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 (meelis@eki.ee), **Indrek Hein** (kiisu@eki.ee),
Mari-Liis Kalvik (mariliis@eki.ee), **Indrek Kiissel** (indrek@eki.ee)

Institute of the Estonian Language

Статья посвящена проблемам восприятия различного темпа речи слепыми и зрячими, а также вопросам моделирования ритмики речи. Выяснилось, что «тренированные слепые» предпочитают гораздо более быстрый темп речи, чем зрячие. При обращении к трем степеням долготы в эстонском языке оказалось, что соотношение долготы гласных в ударном и безударном слоге является важнейшим признаком определения степени долготы.

1. Introduction

The necessity for the two relatively different studies arose in the course of developing an audio system enabling the blind to listen to reference texts and audio books. By means of the on-line system of the Estonian Library for the Blind http://www.epr.ee/kalev the visually impaired can have texts (news, newspapers, magazines and books) read for them and listen to audio books over the Internet. The use of the system revealed that many blind people wish to hear the news and newspaper articles at a considerably higher speech rate than normal. As the system is server-based it cannot afford users tuning the rate smoothly as it would make the system too cumbersome and slow. Hence the need to find some optimal rates to supplement the user menu with two speech rates from the quicker-than-normal range, say, quick and very quick.

Perception of rapid rate of speech and the limits of its temporal compression have been discussed in several studies (Asakawa a. o. 2003 and Moos, Trouvain 2007). It has been found that very rapid speech is preferred and perfectly understood by trained people, i. e. those with an everyday experience of a screen reader and a speech synthesizer. In the course of a joint study by the Universities of Saarland and Tübingen (Moos a. o. 2008) the brains of blind and of sighted subjects were scanned while they were listening to rapid speech; it was found that in the blind a very quick rate was systematically accompanied by activation in the brain zone that in the sighted is used for processing visual information.

Recently the mechanisms of speech rhythmicity have been drawing increasing attention. The focus lies on different aspects of the vowel onset of the stressed syllable (Keller, Port 2007), which play a decisive role in enhancing the naturalness of synthetic speech (Keller 2007). The temporal structure and rhythmicity of speech is partiularly important in Estonian, where foot is the arena for the phonological opposition of three quantity degrees (Q1, Q2, Q3) to realize. In principle, Estonian quantity degrees are defined in the same spirit as the newer approaches to speech rhythmicity (for an overview see Keller, Port 2007): in both the fulcrum is the onset of the rime of a stressed syllable. Quantity degrees are, in essence, suprasegmentals (Lehiste 1997), whose definition, on the acoustic level, relies on the durational relations of the rime of the stressed syllable and the nucleus of the unstressed syllable, plus the F0 contour (Ross, Lehiste 2001), making up a complementary system. In addition, the durational relations of consecutive phones and some other features have been suggested as important (Eek, Meister 2003). The present study compares different parameters of quantity and, by statistical modelling, evaluates their significance.

2. Speech rate perception

Although an on-line audio system is meant for the visually impaired mainly, 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 rate by the visually impaired a myth or not? Is there such a thing as an optimal speech rate?

2.1. 1 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.

2.2. Test material

The stimuli for the test of speech rate perception were generated from two audio books ("American tragedy" by T. Dreiser, male voice, and "Das erste Mal und mehr" by E. Stein-Fischer, female voice, both in Estonian translation) and some news fragments, synthetic voice. The latter was produced by a diphone-based Estonian text-to-speech synthesizer (Mihkla, Meister 2002), 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). This was to test the impression of the blind that in the case of a synthesizer with a statistical prosody module, rate quickening would lower the quality of output speech.

For the female voice the natural reading rate was 135 words per minute, the male voice making 122 words per minute; the synthesizers were tuned to match the female rate. For each voice, eight speech samples of 35–55 sec were generated, each of a different speech rate (see Fig. 1).

