SLOVENIAN SPEECH TIMING AT DIFFERENT SPEAKING RATES

J. Gros, F. Miheli , N. Pave ši
Artificial Perception Laboratory
Faculty of Electrical Engineering
University of Ljubljana
Tr aška 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].

It is to be noted, that stretching and squeezing does not apply to all segments equally. Stop consonants, for example, are much less subject to temporal modification than other types of segments, such as vowels or fricatives.

To apply the two level approach, different aspects of phone and syllable duration have to be measured, especially the influence of speaking rate on phone and syllable duration.

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):

speech rate	number of sentences	number of words	number of phones
Normal	172	1.400	5.433
Fast	172	1.400	5.433
Slow	172	1.400	5.380
Total	516	4.200	16.246

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:

normal rate	7 minutes 32 seconds
fast rate	5 minutes 45 seconds
slow rate	12 minutes 55 seconds

When pronuncing 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 occured in the corpus several times, in the same phonetic environment and in the same type of phrase.

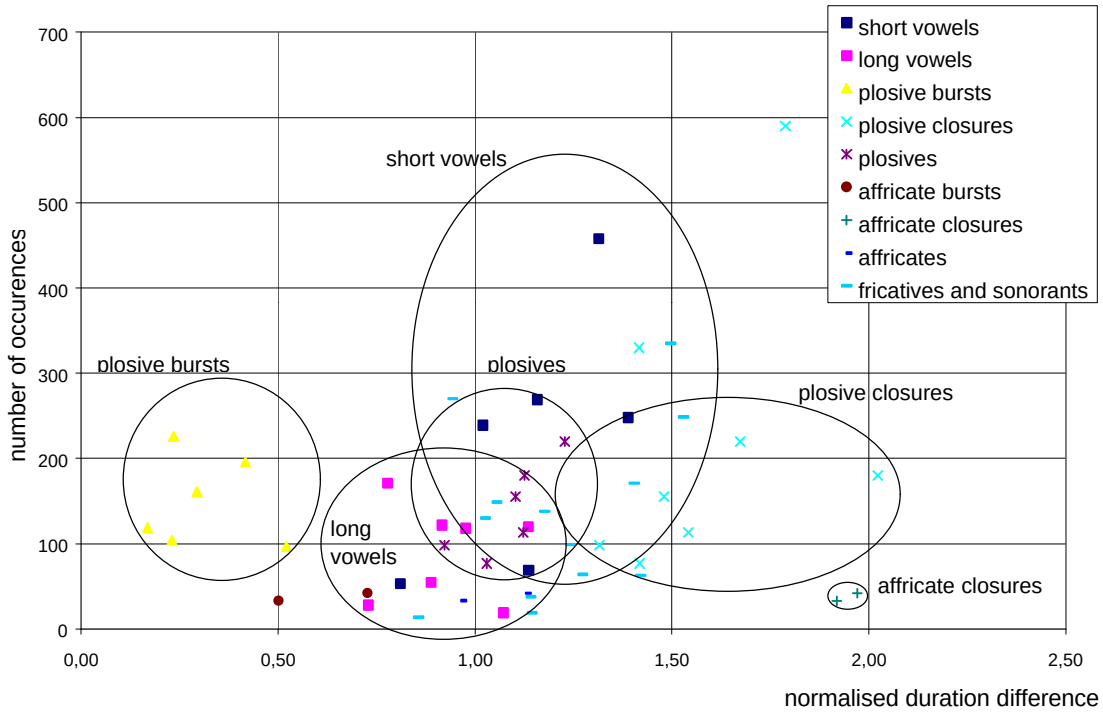

Figure 1. Pair-wise analysis: normal rate - slow rate. Normalised mean duration difference for pairs of phone realisations in the phoneme group context.

