



CONVERSATIONAL PHONOLOGY IN A TEXT-TO-SPEECH SYSTEM

Linda Shockey*

The text-to-speech systems which are currently available commercially are very simple from the point of view of linguistic theory, since they incorporate little of what is known about phonology, morphology, syntax, and discourse analysis. We predict that future systems will have to be much more linguistically sophisticated. Here we will look at the possibility of building a system which is capable of utilising the most common rules of English conversational phonology.

Changes from citation or dictionary-entry pronunciation are common in relaxed speech. Some words (such as "and") are virtually never pronounced in their full form, others show a range of forms, most of which can be described as "reduced" from the full form. That is, the dictionary pronunciation typically is rich in information, and conversational speech processes work to reduce the phonological information in predictable ways.

Several factors add to the probability that a reduction will occur:

1. Lack of stress. Reductions are far more likely to occur in an unstressed syllable than in a stressed one. This is especially true of reductions involving the reduced vowel [ə].
2. Syllable-final position (consonants only). Note that the syllables mentioned here are conversational speech syllables, which can differ from those in citation forms. It is very common, for example, for a syllable-final consonant to move to syllable-initial position in running speech if the following word begins with a vowel. Consider, for example, the sequence "at all" in a sentence like "She's not like that at all." The /t/ in "at" is, for most speakers, noticeably aspirated here, indicating that the syllable division is 'a.tall'.
3. Membership in a consonant cluster. Again, it is best to think of consonant clusters as they relate to the entire phonetic string being produced rather than as instantiated in a single word. The sequence "st" as in "most" will be pronounced in full in the majority of cases when the following element is a vowel. The /t/ in "most of all" will normally be realised, but the /t/ in "in most cases" will not. Obviously, this criterion is interrelated to 2, above, since one might reasonably argue that the "st" in "most of all" has been shifted to syllable-initial position, i.e. 'mo.sto.fall'.
4. Position in the sentence. It has been observed that a very common pattern in English is for the topic to occur near the end of the sentence, where it normally receives primary sentence accent and is assigned the peak of the intonation contour. Information preceding the topic often repeats what is already known in the discourse or provides a grammatical framework for the topic to be expressed.

*Centre for Speech Technology Research, U. of Edinburgh

Segments which precede the topic tend to be shorter than those following it, hence the phenomenon known as sentence-final lengthening. The combination of these two features, i.e. relatively low semantic content and relatively rapid rate of production, bias the beginning of such sentences towards reduction. The portion of the sentence following the topic may show reduction for other reasons, but it will be less marked here than in the pre-topical portion. Sentences where the topic occurs elsewhere show different reduction patterns, but these sentences are in the minority.

5. Semantic content. If an item has been mentioned before in the discourse and especially if it has been used several times, it is more likely to be reduced than a newly-introduced lexical item. For example, in a discussion about welfare benefits, the term "social security" was observed to eventually reduce to [,sos:'kjʒti].

Note that speech rate has been mentioned only indirectly in the above five points. It is a commonly-held belief that conversational speech = fast speech = reduced speech. Behind this simple equation lies the idea that as speech rate increases, it becomes more and more difficult to articulate all the intended segments, so some are left out or incompletely executed, resulting in typical conversational reductions. While a loose connection between high speech rate and phonological reduction can be shown, such a deterministic view is difficult to support. Experimental results (ref 1) suggest that even at very fast rates speakers are capable of producing target segments which they normally reduce at the same rate. The subconscious decision to produce speech containing reduced forms is probably based more on conventions within a language than on the inadequacies of the vocal tract. The fact that characteristic reductions vary considerably from language to language supports this view, since the same mechanical principles hold for all speakers.

Looking at points 1-5 above, we see that most can be implemented in a text-to-speech system which is based on linguistic principles. Determining which syllables are unstressed and which contain [ə] is not difficult, nor is finding the phonetic quality of the segments surrounding the [ə]. Syllabification routines and routines for detecting consonant clusters can be written. Finding the topic word in a sentence is a challenge but one which must be met in any case to assign sentence accent and arrive at an appropriate intonation contour.

Implementation of point 5 is very much more difficult, since it involves something closer to a real understanding of the text. It may be that semantic content will not be a usable concept in text-to-speech systems for some time, but eventually redundancy may be able to play a part in their phonological algorithms.

Using the other four information sources with rate appearing as an added factor, it would be possible to construct rules for when reductions are likely to occur. One would need in the first instance a notion of degree of formality of the speech. Three levels, formal, normal, and informal, would suffice for most purposes. A particular configuration could then be assigned a 'reducibility rating' depending on the number of reduction-inducing factors present. The reduction in question would then either occur or not, depending on the threshold

set by the degree of formality.

Reference

1. Channon, R. and Shockey, L. (eds), For Ilse Lehiste, (Foris, 1987), pp. 217-224