XXVII Fonetiikan päivät 2012

Phonetics Symposium 2012

Proceedings

Tallinn 2012

TTÜ KÜBERNEETIKA INSTITUUT
Institute of Cybernetics at TUT
Electronic publications

XXVII Fonetiikan päivät 2012

Phonetics Symposium 2012

17-18 February 2012, Tallinn, Estonia

Proceedings

Einar Meister (ed.)

TUT PRESS

Tallinn 2012

XXVII Fonetiikan päivät 2012 – Phonetics Symposium 2012
17-18 February 2012, Tallinn, Estonia
Proceedings

Edited by Einar Meister

Cover page photo by Allan Alajaan
Source: Tallinn City Tourist Office & Convention Bureau

Electronic publications
Institute of Cybernetics at Tallinn University of Technology
Akadeemia tee 21, 12618 Tallinn, Estonia
http://www.ioc.ee

TUT Press
Akadeemia tee 1, 12618 Tallinn, Estonia

ISBN 978-9949-23-385-4(PDF)

Sponsored by the Estonian Center of Excellence in Computer Science, EXCS (funded mainly by the European Regional Development Fund), and by the Ministry of Education and Research

EXCS
ESTONIAN COMPUTING

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti tuleviku heaks

Haridus- ja Teadusministeerium
Estonian Ministry of Education and Research

Content

Preface

XXVII Fonetiikan päivät 2012 – Phonetics Symposium 2012 continues the tradition of meetings of Finnish phoneticians started in 1971 in Turku. These meetings, held in turn at different universities in Finland, have been frequently attended by Estonian phoneticians as well. In 1998 the meeting was held in Pärnu, Estonia, and in 2012 it took place in Estonia for the second time.

The meeting was attended by 52 participants – 35 from Finland, 15 from Estonia, 1 from Sweden and 1 from Hungary. There were 22 oral and 12 posters presentations in the program covering a wide variety of topics from basic phonetic research to recent developments in spoken language technology. However, the proceedings contain only 14 full papers, the abstracts of all presentations and the program of the symposium are available at http://www.ioc.ee/fp2012/.

The special opening session was devoted to Ilse Lehiste, Matti Karjalainen and Arvo Eek to commemorate and honor their scientific contributions to Estonian and Finnish phonetics and speech technology.

The symposium was hosted by the Laboratory of Phonetics and Speech Technology, Institute of Cybernetics at Tallinn University of Technology (IoC) and organized in co-operation with the Estonian Centre of Excellence in Computer Science, EXCS, and supported by the Ministry of Education and Research.

Tallinn, October 2012

Einar Meister

Quantity in MOKSHA-MORDVIN

Niina Aasmäe, Karl Pajusalu, Nele Salveste, Tatiana Zirnask

Institute of Estonian and General Linguistics, University of Tartu, Estonia

niina.aasmae@ut.ee, karl.pajusalu@ut.ee, nele.salveste@gmail.com, tatiana@ut.ee

Abstract

This article provides empirical data concerning the prosody of the Moksha language: it evaluates the effect of the position of stress, the characteristics of word structure, and the position of a word in an utterance upon the temporal relationship between the syllable nuclei. Acoustical analyses were carried out using a set of target words embedded in frame sentences, which were read by eight speakers of Central Moksha. As an extension to the main part of the analysis, a comparison of the data for Moksha and results obtained earlier in an analogous study of vowel durations in Erzya is offered.

Index Terms: the Moksha and Erzya languages, stress, intrinsic duration, syllable, vowel reduction

1. Introduction

Written Erzya- and Moksha-Mordvin are mutually intelligible to a fairly high extent. As far as oral communication between Erzyans and Mokshans is concerned, it is hindered because of phonetic and phonological differences between the languages. It has been noted in earlier works on the phonetics of the Mordvin languages, especially following the observations of Heikki Paasonen [1], that the manifestations of stress and the system of vowels in Erzya and Moksha diverge to a considerable extent. Systematic research on the prosodic features of the Mordvin languages in the framework of acoustic phonetics has begun only since the turn of this century; see, for example: [2], [3], [4], [5], [6], [7], [8]. Empirical data are hitherto available mostly on Erzya. It should be noted that the first acoustic data of Moksha were published in the 1960–1970s: [9], [10]. A. Sovijärvi has examined the formant structure of the schwa vowel (*ə*); the theme has been lately treated in [11]. In S. Z. Devayev's works the conditioning role of stress upon vowel durations has been demonstrated in a small set of words read by one informant. Vowel duration data based on a fairly large-sized set of words have been provided of late by T. Zirnask (see [7]); observations have been made on a South-Western variety of Moksha spoken in Novo-Badikovo, a locality in the Zubova Polyana region of the Republic of Mordovia. As material, target words embedded in a frame-sentence were used. Vowel durations in the test words read by two informants showed dependence on the position of stress.

In the case of Moksha, duration can be, a priori, regarded a cue to stressedness for two reasons. Firstly, the length of the segments is not contrastive in the language – there are neither long vowels and diphthongs nor geminates; hence, duration can serve as a phonetic cue to stressedness. Secondly, Moksha is characterized by the presence of vowel reduction, which is considered to be a phenomenon conditioned by duration-based stress; see, for example: [12], [13], [14]. In view of these typological maxims, the aim of the present paper consists in providing evidence to attest to the conditioning role of stress for vowel durations in Moksha. Analyses presented in the paper are based on the core variety of Moksha, namely, the Central dialect group, which is the prototype of the literary language. As an extension to the main part of the analysis, a comparison of data obtained by now on the temporal characteristics of stressed and unstressed vowels for both Moksha and Erzya is offered.

2. Data analysis

2.1. Material and procedures

For scarcity of data on Moksha, only controlled speech was used for observations. Acoustic measurements were made on sub-dialects referred to Central Moksha, the prototype dialect of the written language. Recordings were made of target words read by 6 female and 2 male informants. Below, the initial letters of the informants' names, age and place of residence are indicated.

- Mordovskaya Kozlovka, Kargońžaj, of the Atyuryevo region:

  JM, a female aged 50; NN, a female aged 39; IS, a male aged 56; NM, a male aged 54
- Polskoye Tsibayevo, Pakśań Porańa, of the Temnikov region:

  JS, a female aged 53
- Mordovskiye Parki, Mokšeń Parka of the Krasnoslobodsk region:

  ST, a female aged 41
- Zaitsevo of the Kovylkino region:

  TR, a female aged 40; VT, a female aged 34

The test words were embedded in a frame sentence, in which they occurred in a phrase- and sentence-final position, as shown below:

Märgəĺiń ***maksă***, *af* ***śijä***. 'I would say *liver*, not *silver*'.
Märgəĺiń ***śijä***, *af* ***maksă***. 'I would say *silver*, not *liver*'.

The total number of tokens used for the analysis was 1664. The material was recorded by T. Zirnask in 2008. The acoustic measurement procedures were carried out using the PRAAT-software [15]. In the test words composed of one to six syllables the main structural types (CV, CVC) with variations (like CCVC, CVCC, etc.) occurred. The vowel segments that occurred in the material (*a, o, u, i, e, ä, ə*) correspond to those identified in the inventory of the Central dialects and written Moksha; see, for example: [16], [17]. Prior to the measurement procedures, the location of stress in the test-words was assigned through repeated listening (by T. Zirnask, a native Moksha and N. Aasmäe, a native Erzya). In the majority of cases (nearly 90 percent), stress was marked on first syllables. Cases of non-initial stress included words like *śijä* ‘silver’, *kundasamak* ‘you will catch me’ having a high vowel (*i, u*) in the first syllable and a low vowel (*a, ä*) in a subsequent syllable. The location of stress in such words, however, alternated. A same speaker could pronounce them with initial stress in one utterance and non-initial stress in another utterance. In words with more than two syllables, e.g.: *ajďasajńĕ* ‘I chase them’, *kundańďeŕasamak* ‘if you catch me’, additional stress was present and such polysyllabic words tended to be uttered with stress on either odd- or even-numbered syllables. The phenomena of stress alternation and of additional stress in Moksha have not been the subject of systematic research; some authors ([18], [19], [20]), though, have mentioned variation that occurs in the placement of stress in spontaneous speech and poetry, especially folk songs.

2.2. Results

Duration variation may be conditioned, according to Ilse Lehiste [21], by such factors as “...the phonetic nature of the segment itself (intrinsic duration), preceding and following sounds, other suprasegmental features, and position of the segment within a higher-level phonological unit.” To check the influence of the prosody of an utterance upon the relationship between the vowel durations of a word, the words were embedded in two positions within the frame sentence, as mentioned above. Evaluation of the temporal relationship between stressed and unstressed syllable nuclei in words differing by the number of syllables constitutes the central part of the analysis. Though the conditioning role of consonant environment upon vowel duration in Moksha has not been examined it could be envisaged; to avoid bias in the results of measurement, the word corpus used for analysis was compiled so that varied consonant segments appeared in different positions within the words. In the analysis, the effect of the intrinsic duration of vowels, the effect of the syllable type, and the position of stress in the test words have been also taken into consideration.

2.2.1. Duration of vowels in stressed and unstressed syllables

Table 1 shows data for disyllabic and trisyllabic words marked with stress on first syllables. Disyllabic words of varied structure placed in the phrase-final position, displayed an asymmetry between the duration of the stressed and unstressed syllable nuclei (V1, V2) – longer vowel duration was observed in the stressed syllables. The difference between the values of vowel duration was statistically highly significant (*p*=1E-23, F=106.3, Fcrit. = 3.85). In the sentence-final position the difference between the vowel durations was not significant owing to the effect of pre-boundary lengthening. As it could be expected, the ratios of duration, vowel one to vowel 2 (V1/V2), across the observations in the phrase-final and sentence-final positions significantly differed (*p*=1E-16, F=71.23, Fcrit. = 3.85).

In trisyllabic words, asymmetry between the vowel durations of the stressed and unstressed syllables (V1 and V2) was salient in both phrase- and sentence-final positions (*p*=6E-23, F=123, Fcrit. =3.88; *p*=4E-14, F=64.76, Fcrit=3.88). The mean ratio of duration between the stressed and unstressed syllable nuclei in trisyllabic words was higher than in disyllabic words both in phrase- and sentence-final positions. This difference is statistically highly significant (p = 4.66E-13, F =53.91, Fcrit. =3.852).

It can be noticed that vowels in the final syllable of trisyllabic words, especially in the sentence-final position, are longer than those in the preceding unstressed syllable. This observation allows suggesting that trisyllabic words receiving additional stress, which appears to be duration-based, are likely to constitute two feet – a disyllabic foot and a so-called degenerate foot, the latter having the potential of becoming a di-syllabic foot. As an example, the words *aˈjďămaˈjt’* ‘you chased me’ and *aˈjďăsaˈjńə* ‘I chase them’, with stress on odd-numbered syllables, can be given. The realization of trisyllabic words as two feet in the sentence-final position is even more salient than in the phrase-final position: the length of vowels in the third syllable is comparable to that in the first syllable, to which the effect of pre-boundary lengthening is likely to contribute.

Table 1. *Mean values of vowel durations (V1, V2, V3), in ms, and mean duration ratios between stressed and unstressed syllable nuclei (V1/V2), with values of standard deviation: di- and trisyllabic words with stress on odd-numbered syllables occurring in a phrase- and sentence-final position. Statistically significant differences between the values are starred: *(p<0.05), **(p<0.01), ***(p<0.001).*

disyllabic words	V1(*ms*)	V2(*ms*)	V1/V2	V3(*ms*)
phrase-final n=438	***116.0 28.7	***97,5 24,3	1.2 0.4	
sentence-final n=433	114.2 27.6	112.5 25.6	1.1 0.3	
trisyllabic words				
phrase-final n=110	***106.1 22.4	***74.7 19.6	1.5 0.5	85.9 25.1
sentence-final n=117	***105.8 25.5	***80.4 22.8	1.4 0.4	104.6 24.8

The effect of additional stress upon the relationship between vowel durations was further examined in a subset of words with 4 to 6 syllables, in which stress was also perceived on first and third syllables. In the data for phrase-final words, statistically significant differences were found between vowel durations in the first and second syllables (*p*=0.01, F=6.69, Fcrit.=3.98), as well as those in the third and fourth syllables (*p*=0.02, F=6.13, Fcrit.=3.98). In the sentence-final position, due to pre-boundary lengthening that affected the duration of vowels throughout the word, differences in the vowel durations were not statistically significant. As polysyllabic words are under-represented in the corpus, this issue will have to be revisited on a broader material.

In the case of words with stress on even-numbered syllables, for example, the second syllable of a trisyllabic word, a disyllabic foot could be formed by the second and third syllables, like *kundaˈĺit’* ‘you would catch it/him/her’. As to the initial syllable that receives no stress, in words like *kundaˈĺit’*, it could constitute a foot with the preceding syllable of the word in the carrier sentence, namely: *mäˈrgəĺiˈń kundaˈĺit’*.

The relation between vowel durations in words with stress on the second syllable was asymmetric, as well (see table 2), mean duration ratios being higher (1.9 in the phrase- and 1.7 in the sentence-final positions) compared to those observed in the cases of stress on odd-numbered syllables.

According to the results considered above, the place of stress in a word, the number of syllables constituting a word, and the position of a word in an utterance can be considered to be factors conditioning the relationship between vowel durations in a word.

