SPEECH TIMING IN SLOVENIAN TTS

J. Gros, N. Pavešić, F. Mihelič
Artificial Perception Laboratory
Faculty of Electrical Engineering
University of Ljubljana
Tržaška 25, 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 for duration modelling in the Slovenian text-to-speech system *S5* [1].

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 phonemes and phoneme groups for the Slovenian language [2].

A two-level approach to durational modelling is described. 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.

1. INTRODUCTION

Regardless of whether the duration units are words, syllables or phonetic segments, contextual effects on duration are complex and involve multiple factors [3,4,5]. In terms of speech synthesis, various durational models were proposed, from the traditional sequential rule systems [4,5], neural networks [6] and decision trees to stochastic modelling approaches.

A two-level approach to durational modelling, which we adopted in the recent version of the Slovenian text-to-speech system *S5* [1], is described. The levels correspond to the two levels of durational control [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.

1. MEASUREMENTS

1.1 Phoneme duration

A speech database consisting of isolated words, carefully chosen by phoneticians [3], was recorded in order to study different effects on phoneme duration, which operate on the segmental basis. Vowel duration was studied in different types of syllables: stressed /unstressed, open/closed. Consonant duration was measured in CC and VCV clusters.

Another large continuous speech database was recorded to study the impact of speaking rate on syllable duration and duration of phonemes. A male speaker was instructed to pronounce the same material at different rates: at a normal, fast and slow rate. Thus context, stress and all other factors were kept identical to every realisation of the sentence. As a result, pair-wise comparisons of phoneme duration could be made.

The effect of speaking rate on phoneme duration was studied in a number of ways [9]. An extensive statistical analysis of lengthening and shortening of individual phonemes, phoneme groups and phonemic components, like closures or bursts was performed, the first of the kind for the Slovenian language.

1.2 Articulation rate

Articulation rate, expressed as the number of syllables per second [8], excluding silences and filled pauses, was studied for the different speaking rates. In other studies, articulation rate is usually determined for speech units with the length of individual words or entire phrases. We studied the articulation rate of words along with their associated cliticised words at different positions within a phrase: phrase initial, phrase final and nested within the phrase [2].

2. DURATION MODEL

Similarly to Epitropakis [9], our two-level duration model first determines the words' intrinsic duration, taking into account factors relating to phoneme segmental duration, such as: segmental identity, phoneme context, syllabic stress and syllable type: open or closed syllable [2].

Further, the extrinsic duration of a word is predicted, according to higher-level rhythmic and structural constraints on 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. 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.

Therefore, a method for predicting segment duration was developed, which adapts a word with an intrinsic duration t_i to the determined extrinsic duration t_e, taking into account how stretching and squeezing apply to the duration of individual segments.

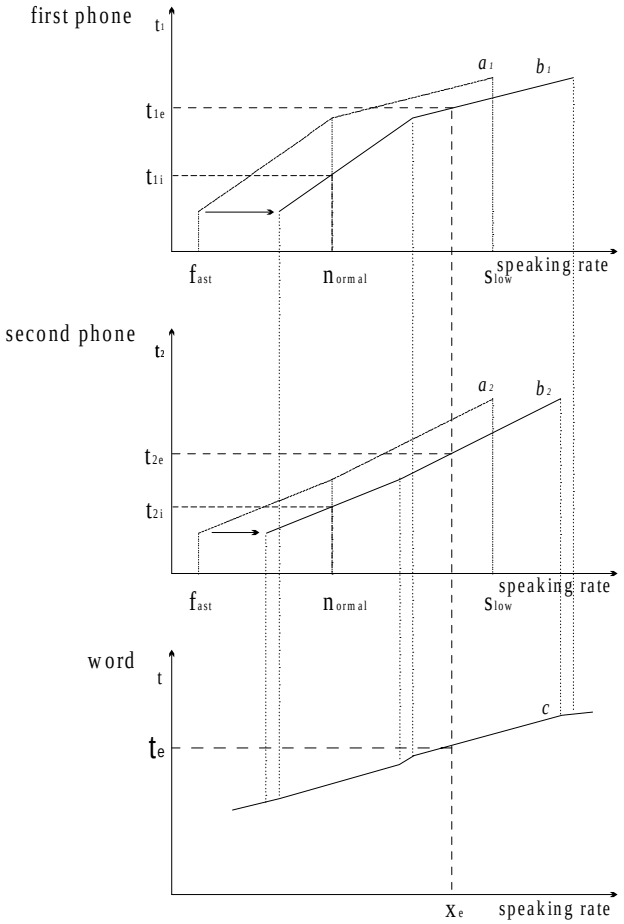

Figure 1: *Adaptation of a word with an intrinsic duration t_i to the predetermined extrinsic duration t_e.*

The basic idea of the method is given in Figure 1, for a word consisting of two phonemes, phoneme$_1$ and phoneme$_2$. Lines a_1 and a_2 are obtained by linear interpolation of the measurements of average segment duration at three different speaking rates: slow, normal and fast.

