Proceedings of The Institute of Acoustics

THE SYNTHESIS OF RHYTHMIC STRUCTURE

STEPHEN ISARD

UNIVERSITY OF SUSSEX

We are still some way from a satisfactory grasp of the principles which govern segmental duration in English speech. We are even further from a knowledge of the ways in which the perceptual system understands and exploits these principles to guide the listener in his task. One way of attacking these problems is to synthesize speech according to different sets of proposed principles and to note the changes in perceptual effects. With these considerations in mind, a speech-synthesis-by-rule system has been set up in which it is relatively easy to incorporate a variety of duration algorithms, and this note reports on some preliminary trials with it.

There have been two main approaches to the problem of assigning durations to phonemes in synthetic speech. One, exemplified by Klatt(3), is to assign "standard" lengths to phonemes, and to modify these for a given utterance, one phoneme at a time, on the basis of the local context in which the phoneme finds itself. Local context for Klatt consists of immediately neighbouring phonemes, position within a syllable, level of stress on the syllable, position of the syllable within a word, and pauses and syntactic boundaries immediately following the word. In particular, a phoneme's length can be affected by the fact that it appears in a polysyllabic word, but not by which specific phonemes the other syllables contain.

This system can be contrasted with that of Witten(6), modified somewhat in Witten and Smith(7), which is based on the isochrony principle, the supposed tendency for stressed syllables in English speech to come at approximately equal intervals. (See Fowler(2) for a review of evidence for and against this principle.) In Witten's algorithms, interstress intervals, known as feet, are assigned equal target durations. These target durations are then shared out "from above", as it were, among the syllables of the foot, according to criteria involving stress, word boundaries, vowel length and the presence of final consonant clusters. Finally, an attempt is made to assign phoneme durations within syllables to achieve these targets. Consideration is given to phoneme type and position within the syllable. The syllable targets are not, in general, met with perfect accuracy, because "standard" durations are consulted, and phonemes prevented from straying from them by too great a proportion. Speech produced in this way thus "tends toward" isochrony, without necessarily achieving it.

It is in fact the case that for a great many sentences, the Klatt procedure also produces a rhythmic effect, because of rules which lengthen stressed syllables and shorten phonemes in polysyllabic words. However, this effect appears as a sort of accidental byproduct of the rules as they are stated, and we can construct cases where the two systems come into conflict. For instance, the Klatt rules will make the final syllable "board" of (1) shorter than that of (2) because it belongs to a polysyllabic word in the first case but not in the second.

(1) We took her aboard.
(2) We took her a board.

The Witten rules, on the other hand, will group the "a" syllable with the previous foot in both cases. In principle, they could take account of the fact that "a" is a word in one case but not the other in sharing out the target duration of this foot, but the influence would then be on the duration of "took her", rather than "board". As the rules stand, the two sentences are pronounced identically.

Thompson(5) looked for a tendency toward isochrony in a sample of natural speech, and concluded that his data were better explained by simply assigning the same basic target duration to all syllables, and then letting the properties "foot initial" and "pre-pausal" each add a fixed amount to this target, thus giving four possible target durations, depending on the presence or absence of the two properties. A variant of the Witten algorithm has been programmed in which sharing out of syllable time among constituent phonemes proceeds as before, but in which syllables receive their target durations according to Thompson's formula, rather than on the basis of what other syllables belong to the same foot. Syllable targets are not necessarily achieved here, any more than in the other version of the algorithm, but this seems to be in the spirit of Thompson's conclusions, since he only claims to account for about forty percent of the variance in phoneme durations and suggests that the bulk of the remainder is due to phoneme type, which is in fact the consideration that prevents syllables from always reaching their targets.

In soliciting judgements about the speech produced by these various duration algorithms, I have avoided general questions about naturalness and intelligibility, which are bound to be confounded by interactions with phonetic quality. The phonemes for the system are produced by a diphone synthesizer, and are reasonably good, but one does not take them for the output of a human vocal tract. I have therefore looked for effects of timing that can initially be isolated from questions of overall comprehensibility and naturalness, although in the long run it is the way in which they will contribute to these that is of interest.