Table 2. *Mean values of vowel durations (V1, V2, V3), in ms, and mean duration ratios between stressed and unstressed syllable nuclei (V2/V3), with values of standard deviation: trisyllabic words with stress on the second syllable occurring in a phrase- and sentence-final position.*
*Statistically significant differences between the values are starred: *(p<0.05), **(p<0.01), ***(p<0.001).*

trisyllabic words	V1(*ms*)	V2(*ms*)	V3(*ms*)	V2/V3
phrase-final n=25	50.72 11.47	***138.3 21.7	***77.0 15.6	1.9 0.6
sentence-final n=27	50.41 10.65	***143.3 22.5	***86.0 16.8	1.7 0.4

2.2.2. *The conditioning role of the intrinsic duration of vowels and of the openness / closedness of a syllable upon the duration of syllable nuclei*

In the subsequent part of the analysis, vowel durations in stressed and unstressed syllables were compared in words with vowels that have different intrinsic duration. In Table 3, values of duration for high and low vowels in monosyllabic words are shown. The difference between the durations of *i* and *a* in the productions of the monosyllabic words was statistically significant (*p*=0.001, F=12.45, Fcrit=4.13 for the phrase-final position; *p*=0.002, F=11.64, Fcrit=4.11 for the sentence-final position).

Table 3. *Mean vowel durations, V (ms) with values of standard deviation in words with high and low vowels as syllable nuclei (stress marked as a dot).*

Phrase-final / Sentence-final:

monosyllabic	*V(ms)*	*V (ms)*

ši, šit';
n=35
	124.1
	27.9
		135.7
		25.1

va, vaj n=38	156.8 23.8	165.9 28.4

This result implies that a high vowel occurring in a stressed syllable might not be longer than a low vowel occurring in an unstressed syllable in Moksha. Data for disyllabic words are well in accordance with this assumption. The duration of *i* and *ä* in the disyllabic word *śijä* 'silver' differed depending on the location of stress. In the cases of stress on the first syllable, *i* was somewhat shorter than *ä* (114.5ms *vs* 120.5ms in the phrase-final position), while in the word marked with stress on the second syllable, the duration of *ä* by far exceeded that of *i* (101.4ms *vs* 172.0ms in the phrase-final position). These data show the effect produced by differences in the intrinsic duration of vowels upon the temporal relationship between the nuclei of stressed and unstressed syllables in a word.

Less transparent is the relation between the duration of syllable nuclei and the openness or closedness of the syllables. Moksha is characterized by the occurrence of two and more intervocalic consonants – a circumstance that posits the question of syllabification, which has not been studied for Moksha.

In the process of analysis it was noticed that, for example, an epenthetic vowel could appear between two consonants in an intervocalic position and, as a result, a disyllabic word was produced as a trisyllabic one (e.g.: *eŕgĕ* > *eŕigĕ* 'a bead', *śeĺmə* > *śeĺəmə* 'an eye', *käd'ga* > *käd'iga* 'from hand to hand'). For the evaluation of overall data in this work, it was important to find out whether the length of vowels in syllables ending in a vowel, on the one hand, and a consonant or a combination of consonants, on the other hand, differs. For this purpose, vowel durations in subsets of monosyllabic and disyllabic words were first analysed. Word-final vowels (number of observations – 57) were longer than those followed by a single consonant or a combination of two consonants (number of observations –156) in both positions of the utterances, mean durations being 154.0*ms vs* 141.5*ms* in the phrase-final position and 161.0*ms* vs 149.5*ms* in the sentence-final position, respectively. Statistically significant difference between vowel durations in the two series of words was observed in the data for the phrase-final position (*p*=0.04, F=4.19, Fcrit.=3.93). Vowels followed by two consonants were just slightly shorter than those followed by a single consonant.

Next, the duration of vowels (V1 and V2) was compared in four series of disyllabic words with different structure. As seen from Boxplots 1, there was an asymmetry in the duration of the stressed and unstressed syllable nuclei. Vowels in the stressed syllable were longer than in an unstressed syllable, the difference being more salient in the words with stress on the second syllable. As to the effect of the openness/closedness of the syllables constituting the words, the current analysis shows that V1 in the series of words with intervocalic combinations of consonants was shorter than in the series of words with a single intervocalic consonant. It might signify that the first syllable was of a closed type; hence, in a syllable ending in a consonant, vowels are shorter than in a syllable ending in a vowel. In a follow-up study, a detailed comparison of word series with controlled types of intervocalic combinations of consonants should be done to gain insights into the question of syllable boundary in Moksha.

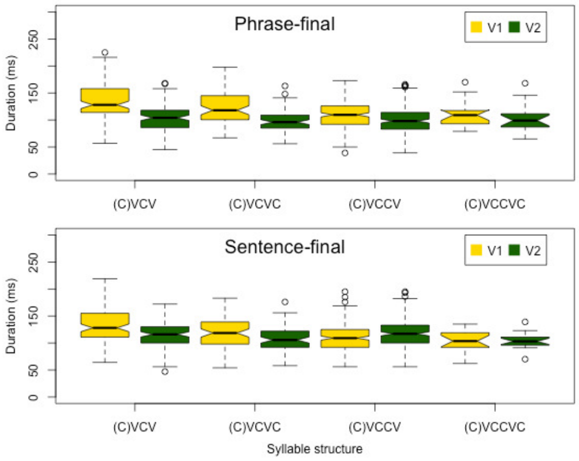


Figure 1. *Duration of V1, V2 in four series of disyllabic words produced with stress on the first syllable.*

3. Interim conclusions

The subsets of data for the Central dialects of Moksha considered above show that the temporal relationship between the syllable nuclei in a word is conditioned by stress; namely, stressed syllable nuclei tend to be longer than unstressed syllable nuclei. Under the influence of the position of a word in an utterance or due to differences between the intrinsic duration of high and low vowels in the stressed and unstressed syllables, asymmetry between the duration of vowels can be reverse, or vowel durations can be equalized. Analysis of data for trisyllabic words shows that asymmetry between the duration of stressed and unstressed syllable nuclei is more salient than in disyllabic words. Words with three or more syllables receive additional stress, which contributes to the rhythmic patterning of a word as a sequence of disyllabic feet. The effect of the openness/closedness of a syllable was apparent in monosyllabic words; in a disyllabic word, or a foot, possible effects of syllable composition seem to be opaque. There was some reduction in the mean vowel duration ratios for words with intervocalic combinations of two consonants, compared to cases of a single consonant at the syllable boundary. The issue of intervocalic combinations of consonants requires further examination for establishing the syllable boundary in different types of syllable composition.

4. Comparison of vowel duration data: dialects of Moksha and Erzya

Vowel duration data of Moksha available at present on the Central dialect (8 speakers) and a South-Western sub-dialect (2 speakers) have been obtained on a same word corpus and proved to be compatible in all the parts of analyses. For this reason, the results presented above will be only referred to in the comparison of data for Moksha and Erzya. As far as evidence on Erzya is concerned, the results of research based on inter-dialect data of vowel durations will be used (see in [4]). The previous work, in which vowel duration data of Erzya have been provided (see in [2]), did not aim at finding out dialect differences in the temporal characteristics of the informants' speech. Regarding the relatedness of duration to stress, the authors' conclusion was: "The results concerning stress are the most interesting – perhaps partly so because they are somewhat ambiguous and point toward directions that future research might take. Neither duration nor pitch serve as reliable stress cues." ([2], p. 85).

In the subsequent research, thus, inter-dialect data obtained from the analyses of both controlled and spoken speech were used, as input; results in the two parts of analysis were found to be compatible. They differentiated between four main varieties of Erzya, as shown in Table 4. The relationship between stress and vowel duration in the variety of Erzya marked in the table as group1 (the prototype of the written language) is different from that found for dialect groups 2, 3, 4. Stress and vowel duration can be regarded as relatively independent in dialect group 1, characterized by lack of vowel reduction and alternation of the position of stress. Mean duration ratios V1/V2 for di- and trisyllabic words did not reveal significant differences. Data for the Erzya dialects in groups 2, 3, 4, characterized by different patterns of vowel reduction and dominance of stress on the first syllable, are analogous to the data of Moksha. Stressed syllable vowels were found to be significantly longer than unstressed syllable vowels in both di- and tri-syllabic words. Vowel duration in the speakers' productions thus depends on stress. In tri-syllabic words, asymmetry between vowel durations in the stressed and unstressed syllable is greater than in di-syllabic ones.

Table 4. *Inter-group data of mean vowel durations of (v1, v2, v3), ms, and mean duration ratios (v1/v2) with values of standard deviation in di- and trisyllabic words (stress on first syllable). Significant differences between v1 and v2 are starred: *(p<0.05), **(p<0.01), ***(p<0.001).*

Dialect group	*n*=	V1 (*ms*)	V2 (*ms*)	V1/V2	*n*=	V1 (*ms*)	V2 (*ms*)	V3 (*ms*)	V1/ V2
1	107	108.67	109.31	1.01	57	89.79	86.53	90.88	1.06
		15.86	15.60	0.16		17.14	13.27	16.94	0.27
2	93	***107.86	***97.37	1.13	38	***97.45	***76.82	83.11	1.30
		18.85	16.66	0.22		21.82	12.70	22.98	0.34
3	54	**121.91	**110.61	1.15	40	**91.68	**75.63	88.40	1.26
		21.08	27.44	0.28		22.41	19.13	25.74	0.34
4	56	***111.55	***94.73	1.21	30	***93.74	***70.33	82.33	1.38
		22.49	17.02	0.28		20.69	13.72	23.14	0.37

5. Conclusion

Vowel duration data for Moksha and Erzya display a continuum of temporal characteristics of stressed and unstressed syllable nuclei. On one side, there are vowel durations that are independent of stress, on the other side there is an asymmetry between vowel durations in a disyllabic foot conditioned by stress. Within this continuum are Erzya dialects that display a mixture of the characteristics of the two sides. Findings presented in this paper attest to some of the well known ideas discussed over time in literature; the data are novel and they allow treating the prosody of Erzya and Moksha on the basis of empirical evidence.

6. References

[1] Paasonen, H., ”Mordvinische Lautlehre”, in Suomalais-ugrilaisen seuran toimituksia XXII, Helsingfors 1903.

[2] Lehiste, I., Aasmäe, N., Meister, E., Pajusalu, K., Teras, P. and Viitso, T.-R., ”Erzya prosody”, in Suomalais-ugrilaisen seuran toimituksia 245, Helsinki, 2003.

[3] Estill, D., ”Diachronic change in Erzya word stress”, in Suomalais-Ugrilaisen Seuran Toimituksia 246, Helsinki, 2004.

[4] Aasmäe, N., “Sources of variability in the duration of stressed and unstressed syllable nuclei in Erzya: inter-idiolect data of spontaneous speech”, in Linguistica Uralica, XVII, 2, 81–93, 2006.

[5] Aasmäe, N., “Duration-dependent undershoot and phonological vowel reduction in Erzya”, in V. Meiliunaite and A. Leskauskaite (Ed.), The sound and aspects of its research: methodology and practice, Vilnius 2007, 17 – 33, Vilnius, 2009.

[6] Aasmäe, N. and Ross, J. “Where is the syllable boundary in Erzya-Mordvin?” in Н. Д. Светозарова (Ed.), Фонетика. Материалы XXXVII Международной филологической конференции, 11-15.03.2008, Санкт-Петербург, 3-10, Факультет филологии и искусств Санкт-Петербургского государственного университета, 2008.

[7] Zirnask, T. “Rõhk ja kestus mokša keele Kesk-Vadi murdes“, in Journal of Estonian and Finno-Ugric Linguistics 1, 99 – 111, 2010.

[8] Aasmäe, N., Pajusalu, K., Zirnask, T., “Variability of stress assignment and vowel durations in Erzya and Moksha”, in Congressus XI Internationalis Fenno-Ugristarum, Piliscsaba 2010, IV, 9 – 17, Piliscsaba, Reguly Társaság, 2011.

[9] Sovijärvi A., “Der mokschamordwinische ə- Vokal im Lichte der Sonagramme“ in Helsingin yliopiston fonetiikan laitoksen julkaisuja 16, 553–566, Helsinki, 1963.

[10] Devayev 1975 = Деваев, С. З., “Словесное ударение в мокша-мордовском языке“, in Congressus Tertius Internationalis Fenno-Ugristarum, Tallinae Habitus 17.–23. VIII 1970, Pars I, 481–483, Tallinn, 1975.

[11] Estill, D., “The enigmatic central vowel in Moksha. How central, how reduced?” in S. Werner and T. Kinnunen (Ed.), XXVI Fonetiikan päivät 2010, 33–37, ISBN 978-952-61-0391-4 (PDF), Joensuu, 2011. http://epublications.uef.fi/

[12] Bybee, J., Chakraborti, P., Jung, D. and Scheibman, J., “Prosody and segmental effect. Some paths of evolution for word stress”, in Studies in Language, 22:2, 267–314, 1998.

[13] Crosswhite, K., “Vowel Reduction”, in Hayes, B., Kirchner, R., Steriade, D. (Ed.), Phonetically-Based Phonology, Cambridge University Press, 2004.

[14] Barnes, J., “Strength and weakness at the interface: positional neutralization in phonetics and phonology”, Berlin/New York: Mouton de Gruyter, 2006.

[15] Boersma, P., Weenik, D., “PRAAT, a system for doing phonetics by computer”, 2007, http://www.praat.org.

[16] Feoktistow, A. P., “Die Dialekte der mordwinischen Sprachen“, in M. Kahla (Ed.), H. Paasonens mordwinisches Wörterbuch, Band I, XXXI–CV, Helsinki, Suomalais-Ugrilainen Seura, 1990.

[17] Ivanova 2006 = Иванова, Г. С., “Система гласных в диалектах мокшанского языка в историческом освещении“, Саранск, 2006.

[18] Azrapkin 1966 = Азрапкин, Ю. И., “Колопинский говор мокша-мордовского языка“, in Очерки мордовских диалектов IV, 251–280, Саранск, 1966.

