Eesti Keele Instituut

ВОСПРИЯТИЕ ТЕМПА РЕЧИ И НЕКОТОРЫЕ НАХОДКИ В СФЕРЕ МОДЕЛИРОВАНИЯ РЕЧЕВОЙ РИТМИЧЕСКОЙ СТРУКТУРЫ ЭСТОНОЯЗЫЧНОЙ РЕЧИ

SPEECH RATE PERCEPTION AND SOME FINDINGS OF MODELLING SPEECH RHYTHMICITY IN ESTONIAN

Meelis Mihkla, Indrek Hein, Mari-Liis Kalvik, Indrek Kiissel

Institute of the Estonian Language, Tallinn, Estonia

Motivation

To improve the quality of synthetic speech and to offer more flexible possibilities in speech mediation.

- Many blind people wish to hear the news and newspaper articles at a considerably higher speech rate than normal

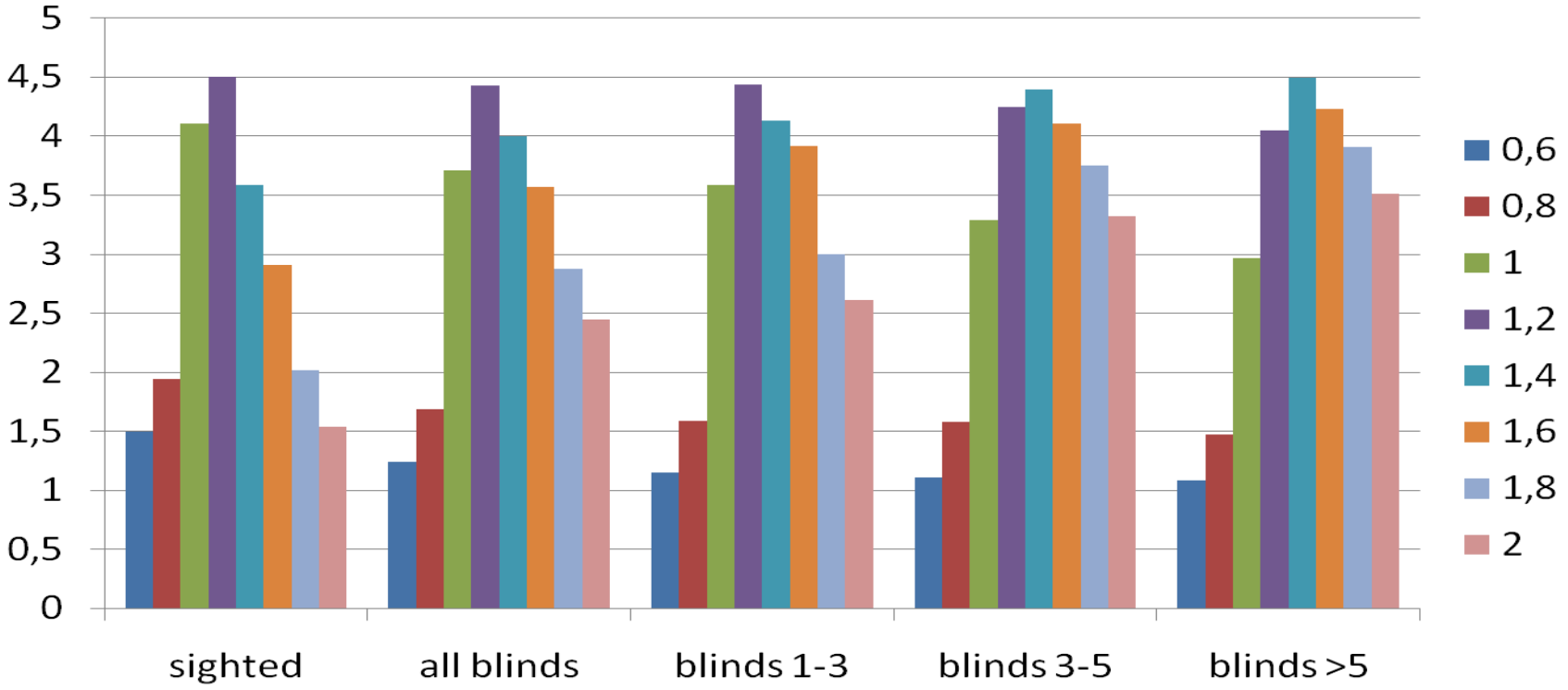
- In Estonian, rather a stress-timing language, the foot is the domain for the phonological opposition of three quantity degrees (Q1, Q2, Q3) to realize

