6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

WORD AND SYLLABLE CONCATENATION IN TEXT-TO-SPEECH SYNTHESIS

Eric Lewis[1] *and Mark Tatham*[2]

[1] Department of Computer Science, Merchant Venturers Building, Woodland Road,
University of Bristol, Bristol, BS8 1UB, UK. email: Eric.Lewis@bristol.ac.uk

[2] Department of Language and Linguistics, University of Essex, Wivenhoe Park,
Colchester, CO4 3SQ, UK. email: Mark.Tatham@essex.ac.uk

10.21437/Eurospeech.1999-158

ABSTRACT

MeteoSPRUCE is a database of 2000 words relating to weather forecasting. While such a database is clearly not large enough to be definitive, its usability can be greatly extended by excising syllables from polysyllabic words in its inventory and recombining them to form new words [1], [2], [3]. The authors believe that it provides sufficient data to start to enable conclusions to be drawn as to how syllables should be modified for concatenation in contexts other than those in which they were recorded. A classification scheme for syllables, based on the class of their initial and final segments, has been defined and used to determine a set of rules for making modifications to syllables so that when concatenated the joins are perceptually not noticeable.

1. INTRODUCTION

In recent years concatenated waveform speech synthesis systems have gained in popularity due to their more natural sounding output. In order to be truly general purpose such systems must have an exhaustive inventory of stored waveforms for rearrangement and concatenation as needed. The size of the waveform units for concatenation differs between systems but the authors have argued [4] that for natural sounding speech the syllable is probably the preferred unit. **MeteoSPRUCE** is a limited domain syllable and word based system which has an inventory consisting of recordings of 2000 monosyllabic and polysyllabic words. Words for which recordings do not exist in the inventory are constructed by extracting syllables from words which **are** in the inventory and recombining them as appropriate. In this paper we describe how such syllables have to be modified for concatenation in contexts other than those from which they were excised.

2. SYLLABLE IDENTIFICATION

In a previous paper [5] the authors showed how to characterise syllables on three representational levels, viz.

- phonological - a phonological syllable is a representation of listener perception. It is therefore abstract and cognitively based, including only the characteristics necessary for perception.
- phonetic - a phonetic syllable is a stretch of actual waveform, including all coarticulations, which triggers a listener's perception of a phonological syllable.
- synthetic - a synthetic syllable is a normalised syllable model based on a phonetic syllable and is used in forming new words.

Although this suggests that the syllable boundaries in the lexicon should be marked phonologically we have decided that wherever possible morphemic boundaries should be used, provided the separated morpheme is syllabic. Therefore *windy* is expressed as *wind-ey* rather than *win-dey* but the word *winds* is expressed as *windz* rather than *wind-z.* If there is a syllable boundary which is not a morpheme boundary, then segmentation occurs on the basis of the phonology. For example *afternoon* is *arf-ta-nuun* - the first division is phonological while the second is morphemic. The reason for marking up the syllables in this way is that since we will be using the syllables for making new words it is more likely that the new words will be built-up on a morphemic basis rather than on a phonological basis.

In SPRUCE a syllable waveform in the inventory can exist as a phonetic syllable, for example the recording of the monosyllabic word *rain*. However, in order to produce the word *raining* the phonetic syllable is transformed by a normalisation procedure into a synthetic syllable, one which the speaker may not normally produce but which triggers the same perceptual response in the listener as the phonetic syllable.

3. SYLLABLES TYPES AND CONTEXTS

All syllables can be written in the form

$$C_{0,3} + V + C_{0,4}$$

where $C_{0,n}$ indicates 0 to n consonants and V indicates a vowel. Although not all combinations of consonants are allowed the possible number of syllables is still large. However, by classifying syllables in terms of their initial and final segments it becomes necessary to consider the concatenation of syllables only in terms of their class, and the number of these is fairly small.

The relevant classes are

- vowels
- diphthongs
- liquids
- nasals
- voiced fricatives
- voiceless fricatives
- voiced plosives
- voiceless plosives

A syllable inventory is created consisting of all syllables in the lexicon where each entry also includes information about the syllable stress, the classes of the initial and final phonemes as well as the classes of the phonemes which immediately precede and follow the syllable. Since the same syllable can occur in different contexts in the lexicon the syllable inventory contains multiple entries for each syllable.

4. RE-COMBINING RULES

Ideally, a syllable required for concatenation is extracted from the inventory such that its context in the sentence under construction is the same as that in which it was recorded. With a sufficiently large inventory of recordings it is highly likely that such a situation will be the norm rather than the exception. In the current situation the inventory of recordings consists of nearly 2000 mono and polysyllabic words relating to weather forecasting so the need for re-combining rules is very much higher. In either case a strategy is required for recombining syllables in contexts other than those from which they were extracted.