[19] Lomakina 1966 = Ломакина, Т. И., “Городищенский диалект мокша-мордовского языка“, in Очерки мордовских диалектов IV, 289–329, Саранск. 1966.

[20] Feoktistov 1979 = Феоктистов, А. П. “О характере словоупотребления в начальной стопе поэтической строки (к проблеме становления силлабо-тонической версификации в мокшанской поэзии)“, in Финно-угристика, Вып. 2, 124–149, Саранск, 1979.

[21] Lehiste, I., “Suprasegmental Features of Speech” in Contemporary Issues in Experimental Phonetics, 225-239, New York, San Francisco, London, Academic Press, 1976.

The acoustic characteristics of monophthongs and diphthongs in the Kihnu variety of Estonian

Eva Liina Asu, Pärtel Lippus, Ellen Niit, Helen Türk

Institute of Estonian and General Linguistics, University of Tartu

eva-liina.asu@ut.ee, partel.lippus@ut.ee, ellen.niit@ut.ee, helen.tyrk@gmail.com

Abstract

This paper presents the first acoustic study of monophthongs and diphthongs in the Kihnu variety of Estonian. The focus is on six diphthongs which have arisen after the diphthongization of long open and mid vowels. The comparison of their components with monophthongs revealed that the target values of the diphthongs are close to the corresponding monophthongs. There was some variation due to coarticulatory influences between the diphthong components. It was shown that the duration of both target vowels in the diphthong is longer in the third quantity than in the second quantity. The analysis also revealed that there is an acoustic basis for postulating the existence of two triphthongs in Kihnu. Formant trajectory length proved to be a useful measure for comparing monophthongs, diphthongs and triphthongs.

Index terms: vowel quality, diphthongs, triphthongs, formant trajectory length, Kihnu, Estonian dialects

1. Introduction

Regional varieties of Estonian exhibit considerable variation as to the realization of their vowel systems. One of the richest inventories is that of the variety of Kihnu, which belongs to the Insular dialects of the North Estonian dialect group, and is spoken by about 600 people on the Island of Kihnu. Due to the relative isolation and smallness of the variety several features that are rare or not present elsewhere in the North Estonian dialect area have been preserved. Kihnu is, for instance, unique among the North Estonian dialects in that it has retained vowel harmony involving four different vowels: /æ/ (e.g. /kylæ/ 'village'), /ø/ (e.g. /pøgø/ 'ghost'), /y/ (e.g. /sygyse/ 'in autumn') and /ɤ/ (e.g. /pɤlɤ/ 'is not') [1]. Another characteristic, which does not occur in other North Estonian dialects (although is common in South-East Estonian dialects), is the change of the front vowels /i/ and /æ/ in the first syllable into /ɤ/ and /ɑ/ respectively, with the accompanying /j/-glide, e.g. /linɑ/ > /ljɤnɑ/ 'flax', /nægu/ > /njɑgu/ 'face' [2, 3]. This phenomenon only takes place in words which have a back vowel in the second syllable.

Kihnu vowel system is also characterized by the diphthongization of long vowels. This general feature of the North Estonian dialect group is rare among Insular varieties, appearing only in Eastern Saaremaa and Muhu. Diphthongization in Kihnu involves long open and mid vowels /æ, ɑ, e, ø, ɤ, o/ resulting in diphthongs which have mostly been written down as *iä, ua, ie, üe, õe, uõ* (e.g. /sæːːr/ > /siæːr/ 'leg', /mɑːː/ > /muɑː/ 'land', /keːːl/ > /kieːl/ 'tongue', /tøːː/ > /tyeː/ 'work', /vɤːrɑs/ > /vɤerɑs/ 'stranger', /koːːli/ > /kuɤːli/ 'to school'). Long close vowels /i, y, u/ have been preserved in all words [4]. Despite widespread diphthongization long open and mid vowels occur in certain words and word forms, in particular newer vocabulary and loan words [5]; least complete is the diphthongization of long /ø/ and /ɤ/ [1]. Thus, the Kihnu variety has both long monophthongs and diphthongs.

Of particular interest are the diphthongs that have been formed as a result of the diphthongization of long /ɑ/ and /o/. While for other diphthongs dialectal transcriptions are uniform, there is notable variation for these two. The diphthong /uɑ/ corresponding long /ɑ/ has also been transcribed as *uä* or *ua*^{c} (e.g. *puät, pua*^{c}*t* 'boat' sg. nom.) or *oa* (e.g. *puadi~poadi* 'boat' sg. gen.), and the diphthong /uɤ/ corresponding long /o/ as *ue* or *uõ*^{c} (e.g. *kuel, kuõ*^{c}*l* 'school' sg. nom.) or *uo* (e.g. *kuoli* 'school' sg. gen.) [5, 6]. The transcription *uä* was first used by Saareste [7], and is in itself curious because this diphthong does not occur in any other Estonian variety [3]. It is obviously not easy to judge the quality of a sound solely by auditory impression, but in this case an additional reason for varying transcriptions lies in palatalization [5, 2, 8]. In words where a diphthong is followed by a palatalized consonant, an epenthetic vowel is inserted before the consonant after the diphthong [2]. The result is a triphthong. Alternatively, the formation of triphthongs has been explained by consonant fission [9]. According to Sang [8] the phonological system of Kihnu contains the triphthongs /uɑe/ and /uɤe/. This is a generally accepted view among the speakers of Kihnu. The textbook on Kihnu morphology uses three adjacent vowel letters to spell the words with triphthongs (e.g. *puaek* 'lighthouse') [10]. In other publications in Kihnu, also variants with the superscript *e* are used. Triphthongs have also been transcribed in the dialectal texts for Muhu [1] and Leivu [11].

The existence of triphthongs has not been studied acoustically. In fact, the vowel system of Kihnu has only been described in the accounts of traditional dialectological studies (e.g. [5, 1]). The only acoustic analysis of Kihnu vowels is a small-scale study focusing on the quality of short vowels in stressed and unstressed syllables based on data from one elderly female subject [12].

The present paper aims to address this gap in our knowledge and focus on the acoustic study of the six diphthongs which have formed due to the diphthongization of long open and mid vowels. We are interested in both the quality and durational characteristics of diphthong components. Our hypotheses are derived from earlier acoustic studies on Standard Estonian diphthongs, of which there are only a handful. Firstly, based on Lehiste [13] we hypothesize that the diphthong components are similar in their quality to the corresponding monophthongs, although following Piir [14] some variation is likely to occur. Secondly, also based on Lehiste [13] we would expect the duration of both vowels in a diphthong to be longer in the third quantity (Q3) than in the second quantity (Q2). There are, however, contradicting views on this matter. It has also been suggested that it is the duration of the second element of the diphthong that differentiates between Q2 and Q3, whereas the first element is not influenced by the phonological quantity and is similar in both quantities [2]. Piir's [14] results imply that the durations vary depending on the diphthong, but the second components are for the most part longer than the first components. Finally, we hypothesize that there is an acoustic basis for postulating the existence of two triphthongs in Kihnu.

2. Materials and method

The data set for the present analysis comprised disyllabic test words containing 9 monophthongs /ɑ, e, i, o, u, ɤ, æ, ø, y/ represented by the letters *a, e, i, o, u, õ, ä, ö, ü* and 6 diphthongs /iæ, uɑ, ɤe, uɤ, ie, ye/ spelled as *iä, ua, õe, uõ, ie, üe*. The test words for monophthongs were in all three quantity degrees (short (Q1), long (Q2) and overlong (Q3)), and those for diphthongs in Q2 and Q3 (as diphthongs cannot be in the short quantity degree). In order to elicit triphthongs, monosyllabic words where the triphthongs /uɑe, uɤe/ were expected to occur were included in the material, and spelled with *uae* and *uõe* respectively. As monosyllabic words in Estonian are considered to be in Q3, the triphthongs in the present study are only in Q3.

All test words were embedded in utterance initial and final positions of read sentences, which were written down in the Kihnu variety, e.g. *Kuõli läksid poesid* 'To school went the boys', *Poesid läksid kuõli* 'The boys went to school'. Each sentence was printed on a separate card. In total there were 153 sentences which contained 165 test vowels.

The data was recorded from six female speakers aged 23–42 years (average age 37) who are all native speakers of the Kihnu variety. All subjects were born in Kihnu and have at least one parent from Kihnu. Five speakers have higher education and one was a university student at the time of recording. Mainly in connection with their studies they have all lived outside Kihnu. The time spent elsewhere in Estonia varied from a couple of years to 10 years.

The subjects were presented with the pack of test cards where the sentences appeared in random order. The recordings were carried out in quiet settings on the Island of Kihnu using an Edirol R-09HR digital recorder. Not all test tokens could be used in the analysis because of the background noise, whisper, or other reasons. In total, the final set of materials consisted of 939 test words, the distribution of which can be seen in Table 1.

Table 1. *The distribution of the analyzed tokens according to the three quantities (Q1, Q2, Q3)*

	Q1	Q2	Q3
Monophthongs	225	163	121
Diphthongs	-	176	224
Triphthongs	-	-	30

The data were analyzed with the speech analysis software *Praat* [15]. In this paper, for the first time for Estonian, the method of formant dynamics was used, i.e. rather than measuring a single point in the middle of the steady state of the vowel, equidistant temporal points were used to characterize the formant trajectory within the whole vocalic part. The boundaries of all vocalic segments were marked by hand. A script was used to calculate the total duration of each vocalic segment and divide it into ten equal intervals (see Figure 1). The frequencies of F1 and F2 were automatically measured and checked manually for inconsistencies. Thus, for each formant a contour consisting of nine values was obtained. For plotting the trajectories, the values of the first and last point were left out so as not to include formant transitions from the preceding and following consonant.

Additionally, in order to track more closely the formant frequency change over the course of vowels' duration in both monophthongs and diphthongs, trajectory length (TL) for each separate vowel section was calculated (cf. [16]).

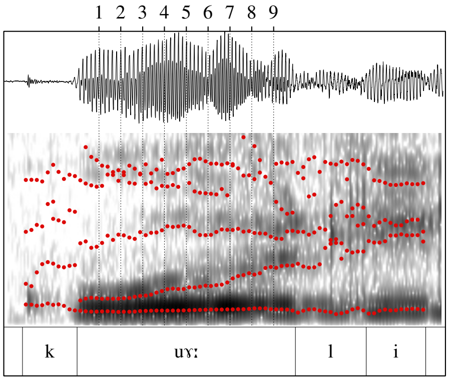


Figure 1. *Spectrogram of the test word /kuɤːli/ 'to school'.*

3. Results and discussion

3.1. The acoustics of monophthongs

We will start the presentation of our results by looking at the quality of Kihnu monophthongs. We compared their values to earlier data from Standard Estonian based on spontaneous speech from 8 women in a comparable age group [17]. Figure 2 plots all Kihnu monophthongs in the three quantity degrees in F1–F2 space. Red letters mark the mean values of 6 Standard Estonian monophthongs. The general direction of the formant frequency change is indicated by arrows.

It can be observed that the trajectories for monophthongs are very short and are clustered around the target value of the vowel. It is perhaps striking that in Q1 the vowels /ɤ/ and /ø/ coincide in Kihnu. Our comparative Standard Estonian data did not include these vowels, but a similar result has also been shown for Standard Estonian where a short stressed /ɤ/ has shifted into the space of /ø/ [18]. Overall, Kihnu monophthongs are similar to those of Standard Estonian. In both varieties, short monophthongs are more centralized than the long ones, and there is not much difference between the monophthongs in Q2 and Q3. The two vowel systems differ only in the realization of /æ/ and /i/. In Q1 and Q3, the Kihnu /æ/ is more reduced and higher than the Standard Estonian /æ/. Long /i/ in Standard Estonian Q2 and Q3 is more peripheral than its Kihnu counterpart.

3.2. The quality of diphthong components

In order to study the quality of Kihnu diphthongs, the target values of the diphthong components were compared to the corresponding monophthongs. Figure 3 displays the six diphthongs in Q2 and Q3 in F1–F2 vowel space. The mean values of Kihnu monophthongs are marked in red. It can be seen that the diphthongs have longer trajectories than the monophthongs in Figure 2. All diphthongs are acoustically clearly realized. The trajectories for diphthongs in Q2 and Q3 are very similar implying that there is not much difference between the quality of the diphthongs in the two quantities.