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)

The limits of temporal compression had been first agreed upon with some "trained persons", who have everyday practice of listening to synthetic speech. In their opinion the maximal rate for listening to prose texts is twice the usual rate. During test preparation a few older representatives of the visually impaired suggested that they might perhaps wish to listen to some paragraphs at a slower pace. Thus we added two slower samples (0.8 and 0.6 times the natural rate). The rate of the samples from audio books was regulated by means of the signal processing program Adobe Audition 3 for high precision time compression with time stretch (preserves pitch).

The subjects were exposed to the speech rate stimuli by voice series presented in a random order. The appropriateness of the speech rate was asked to be evaluated in a five-point system (5 — the best, 4 — good, 3 — tolerable, 2 — uncomfortable, 1 — unsuitable , i. e. unintelligible, too quick or too slow)

2.3. Results

Fig. 2 presents the average blind vs. sighted scores for different speech rates. The left diagram shows the scores given for the female voice, the right one for the synthetic voice SYNT1. According to the diagrams the blind prefer the speech rates 1.2 and 1.4, which are 20 % and 40 % quicker than natural, respectively. The sighted think highly of natural speed, but 1.2 and 1.4 are not considered bad either.

The following figure (3) demonstrates the points scored by the male and the synthetic voice SYNT2 (the left and right diagrams, respectively). For a male voice 1.2 was considered the best speech rate both by the sighted and the blind. The results are probably due to the natural speech rate for a male voice, which is about 10 % slower than the rest (122 s/min *versus* 135 s/min).

Although the subjects were asked to evaluate the suitability of the speech rate only, not the pleasantness of the voice, synthetic speech scored almost a point lower, on average, than human speech. However, the prosody module of the synthesizer (SYNT1 vs. SYNT2) does not seem to have influenced the score significantly at quicker speech rates. Thus the results fail to support the idea that the quality of the SYNT2 voice might deteriorate at quicker rates.

The test proved that the ratings of the blind and the sighted differ far less than first believed. Figures 2 and 3 present the average scores from all visually impaired subjects. However, the blind include many who seldom use a computer, if at all, and who thus lack the experience of listening to synthetic speech at different speech rates. Figure 4 presents 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. The results reveal an obvious tendency that in the visually impaired, longer practical "training", i. e. experience of using the above devices causes an increase in the ability of understanding rapid speech.