There are several quite different types of concatenation which can be categorised as follows:

- the syllable for insertion is a monosyllable.
- the syllable for insertion has the correct context and comes from a polysyllabic word.
- the syllable for insertion has the wrong context and comes from a polysyllabic word.

In each of the above cases it is also possible that the stress of the inserted syllable may not be the required stress. And, of course, it is possible that in many cases one half of the syllable context will be correct but the other half wrong. A monosyllable almost always has the wrong context in as much as it has been recorded within a standard frame of a sentence which always has the same words preceding and following it, and which is specially designed to minimise coarticulatory and rhythm effects.

A sentence is built up linearly from the start by successively concatenating words and/or syllables. Therefore, the concatenation phase consists essentially of having rules for joining two syllables depending on the final class of the first syllable and the initial class of the second syllable. We shall henceforth refer to these two syllables as syllable-F and syllable-I respectively. The position of the syllable within the word is not important except when it's at the start or end of a phrase.

We have, therefore, the situation of concatenating two syllables where each syllable is one of the following:

a monosyllable - a syllable which is also a word

an onset-syllable - that is, one extracted from the start of a polysyllabic word

a medial-syllable - that is, one extracted from within a polysyllabic word

a coda-syllable - that is, one extracted from the end of a polysyllabic word.

The problem in replacing a syllable by one from a different context is that

- the coarticulation of the adjacent segments can be wrong. For example, the plosive /b/ at the beginning of the monosyllable *broke* has its normally devoiced stop period replaced by a voiced stop in the polysyllable word *unbroken,* see Fig. 1.
- the timing of the syllable can be wrong. All the stressed monosyllabic words are recorded to occupy one rhythm unit, or *foot*, [3] so when

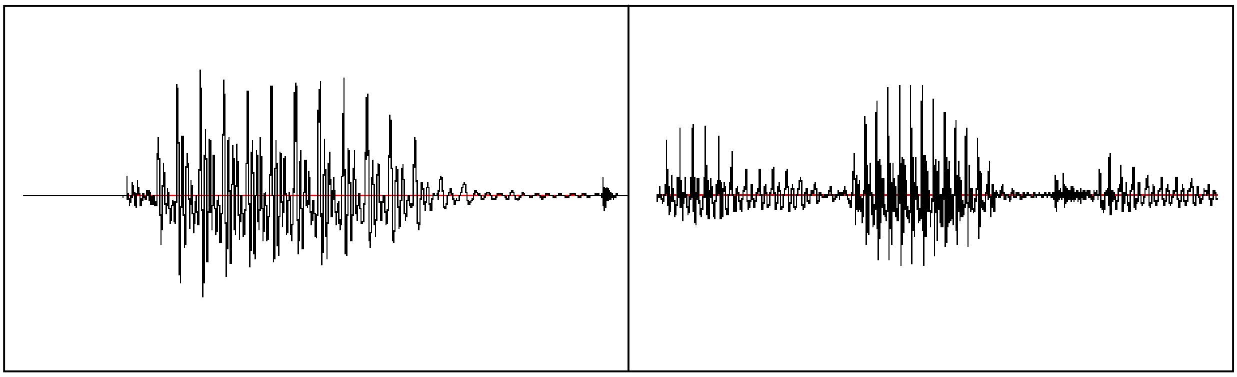

Fig. 1. Waveforms of *broke* and *unbroken.*

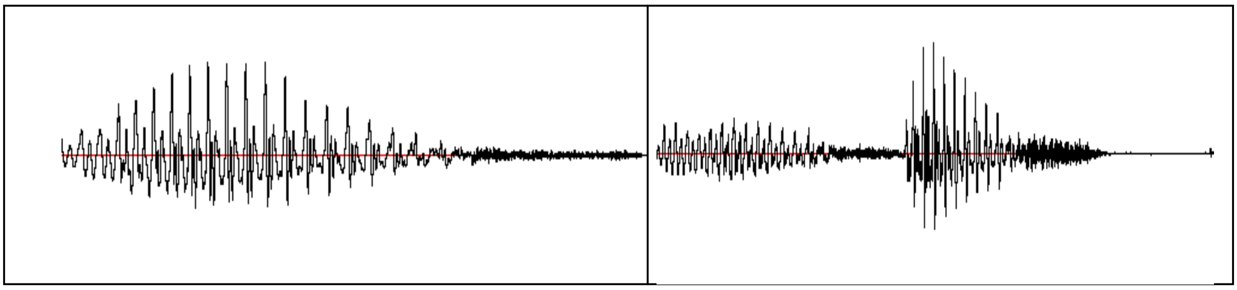

Fig 2. Waveforms for *north* and *northeast*.