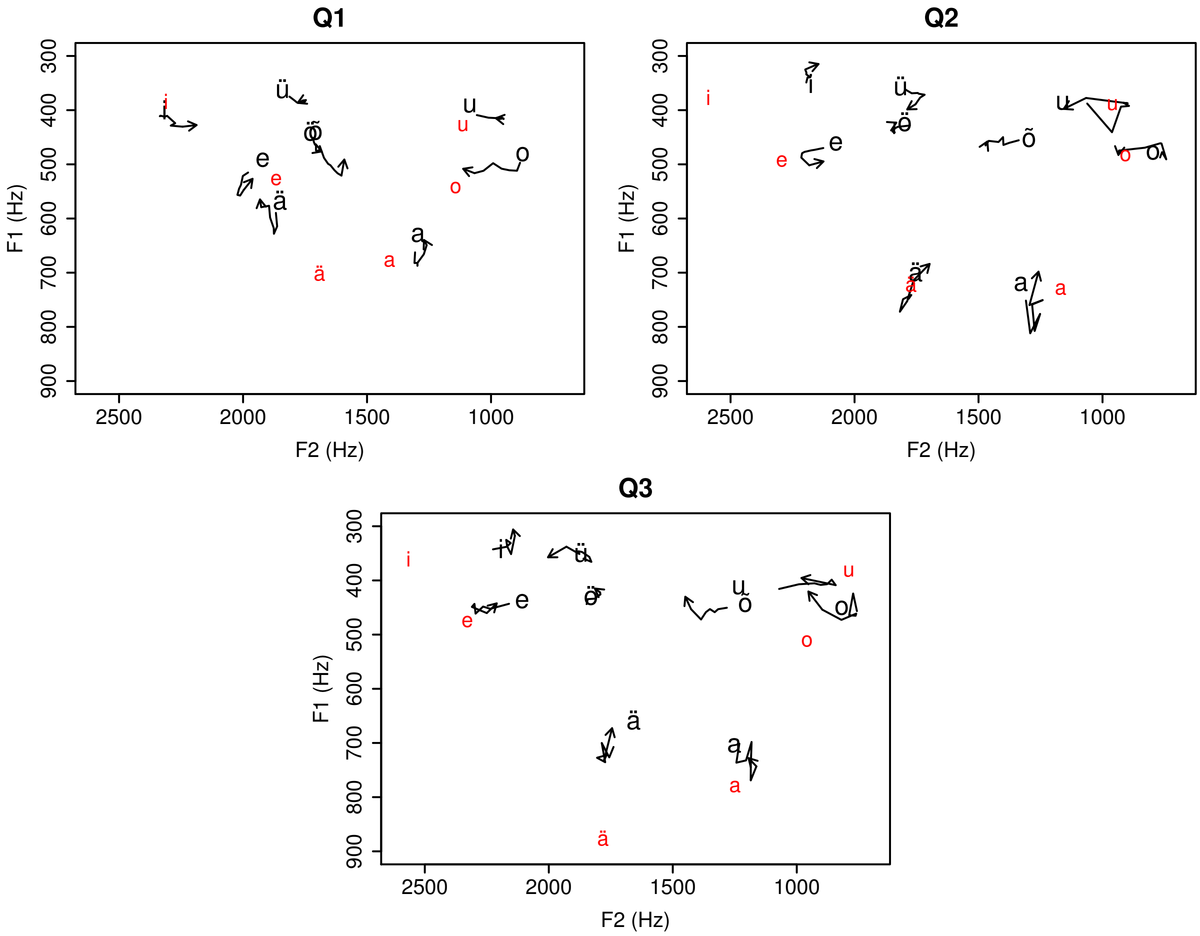


Figure 2. *Monophthongs in three quantity degrees (Q1, Q2, Q3) in F1–F2 vowel space. Mean values of six Standard Estonian vowels are marked in red.*

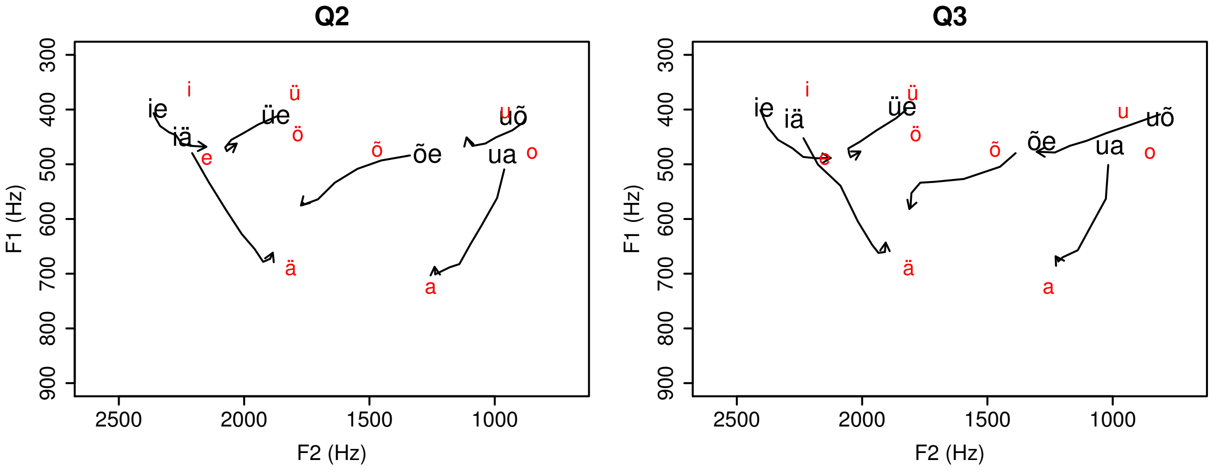


Figure 3. *Diphthongs in two quantity degrees (Q2, Q3) in F1–F2 vowel space. Mean values of Kihnu monophthongs are marked in red.*

The comparison of the diphthong components with monophthongs shows that the target values of the diphthongs are close to the corresponding monophthong values, which was expected in our first hypothesis. The realization of similar diphthong components is, however, affected by the other component. Thus, /e/ in /ye/ and /ɤe/ is realized differently, being lower and further back in /ɤe/. Also, /i/ in /ie/ and /iæ/ differs in quality and is realized closer to /e/ in /iæ/. The first component of /uɤ/ and /uɑ/ is closer to /o/ in the latter diphthong. Analogous results for Standard Estonian were obtained by Piir [14] who showed that similar first or second components have different values. Also, a study of the diphthongs /ai, ei, ui/ in the second syllable of disyllabic Standard Estonian words [19] demonstrated strong coarticulatory influences between the diphthong components.

3.3. The acoustics of triphthongs

Before examining the durational characteristics of the diphthongs we will test our third hypothesis concerning the realization of triphthongs. The trajectories for the triphthongs are plotted in Figure 4. It can be seen that the two postulated triphthongs in Kihnu Estonian are indeed acoustically realized as such. They have visibly longer trajectories, traversing clearly three different vocalic qualities, as compared to the diphthongs in Figure 3. It can be seen, though, that the trajectory for /uɑe/ is lower than what we would have expected based on dialectal transcriptions of this triphthong. Its first component starts below /o/, moves to /ɑ/ and ends just above /æ/ without reaching /e/. Thus, the realization of this triphthong is closer to /oɑæ/ than /uɑe/, which means that the controversial transcription *uä* used since Saareste [7] captures its actual quality relatively well.

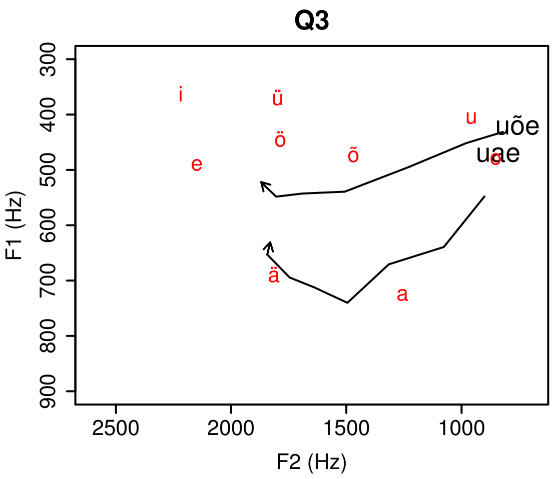


Figure 4. *Triphthongs in F1–F2 vowel space. Mean values of Kihnu monophthongs are marked in red.*

3.4. Durational characteristics

3.4.1. Trajectory length

We will now look more closely at the measures trajectory length (TL) and vowel section length (VSL). For calculating TL we adjusted the formula from Jacewicz et al. [16]. This was done because the calculation of TL using the original formula (based on the sum of vowel sections using all 9 measurement points in the vocalic segment) did not separate monophthongs and diphthongs as clearly as would be expected from Figures 2 and 3. Long monophthongs, in particular /o/ and /u/ and sometimes /a/, have a relatively long trajectory, which spans about 2-3 Barks. For other vowels, this trajectory is within 1 Bark. The movement of the trajectory is, however, different in monophthongs and diphthongs. Diphthongs (and also triphthongs) have trajectories which move from one target value to another, but for monophthongs the trajectory is not unidirectional but zigzags around the target value of the vowel. Thus, it was considered optimal to use only two points in each vocalic segment, and the trajectory length was calculated on the basis of formant values at 20% and 80% points of the vocalic segment using the formula:

$$TL = \sqrt{(F1_{20\%} - F1_{80\%})^2 + (F2_{20\%} - F2_{80\%})^2} \qquad (1)$$

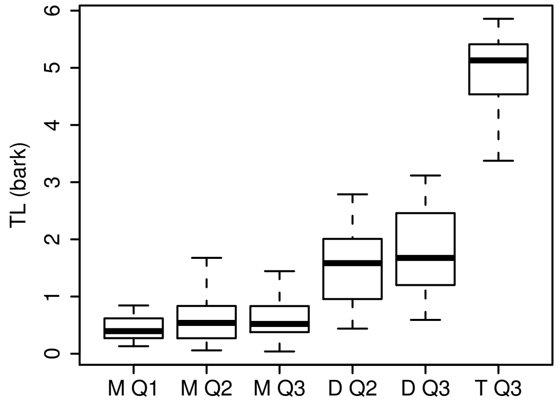


Figure 5. *Boxplots for the trajectory length (TL) measured on the basis of formant values at 20% and 80% points of the vocalic segment.*

Figure 5 presents the boxplots for the trajectory length of Kihnu monophthongs, diphthongs and triphthongs. The three types are clearly separated [F(2,238)=302.61; p<0.001]. An ANOVA also showed a significant main effect of quantity [F(2,238)=67.249; p<0.001], although this could be explained by unbalanced data: diphthongs cannot be in Q1 and triphthongs only occurred in Q3. Thus, the post-hoc test showed no significant difference between the realization of monophthongs in the three quantity degrees or diphthongs in the two quantity degrees.

The boxplots for vowel duration are presented in Figure 6. The trajectory length is not connected with the duration of the vowel, which is dependent on the phonological quantity of the word rather than whether the vowel is a monophthong, diphthong or a triphthong. An ANOVA showed significant main effects of both quantity [F(2,238)=224.65; p<0.001] and vowel type [F(2,238)=57.011; p<0.001], but post-hoc tests confirmed that the difference was between short (Q1) and long (Q2, Q3) vowels (p<0.001) on the one hand, and between Q2 and Q3 vowels (p<0.05) on the other hand.

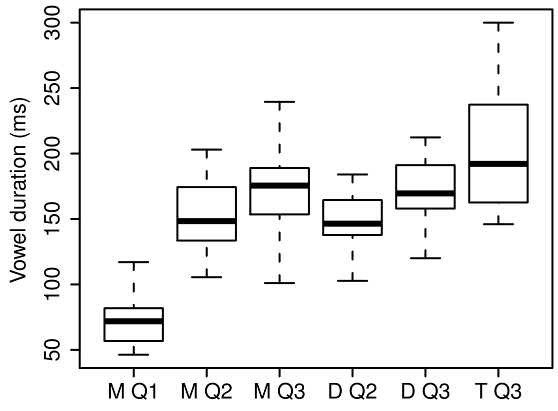


Figure 6: *Boxplots for vowel duration in monophthongs, diphthongs and triphthongs.*

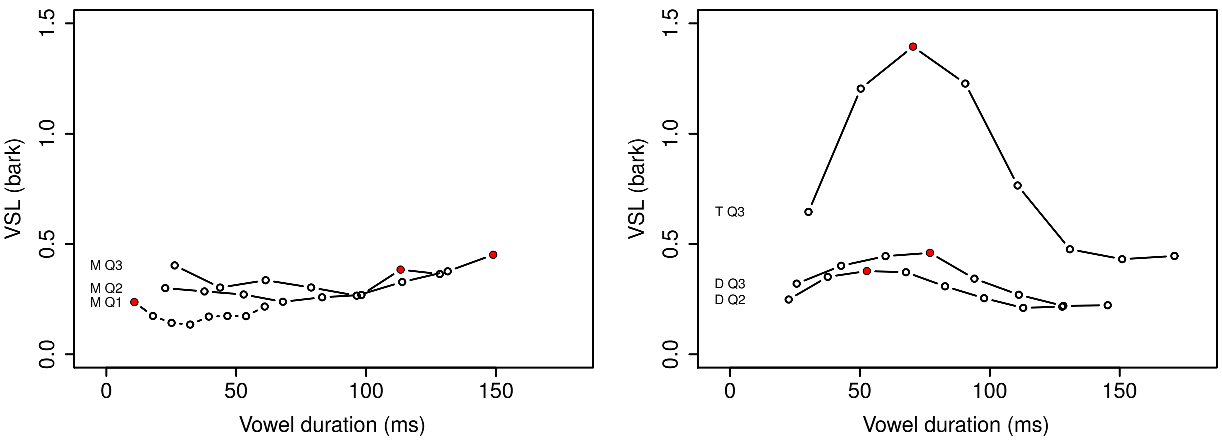


Figure 7. *The VSL contours for monophthongs (left panel), and diphthongs and triphthongs (right panel). The vowel sections with the biggest change are marked with filled red circles.*

There was no difference between monophthongs and diphthongs in the same quantity degree, but the triphthongs were longer than the monophthongs and diphthongs in Q3 words ($p<0.05$). The latter might, however, be an effect of word structure as the triphthongs occurred only in monosyllabic words.

3.4.2. Vowel section length

The length of each vowel section (VSL) was calculated using the following formula where n=8:

$$VSL_n = \sqrt{(F1_n - F1_{n+1})^2 + (F2_n - F2_{n+1})^2} \qquad (2)$$

Figure 7 displays the VSL values for monophthongs (left panel), and diphthongs and triphthongs (right panel). Each contour consists of 8 points corresponding to 8 sections calculated on the basis of 9 measured segments. We can see that in the case of monophthongs the VSL contour is bowl-shaped, which implies that bigger changes take place in the beginning and end of the vowel (i.e. transitions from and to the consonant), while the middle part is relatively stable. The VSL contours for diphthongs and triphthongs are hat-shaped; the more stable parts are in the beginning and in the end, and the transition from one target value to another takes place in the middle.