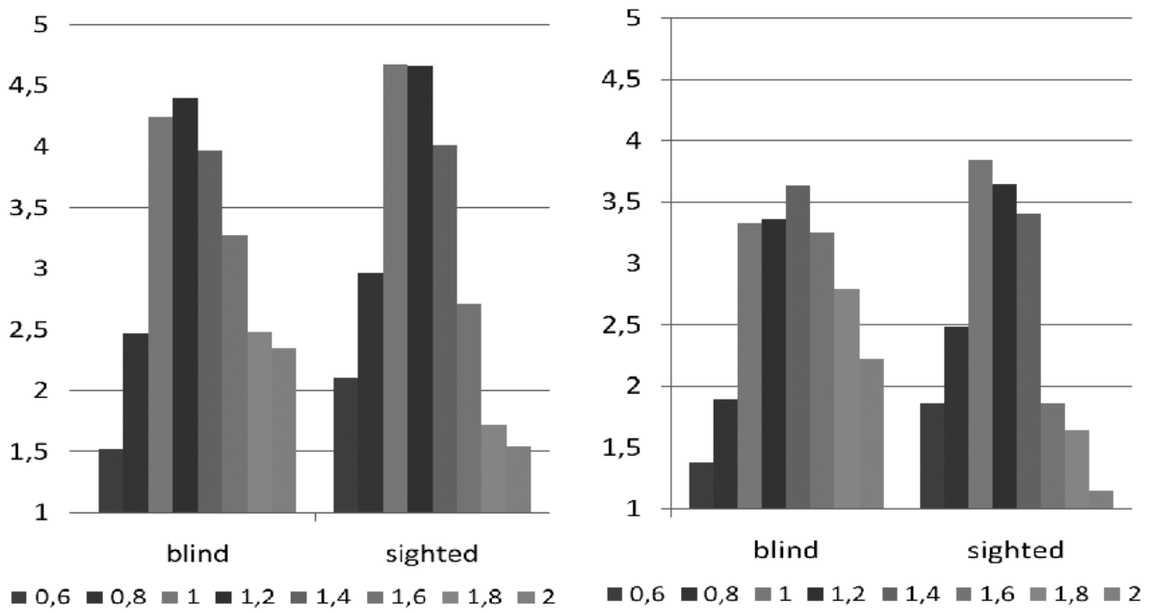

Figure 2. The average points from the blind and the sighted for the female and the synthetic (SYNT1) voice

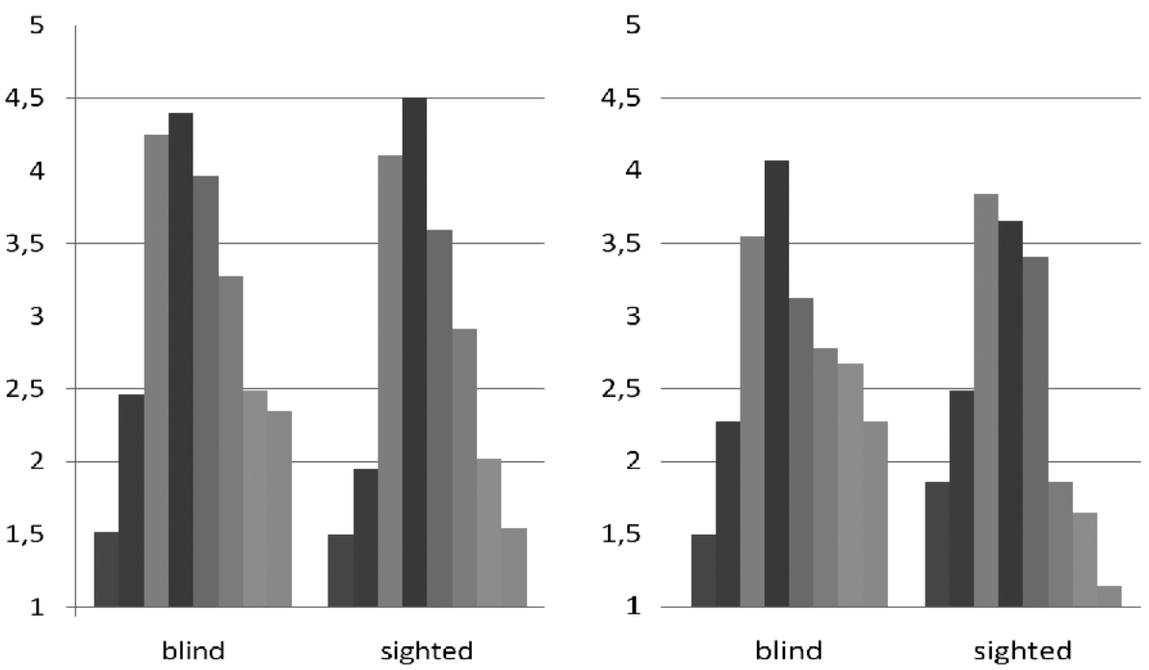

Figure 3. Average points from the blind and the sighted for the male and the synthetic (SYNT2) voice