- inserted in a polysyllabic word room must be made for the accompanying unstressed syllables. In the above example the length of *broke* is 34ms compared with 53ms for the length of *unbroken.*
- the amplitude of the syllable can be wrong. This would be the case when trying to use a stressed syllable as a substitute for its unstressed form, or vice versa. Compare the two waveforms in Fig. 2 for *north.*

The **MeteoSPRUCE** database contains very few examples of syllables ending or beginning with diphthongs so syllables belonging to this class are not considered in any of the following rules.

3-period rule

One rule, which is applied frequently to syllables whose onsets or codas are periodic, is that of cutting three periods from the start or end of such syllables. It can apply to monosyllables at either end, as well as to onset-syllables at the start and to coda-syllables at the end. It is **not** applied if the length of the vowel is less than 12 periods. This rule will henceforth be referred to as the 3-period rule.

Using the classification defined in section 3 rules for word and syllable concatenation are given in Table 1. In addition to these rules it is also necessary to consider the situation of being forced to use a stressed syllable in an unstressed position and vice versa. In the former case the authors believe that the best strategy is to reduce the length of the vowel and the overall amplitude of the syllable. Experimentation is still taking place as to how this may best be implemented.

It is also important to remember that these rules have been derived for a particular speaker's voice and that the specified quantification may not be appropriate for other voices and recording rates.

5. CONCLUSIONS

Rules for word and syllable concatenation have been derived using a 2000 word database, using a classification system based on the classes of initial and final syllable segments. The size of this database necessarily limits the possible combinations of syllables that may occur and which may be used for deriving and testing the appropriate concatenation procedures. However, the authors have developed a strategy that they believe can be extended to process much larger databases and provide the basis for an unrestricted text-to-speech synthesis system.

Syllable-F	Syllable-I	Rule
Vowel	Vowel	Apply the 3-period rule to syllable-F and syllable-I.
	Nasal Liquid Fricative Plosive	Apply the 3-period rule to syllable-I.
Nasal	Vowel	Apply the 3-period rule to syllable-I.
	Nasal Liquid Fricative Voiceless Plosive	Do nothing
	Voiced Plosive	If syllable-I is a monosyllable or onset-syllable cut that part of the signal prior to the release of the plosive. Syllables whose initial context is classified as voiced plosive cannot be used in any other context because there is no release for the plosive.
Liquid	Vowel Liquid	Apply 3-period rule to syllable-I.
	Voiced fricative Voiced plosive	Remove the silence before the release at the start of syllable-I
	Voiceless fricative Voiceless plosive Nasal	Do nothing
Fricative	Vowel	If syllable-F is a monosyllable or coda-syllable ending with segment /v/ then detect the start and end of the voicing for that segment. If these positions are A and B, respectively, then delete the remainder of the signal after B and double the length of AB. Apply the 3-period rule to syllable-I If syllable-F is a monosyllable or coda-syllable ending with segments /s/, /sh/, /z/ or /zh/ then apply the 3-period rule to syllable-I.. For syllable-F ending in segments/f/, /th/ or /dh/ there is insufficient data in the inventory to draw any conclusions.
	Fricative	Not enough examples to make any strong recommendation but a promising interim rule is to trim both fricatives are trimmed back by 25%.
	Nasal Liquid Plosive	Do nothing
Plosive	Vowel Liquid Fricative Voiced plosive	If syllable-F is a monosyllable or coda-syllable, then determine the length of the signal between the last marked period and the release of the plosive. Preserve 50ms of this signal following the marked period and delete the remainder. Syllable-I's which have been extracted with a pre-context of a voiced plosive must not be used in any other context.
	Nasal	Do nothing
	Voiceless plosive	Remove the release of the plosive of syllable-F and preserve 100ms of the signal after the last marked period.

Table 1. Rules for concatenating syllables.

6. REFERENCES

[1] Boeffard, O. Miclet, I. and White, S. (1992) Automatic generation of optimized unit dictionaries for text-to-speech synthesis. *Proceedings of the International Conference on Spoken Language Processing, Banff, pp 1211-1214.*

[2] Campbell, N. and Black, A. (1995) Prosody and the selection of source units for concatenative synthesis. In, J. van Santen, R. Sproat, J. Olive and J. Hirshberg (eds) *Progress in Speech Synthesis, Springer Verlag, New York.*

[3] Hunt, A.J. and Black, A. (1996) Unit selection in a concatenative speech synthesis system using a large speech database. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Atlanta.*

[4] Lewis, E. and Tatham, M. (1991) SPRUCE - a new text-to-speech synthesis system. *Proceedings of Eurospeech '91. ESCA Genova.*

[5] Tatham, M. and Lewis, E. (1999) Syllable reconstruction in concatenated waveform speech synthesis. In J. Ohala [ed.] *Proceedings of the International Congress of Phonetic Sciences*, San Francisco