We were interested in the duration of diphthong components in Q2 and Q3. Based on earlier findings from Lehiste [13] we expected the duration of both components to be longer in Q3. Rather than trying to divide the diphthongs into components by hand, we located the vowel section with maximal change in formant values, taking it to mark the most objective boundary between the two diphthong components. In Figure 7, the sections with maximal change are marked with filled circles. It can be seen that the transition from one target value to the next occurs later in Q3 diphthongs, which means that also the first component of a diphthong is longer in Q3. Thus, our results lend support to Lehiste's work [13], and are not in line with the view expressed in [2] according to which the first component is similar in both quantities. Similar to [14] our results show that both in Q2 and Q3 diphthongs, the first component has a shorter duration than the second one.

In the case of triphthongs we would expect to see a two-peaked VSL contour which would traverse three target values and two transitions. In Figure 7 we can see that this not the case, as there is only one clear peak. Nevertheless, the trajectory of the triphthongs is different from those of the diphthongs; the triphthongs are considerably more dynamic than the diphthongs.

4. Conclusions

The main aim of this paper was to study the acoustics of six Kihnu diphthongs which have arisen as a result of diphthongization of long open and mid vowels. The analysis used the method of formant dynamics. Three hypotheses were tested and proven right.

Firstly, the results showed that the diphthong components are close in their quality to the corresponding monophthongs. The realization of similar diphthong components was, however, shown to be affected by the other component. Coarticulatory influences between the diphthong components were evident in the case of /uɑ/ and /iæ/ where the first components were lower than the target values. Also, the second component /e/ was lower and further back in /ɤe/ than in /ie/. A comparison of Kihnu monophthongs with Standard Estonian showed that in Q1 and Q3, the Kihnu /æ/ is more reduced and higher than the Standard Estonian /æ/. Long /i/ in Standard Estonian Q2 and Q3 is more peripheral than its Kihnu counterpart.

Secondly, using the measure vowel section length we showed that the first diphthong component is affected by the phonological quantity. Both components were longer in Q3 diphthongs than in Q2 diphthongs.

Thirdly, the results of the present study demonstrated that there is an acoustic basis for postulating two triphthongs in Kihnu. The trajectory length measure proved to be useful in separating monophthongs, diphthongs and triphthongs.

5. Acknowledgements

We would like to thank Külli Laos for her help with test sentences and all our subjects. Research presented in this paper was supported by Estonian Science Foundation grant no. 7904.

6. References

[1] Lonn, V., Niit, E., ”Saarte murde tekstid. Eesti murded VII”, Tallinn: Eesti Keele Instituut, 2002.

[2] Eek, A., ”Eesti keele foneetika I”, Tallinna Tehnikaülikooli Küberneetika Instituut: TTÜ Kirjastus, 2008.

[3] Pajusalu, K., Hennoste, T., Niit, E., Päll, P., Viikberg, J., ”Eesti murded ja kohanimed”, (2nd ed). Tallinn: Eesti Keele Sihtasutus, 2009.

[4] Saar, T., Valmet, A., ”Kihnu murrakust”, Kõiva, O., Rüütel, I. (Eds.), Vana kannel VII:1. Kihnu regilaulud, Tallinn: Eesti Kirjandusmuuseum. Eesti Rahvaluule Arhiiv. Eesti Keele Instituut. Folkloristika osakond, 35–40, 1997.

[5] Saar, T., ”Kihnu murde häälikud”, Eesti Keele Instituudi eesti murrete ja soome-ugri keelte arhiiv, (MS), 1958.

[6] Grigorjev, P., Keevallik, L., Niit, E., Paldre, L., Sak, K., Veismann, A., ”Kihnu murde assimileerumise mustreid Manilaiul”, Pühendusteos Huno Rätsepale. Erelt, M., Sedrik, M., Uuspõld, E. (Eds.), Tartu Ülikooli eesti keele õppetooli toimetised 7. Tartu: Tartu Ülikooli Kirjastus, 26–44, 1997.

[7] Saareste, A., ”Häälikajalooline uurimus Kihnu murdest. Konsonandid”, H-98, Tartu Ülikooli eesti murrete ja sugulaskeelte arhiiv, (MS), 1920.

[8] Sang, J., ”Ühest fonotaktilisest kollisioonist (Kihnu näitel)”, Keel ja Kirjandus LII, 11, 809–817, 2009.

[9] Viitso, T-R., “Rise and development of the Estonian language. – Estonian Language”, Mati Erelt (Ed.), Linguistica Uralica, Supplementary Series 1, Tallinn: Estonian Academy Publishers, 130–230, 2003.

[10] Laos, K., ”Kihnlasõ emäkiel”, Pärnu: SA Kihnu Kultuuri Instituut, 2010.

[11] Ariste, P., ”Eesti keele foneetika”, Tallinn: Eesti Riiklik Kirjastus, 1953.

[12] Türk, H., ”Kihnu murraku vokaalidest”, Bakalaureusetöö, Tartu Ülikool (MS), 2010.

[13] Lehiste, I., “Diphthongs versus vowel sequences in Estonian”, Proceedings of the Sixth International Congress of Phonetic Sciences, Held at Prague 7–13 September 1967, B. Hála, M. Romportl, P. Janota (Eds.), Prague: Academia Publishing House of the Czechoslovak Academy of Sciences, 539–544, 1970.

[14] Piir, H., “Acoustics of the Estonian diphthongs”, Estonian Papers in Phonetics, EPP 1982–1983, Tallinn, 5–103, 1985.

[15] Boersma, P., Weenink, D., “Praat: doing phonetics by computer” [Computer program], Version 5.3, retrieved October 2011 from http://www.praat.org/.

[16] Jacewicz, E., Fox, R.A., Salmons, J., “Vowel change across three age groups of speakers in three regional varieties of American English”, Journal of Phonetics 39, 683–693, 2011.

[17] Lippus, P., “The acoustic features and perception of the Estonian quantity system”, Dissertationes Philologiae Estonicae Universitatis Tartuensis 29, Tartu: Tartu University Press, 2011.

[18] Eek, A., Meister, E., “Quality of Standard Estonian Vowels in Stressed and Unstressed Syllables of the Feet in Three Distinctive Quantity Degrees”, Linguistica Uralica XXXIV, 3, 226–233, 1998.

[19] Teras, P., ”Eesti keele teise silbi diftongide akustikast”, Lõputöö. Tartu Ülikool (MS), 1996.

The phonetic and phonological analysis of the fall-rise intonation pattern in the Kihnu variety of Estonian

Eva Liina Asu, Nele Salveste

Institute of Estonian and General Linguistics, University of Tartu

eva-liina.asu@ut.ee, nele.salveste@ut.ee

Abstract

The paper presents an analysis of fall-rise intonation patterns in the Kihnu dialect of Estonian. The study is based on recordings of spontaneous speech from three generations of female speakers. Fall-rises occurred almost five times less frequently in the data of the speakers born around 1935 as compared to those born around 1900. The data of the speakers born after 1970 did not contain any fall-rises at all, which could be interpreted as a sign of prosodic levelling of the traditional variety. An investigation of F0 alignment revealed a systematic variation in the location of high and low tonal targets. The peaks were aligned significantly later in the case of longer phrases. The first low 'elbow' was aligned either with the accented syllable (in monosyllabic and disyllabic words) or with the syllable following the accented syllable (in longer words and phrases), and the second low 'elbow' with the final syllable of the phrase just before the high boundary. These findings lend support to the treatment of the fall-rise accent as H*L H% in accordance with the autosegmental-metrical analysis of Estonian intonation.

Index terms: intonation, fall-rise accent, peak alignment, Kihnu Estonian, dialect levelling

1. Introduction

Prosodically, one of the most interesting areas in Estonia is constituted by the islands off the west coast of the Estonian mainland. Distinctive intonational variation of insular dialects has been attributed to influences from Swedish, which used to be an important contact-language in this language convergence area [1, 2]. The two Estonian varieties usually singled out for their intonation, and often impressionistically characterised as 'sing-songy' are those of the Islands of Saaremaa and Kihnu. It has been shown that in the Saaremaa variety, the impression of 'singing' intonation is phonetically given by systematically later peak alignment as compared to Standard Estonian [3, 4]. The Kihnu variety, on the other hand, is noted for the frequent occurrence of rising intonation, which has already been mentioned in older dialectal descriptions [5, 6]. On closer examination the rises actually appeared to be mostly fall-rises occurring in phrase-final position [7]. Such accents have not been encountered in the intonational inventory of Standard Estonian [8], and are therefore particularly interesting from the typological point of view. The present paper focuses on their phonetic and phonological analysis in spontaneous speech.

Fall-rise accents occur in many other European languages, including English, Dutch and German. The most common analysis of such accents within the autosegmental-metrical theory of intonation is a fall followed by a high boundary tone: H*+L H% [9, 10, 11, 12]. In the transcription system for German (GToBI), fall-rises are generally represented using the L- phrase accent as (L+)H* L-H% [13]. As intermediate phrases are not used in the intonational transcription for Estonian [8], the optimal phonological analysis for the fall-rise accent of Kihnu would be H*L H%. We hope to find further confirmation to the appropriateness of this label on the basis of the following analysis.

The present paper has two broader research aims. Firstly, our goal is to investigate the general occurrence of the fall-rise accent in the speech of women representing different age groups with a view to establish whether there are any differences. This research question is relevant from the point of view of dialect levelling. Already back in the 1930s, it was claimed that the characteristic quality of Kihnu speech melody was disappearing in the speech of younger islanders [5]. Levelling in the Kihnu variety has so far only been addressed with respect to some grammatical and phonological aspects [14]. If we assume that the fall-rise accent is (at least) one of the manifestations of the distinct Kihnu melody, the study of the distribution of this pattern in different generations of speakers would shed some more light on this matter. As a result of prosodic dialect levelling we would expect the fall-rise to appear less frequently in the data of younger informants.

The second main goal of the paper is to investigate the phonetic realisation of the fall-rise accent. We will focus on the study of the alignment of H and L targets in nuclear fall-rises of varying length. The alignment of tonal targets has been shown to depend on a variety of factors, such as for instance, the position of the accent in the utterance (sentence-medial vs. sentence-final) [15], the phonological length of the accented vowel [16, 15], or the number of unstressed syllables following the accented syllable [17, 18, 19]. Studies on the influence of syllable structure on peak alignment have yielded contradicting results. For instance, it was shown for Spanish that peaks were aligned earlier in closed syllables than in open syllables [20], while no effect of syllable structure on alignment was found in Dutch [21].

In Estonian, the F0 peak placement is additionally affected by the phonological quantity of the foot. Broadly speaking, in the short (Q1) and long (Q2) quantity degree, peaks are aligned later than in the overlong quantity degree (Q3) that is characterised by a steep F0 fall early in the accented syllable. It has been shown that these tonal differences are also present in spontaneous speech [22].

Thus, in our data we would expect H* to be aligned later in the case of longer unstressed material following the accented syllable. Despite the number of syllables in the phrase we expect the location of the peak to be dependent on the phonological quantity. Possible differences in alignment due to syllable structure will also be investigated.

2. Materials and method

For the present analysis we used recordings of spontaneous speech from three generations of Kihnu women: (1) those born around 1900, (2) around 1935, and (3) after 1970. The data was drawn from the University of Tartu Archives of Estonian Dialects and Kindred Languages based on the birth-year of the informants. The speech files consisted mainly of dialogues between a field-worker and an informant, and in the case of some newer recordings between two speakers of Kihnu Estonian. In total we analysed 10 h 47 min of speech by 21 informants. The summary of the data is presented in Table 1. It can be seen that there was least data from the youngest generation, which is due to there being on the whole less data for younger speakers in the database. The division of speakers into three groups according their birth-year is in accordance with the speakers' age at recording. With only one exception (the 80 year old informant in the middle group) the informants in the oldest age group, as classified by their birth-year, were also oldest at the time of recording.

Table 1. *Speech data used for the analysis.*

Year of birth	Number of informants	Age at recording	Year of recording	Duration of the data (h:min:sec)
1889-1902	6	68-91	1961-1988	3:36:35
1930-1943	8	61-68, 80	1992-2011	5:17:39
1971-1989	7	20-38	2009	1:53:28

The recordings were made at different times during dialectal field-work on the Kihnu Island. Therefore, the quality of the sound files varies considerably depending on the recording equipment used and the nature of the recording environment. Despite this variation the data was deemed suitable for the present purposes.

We located all instances of fall-rise accents in the recordings by combined listening and visual inspection of the F0 traces. In total 282 tokens were found (275 in intonational phrase-final and 7 in non-final position). The present analysis focuses on the phrase-final (nuclear) accents. The distribution of the tokens according to the phonological quantity of the accented syllable was the following: Q1 - 84, Q2 - 49, Q3 - 142.

The fall-rise accents were transcribed on four annotation tiers: word, syllable, segment and intonation, using the speech analysis software Praat [23]. On the intonation tier, the F0 contour of each phrase was labelled at 5 events: (1) the beginning of the word at the start of fall-rise, which often coincided with the phrase boundary and was therefore marked as %, (2) the F0 maximum in the accented syllable (H*), (3) the first low 'elbow' after the fall (L_1), (4) the second low 'elbow' before the rise (L_2), and (5) the final intonational boundary (H%). We chose to identify two low pitch elbows rather than the lowest F0 point in order to render the shape of the accent contour better particularly in the case of longer phrases [cf. 24]. Figure 1 presents an F0 contour with the H*L H% accent, exemplifying the annotation and measurement points.