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 intial 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.

speech rate	average absolute phone duration difference [ms]	standard devation [ms]
normal	5.3	8.2
fast	4.0	6.4
slow	13.8	20.6

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 devations 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.

	normal speech rate: articulation rate [syllable/s]							
number of syllables	1	2	3	4	5	6	7	8
isolated word	3.03	4.27	5.44	6.02	7.76			
phrase initial word	4.26	6.33	7.11	7.18	8.76			
word within a phrase	5.42	5.80	6.41	6.78	7.20	6.73	7.16	7.25
phrase final word	3.19	4.47	5.05	5.65	5.94	6.69	6.03	6.11

	slow speech rate: articulation rate [syllable/s]							
number of syllables	1	2	3	4	5	6	7	8
isolated word	1.39	1.87	2.48	2.43	3.08			
phrase initial word	1.86	2.64	3.16	3.30	4.02			
word within a phrase	2.25	2.72	3.25	3.55	3.75	3.76	4.34	3.82
phrase final word	1.56	2.13	2.59	2.92	3.02	3.25	3.46	3.92

	fast speech rate: articulation rate [syllable/s]							
number of syllables	1	2	3	4	5	6	7	8
isolated word	5.24	6.08	6.60	7.38				
phrase initial word	6.02	8.16	8.75	8.86	9.63			
word within a phrase	6.90	6.93	7.49	7.78	8.46	8.11	7.80	8.12
phrase final word	3.89	5.30	6.08	6.65	7.15	8.39	7.07	7.29

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.

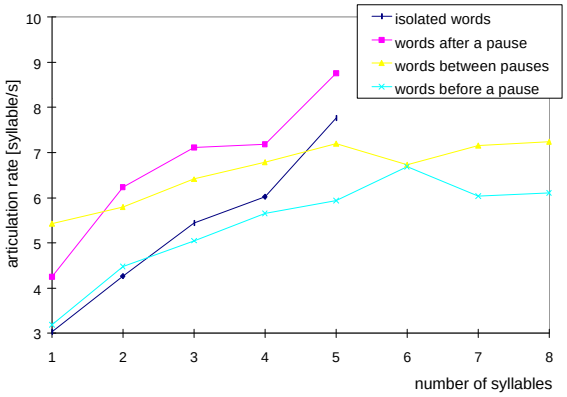

Figure 2. Articulation rate in number of syllables per second is shown for different word positions within a phrase.

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

[1] Gros, J., Paveši , N. and Miheli , F. 1996. A text-to-speech system for the Slovenian language. *Proceedings of the EUSIPCO'96*. Trieste. 1043-1046.

[2] Srebot Rejec, T. 1988. *Word Accent and Vowel Duration in Standard Slovene: An Acoustic and Linguistic Investigation*. Slawistische Beiträge. Band 226. Verlag Otto Sagner, München.

[3] van Santen, J.P.H. 1993. Timing in text-to-speech systems. *Proceedings of the EUROSPEECH'93*. Berlin. 1397-1404.

[4] Klatt, D.H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*. Vol. 59. 1209-1221.

[5] Ohno, S. and Fujisaki, H. 1995. *A method for quantitative analysis of the local speech rate*. Proceedings of the EUROSPEECH'95. Madrid, 421-424.

[6] Ferreira, F. 1993. Creation of prosody during sentence production. *Psychological Review*. No. 2. 233-253.

[7] Epitropakis, G., Tambakas, D., Fakotakis, N. and Kokkinakis, G. 1993. Duration modelling for the Greek language. *Proceedings of the EUROSPEECH'93*. Berlin. 1995-1998.

[8] Gros, J., Paveši , N. and Miheli , F. 1997. Speech timing in Slovenian TTS. *Proceedings of the EUROSPEECH'97*, Rhodes.

[9] Gros, J. 1997. Converting Slovenian Text into Speech. *PhD Thesis*. University of Ljubljana. 1997. (in Slovenian).

[10] Dobrišek, S., Gros, J., Miheli , F., Pepelnjak, K. and Ipši , I. 1996. GOPOLIS: Slovenian speech database of spoken flight information queries. *Proceedings of the 2nd SDRV Workshop on Speech and Image Understanding*. Ljubljana. 37-46.

[11] Gros, J., Ipši , I., Paveši , N., Miheli , F. and Dobrišek, S. 1996. Automatic segmentation of Slovenian diphone inventories. *Proceedings of the COLING'96*. Copenhagen. 298-303.

[12] O'Shaughnessy, D. 1995. Timing patterns in fluent and disfluent spontaneous speech. *Proceedings of the ICASSP'95*. Detroit. 600-603.

[13] Bakran, J. 1994. A model of the temporal organisation of the Standard Croatian language. *PhD Thesis*. University of Zagreb. (In Croatian.)

[14] ITU. 1994. *A method for subjective performance assessment of the quality of speech voice output devices*. ITU-T Recommendation P.85. International Telecommunication Union.