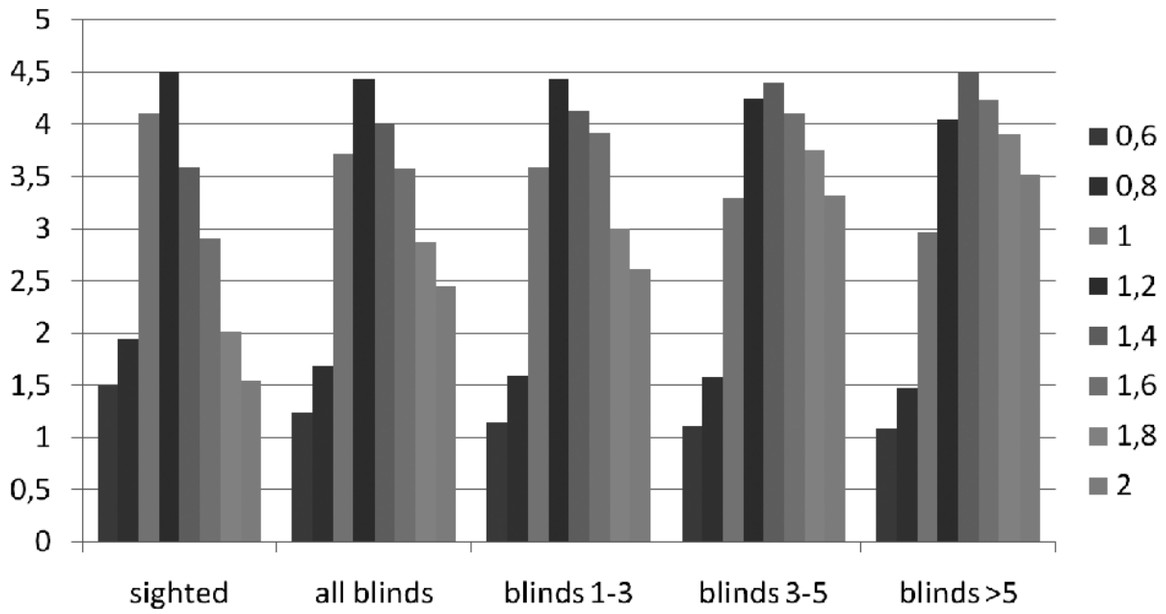

Figure 4. The average scores given to different speech rates vs. the previous experience, in years, of the visually impaired subjects with a computer and screen reader

3. Recognition and modelling of speech rhythmicity

3.1. Introduction

For quite a while now the problem of the distinctive features of the three quantity degrees — short, long and overlong (Q1, Q2, Q3) — observed in standard Estonian have been subject to discussion among phoneticians. Up to the present the duration ratio between the stressed and unstressed syllable in a foot and, in particular for Q2 and Q3, a difference in their pitch curves have been considered the most important parameters to describe and analyse the quantity degrees of words from differently structured speech, lab-generated as well as spontaneous. As a result of several studies it has been found that the duration ratio between the first and second syllables is 2:3 for Q1, 3:2 for Q2, and 2:1 for Q3 (Lehiste 1960, Ross, Lehiste 2001).

The present study investigates, in addition to the traditional duration ratios presented, the ratio of adjacent segments and discusses the possible role of intensity. Manifestation and perception of phonetic quantity make up a complicated system, where different conditions may evoke different combinations and different salience of parameters. In this study we seek additional parameters possibly depending on phonetic quantity, by weighing their relevance with statistical methods. As our aim in modelling speech temporal structure is high quality synthetic speech we need the best possible parameters to describe and discriminate the three Estonian degrees of phonetic quantity, so that each degree could get its own model.

In the present study we test some of the ideas suggested by Arvo Eek. In the first place, Eek and Meister have, on the basis of perception tests, created a theory focused on adjacent phones within the main stress syllable and the successive syllable. In a two-syllable word with a vowel-centred structure CV(::)CV, the duration ratio of the vowel (V1) to the consonant (C1) of the first, stressed syllable is supposed to discriminate Q1 words, which have a short V1, from Q2 and Q3 words, whose V1 is long. Q3 words can purportedly be distinguished from Q1 and Q2 words by the duration ratio of the vowel (V2) to the inter-vowel consonant (C2) of the second, unstressed syllable. Starting from a belief that the perception of duration difference between two adjacent phones is not possible unless one is 20–25 % longer than the other they have calculated 1.4 as the limit of duration difference. Words have a short V1 (thus qualifying as Q1 words) if their V1:C1 ratio is less than 1.4, while for a long V1 (signalling of Q2 or Q3) the ratio needs be equal to or higher than 1.4. A similar ratio computed for the unstressed syllable (V2:C2) supposedly signals of Q3 if it is less than 1.4, while its values equalling or exceeding 1.4 indicate Q1 or Q2, without, however, discriminating between the two. (Eek, Meister 2003). Second, Eek has pointed out that the first syllable of a Q3 word should be distinguishable by its mean intensity of the first syllable being higher than that of the successive syllable. For Q2 the intensity of the two syllables is suggested as equal (Eek, Meister 1997).