Since intrinsic durations are calculated for normal rate values, lines a_1 and a_2 are translated horizontally in order to cross the line of normal speaking rate at the phoneme intrinsic duration t_{1i} and t_{2i}: so lines b_1 and b_2 are formed. Line c represents the sum over all the individual phoneme lines constituting the word in each of the new breaking points, introduced by the translation of lines a_1 and a_2.

The desired extrinsic phoneme duration t_{1e} and t_{2e} is determined by the speaking rate at which the extrinsic word duration t_e crosses line c.

3. RESULTS

The reliability of our two-level prediction method was evaluated on a speech corpus consisting of:

speech rate	number of sentences	number of words	number of phonemes
normal	172	1400	5433
fast	49	607	2351
slow	60	800	2900

The predicted durations were compared to those in the same position in natural speech. Additionally, duration predicted by the simple proportional method where no lengthening or shortening of phonemes was considered was compared to phoneme duration taken from natural speech.

Natural duration variation was evaluated by averaging phoneme duration differences for words, which occured 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 strandard deviation of 8.2 ms was obtained for different realisations of the intial part of the phrase *Ob kateri uri ...*, meaning *At which hour ...* for the three speaking rates:

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

As we can see from Table 1, the two-level approach predicts extrinsic phoneme duration quite accurately and performs better in comparison to the simple proportional adaptation method. It performs even better when predicting duration of stressed phonemes. The mean difference lies much under the JND or the just noticeable difference, as defined by Huggins [10].

Standard deviation of the difference between natural and predicted duration difference is 15.4 ms for normal speaking rate, and even less for stressed phonemes the duration of which is of crucial importance to the perception of naturalness of synthetic speech. In the well known MITalk system the achieved standard deviation was even larger, 17 ms [11].

An example of different methods of phoneme duration prediction for the phrase *Dobro jutro*., meaning *Good morning*., is given in Figure 2. Natural phoneme duration is compared to those predicted by the translation and proportional adaptation method.

		t_i to t_e duration adaptation method		
		translation	proportional	natural speech
normal	average absolute duration difference [ms]	10.97	30.20	5.3
	- stressed vowels [ms]	6.89	33.67	
	standard deviation [ms]	15.24	26.41	8.2
	- stressed vowels [ms]	13.18	28.07	
slow	average absolute duration difference [ms]	37.67	62.26	13.8
	- stressed vowels [ms]	23.30	87.41	
	standard deviation [ms]	40.52	65.79	20.6
	- stressed vowels [ms]	36.65	100.37	
fast	average absolute duration difference [ms]	17.67	19.39	4.0
	- stressed vowels [ms]	21.64	29.18	
	standard deviation [ms]	66.49	66.61	6.4
	- stressed vowels [ms]	16.80	19.86	

Table 1: *Statistical evaluation of differences of natural and predicted phoneme duration.*

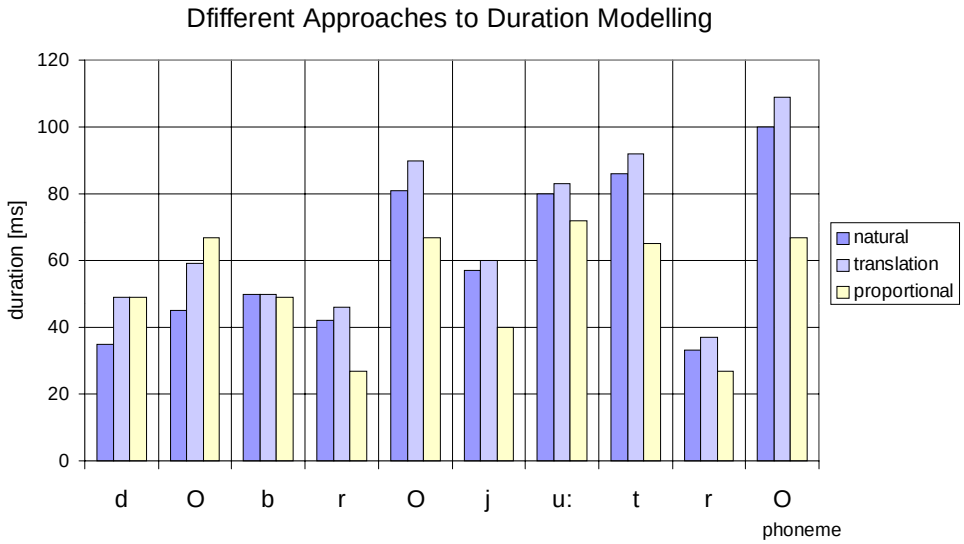

Figure 2: *Comparison of phoneme duration in the phrase* Dobro jutro, *meaning* Good morning. *Natural phoneme duration is compared to those predicted by the translation and proportional method.*

4. PERCEPTUAL EVALUATION

The adequacy of the TTS system was tested in terms of acceptability and in terms of intelligibility. The experiment was performed in laboratory conditions with 21 subjects within the age span between 19 and 45 years, ten of them being female.