Audio system of the Estonian Library for the Blind http://www.epr.ee/kalev

Enables the visually impaired people

- To listen electronic texts (news, newspapers, magazines, books) and audio books over Internet
- To change speech rates (normal, fast and very fast) and synthetic voices

The average scores given to different speech rates vs. the previous experience, in years, of the visually impaired subjects with a computer and screen reader.



Speech rhythmicity

- The vowel onsets in the stressed syllables play a decisive role in enhancing the naturalness of synthetic speech
- In Estonian the foot is the arena for the phonological opposition of three quantity degrees (Q1, Q2, Q3)
- In principle, Estonian quantity degrees are defined in the same spirit as the newer approaches to speech rhythmicity
- The present study compares different parameters of quantity and, by statistical modelling, evaluates their significance

ВОСПРИЯТИЕ ТЕМПА РЕЧИ И НЕКОТОРЫЕ НАХОДКИ В СФЕРЕ МОДЕЛИРОВАНИЯ РЕЧЕВОЙ РИТМИЧЕСКОЙ СТРУКТУРЫ ЭСТОНОЯЗЫЧНОЙ РЕЧИ

SPEECH RATE PERCEPTION AND SOME FINDINGS OF MODELLING SPEECH RHYTHMICITY IN ESTONIAN

Institute of the Estonian Language

Meelis Mihkla, Indrek Hein, Mari-Liis Kalvik, Indrek Kiissel

Institute of the Estonian Language, Tallinn, Estonia

Background

The necessity for the two relatively different studies arose in the course of developing an audio system enabling the blind to listen to electronic texts and audio books over the Internet http://www.epr.ee/kalev. The use of the system revealed that many blind people wish to hear the news and newspapers at a considerably higher speech rate than normal. Hence the need to find some optimal speech rates: normal, quick and very quick. The speech rhythmicity plays a decisive role in enhancing the naturalness of synthetic speech. The rhythmicity of speech is particularly important in Estonian, where foot is the arena for the phonological opposition of three quantity degrees (Q1, Q2, Q3) to realize. The present study compares different parameters of quantity and, by statistical modelling, evaluates their significance.

Speech rate perception

Our test of speech rate perception was also applied to a group of sighted subjects, for comparison. This was meant to answer such questions as: What speech rates are preferred by the blind vs. the sighted? Is the preference of very quick speech rates by the visually impaired a myth or not? Is there such a thing as an optimal speech rate?

Structuring the speech corpus

In order to optimize the unit selection algorithm and to guarantee the necessary quality of the synthetic speech the whole speech database as well as the utterance to be synthesized is represented as a phonological tree. The figure below represents a fragment of the database tree. Our phonological tree has the following levels: phoneme, syllable structure, syllable, foot, word, phrase, and sentence. Search for appropriate speech units involves all those levels, beginning from the higher ones, e.g. preferring longer units.

Subjects

The test was taken by 58 blind or heavily visually impaired subjects (29 female and 29 male, aged 14-79) and by 56 sighted subjects (41 female and 15 male, aged 18-58). For all subjects, Estonian was the mother tongue.

Test material

The stimuli for the test of speech rate perception were generated from two audiobooks and some news fragments, synthetic voice. The latter was produced by a diphone-based Estonian text-to-speech synthesizer, using an MBROLA synthesis motor. The synthetic voice was generated in two variants, one using a rule-based prosody model (SYNT1) the other a statistical one (SYNT2).

81	108	**135**	162	189	216	243	270	words/min
60	80	**100**	120	140	160	180	200	%

Figure 1. Speech samples as stimuli of different speech rates (natural speech rate 100 = 135 s/min).