3.2. Material and method

The material consisted of 485 words (including words of all three quantity degrees) read, in sentences, by 12 male and 13 female speakers of standard Estonian. Most of the samples analysed belongs to the Babel linguistic corpus in possession of the Institute of Cybernetics at Tallinn University of Technology, some additional samples had been read by two announcers from the Estonian Broadcasting Company. The recorded material was segmented and phonetically analysed by means of the PRAAT program (Boersma, Weenink 2008).

The research focus lies on vowel-centred Q1, Q2 and Q3 words where both the main stress syllable and the successive one have the structure CV(::)CV. In Q1 words the first-syllable vowel is short (e. g. *pole* [pole] 'is, are not'), whereas in Q2 and Q3 words the first-syllable vowel is long (e. g. *poole* [po:le] 'half GenSg' and *poole* [po::le] 'towards', respectively). Thus, the ratio of the stressed and unstressed syllables is found from the ratio of their vowel durations (V1:V2), while the ratio of adjacent phones in the first and second syllables can be written as V1:C1 and V2:C2, respectively. Most of the words are disyllabic, but longer words can also be found. A small number of the words begin either with a vowel or with a consonant cluster. The material also includes some principal and attributive components of compound words, and some foreign words pronounced like genuine ones; the total share of such words is less than a third of the whole bulk. The analysed material contains stressed as well as unstressed words from different positions (initial, middle, final) in the sentence or phrase. For each word its sound durations (ms) and the mean intensity of V1 and V2 (dB) were measured. The results were averaged, adding the standard deviation (SD).

3.3. Results

3.3.1. Duration ratios

The results of sound measurement and the ratios computed have been summarized in Table 1. (The total means and standard deviations have been calculated from the whole bulk of data, not from mean values.) In total the material contained 234 Q1 words, 150 Q2 words and 101 Q3 words.

V1:V2 is the classical ratio to be examined. The mean durations easily reveal that in Q1 words V1 is about twice as short as in Q2 and Q3 words, which is generally considered sufficient to perceive the short/long opposition between Q1 and the rest (the short Q1 vs. the long Q2 and overlong Q3).

Comparing our results with those received on laboratory speech earlier we find that our ratios are realistic, as far as quantity degrees go, albeit a little higher than expected for Q1 and Q2. In general, a typical duration ratio of Q1 should fall in the interval 0.6–0.7 (Lehiste 1960, Liiv 1961, Eek 1983, Eek, Meister 1997). Like the authors of the present study, G. Liiv as well as D. Krull analysed vowel-centred words. G. Liiv adds that a Q1 ratio can range from 0.50–1.00, while for Q2 the range is 1.00–2.00 (Liiv 1961). Most likely the reason for our slightly higher duration ratios for Q1 and Q2 words lies in that our research material of those two quantity degrees contains more foreign words and compound components.

For Q2 the ratios of V1:V2 vary more across different studies, ranging from 1.2–1.60; for Q3 words the range is 2.4–2.6 (Liiv 1961, Krull 1991, Eek ja Meister 1997). The vowel-centred structure has also been the research object for E. L. Asu and others, who study spontaneous speech. According to their results the average duration ratio is about 0.7 for Q1, 1.7 for Q2 and over 2.0 for Q3 (Asu a. o. 2009). Those numbers do not contradict ours either.