A Praat script was used to extract the values of pitch (Hz) and time (ms) at the labelled locations, as well as to calculate the duration of each transcribed phrase and syllable. All F0 measurements were subsequently manually checked and corrected.

The H*L H% accent occurred on material of varying length, from monosyllabic words to words or phrases consisting of six syllables. Table 2 shows the distribution of the data according to the quantity and structure of the first syllable, and the number of syllables in the phrase.

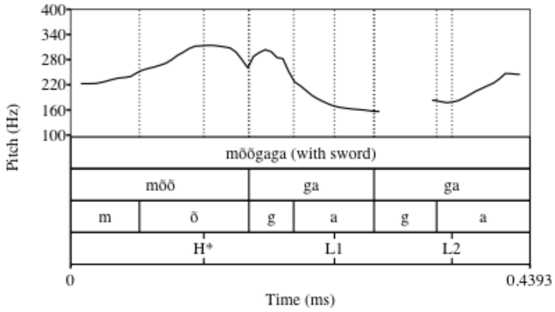


Figure 1. *An F0 contour with the H*L H% accent on a trisyllabic utterance showing the annotation and measurement points.*

In order to study the alignment of tonal targets the proportional time of H* was calculated using the formula:

$$H^*\text{-}S_0\,/S_B\text{-}S_0 \qquad (1)$$

where H* denotes the time for the peak, S_0 for the beginning of the stressed syllable, S_B for the end of the stressed syllable (syllable boundary).

Table 2. *The distribution of the data according to the quantity (Q1, Q2, Q3) and structure of the first syllable, and the number of syllables in the phrase.*

Number of syllables in the phrase	(C)V	(C)VV		(C)VC		(C)VVC	
	Q1	Q2	Q3	Q2	Q3	Q3	Total
1	-	-	11	-	6	13	30
2	27	7	23	5	26	-	88
3	30	7	16	7	12	-	72
4	11	4	11	11	12	-	49
5	11	5	4	3	5	-	28
6	5	-	1	-	2	-	8
Total	84	23	66	26	63	13	275

3. Results and discussion

3.1. Distribution of the fall-rise accent in three generations of speakers

In order to test the hypothesis about prosodic dialect levelling we looked at the distribution of the fall-rise accent in the three age groups. The frequency was calculated on the basis of the number of times the accent occurred in every sound file. Admittedly, this is a relatively simplified way of analysis but nevertheless provides us with a clear answer. Figure 2 presents an exponential curve for the frequency of the accent in relation to the birth-year of the speaker. It can be seen that there is a strong correlation (R2 = 0.752). In the older group, the fall-rise accent occurred on average 1.2 times/min, and in the middle group on average 0.2 times/min, or in other words, the speakers in the older group used the fall-rise accent on average every 1.3 min, whereas those in the middle group every 6 min.

The accent did not occur at all in the data of the youngest informants. It can be argued that this might be due to there being least data for this group of speakers, but on the other hand the data was similar in nature and by a comparable number of informants as for the other two age groups. Therefore, we can infer that the hypothesis about the gradual disappearance of the characteristic speech melody with regard to the occurrence of the H*L H% accent was borne out.

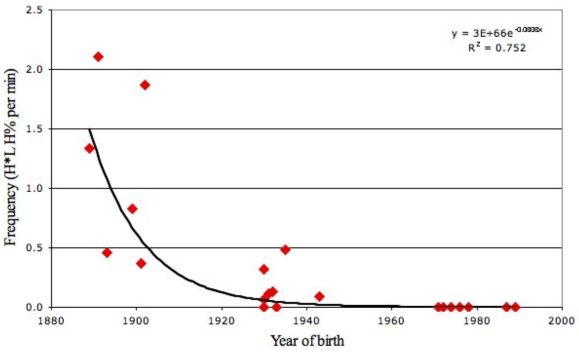


Figure 2. *Exponential curve for the frequency of H*L H% vs. year of birth of the speaker.*

The lack of the fall-rise accent in the speech of younger informants can be interpreted as a reflection of the levelling of traditional Kihnu speech melody. Yet, the general infrequency of the H*L H% pattern in the present data (only 275 phrase-final tokens in nearly 11 hours of speech) can denote its limited discourse functions. It is likely that the accent only occurs in certain contexts. Preliminarily, we observed that fall-rises were often used in short (one-word) answers, and echo-questions. It was also clear that the fall-rise accent was much more common in phrase-final than in non-final position. A more thorough investigation of its discourse functions is needed.

3.2. Alignment of H and L targets

Peak alignment was analysed with respect to the number of syllables in the phrase, and the phonological quantity and syllable structure of the accented syllable. The phrases in the data were maximally 6 syllables long, but due to there being no tokens for Q2 for the longest material, we included maximally pentasyllabic phrases in the analysis. Firstly, we tested whether the peak alignment was influenced by syllable structure, separating open ((C)V(V)) and closed ((C)(V)VC) syllables. As we did not find any significant effect of syllable structure on peak location, the results will be presented based on the average of all syllable structures.

Figure 3 shows the proportional location of the peak relative to the beginning of the accented syllable in phrases of different length. As expected, the peak was aligned later in the case of more unstressed syllables following the stressed syllable. There was a significant effect of the number of syllables on the alignment of H* (F(4, 254) = 15.29, p<0.001). This finding is line with earlier work on Estonian [19] and other languages [e.g. 17, 15, 21].

Various explanations for the observed systematic differences in peak alignment have been offered. The most compelling reason for earlier peaks in the case of shorter phrases in the present data seems to be so called 'tonal crowding', where the variable spacing of tones is influenced by the preceding and following tones [cf. 17, 25]. In the case of shorter material, there is less time for the realisation of the fall-rise contour before the phrase boundary, which means that the whole tune has to be compressed. This explains the significantly earlier location of H* in monosyllabic utterances, and is a strong argument for the analysis of the contour as one pitch accent rather than composed of different tunes.

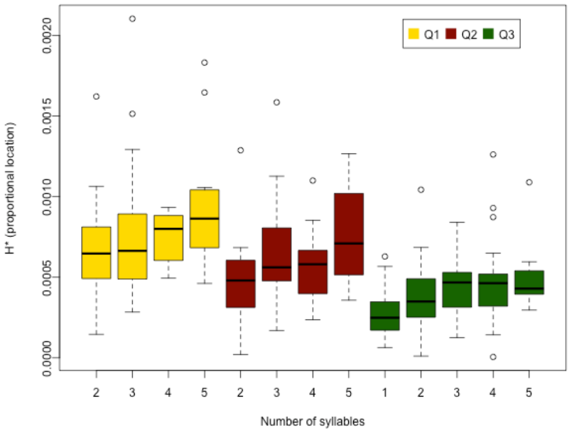


Figure 3. *Boxplots showing the proportional location of H* relative to the duration of the accented syllable in three quantities (Q1, Q2, Q3) in phrases of different length (1-5 syllables).*

Despite the influence of the length of the post-accentual material, the quantity-dependent effect on peak alignment was also clearly evident and statistically significant (F(2, 254) = 29.34, p<0.001). As expected, the F0 peak occurred earlier in Q3 than in Q1 and Q2 words. A similar result was shown for Standard Estonian in [19] on the basis of tightly controlled read sentences.

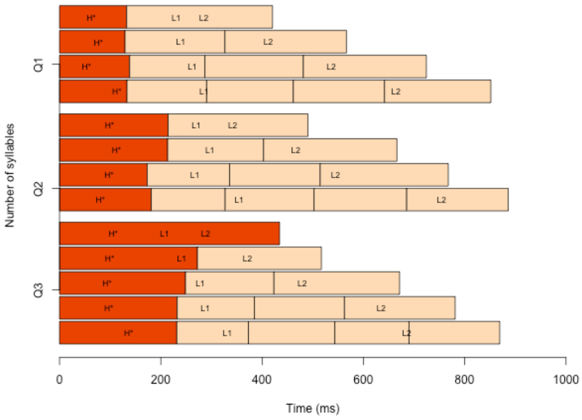


Figure 4. *The alignment of H* relative to the onset of the accented syllable (red) and the location of low targets (L_1 and L_2) in phrases of different length (1-5 syllables) in three quantities (Q1, Q2, Q3).*

Subsequently, the alignment of low targets was explored. Figure 4 presents both the timing of H* and the two low elbows in phrases varying from 1 to 5 syllables. The accented syllables are marked in red. It can be seen that a systematic pattern also emerges for the alignment of low elbows. In monosyllabic and disyllabic words in Q3, the first low elbow

(L_1) was located in the first syllable; in longer phrases it was always aligned with the second syllable with only one exception (pentasyllabic phrases in Q2). This relatively invariant location of L_1 in relation to H* denotes that the tones ‘belong together’, which gives support to the analysis of the first component of the fall-rise accent as a bi-tonal pitch accent, a fall (H*L). The second low elbow (L_2) was always aligned with the final syllable of the phrase (with only one exception). It can be argued that the stable alignment of L_2 might be an evidence for the existence of a low phrase accent (L-), or a bi-tonal boundary tone (LH%). It is, however, impossible to solve the issue on the basis of the present data. Therefore, until further evidence, there is no reason not to analyse the pattern as H*L H%, which is in line with the treatment of similar patterns in comparable intonational transcription systems such as IViE for British English [9] and ToDI for Dutch [10].

In sum, the general shape of the tonal contour is realised similarly in phrases of different length independent of the structure or quantity of the accented syllable. There is a fall in pitch immediately after the H* after which the F0 contour stays low until the rise starts just before the high boundary. Figure 5 presents boxplots showing pitch at the 5 locations in the fall-rise contour. The data has been averaged over all syllable structures and quantity degrees as well as phrases of different length, as there were no significant differences in the range of the fall from H* to L_1 and the span of the rise from L_2 to H% between these different conditions.

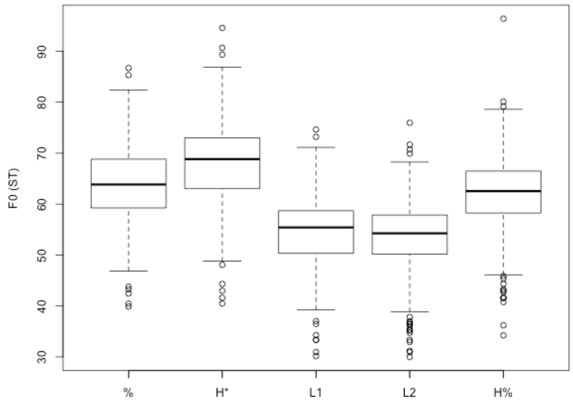


Figure 5. *Boxplots of the pitch (st) at the initial boundary (%), High target (H*), Low targets (L_1, L_2) and final intonational boundary (H%).*

4. Conclusions and further work

The paper studied the fall-rise pitch accent in Kihnu Estonian. Firstly, we explored its distribution in the speech of three generations of female speakers. It was hypothesised that due to the levelling of the traditional dialect, fall-rise patterns (as one of the manifestations of the distinct intonation of the variety) would occur less frequently in the speech of younger informants. This hypothesis was borne out. The speakers born around 1935 used the fall-rise accent almost five times less frequently than those born around 1900, and the data of the speakers born after 1970 did not contain any fall-rises at all.

Secondly, we investigated the alignment of tonal targets in the fall-rise contour. The analysis revealed a systematic variation in the location of the peak depending on the number of syllables in the phrase. The H* was aligned significantly later when more unstressed material followed the accented syllable. This finding is in line with work on other languages. It was also shown that the previously attested tonal alignment characteristics of the Estonian three quantity degrees were realised in the present spontaneous data: the peak was aligned earlier in Q3 than in Q1 and Q2.

Two low pitch elbows were located in order to characterise the low valley between the end of the fall and the beginning of the rise in the fall-rise contour. It was shown that the first low elbow (L_1) was aligned after the accented syllable (except in monosyllabic and disyllabic words in Q3, where it was located in the accented syllable), and the second low elbow (L_2) with the final syllable of the phrase. Further work is needed to establish whether the stable alignment of L_2 just before the high boundary tone might be evidence of a low phrase accent (L-), or alternatively a bi-tonal boundary tone (LH%). These questions have to be tested on more tightly controlled materials.

The domain of the fall-rise varied from monosyllables to words and phrases consisting of six syllables. The realisation of the tonal contour was not dependent on the length of the material, which implies that the pattern is not composed of different tunes but constitutes one pitch accent plus a boundary tone. This finding lends support to its analysis as H*L H% following the current framework for the analysis of Estonian intonation [8].

We are aware that the characteristic intonation of Kihnu may also have other acoustic correlates in addition to the H*L H% pattern. As has been shown for Estonian [3] and many other languages [e.g. 26, 27, 15], different varieties of the same language can vary in the realisation of the same F0 contour with respect to peak alignment. A comparison of Kihnu intonation with other varieties of Estonian might therefore reveal further important characteristics of the prosody of this variety.

5. Acknowledgements

We would like to thank Pärtel Lippus for his assistance with the statistical analysis, and Francis Nolan for discussing aspects of the paper. This research was supported by the Estonian Science Foundation grant no. 7904.

6. References

[1] Niit, E., “The structure of the Baltic prosodic area and the place of Estonian dialects in it”, Tallinn: Academy of Sciences of the Estonian SSR. Preprint KKI-17, 1980.

[2] Kalits, V., ”Kihnlaste elatusalad XIX sajandi keskpaigast XX sajandi keskpaigani”, Kihnu Kultuuriruum, 2006.

[3] Asu, E. L., “Intonational contour alignment in Saaremaa and Standard Estonian”, Linguistica Uralica XLI 2005, 2, 107-112, 2005.

[4] Asu, E. L., “Tonal alignment in two varieties of Estonian”, In G. Bruce and M. Horne (Eds.) Nordic Prosody. Proceedings of the IXth Conference, Lund 2004. Frankfurt am Main: Peter Lang, 29-35, 2006.