A phenomenon that arises when speech is synthesized using fixed standard phoneme durations is that speech rate appears to change in the course of any reasonably long utterance. This is apparently due to stressed syllables not being given a sufficient advantage over unstressed ones, creating a rushed effect on the shortest stressed syllables and a drawn out effect on the longest of the others. Scaling the durations down to normal speaking rates generally produces an effect of more rushing and less drawing out. All three of the algorithms just described banish this effect to a great extent, but the Thompson formula does not seem to give sufficient extra length at the ends of sentences, which come out sounding hurried. This is possibly due to the fact that Thompson has lumped together a variety of different kinds of pauses in constructing his "pre-pausal" category, some of which do not cause as much lengthening as others. When he introduces a "pre-tone-group-boundary" category of syllable, his analysis gives it an extra thirty-four milliseconds over other prepausal syllables. This is still not exactly the same thing as sentence final, but makes some improvement, to my ear, in the synthesis. However, Thompson ultimately rejects the new

category on grounds of simplicity, since it does not account for much extra variance in his data. The isochronous algorithm also fails to lengthen the ends of sentences, and rushes them somewhat as a result, but Witten and Smith raise the possibility of assigning greater duration to certain feet, and when the target duration of the final foot is doubled, the rushed effect goes away.

Note that this tactic of slowing down an entire foot at the end of a sentence distributes extra length more widely than does the Klatt procedure, or the Thompson one with the simple addition of extra time for sentence-final syllables. Some evidence exists for the occurrence of such distribution of time over larger units of speech, as opposed to just lengthening of single syllables or insertion of pauses. Cutler and Isard(1) found that subjects asked to disambiguate a sentence such as

(3) I'm allergic to ripe squashes, melons and cucumbers.

to give the reading

(4) I'm allergic to (ripe squashes), melons and cucumbers.

would slow down the two units "melons" and "cucumbers" to a length approximately matching that of "ripe squashes". When they were supposed to produce "ripe (squashes, melons and cucumbers)", the three vegetables tended to get similar length. Scott(4) found that extra length spread throughout the foot "Pat and An-" in

(5) Pat and Antonia or Dave will take charge.

disambiguated in favour of the reading "Pat and (Antonia or Dave)" as effectively as an equal length pause after "Pat". Constructing examples such as

(6) The climate of the region is acceptable enough to the people inhabiting it.

and adding extra length either throughout the final foot or to the final syllable alone, we get a slight tendency toward the impression of speeding up on the final stressed syllable "hab-" in the case where only the final word "it" is lengthened.

Another difference, initially unexpected, is that first-time listeners judge that there is "more separation between words" in the Witten and Thompson systems than the Klatt, although this is not associated with any particular preference for their output. This may be an effect of the more regular syllable lengths in the former systems, though why it should have this effect is unclear. It might also be a phonetic effect attributable to the way that phoneme durations are distributed within syllables. An analysis of this possibility is underway.

References

1. A. CUTLER and S.D. ISARD 1980 in B. Butterworth(ed.) Language Production, vol 1. Cambridge: Cambridge University Press. The production of prosody.

2. C.A. FOWLER 1977 Indiana University Linguistics Club, Bloomington.
Timing control in speech production.

3. D.H. KLATT 1979 Proc. 9th International Congress of Phonetic Sciences, Copenhagen.
Synthesis by rule of segmental durations in English sentences.

4. D.R. SCOTT 1982 J. Acoust. Soc. Am. 71(4), 996-1007.
Duration as a cue to the perception of a phrase boundary.

5. H.S. THOMPSON 1980 Palo Alto: Xerox Palo Alto Research Center.
Stress and Salience in English: Theory and Practice.

6. I.H. WITTEN 1977 Language and Speech 20, 240-260.
A flexible scheme for assigning timing and pitch to synthetic speech.

7. I.H. WITTEN and A. SMITH 1977 Edinburgh University Department of Linguistics Work in Progress 10, 33-44.
Synthesizing English rhythm: A structured approach.