Next, let us consider the ratios of adjacent phones for the stressed vs. unstressed syllables. Our results for a stressed syllable confirm that V1 is short (as expected for Q1) if and only if V1:C1 is less than 1.4 (1.05 in Table 1). For Q2 and Q3 words the respective ratios are 1.99 and 2.59, respectively, which are both considerably higher than 1.4. In an unstressed syllable the ratio of the vowel V2 and its preceding consonant C2 is perhaps a little less unambiguous. Still, the theoretical ratio for Q1 and Q2 words being 1.4 or higher, our result for Q1 (1.8) should do well, and so does the ratio for Q2 as it can be rounded to 1.4 easily. Also, our V2-to-C2 ratio for Q3 words supports the theory as 1.17 is clearly lower than 1.4. Thus, our material indeed seems to corroborate Eek's theory.

Next, statistical methods will be used to find out which duration ratios, the traditional V1:V2 or the two-step system of V1:C1 → V2:C2 suggested by Eek, are

	C1	SD	V1	SD	V1:C1	SD	C2	SD	V2	SD	V1:V2	SD	V2:C2	SD
Q1	69	19	68	15	**1.05**	0.3	52	12	87	24	**0.82**	0.2	**1.74**	0.5
Q2	64	17	120	28	**1.99**	0.7	52	13	69	17	**1.80**	0.4	**1.37**	0.4
Q3	66	16	165	35	**2.59**	0.8	59	16	66	19	**2.59**	0.6	**1.17**	0.4

Table 1. Mean durations (ms), duration ratios and standard deviations (SD) of the first-syllable consonant (C1) and vowel (V1), the inter-vowel consonant (C2) and the vowel of the successive syllable (V2) in Q1, Q2, and Q3 words

vital for classifying quantity degrees and, at the same time, more important for modelling speech temporal structure. Figures 5 and 6 present two CART-generated decision trees for quantity degree classification.

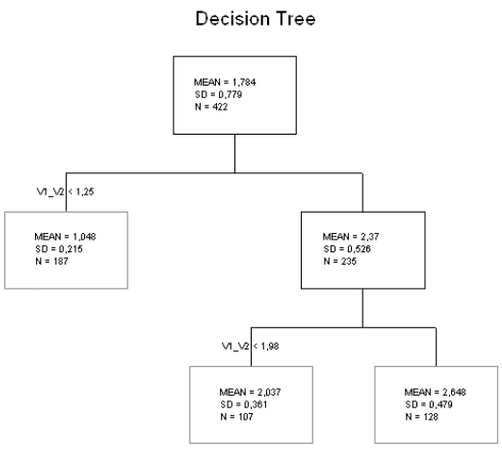

Figure 5. Decision tree based on the classical duration ratio of V1:V2

From Figure 5 we can see that for Q1 the duration ratio V1:V2 is less than 1.25, while for Q2 the values of the ratio range from 1.25–1.98 and the criterion for recognizing Q3 is a V1-to-V2 ratio that exceeds 1.98. Figure 5 reveals that a decision tree based on two duration ratios (V1:C1 and V2:C2) actually manages to classify the quantity degrees by using only one of them (V1:C1) as the sufficient criterion. This points to the relative length (short or long) of the first syllable vowel V1 and thus, to the primary (durational) division of the quantity degrees into short (Q1) and long ones (Q2, Q3). Consequently, for our material the V2:C2 criterion has proved marginal after all. This brings back memories of M.Hint's theory of syllabic quantity degree, arguing that although phonetic quantity is manifested in the foot, its degree depends on certain parameters of the stressed syllable (Hint 2001).