[5] Saar, T., “Murdeülevaade”, EKI eesti murrete ja soome-ugri keelte arhiiv, 1934.

[6] Tanning, S., “Murdepäevik”, EKI eesti murrete ja soome-ugri keelte arhiiv, 1948.

[7] Asu, E. L., Niit, E., ”Prosodic features of the West Estonian dialect of Kihnu”, Congressus XI Internationalis Fenno-Ugristarum, Piliscsaba 2010, Pars III. Summaria acroasium in symposiis factarum, 7, 2010.

[8] Asu, E. L., “The Phonetics and Phonology of Estonian Intonation”, Doctoral dissertation. University of Cambridge, 2004.

[9] Grabe, E., Post, B., Nolan, F., “Modelling intonational variation in English. The IViE system”, In Puppel, S. and Demenko, G. (Eds.). Proceedings of Prosody 2000. Adam Mickiewitz University, Poznan, Poland, 51-57, 2001.

[10] Gussenhoven, C., “Transcription of Dutch intonation”, In S-A. Jun (Ed.) Prosodic Typology. The phonology of intonation and phrasing. Oxford University Press, 118-145, 2005.

[11] Uhmann, S., “Fokusphonologie. Eine Analyse deutscher Intonationskonturen im Rahmen der nicht-linearen Phonologie”, Tübingen: Niemeyer, 1991.

[12] Féry, C., “German Intonational Patterns”, Tübingen: Niemeyer, 1993.

[13] Grice, M., Baumann, S., Benzmüller, R., “German intonation in Autosegmental-Metrical phonology”, In S-A. Jun (Ed.) Prosodic Typology. The phonology of intonation and phrasing. Oxford University Press, 55-83, 2005.

[14] Grigorjev, P., Keevallik, L., Niit, E., Paldre, L., Sak, K., Veismann, A., “Kihnu murde assimileerumise mustreid Manilaiul”, Pühendusteos Huno Rätsepale. Tartu Ülikooli eesti keele õppetooli toimetised 7. Tartu: Tartu Ülikooli Kirjastus, 26–44, 1997.

[15] Ladd, D. R., Schepman, A., White, L., Quarmby, L. M., Stackhouse, R., “Structural and dialectal effects on pitch peak alignment in two varieties of British English”, Journal of Phonetics 37(2), 145-161, 2009.

[16] Ladd, D. R., Mennen, I., Schepman, A., “Phonological conditioning of peak alignment of rising pitch accents in Dutch”, Journal of the Acoustical Society of America 107: 2685-2696, 2000.

[17] Silverman, K. and Pierrehumbert, J., “The timing of prenuclear high accents in English”, Papers in Laboratory Phonology I, 72-106, 1990.

[18] Prieto, P., van Santen, J., and Hirschberg, J. 1995. Tonal alignment patterns in Spanish. Journal of Phonetics, 23, 429–451.

[19] Plüschke, M., “Peak alignment in falling accents in Estonian”, Proceedings of the ICPhS XVII, 17-21 August 2011, Hong Kong, 1614-1617, 2011.

[20] Prieto, P., Torreira, F., “The segmental anchoring hypothesis revisited. Syllable structure and speech rate effects on peak timing in Spanish”, Journal of Phonetics 35.4, 473-500, 2007.

[21] Schepman, A., Lickley, R., Ladd, D. R., “Effects of vowel length and “right context” on the alignment of Dutch nuclear accents”, Journal of Phonetics 34, 1-28, 2006.

[22] Asu, E. L., Lippus, P., Teras, P., Tuisk, T., “The realization of Estonian quantity characteristics in spontaneous speech”, In M. Vainio, R. Aulanko and O. Aaltonen (Eds.) Nordic Prosody. Proceedings of the Xth Conference, Helsinki 2008. Frankfurt: Peter Lang, 49–56, 2009.

[23] Boersma, P., Weenink, D.,” Praat: doing phonetics by computer” [Computer program], 2011.

[24] Lickley, R., Schepman, A., Ladd, D. R., “Alignment of “Phrase Accent” Lows in in Dutch Falling Rising Questions: Theoretical and Methodological Implications”, Language and Speech 48 (2), 157-183, 2005.

[25] Arvaniti, A., Ladd, D. R., Mennen, I., “Phonetic effects of focus and ‘‘tonal crowding’’ in intonation: Evidence from Greek polar questions”, Speech Communication, 48, 667–696, 2006.

[26] Atterer, M., Ladd, D. R., “On the phonetics and phonology of "segmental anchoring" of F0: Evidence from German”, Journal of Phonetics 32, 177-197, 2004.

[27] Arvaniti, A., Garding, G., “Dialectal variation in the rising accents of American English”, In J. Cole and J. H. Hualde (Eds.) Papers in Laboratory Phonology 9. Berlin, New York: Mouton de Gruyter, 547-576, 2007.

Production of short and long Finnish vowels with and without noise masking

Osmo Eerola [1,2], *Janne Savela* [3]

[1] Faculty of Telecommunication and e-Business, Turku University of Applied Sciences, Turku, Finland
[2] Centre for Cognitive Neuroscience, University of Turku, Turku, Finland
[3] Department of Information Technology, University of Turku, Turku, Finland

osmo.eerola@tut.fi, jansav@utu.fi

Abstract

In order to further examine the possible quality differences between produced short and long Finnish vowels, we studied the formant frequencies F1–F4 and duration of the eight Finnish vowels /a/, /e/, /i/, /o/, /u/, /y/, /æ/ and /ø/ when uttered in carrier words (e.g., /tili/ - /tiili/) in two different masking conditions and without a noise mask. Babble noise at 92dB SPL was used to simulate a loud, crowded cocktail party, and pink noise at 83dB SPL an environment with the maximum noise level allowed for continuous working. Minor quality differences were found between the short and long vowels. Noise masking caused a significant prolongation of produced short vowels, and a significant increase in the F1 frequency.

Index Terms: vowel production, vowel quality and quantity, noise masking

1. Introduction

The Finnish vowel system includes eight vowels: /a/, /e/, /i/, /o/, /u/, /y/, /æ/ and /ø/, which all can occur as short (single) or long (double) in any position of a word. Generally, the two durational variants are regarded as being similar in perceived quality, and also the Finnish orthography reflects the interpretation that the long vowels consist of two successive and identical short vowels. Karlsson [1] refers to this interpretation as the identity group interpretation. Eerola et al. [2] investigated the perception of short and long Finnish /y/ and /i/ vowels, and found that the location of the category boundary between /y/ and /i/ on the F2 formant frequency axis, the width of the category boundary on the F2 formant frequency axis, the goodness rating value of the prototypical /i/, and the location of the prototypical /i/ on the F2 formant frequency axis were all independent of the stimulus duration.

However, the results of some earlier studies on the pro-duction of Finnish vowels suggest that there exist minor spec-tral dissimilarities in the formant frequencies F1–F3 of the produced short and long vowels. For example, based on five informants, Wiik [3] reported clear differences in the variability ranges of Finnish single and double /y/ and /i/ vowels, as measured in terms of F1, F2 and F3, stating that F1 is 40 Hz higher and F2 is 75 Hz lower in [y] than in [y:], and, corre-spondingly, F1 is 65 Hz higher, F2 is 140 Hz lower, and F3 is 265 Hz lower in [i] than in [i:]. The results indicate that the produced single vowels are more centralized than the double vowels are. In a later study on vowel production by Kukkonen [4], differences of similar type but smaller magnitude were reported in a normal Finnish-speaking control group (N=4): F1 was 16 Hz higher, and F2 and F3 were 63 Hz and 32 Hz lower in single than in double /i/ vowel. Correspondingly for single and double /y/ vowels, the differences were as follows: F1 was 19 Hz higher, F2 was 75Hz lower, and F3 was 20 Hz lower in the single vowel. However, only differences in F1 were statistically significant. In our earlier studies [5], a non-significant difference of 108 Hz was found for F2 between the short /i/ (F2=2391 Hz, SD=194 Hz) and long /i:/ (F2=2500 Hz, SD=212 Hz) produced by 26 informants in the first syllables of the words tikki and tiili. In a more recent study by Eerola and Savela [6], a significant difference (paired t-test, $p<0.01$, N=14) of 104 Hz was found for F2 between the short /i/ and long /i:/ in an uttered word pair tili/tiili. Iivonen and Laukkanen [7] studied the qualitative variation of the eight Finnish vowels in 352 bisyllabic and trisyllabic words uttered by a single male speaker. They found a clear tendency for the short vowels to be more centralized in the psychoacoustic F1–F2 space, as compared to the long ones. However, except for the /u/–/u:/ pair, this difference was smaller than one critical band, and thus auditorily negligible. In a comparative study of the monophthong systems in the Finnish, Mongolian and Udmurt languages, Iivonen and Harnud [8] report on minor spectral differences in the short/long vowel contrasts in stressed (e.g., [sika] / [si:ka]) and non-stressed (e.g., [etsi] / [etsi:]) syllables in Finnish words uttered by a single male speaker. The biggest differences between short and long vowels were found in /u/. As in the study by Iivonen and Laukkanen, [u] is more centralized and does not overlap with [u:]. Also for /y/ and /i/, the short vowels are more centralized than their longer counterparts, but the short and long vowel versions overlap on the F1 axis. Interestingly, the /y/ and /i/ vowels, both short and long, also overlap on the F2 axis instead of being clearly separate phoneme categories. To summarize, minor spectral differences have been reported in the F1 and F2 formant frequencies of the produced short and long Finnish vowels, and the biggest difference occurs between the high back vowels [u] and [u:].

In this study, we further examine the reported quality differences between produced short and long variants across the entire Finnish vowel system in two different noise masking conditions and without any noise mask. It was assumed that noise masking may cause hyperarticulation, and possibly accentuate the reported minor quality differences between short and long Finnish vowels. Since speakers are known to alter their vocal production in noisy environments (the Lombard effect) [9], such as a loud restaurant or a noisy factory, we included two different types of masking noise to simulate these conditions. Multi-talker babble noise at 92 dB SPL (sound pressure level) was used to simulate a loud, crowded cocktail party, and pink noise at 83 dB SPL an environment with the maximum noise level allowed for continuous working. The Lombard effect has been reported to cause measureable differences in vowel intensity and duration, and also in formant frequencies: ambient noise elevates the speech amplitude by 5–10 dB, increases word durations by 10–20%, and increases significantly the F1 and F2 frequencies, thus causing a shift in the vowel space [10,11,12].

2. Materials and methods

2.1. Subjects

Fourteen (14) normally hearing young adults (7 male, 7 female, aged 17–31 years) and speaking the modern educated Finnish of South-West Finland volunteered as subjects. All subjects were screened for hearing impairments by means of an audiometer (Amplivox 116). Not all subjects participated in the noise masking experiments.

2.2. Procedure and analysis

The articulation of the eight Finnish vowels /a/, /e/, /i/, /o/, /u/, /y/, /æ/ and /ø/ when uttered in different carrier words and non-words (e.g., /tili/ - /tiili/, see Table 1) was recorded in two different masking conditions and without a noise mask. The subjects were asked to utter each word five times successively using their normal speech style, first without the noise mask, and then in the masking conditions. The recordings were carried out in an acoustically dampened room (27 dBA SPL) by using a high quality microphone (AKG D660S) that was connected via an amplifier to a PC. The recordings were made at a sampling rate of 44.1 kHz, and saved as sound files for later analysis. Praat software was used for both the recordings and analysis.

The sound samples were automatically analyzed using a text grid in which the steady-state part of each target vowel was windowed varying between utterances. The f0, formants F1 – F4, and vowel durations were analyzed by using the Burg method in which short-term LPC coefficients are averaged for the length of an entire sound. The Praat formant analysis settings were 0.025 s for Window length, and 5000 Hz (male) and 5500 Hz (female) for Maximum formant. The analysis results of the five repetitions were averaged for individual results..

Table 1. *Carrier utterances used in the experiments.*

Short		Long	
IPA, *Finnish*	Meaning	IPA, *Finnish*	Meaning
[tali], *tali*	'tran'	[ta:li], *taali*	non-word
[teli], *teli*	'twin axle'	[te:li], *teeli*	non-word
[tili], *tili*	'account'	[ti:li], *tiili*	'brick'
[toli], *toli*	non-word/NA	[to:li], *tooli*	non-word
[tuli], *tuli*	'fire'	[tu:li], *tuuli*	'wind'
[tyli], *tyli*	non-word	[ty:li], *tyyli*	'style'
[täli], *täli*	non-word	[tä:li], *tääli*	non-word
[tøli], *töli*	non-word	[tø:li], *tööli*	non-word

2.3. Noise masks

Multi-talker babble noise at 92 dB SPL was used to simulate a loud, crowded cocktail party, and pink noise at 83 dB SPL an environment with the maximum noise level allowed for continuous working. Being difficult to synthesize, recorded babble noise was used. Pink noise was selected because of its good speech masking properties [13]. Its spectral envelope follows the spectral properties of speech signals: the peak intensity in the f0–F1 range and an even roll-out of 6 dB per octave at the higher frequencies of F1–F5 formants. Masking was on throughout the recording of each utterance, and the noise masks were presented via Sennheiser PC161 headphones, which were calibrated in the beginning of each session by Brüel & Kjaer Type 2235 SPL meter to deliver 83 +/- 0.5 dB$_A$ SPL at the pink noise mask.

