extremely fast. Each parse operation is permitted only a single solution, and a parse failure results in a statistically based choice between default solutions. Thus every sentence is parsed quickly, and the system has been tuned iteratively by experiment to fail with minimal damage when it does fail. The error level was determined by researching the repair potential of the rule system for assigning prosodies, and, importantly, by experiments evaluating listener judgement of errors. In its final version the system will automatically flag potential errors, attempt a repair, and if necessary fail to an acceptable though not necessarily optimally correct assignment.

As the parse is taking place words (or the stressed syllables in the case of polysyllabic words) are marked with a notation based on that proposed by Pierrehumbert [1981]. The notation reflects the abstract intonation contour. At the same time rhythm is marked, with adjustments made with reference to the intonation assignment. Rhythm and intonation assignment are processes which are parallel and complementary.

The output of stage one for generating prosodic contours for input text is a series of abstract phonological strings which have been annotated with rhythm and intonation markers. At this stage there is as yet no indication of the final acoustic parameters of timing and fundamental frequency. These derive from implementations of the abstract markers in the next stage of the prosodic process.

b. Stage two – acoustic implementation

The second stage of prosodics in SPRUCE, which we call acoustic implementation, derives the final timings of syllables (and elements within a syllable) as well as actual f0 values for insertion into the final output file.

As we move to stage two in generating the prosodies the inventory containing the normalised acoustic syllables (in the standard system – allophones, diphones, words, etc. in the derivative systems) is now accessed, and units matching the syllables of the phonological representation are retrieved in sequence. Syllable conjoining rules are applied to smooth the boundaries between segment units. The representation is now in the form of a string of parametrically coded and properly conjoined syllable sized units expressed in 10ms frames, each of which contains values for all parameters required by the subsequent external process of generating the output waveform.

At this interpretive or implementational level the abstract markers representing rhythm are converted to the acoustic parameters of syllable duration and overall sentence timing. The process uses specially developed algorithms for dealing with rhythm changes during the course of a sentence or between sentences. For example, whole-word prominence can be achieved by a combination of pause insertion to disrupt the rhythmic flow and pitch peaking on the main stress within the word; rallentando effects following pitch peaking are included. Some effects are difficult to implement; for example, embedded phrases, which for some pragmatically determined reason have a rhythm and intonation quite distinct from that of the main sentence (see Morton 1992c). These generally require information not available from the input text to be properly implemented.

The interpretation of abstract intonation markers is based on a concept of tunnels (see Fig.2). Changes of fundamental frequency are normally held within these tunnels as clauses and sentences progress. The floor and ceiling of a tunnel are carefully chosen to reflect the normal speech of the person used for recording the database from which the units in the syllable inventory have been derived. The system selects the tunnel appropriate to a particular sentence from an inventory of default tunnels on the basis of sentence type as determined during the earlier parse. The tunnel model has been chosen because it permits *relative* intonation interpretation: a few simple parameters can be altered to enable SPRUCE to synthesise different voices (in conjunction with alternative acoustic syllable inventories).

The micro-prosodic effects which were carefully preserved in the syllable inventory are combined with the fundamental frequency contour as determined by the abstract intonational contour. The inclusion of these micro-prosodic effects (sometimes called micro-intonation) has a marked effect on the perceived naturalness of the final synthesised speech.

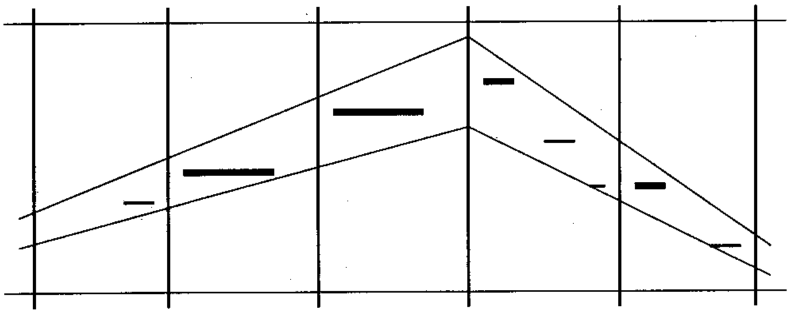

Fig.2 Pitch and rhythm markers fitted within a default ‘tunnel’. Vertical lines are rhythm markers, horizontal lines are pitch markers.