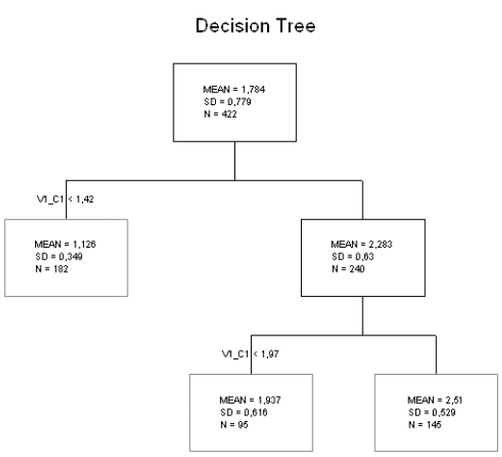

Figure 6. Decision tree based on a two-step system of the ratios V1:C1 and V2:C2

For weighing the relevance of duration ratios in the model of phonetic quantity some simple equations of linear regression were generated. For classical duration ratios the linear model yielded quite a strong correlation between the input and output (correlation coefficient r=0.867). Consequently the model explains over 75 % of the data variation (coefficient of determination r^2=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.

According to the above results, the classical duration ratio (V1:V2) is the most relevant parameter to be considered in modelling the temporal structure of Estonian speech. However, although the above parameter guarantees quite a close correlation between the model input and output, similar tests should be run for some other physical factors, such as, for example, intensity and pitch, to find out their possible role in the formation of important phonological oppositions. In the present study the line will be drawn at intensity.

3.3.2. Intensity

Our analysis of intensity did not reveal anything remarkable. In all Q1 words the mean intensity of articulation is 73 dB, the only exception being the first vowel of Q1 words, which is pronounced at 74 dB. on average. The results for Q2 and Q3 offer more material for discussion and even confusion as according to our measurements the average intensity for Q2 words is 74 dB for V1 and 71dB for V2, while for Q3 words the respective readings are 74 and 72 dB. According to Eek and Meister, however, the average intensity of articulating the V1 and V2 in Q2 words is 73 dB throughout, whereas in Q3 words the respective intensities are 75 and 69 dB (Eek, Meister 1997). Our study has not detected so big a difference in any comparison; neither does it support the argument that V1 is articulated with most intensity in the initial syllable of Q3 words. True, there is a difference between the V1 and V2 as pronounced in Q2 and Q3 words (2–3 dB), respectively, observed in total as well as when comparing different word groups (stressed/unstressed, male/female), but the difference is not salient enough. Obviously intensity may help perceive the difference between Q2 and Q3 words, serving as a supportive feature if the duration ratio and the pitch curve, for some reason or other, fail to define unambiguously whether the word is a Q2 or Q3 one. Such cases are obviously very few. M. Parve, who analysed spontaneous dialect speech (Parve 2003) also reached the conclusion that although the intensity difference between Q2 and Q3 words can be distinguished by phonetic criteria, it is dubious or too vague for perception.

4. Conclusions

The conducted test of speech rate perception did not provide an unambiguous answer to all set questions. There are, indeed, certain differences observable between the speech rates preferred by the blind and the sighted, but the level of the difference depends not on the visual impairment but rather on the subjects' experience with using a computer and a screen reader. The ability to understand rapid speech appears after about three years of everyday practice of using of a computer and listening to synthetic speech. No so-called optimal speech rate can be established either for the blind or the sighted, as the preferred speech rate is very individual and depends on many circumstances.

The aim of the study was to find out whether Estonian quantity degrees could be distinguished by any other features but the traditional duration ratio of V1:V2. Our analysis of copious data proved that neither intensity nor adjacent sound ratio are as relevant as the ratio of the first and second syllable sounds. Possibly, the actual role of intensity in the quantity degree model could be approached better by pitch analysis. When modelling speech temporal structure one should keep in mind that standard Estonian is characterized by an alternation of words of three different quantity degrees, based on a natural alternation of stressed and unstressed syllables. The quantity degrees can be distinguished by a comparison of the duration ratios of those syllables, but obviously this is not all there is to it. The next object of research relevant in this respect should be the manifestation and role of pitch.

Acknowledgements

This work has been supported by the National Programme for Estonian Language Technology, grant ETF7998 and project SF0050023s09.