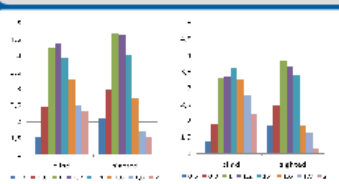


Figure 2. The average points from the blind and the sighted for the female and the synthetic (SYNT1) voice.

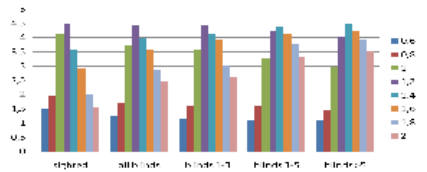


Figure 3. The average scores given to different speech rates vs. the previous experience, in years, of the visually impaired subjects with a computer and screen reader.

Recognition and modelling of speech rhythmicity

Three Estonian quantity degrees – short (Q1), long (Q2) and overlong (Q3) – are a complex phenomenon which has an important role of forming Estonian speech rhythmicity. Q1, Q2 and Q3 (e.g. *pole* [pole] 1s, are *not*', *poole* [po:le] 'half GenSg', *poole* [po:˙le]'towards') occur in disyllabic foot (which contains stressed and unstressed syllable) and are differentiated by both durational and tonal features. The main and most constant feature carrying the opposition of quantity degrees is the duration ratio of the syllables in foot. In this study we observe durational ratios both inside the foot (classical ratio) and it's syllables (according to the theory of adjacent segments) in the standard Estonian words with structure CV(:)CV.

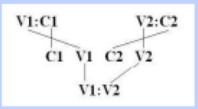


Figure 4 shows all the durational ratios investigated. V1:V2 represents the durational ratio of the foot, two-stepped V1:C1, V2:C2 represents the durational ratio of adjacent segments of each syllable.

	C1	V1	V1:C1	C2	V2	V1:V2	V2:C2
Q1	69	68	1,05	52	87	0,82	1,74
Q2	64	120	1,99	52	69	1,80	1,37
Q3	66	165	2,59	59	66	2,99	1,17

Table 1 presents the mean durations (ms) and duration ratios of the segments in Q1, Q2, Q3 words. The results are comparable to the data of earlier studies and seem also to corroborate the theory of adjacent segments.

For weighing the relevance of different duration ratios equations of linear regression were generated. For classical duration ratios the linear model yielded strong correlation between the input and output (correlation coefficient r=0.867) and the model explains over 75% of the data variation (coefficient of determination r2=0.752). The alternative model generated from the other two duration ratios, however, yielded a correlation coefficient equal to 0.759, which means that it explains only 58% of the variation in the data analysed.

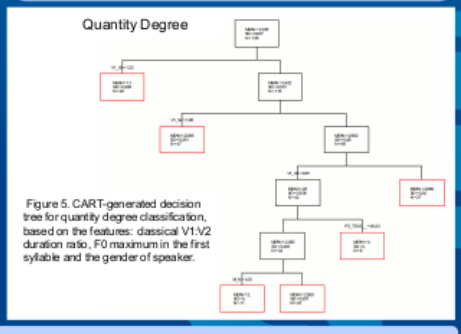


Figure 5. CART-generated decision tree for quantity degree classification, based on the features: classical V1:V2 duration ratio, F0 maximum in the first syllable and the gender of speaker.

Conclusion

The perception test did not provide an clear answer to all above questions. An "optimal speech rate" could not be established either for the blind or the sighted as the speech rate preferences were extremely individual, depending on many factors. Considering the results of the tests the audio system was supplemented, in addition to the normal speech rate, with a fast rate (1.3 times, i.e. 30%, faster than normal) and a very fast rate (1.6 times, i.e. 60%, faster than normal) for an advanced computer user.
According to this study classical duration ratio (V1:V2) is the most relevant parameter to be considered in modelling the temporal structure of Estonian speech.

Dialog 2010, Conference on Computational Linguistics and Intellectual Technologies, Bekasovo, May 26-30, 2010

29.03.2016