ADDITIONAL PROSODY TO A PHONEME SYNTHESIZER

GEORGE E. OTTEREN
Acoustics Research Center, ELAB
N-7534 Trondheim-NTH, NORWAY

ABSTRACT

This work investigates the possibility of increasing methods of sentence synthesis by adding some prosodic information. The prosodic syllables are typed onto the keyboard of a computer. The vowel and consonant durations are modified by rules fit to 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 performed on Norwegian sentences synthesized in whisper.

INTRODUCTION

Text-to-speech synthesizers and phoneme synthesizers are not used for public services in Norway. They are judged to sound very unusual. The main shortcoming is the lack of natural prosody. Norwegian has a complex prosodic system. There are two distinct word tones, the timing pattern is mainly stress based, and the overall sentence prosody is strongly dependent on the syntactic and semantic structure. The natural prosody cannot be derived automatically from text without advanced methods of prosody 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 synthesized 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 the vowel in each syllable. Using a timing is a speech synthesizer tapping the rhythm of the syllables onto the keyboard of the computer. Simple rules are given to align the rhythmic speech with the keystroke sequence.

The synthesizer is a 4-formant phonetic synthesizer with Norwegian phonemes. Each phoneme corresponds to 1, 2 or 3 phonetic elements in the synthesizer. Vowels consist of one element, diphthongs of two, and unvoiced glides 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 better timing than tapping with one finger only. The sentence is spoken by the person simultaneously with the tapping. Each tap corresponds to the starting point of the vowel in the syllable. The vowel or diphthong and the following consonants make up the time interval between two keystrokes.

Initial experiments show that there is slight scaling of the elements of each syllable not accepted. The burst of the unvoiced vowels /p, t, k/ cannot be stretched significantly without losing the impression of a burst. A major limitation is that tapping in the duration of the short vowels creates confusion between short and long vowels, as in the words /kaem/ and /kaen/. 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. A factor of the form \( T_{\text{vowel}} / T_{\text{actual}} \) is called the time warping sensitivity, \( s_v \), of the phonetic element. Each phonetic element is given an additional length, \( T_{\text{vowel}} \), which is proportional to the default duration, \( T_{\text{vowel}} \), to the time warping sensitivity, \( s_v \), and to the difference of actual syllable length, \( T_{\text{syll}} \), and the sum of the default lengths:

\[
T_{\text{vowel}} = s_v T_{\text{vowel}} (T_{\text{vowel}} - T_{\text{syll}}) / T_{\text{syll}}.\]

Figure 1 gives an illustration of this alignment.

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