ADDING NATURAL PROSODY TO A PHONEME SYNTHESISER

GEORG E. OTTESEN

Acoustics Research Center, ELAB
N-7034 Trondheim-NTH, NORWAY

ABSTRACT

This work investigates the possibility of increasing the quality of synthetic speech by adding some timing information. The rhythm of the syllables is tapped onto the keyboard of a computer. The vowel and consonant durations are modified by rules to fit the rhythmic pattern given. The result is compared to speech synthesis based on standard phoneme lengths and to synthesis with phoneme lengths aligned with actual speech. Listening tests are performemd on Norwegian sentences synthesised in whisper.

INTRODUCTION

Text-to-speech synthesisers and phoneme synthesisers are not used for public services in Norway. They are judged to sound very unnatural. The main shortcoming is the lack of natural prosody. Norwegian has a complex prosodic system. There are two distinct word tonemes, the timing pattern is mainly stress based, and the overall sentence prosody is strongly dependant on the syntactic and semantic structure. The natural prosody can not be derived automatically from text without advanced methods of sentence analysis.

In many applications the text can be prepared by adding some prosodic information. This work investigates a method for adding the timing information to sentences by tapping the rhythm of the sentences onto the keyboard of a computer. The sentences are synthesised in whisper, which isolates the timing information from the pitch information. The synthesis of pitch contours will, therefore, not be discussed in this paper.

METHOD OF TIMING

Experiments on the rhythm of German speech indicate a close relationship between the perceived rhythm of speech and the point of onset of the vowel in each syllable [1]. This fact is used to add a natural timing to a speech synthesiser by tapping the rhythm of the syllables onto the keyboard of the computer. Simple rules are given to align the synthetic speech with the keystroke sequence.

The synthesiser is a 4-formant phonetic synthesiser with Norwegian phonemes. Each phoneme corresponds to 1, 2 or 3 phonetic elements in the synthesiser. Vowels consist of one element, diphthongs of two, and unvoiced plosives of three elements. The durational resolution of each element is 10 ms. Each phonetic element has a default duration which is the mean duration of that element in actual speech.

The rhythm of the syllables is tapped on the keyboard of a personal computer by using two fingers to operate two keys; this method gives a better timing than tapping with one finger only. The sentence is spoken by the person simultanously with the tapping. Each tap corresponds to the starting point of the vowel in the syllable. The vowel or diphtong and the following consonants make up the time interval between two keystrokes.

Initial experiments show that a linear scaling of the elements of each syllable is not acceptable. The burst of the unvoiced plosives /p,t,k/ cannot be stretched significantly without losing the impression of a burst. A major increase in the duration of the short vowels creates confusion between short and long vowels, as in the words /kane/ and /ka:ne/. Other elements can be prolonged by a factor of 3 between slow and fast speaking rate. These observations lead to the introduction of a stretching factor for each phonetic element. This factor is called the time warping sensitivity, s_i, of the phonetic element, i. Each phonetic element is given

an additional length, ΔT_i, which is proportional to the default duration, T_i, to the time warping sensitivity, s_i, and to the difference of actual syllable length, T, and the sum of the default lengths:

$$\Delta T_i = \frac{s_i\, T_i\, (T - \Sigma T_i)}{\Sigma s_i T_i}$$

Figure 1 gives an illustration of this alignment.

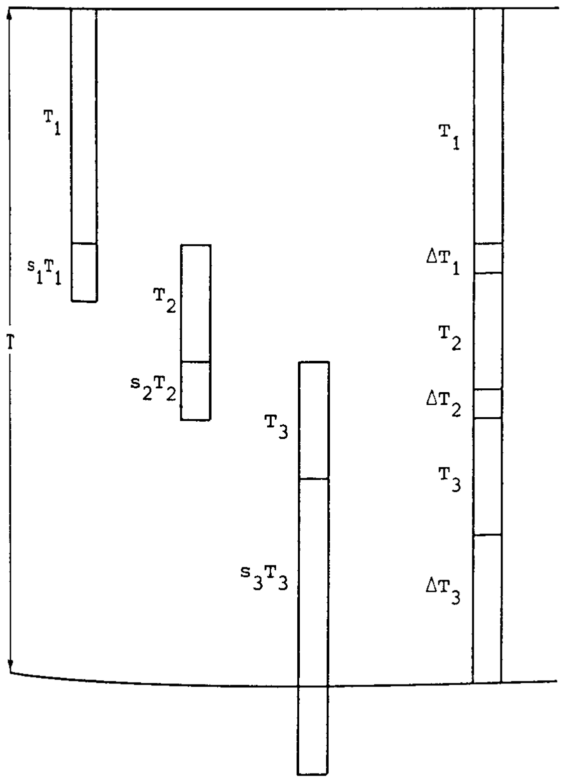

Figure 1. Alignment of phoneme durations.
T - time gap between vowel onsets
T_i - phoneme default duration
s_i - time warping sensitivity
ΔT_i - additional duration

The numerical range of the warping sensitivities is arbitrary; a constant factor does not change the timing. The range is chosen to be [0, 100]. A warping sensitivity of zero means that the element can not be stretched, one of 100 means that the element is maximally stretched. Phonemes with similar properties are ordered in groups with the same sensitivity. Long vowels have maximum time warping sensitivity. Nasals and the closure phase of unvoiced plosives have medium sensitivity, short vowels have low sensitivity and the burst of the unvoiced plosives have minimum sensitivity. The values are adjusted to give an acceptable pronounciation of single words spoken slowly and quickly.

AUTOMATIC TIMING

The work on manual timing of synthetic speech is a step towards an automatic timing procedure. The next step will be to mark the stressed syllables of the sentence and let that form the basis of the synthesis of prosody.

The duration of consonant clusters presents an additional problem. It is well known that consonants in clusters normally are shorter than single consonants. Our first algorithm for automatic timing assigns a syllable interval which is a linear combination of the sum of the element default durations, and a syllable default duration. Using a syllable default duration only would mean making the intervals between the vowel onsets equal. A simple listening test was performed to find the best balance between these two factors. A Norwegian sentence with large consonant clusters was randomly presented with different timing parameters, and the listeners were asked to choose the sample with the most natural rhythm. The test was repeated with several sentences and several listeners. The preferred timing consists of 80% of the element default durations and 20% of a syllable default duration.

LISTENING TEST

As this work only considers the timing of sentences, all the synthesis is made in whisper. Three different methods of sentence timing are compared:

1) Manual timing of vowel onsets,
2) Automatic timing based on element default durations combined with a syllable default duration,
3) Element timing aligned with human speech.

The test is designed to measure the perceptual distance between these three methods. The listeners are asked to pick the sentence with the most natural rhythm from a presented pair. Results will be presented at the conference.

REFERENCES

[1] W. Heinback:
Rhythmus von Sprache: Untersuchung methodischer Einflüsse.
Proceedings of DAGA´85, Stuttgart