References

1. *Asakawa, C., Takagi, H., Ino, S., Ifikube, T.* 2003. Maximum Listening Speeds for the blind. Proceedings of the 20003 International Conference on Auditory Display. Boston, MA, USA, 276–279.
2. *Asu, E. L., P. Lippus, P. Teras, T. Tuisk*. 2009. The Realization of Estonian Quantity Characteristics in Spontaneous Speech. *Nordic Prosody — Proceedings of the Xth Conference*, Helsinki 2008. Editors Aaltonen, O., Aulanko, R., Vainio, M. Frankfurt: Peter Lang Verlag, 49–56.
3. *Boersma, P., D. Weenik*. Praat: doing Phonetics by computer (http://www.fon.hum.uva.nl/praat/)
4. *Eek, A.* 1983. Kvantiteet ja rõhk eesti keeles (I). Fonoloogiliste tõlgenduste kriitikat. — *Keel ja Kirjandus 9*, 481–489.
5. *Eek, A., E. Meister,* 1997. Simple Perception Experiments on Estonian Word Prosody: Foot Structure vs. Segmental Quantity. — *Estonian Prosody: Papers from a Symposium*. Editors Lehiste, I., Ross, J. Tallinn: Institute of Estonian Language, 77–99.
6. *Eek, A., Meister, E.* 2003. Foneetilisi katseid kvantiteedi alalt. — *Keel ja Kirjandus* 11–12, 815–837; 902–916.
7. *Hint, M.* 2001. Prosoodiaväitlustes läbimurdeta. — *Keel ja Kirjandus* 3–5, 164–172, 252–258, 324–335.
8. *Keller, Eric* 2007. Waves, beats and expectancy. — Proceedings of the 16th International Congress of Phonetic Sciences (eds. Jürgen Trouvain, William J. Barry). Saarbrücken, 6–10 August 2007. Saarbrücken, 355–360.
9. *Keller, Eric, Port, Robert* 2007. Speech timing: approaches to speech rhythm. — Proceedings of the 16th International Congress of Phonetic Sciences (eds. Jürgen Trouvain, William J. Barry). Saarbrücken, 6–10 August 2007. Saarbrücken, 327–329.
10. *Krull, D.* 1991. Stability in some Estonian duration relations. — Experiments in speech processes. PERILUS (Phonetic Experimental Research, Institute of Linguistics, University of Stockholm) No XIII, 57–60.
11. *Lehiste, I.* 1960. Segmental and Syllabic Quantity in Estonian. *American Studies in Uralic Linguistics Vol 1*. Bloomington, Indiana University, 21–28.
12. *Lehiste, I.* 1997. Search for phonetic correlates in Estonian Prosody. — Estonian Prosody: Papers from a Symposium, Proceedings of the International Symposium on Estonian Prosody, Lehiste, I.; Ross, J. (eds.). Tallinn, Estonia, October 29–30, 1996. Institute of the Estonian Language and Authors, Tallinn, 11–35.
13. *Liiv, G.* 1961. Eesti keele kolme vältusastme kestus ja meloodiatüübid. — *Keel ja Kirjandus* 7–8, 412–424, 480–490.
14. *Mihkla, M.; Meister, E.* 2002. Eesti keele tekst-kõnesüntees. Keel ja Kirjandus, 45(2); 88–97 ja 45(3); 173–182.
15. *Moos, A., Trouvain, J.* 2007. Comprehension of Ultra-Fast Speech — Blind vs „Normally Hearing" Persons., 677–680.
16. *Moos, A., Hertrch, I., Dietrich, S., Trouvain, J., Ackermann, H.* 2008. Perception of Ultra-Fast Speech by a Blind Listener — Dos He Use His Visual System? Proceedings of the 8th Seminar on Speech Production, ISSP 2008, 297–300.
17. *Parve, M. 2003*. Välted lõunaeesti murretes. Doktoritöö. Dissertationes philologiae estonicae universitatis tartuensis 12. Tartu: Tartu Ülikooli Kirjastus.
18. *Ross, J. Lehiste, I.* 2001. The temporal Structure of Estonian Runic Songs. *Phonology and Phonetics 1*. Editor A. Lahiri. Berlin; New York: Mouton de Gruyter.