The test was performed according to the ITU-T Recommendation P.85 [12], describing a method for subjective performance assessment of the quality of speech voice output devices.

In one part of the test, different versions of the TTS system were compared, including different types of duration modelling. Segment duration as predicted by our two-level approach and segment duration taken from natural speech were passed to the synthesiser.

Both types of sentences were presented to the test subjects, which were asked to decide on which synthetic sound they preferred.

The results are given in Figure 3. For the major part of the sentences, the test subjects said there was no perceptible difference between synthetic speech obtained by natural durational parameters and synthetic speech controlled by the two-level approach durational parameters.

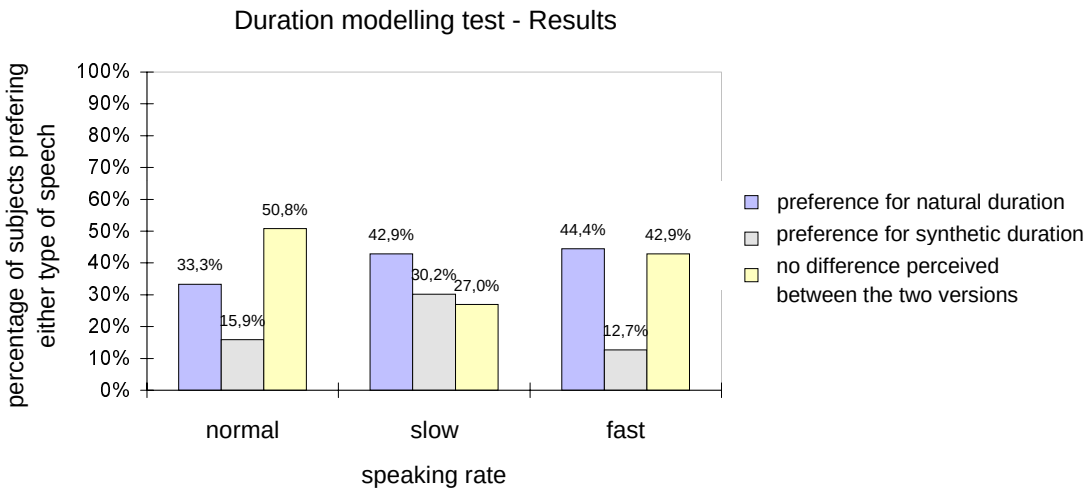

Figure 3: *Test subjects were asked which type of synthetic speech they preferred: the one modelled with durations taken from natural speech or the one with automatically predicted durations*.

We enclose four synthetic sentences, all of them derived from the same text: *A je kakšen avion jutri po petnajsti uri za Skopje?* meaning *Is there a flight to Skopje tomorrow after 3 o'clock?* The first two sentences are synthesised at the normal speaking rate, the second two at the fast speaking rate. In each group, the first sentence was produced by durations taken directly from natural speech (A1352S01.wav and A1352S03.wav). Duration in the second sentence of each group was modelled by the two-level approach (A1352S02.wav and A1352S04.wav).

5. CONCLUSION

The proposed duration prediction method yields accurate phoneme duration predictions for the normal speaking rate. In case of fast and slow speaking rate, the intrinsic duration prediction should be derived from measurements for intrinsic phoneme duration for phonemes in different contexts at these two speaking rates.

ACKNOWLEDGEMENT

This work was funded by the European project Copernicus under contract No. COP-1634 and by the Slovenian Ministry of Science and Technology.

The authors wish to thank dr. Tomaž Erjavec for his proofreading of the text.

REFERENCES

[1] J.Gros, N.Pavešić, F.Mihelič, "*A text-to-speech system for the Slovenian language*", Proc. EUSIPCO'96, Trieste, pp. 1043-1046, 1996.

[2] J.Gros, N.Pavešić, F.Mihelič, "*Syllable and segment duration at different speaking rates in the Slovenian language*", Proc. EUROSPEECH'97, Rhodes, 1997.

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

[4] J.P.H. van Santen, *"Timing in text-to-speech systems*", Proc. EUROSPEECH'93, Berlin, pp. 1397-1404, 1993.

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

[6] W.N. Campbell, "*Predicting syllable durations for accomodation within a syllable-level timing network*", Proc. EUROSPEECH'93, Berlin, pp. 1082-1085, 1993.

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

[8] D.O'Shaughnessy, "*Timing patterns in fluent and disfluent spontaneous speech*", Proc. ICASSP'95, Detroit, pp. 600-603, 1995.

[9] G.Epitropakis, D. Tambakas, N. Fakotakis, G. Kokkinakis, *"Duration modelling for the Greek language*", Proc. EUROSPEECH'93, Berlin, pp. 1995-1998, 1993.

[10] A.W.F. Huggins, "Just noticeable differences fot r segment duation in natural speech", *Journal of the Acoustical Society of America*, Vol. 51, pp. 1270-1278, 1971.

[11] J. Allen, M.S.Hunnicutt, D.H.Klatt, "From Text to Speech: The MITalk System", Cambridge University Press, Cambridge, 1987.

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