SLOVENIAN SPEECH TIMING AT DIFFERENT SPEAKING RATES

J. Gros, F. Miheli, N. Pavesi
Artificial Perception Laboratory
Faculty of Electrical Engineering
University of Ljubljana
Traška 25, SI-1000 Ljubljana, Slovenia
e-mail: nejka@fe.uni-lj.si

ABSTRACT
Speech timing at different speaking rates was studied for the Slovenian language and the results were applied in the two level duration prediction model in the SQEL Slovenian text-to-speech system S5 [1].

In order to provide the synthesiser with the possibility to pronounce input text with several speaking rates, tests were made to study the impact of speaking rate on syllable duration and duration of individual phones and phoneme groups for the Slovenian language.

1. INTRODUCTION
Regardless of whether the duration units are words, syllables or phonetic segments, contextual effects on duration are complex and involve multiple factors [2,3,4].

Results of perception experiments show that tempo variation contributes significantly to the perceived naturalness of speech [5]. In order to enable the synthesiser to pronounce input text with several speaking rates, tests were made to study the impact of speaking rate on syllable duration and duration of individual phones and phoneme groups.

For prosody prediction in the SQEL Slovenian Speech Synthesis System S5, we use a two-level approach to duration modelling. The levels correspond to the two levels of durational control [6,7]: the extrinsic and the intrinsic one. Units of word length are said to have a set of intrinsic durations, stored in our mental lexicon. As these units are integrated into larger entities, such as phrases, they get stretched and squeezed in accordance to larger speech demands, which correspond to an extrinsic level of durational control.

We first determine the words’ intrinsic duration, taking into account factors, relating to phone segmental duration, such as: segmental identity, phone context, syllabic stress and syllable type: open or close syllable.

Further, the extrinsic duration of a word is predicted, according to higher-level rhythmic and structural constraints of a phrase, operating on the syllable level and above. Here the following factors are considered: the chosen speaking rate, the number of syllables within a word and the word’s position within a phrase, which can be phrase initial, phrase final or nested within a phrase.

Finally, the intrinsic segment duration is modified, so that the entire word acquires its predetermined extrinsic duration. A method for segment duration prediction was developed, which adapts a word with an intrinsic duration to the desired extrinsic duration, taking into account how stretching and squeezing apply to duration of individual segments [8,9].

2. SPEECH CORPUS
A large continuous speech database was recorded to study the impact of speaking rate on syllable duration and duration of phones and phoneme groups in the Slovenian language.

When reading the same text at different speaking rates, it is possible to obtain phone realisations that differ only in duration. Thus context, stress and all other factors are kept identical to every realisation of the sentence. As a result, pair-wise comparisons of phone duration can be made.

We opted for a relatively long text of 172 sentences derived from the Slovenian speech database GOPOLIS [10], covering the domain of air timetable information retrieval (Table 1):

<table>
<thead>
<tr>
<th>speech rate</th>
<th>number of sentences</th>
<th>number of words</th>
<th>number of phones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>172</td>
<td>1.400</td>
<td>5.433</td>
</tr>
<tr>
<td>Fast</td>
<td>172</td>
<td>1.400</td>
<td>5.433</td>
</tr>
<tr>
<td>Slow</td>
<td>172</td>
<td>1.400</td>
<td>5.380</td>
</tr>
<tr>
<td>Total</td>
<td>516</td>
<td>4.200</td>
<td>16.246</td>
</tr>
</tbody>
</table>

Table 1. GOPOLIS database. Number of sentences, words and phones for the three speaking rates.

A male speaker was instructed to pronounce the same material at different rates: at a normal rate, very slow and as fast as possible. Reading the text took:

<table>
<thead>
<tr>
<th>rate</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal rate</td>
<td>7 minutes 32 seconds</td>
</tr>
<tr>
<td>fast rate</td>
<td>5 minutes 45 seconds</td>
</tr>
<tr>
<td>slow rate</td>
<td>12 minutes 55 seconds</td>
</tr>
</tbody>
</table>

When pronouncing the text, the speaker kept the speaking rate rather constant, as it can be seen from Table 2, showing the phone duration variation. Phone duration variation was evaluated for a given speaking rate by averaging phone duration differences for words, which occurred in the corpus several times, in the same phonetic environment and in the same type of phrase.
An average absolute phone duration difference of 5.3 ms with a standard deviation of 8.2 ms was obtained for different realisations of the initial part of the phrase *Ob kateri uri* ... , meaning *At what time* ... for the normal speaking rate (Table 2).