The prosodic procedures within the standard SPRUCE system result in a speech output which can be described as ‘neutral sounding’. We recognise that in fact human beings do not usually produce such speech: truly natural speech always has overlays of pragmatic origin reflecting the speaker's attitudes and beliefs, and so on. Hooks are inserted at this point in the prosodic process to enable these emotive overlays to be added by a separate post-processing procedure. Although not yet fully incorporated into the standard SPRUCE, the principles of reprocessing timing and fundamental frequency are described in Morton [1992b, 1992c].

4. CONCLUSION

The strategies we have described here for generating prosodies within the SPRUCE text-to-speech system have resulted in a significant improvement in the naturalness of the synthetic output.

Of course, like all speech synthesis systems, SPRUCE still suffers from the fact that it doesn’t know what it's talking about – that is, it does not incorporate full understanding of the input text – and does not have feelings about the content of what is being said. For this reason the system has therefore to work in terms of neutral sounding prosodic contours which will consequently be in error from time to time. The system is however designed to incorporate pragmatic markers inserted in the text; these can trigger post-processing leading to a further improvement in naturalness. Especially in dialogue applications this aspect of naturalness is likely to become increasingly important, though until more understanding can be incorporated into synthesis systems errors will continue to occur. Dialogue systems however will eventually dispense with text altogether, outputting some other representation for conversion to speech. Such a representation could easily incorporate exactly the kind of information necessary to take synthesis systems such as SPRUCE significantly further in the direction of naturalness. We have made provision for an input of this type, though at present there are no dialogue systems supplying it.

In the SPRUCE text-to-speech system we are taking care to marginalise the errors that are inevitable, given a plain text input, such that a listener might be aware more of inappropriateness rather than outright error.

There is a rapidly increasing demand for high quality speech synthesis. We believe that for the moment, despite significant advances in signal processing techniques and the available component technology, there is still room for engineering a system based on a sound theoretical foundation.

REFERENCES

Holmes, J.N. (1985) A parallel-formant synthesizer for machine voice output. In *Computer Speech Processing* (eds. F. Fallside and W.A. Woods), pp. 163-187. London: Prentice Hall International

Le Faucheur, L, Boeffard, 0., Cherbonnel, B. and White, S. (1991) Un algorithme de synthèse de parole de haute qua1ité. *Séeminaire SFA/GCP*, Le Mans, pp. 104-107

Lewis, E. and Tatham, M.A.A. (1991) SPRUCE – a new text-to-speech synthesis system. *Proc. 2nd European Conference on Speech Communication and Technology*, Vol. 3, pp. 1235-1238

Morton, K. (1992a) Pragmatic phonetics. In *Advances in Speech, Hearing and Language Processing* (ed. W.A. Ainsworth), Vol. 2, pp. 17-55. London: JAI Press

Morton, K. (1992b) Adding emotion to synthetic speech dialogue systems. *Proc. ICSLP* '92, Banff, Canada

Morton, K. (1992c) PALM: psychoacoustic language modelling. *Proc. Institute of Acoustics*, this volume

O'Shaughnessy, D. (1987) *Speech Communication: Human and Machine.* Reading, Mass.: Addison-Wesley

Pierrehumbert, J. (1981) Synthesizing intonation. J. *Acoustical Society of America,* Vol. 70, pp. 985-995

Tatham, M.A.A. and Lewis, E. (1992) Prosodics in a syllable-based text-to-speech synthesis system. *Proc. ICSLP* '92, Banff, Canada