As in [11], the speech material was initially labelled using a Hidden Markov model speech recogniser for the Slovenian language in forced segmentation mode. The obtained labels were manually corrected using a special visual interface we developed for viewing, editing and labelling speech signals.

<table>
<thead>
<tr>
<th>speech rate</th>
<th>average absolute phone duration difference [ms]</th>
<th>standard deviation [ms]</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>5.3</td>
<td>8.2</td>
</tr>
<tr>
<td>fast</td>
<td>4.0</td>
<td>6.4</td>
</tr>
<tr>
<td>slow</td>
<td>13.8</td>
<td>20.6</td>
</tr>
</tbody>
</table>

Table 2. Phone duration variation for the phrase *Ob kateri uri* given in form of average absolute phone duration difference and the standard deviation.

3. PHONE DURATION

The effect of speaking rate on phone duration was studied in a number of ways.

An extensive statistical analysis of lengthening and shortening of individual phones, phoneme groups (nasals, liquids, plosives, fricatives) and phone components (closures, bursts) was made, the first of this kind for the Slovenian language.

Pair-wise comparisons of phone duration were calculated. Average mean duration differences and standard deviations were calculated for pairs of phones pronounced at different speaking rates.

Prior to the comparison, phone duration was normalised to the corresponding normal rate phone duration. Pairs were first composed of normal and slow rate phones, and later of fast and normal rate phones. Figures 2 and 4 give the results of these pair-wise comparisons and show in what extent the average phone duration when speaking or slow or fast increases or reduces with respect to its normal rate duration.

Closures of plosives change but slightly and maintain almost the same duration regardless of the speaking rate. Affricate closures exhibit an interesting behaviour, since they lengthen considerably, whereas they do not shorten at all.

The opposite holds for affricate bursts, together with their corresponding fricatives. In the fricative group, voiced fricatives change more than unvoiced ones. Short vowels, contrary to long vowels, increase more in duration when speaking slower than they shorten when speaking faster.
Table 3. Articulation rate expressed in syllables per second, given for speech units in different phrase positions, different lengths (number of syllables in the word) and for three speaking rates: normal, fast and slow.

From these observations we may draw a conclusion: phones or phone components, which are considered as short by nature (except for bursts of plosives), increase more in duration at a slow rate than they shorten at a fast rate. The opposite holds for affricates and long vowels.

4. SYLLABLE DURATION AND ARTICULATION RATE
Articulation rate, expressed as the number of syllables per second [12], excluding silences and filled pauses, was studied for the three different speaking rates for different word positions within a phrase: isolated, phrase initial, phrase final and nested within a phrase.

Figure 2 shows articulation rate, given in number of syllables per second, plotted as a function of word length, given in number of syllables, and the word position in a phrase. The shown values apply for normal speaking rate.

The articulation rate immediately after pauses is higher than the one prior to pauses. This prepausal lengthening may be attributed as a slowing down of the speech in anticipation of a pause [12]. The articulation rate increases with longer words as average syllable duration tends to decrease with more syllables in a word. The same observations hold for fast and slow speaking rate (Table 3) [9].

We observed that in case atona are associated to their neighbouring words, articulation rate adopts a quasi-logarithmic contour, which can be described parametrically, as in [13]. Isolated words and those following a pause differ from this rule for words with more than four syllables, of which only a few realisations were available.

5. CONCLUSION
Measurements of different durational parameters of Slovenian phones and syllables are presented and discussed. The results were directly applied in the two level approach for duration prediction in the Slovenian speech synthesiser S5.

A comprehensive perceptual evaluation of the quality of the resulting synthetic speech was performed, according to ITU-T
Recommendation P.85 [14], describing a method for subjective performance assessment of the quality of speech voice output devices. In the evaluation different duration modelling methods were compared. The two-level duration prediction method based on the measurements discussed in this paper proved to yield the most natural sounding synthetic speech [9].

ACKNOWLEDGMENTS
This work was funded by the Commission of the European Community under COP-94 contract No. 01634 (SQEL) and by the Slovenian Ministry of Science and Technology.

REFERENCES