3. Results

3.1. Short versus long vowels

The individual results of articulated Finnish vowels in the F1–F2 space are illustrated in Figure 1. As can be seen from the figure the /y/ and /i/, and correspondingly, /ø/ and /e/ categories overlap clearly with each other. The short and long vowels differ in terms of F1 and F2 between the categories with the differences being largest between /u/ and /u:/. Except for /y/ and /ø/, the other vowel categories show a pattern where short vowels are more centralized than long vowels. This is in accordance with the results of Iivonen and Laukkanen [7]

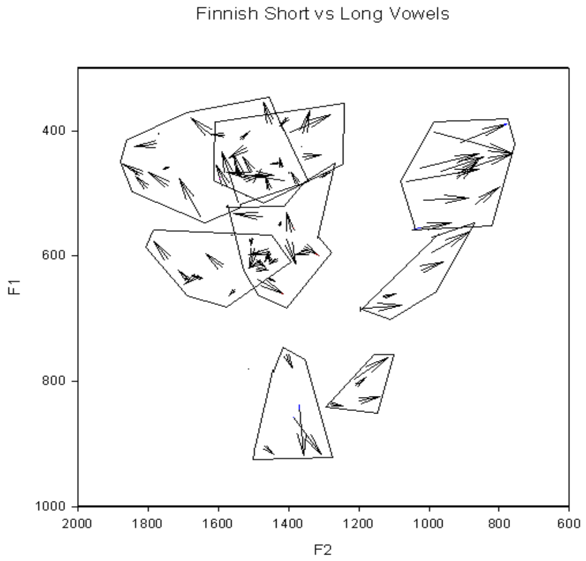


Figure 1. *Individual articulations of the short and long Finnish vowels in the F1–F2 space (in mel). Vector starting points represent the short vowels and end points the long vowels.The number of subjects varies in different categories.*

The mean values of the five repetitions of all subjects for the short and long Finnish vowels are shown in Table 2 and Table 3, respectively, and illustrated in Figure 2. The mean duration was 118 ms (SD 35 ms) for the short vowels and 324 (SD 79 ms) for the long vowels. These results are in line with the earlier reports on the durational variation of the Finnish short and long vowel quantities (for a review, see [2]).

The averaged results confirm the earlier findings that there are minor quality differences of 29–138 mel between short and long vowels in Finnish (Table 4). The mean individual distance in the F1–F2 plane between the long and short vowels without noise masking was 63 mel over all vowel categories. Variation was found between vowel categories: /e/, /y/ and /ø/ had distances of the order of 30 mel and no centralization tendency was observed, whereas /o/, /u/ and /æ/ showed clearly larger distances, up to 138 mel, and noticeable centralization of the short vowels, especially in /i/, /u/, /o/, /a/ and /æ/(Figure 2).

3.2. The effect of a masking noise

Interestingly, both types of noise masking caused a significant prolongation in the duration of the short vowels, but not of the long vowels. With babble noise, the mean durations were 141 ms (SD 36 ms) and 347 ms (SD 75 ms), and correspondingly with pink noise, 129 ms (SD 32 ms) and 339 ms (SD 77 ms). Using paired t-tests for each category, the difference between the non-masking (mean 118 ms, SD 35 ms) and babble noise conditions was highly significant (p<0.001), between the non-masking and pink noise conditions significant (p<0.05), and between the two different masking conditions highly significant (p<0.001). Since the durations increased along with increasing sound pressure level, the phenomenon may rather be explained by the amplitude of the mask than its type. However, when using a low pass filtered white masking noise, Summers et al. [10] did not find any significant differences between the effects of the 80 dB and 90 dB SPL masks on durations, but instead, they found a highly significant (p<0.0001) difference between non-masking and masking conditions. On the other hand, Beckford Wassink et al. [11] did not find significant differences in segment durations between Lombard speech and (non-mask) citation speech. Our finding that only the short vowels are prolonged with Lombard speech is interesting and motivates further investigation.

Table 2. *Mean values (and standard deviations) of the durations (in ms) and formants F1–F4 (in mel) for the produced short Finnish vowels.*

Vowel	Duration	F1	F2	F3	F4
ɑ	100 (27)	471 (37)	1708 (107)	1902 (63)	2135 (104)
æ	118 (34)	617 (27)	1608 (93)	1862 (62)	2129 (111)
e	140 (42)	840 (61)	1408 (43)	1786 (14)	2010 (52)
i	114 (39)	452 (33)	1452 (60)	1748 (88)	2037 (48)
o	121 (27)	599 (33)	1448 (46)	1805 (69)	2093 (88)
u	109 (30)	475 (45)	970 (67)	1768 (122)	2017 (109)
y	139 (43)	642 (41)	1083 (92)	1803 (62)	2032 (85)
ø	140 (43)	818 (19)	1225 (37)	1801 (33)	2054 (45)
Mean	**118** (35)	570 (127)	1399 (260)	1813 (91)	2073 (98)

Table 3. *Mean values (and standard deviations) of the durations (in ms) and formants F1–F4 (in mel) for the produced long Finnish vowels.*

Vowel	Duration	F1	F2	F3	F4
ɑ	294 (61)	449 (29)	1749 (108)	1946 (66)	2147 (113)
æ	305 (61)	617 (30)	1630 (99)	1872 (59)	2142 (109)
e	387 (96)	883 (69)	1374 (57)	1797 (39)	2078 (67)
i	316 (74)	436 (41)	1449 (88)	1732 (92)	2044 (77)
o	318 (80)	603 (42)	1444 (62)	1791 (85)	2110 (101)
u	319 (87)	459 (45)	834 (57)	1782 (113)	2059 (107)
y	396 (94)	628 (53)	1004 (98)	1818 (51)	2032 (74)
ø	366 (95)	805 (34)	1170 (56)	1801 (37)	2055 (69)
Mean	**324** (79)	563 (140)	1375 (318)	1821 (102)	2091 (101)

The effect of noise on the produced vowel quality was similar in both two masking conditions, and no major differences between babble and pink noise were found (Figure 3). Both noise types seem to cause higher F1 and F2 frequencies in the production of the mid-high vowels: On the average, the formant values of the short and long vowels produced in the masking conditions are about 50 mel (in Euclidean distances) higher than without masking. No similar effect was found for the low vowels /a/ and /æ/. The largest individual F2 difference between the masking and non-masking conditions was 70 mel for /u/. The results indicate that noise masking causes a systematic shift of F1–F2 values in the production of mid-high Finnish vowels, as illustrated in Figure 3. Using Wilcoxon test for related samples, a significant difference (p<0.05) was found in F1 between the vowels produced without masking and in the babble noise and pink noise conditions. No significant differences in the Euclidean distances between the short and long vowels were found for the different conditions.

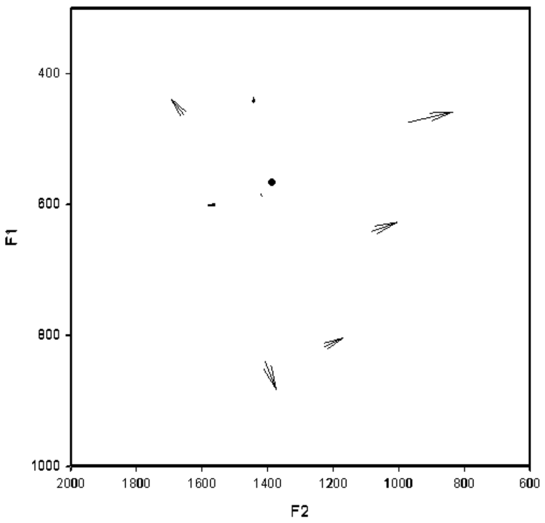


Figure 2. *The grand averages of short and long Finnish vowels in the F1–F2 space (in mel). Vector starting points represent the short vowels and end points the long vowels.*

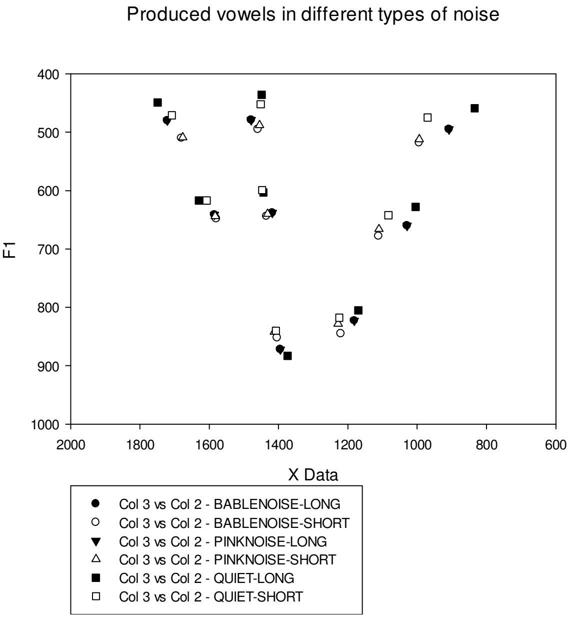


Figure 3. *The grand averages of short and long Finnish vowels in the F1–F2 space (in mel) in the two different masking conditions and without noise masking.*

Table 4. *Mean Euclidean distances (and standard deviations) in mels between the produced short and long Finnish vowels without noise masking (column S-L), and between the short vowels without and with babble (SBN) and pink noise (SPN) masking, and between the long vowels without and with the babble (LBN) and pink noise (LPN) masking.*

Vowel	S-L	SBN	SPN	LBN	LPN
ɑ	57 (32)	46 (20)	44 (14)	54 (37)	50 (45)
æ	59 (36)	50 (11)	33 (11)	93 (78)	47 (19)
e	29 (16)	53 (33)	50 (30)	53 (33)	48 (26)
i	49 (22)	59 (33)	58 (35)	44 (22)	44 (19)
o	80 (37)	55 (24)	38 (16)	63 (43)	63 (36)
u	138 (52)	58 (31)	57 (30)	83 (61)	83 (39)
y	56 (48)	59 (34)	56 (37)	58 (37)	63 (55)
ø	39 (23)	61 (30)	53 (23)	65 (44)	66 (36)
Mean	63 (49)	56 (29)	51 (28)	62 (44)	59 (37)

4. Discussion and conclusions

The results of this study on the production of the short and long Finnish vowels confirmed, first, the earlier findings that the short vowels /i/, /u/, /o/, /a/ and /æ/ are more centralized in the F1–F2 space than their longer counterparts. Second, the Lombard effect induced by the two different noise masks caused the duration of the short vowels, but not the long ones, to increase significantly. The increase was larger with the louder babble noise than with the pink noise. Whether this difference was due to the higher amplitude of the babble noise or due to the noise type itself is a subject for further studies.

Third, the Lombard effect resulted in an increase in the F1 of the mid-high vowels, but had no effect on the Euclidean distances of the short and long vowels. These results in terms of the F1 value and the Euclidean distances are in line with the findings of Summers et al. [10], and Beckford Wassink et al. [11]. The latter study among Jamaican speakers is particularly interesting, since Jamaican Creole utilizes the phonemic vowel length in a similar manner as Finnish, which, however, is a distinctive quantity language. The vowel quality (in terms of F1 and F2) was affected similarly by the Lombard speech in both these languages, but a clear durational prolongation of short vowels was only found in Finnish.

5. Acknowledgements

The study was supported by a grant from the Finnish Cultural Foundation. We wish to thank Lea Heinonen-Eerola, M.A., for revising the English language of the manuscript.

6. References

[1] Karlsson, F., "Suomen kielen äänne- ja muotorakenne" [Sound and Form Structures in Finnish], Werner Södesrstöm Oy, Porvoo, Finland, 1983.

[2] Eerola, O., Savela, J:, Laaksonen, J.P., and Aaltonen, O., "The effect of duration on vowel categorization and perceptual prototypes in a quantity language", Journal of Phonetics, 40(2), 315-328, 2012.

[3] Wiik, K., "Finnish and English vowels", (Doctoral Thesis, University of Turku). Annales Universitatis Turkuensis, Series B (94), 1965.

[4] Kukkonen, P., "Patterns of phonological disturbances in adult aphasia", (Doctoral Thesis, University of Helsinki), Suomalaisen Kirjallisuuden Seuran Toimituksia, (529), 1-231, 1990.

[5] Eerola, O., Laaksonen, J.P., Savela, J., and Aaltonen, O., "Perception and production of the short and long Finnish [i] vowels: Individuals seem to have different perceptual and articulatory templates," Proceedings of the 15th International Congress of Phonetics Sciences, Barcelona, 2003.

[6] Eerola, O., and Savela, J., "Differences in Finnish front vowel production and weighted perceptual prototypes in the F1-F2 space", Proceedings of the 17th International Congress of Phonetics Sciences, Hong Kong, 2011.

[7] Iivonen, A. and Laukkanen, A., "Explanations for the qualitative variation of Finnish vowels", Studies in Logopedics and Phonetics, 4, 29-55, 1993.

[8] Iivonen, A. and Harnud, H., "Acoustical comparison of the monophtong systems in Finnish, Mongolian, and Udmurt", Journal of the International Phonetic Association, 35(1), 59-71, 2005.

[9] Lane, H.L., and Tranel, B., "The Lombard sign and the role of hearing in speech", J. Speech Hear. Res., 14, 677-709, 1971.

[10] Summers, W.Van, Pisoni, D., Bernacki, R., Pedlow, R., and Stokes, M., "Effects of noise on speech production: Acoustic and perceptual analyses", J. Acoust. Soc. Am., 84(3), 1988.

[11] Beckford Wassink, A., Wright, R., Franklin, A., "Intraspeaker variability in vowel production: An investigation of motherese, hyperspeech, and Lomabrd speech in Jamaican speakers", Journal of Phonetics, 35, 363-379, 2007.

[12] Castellanos, A., Benedi, J-M., Casacuberta, F., "An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect", Speech Communication, 20(1-2), 23-35, 1996.

[13] Rao, M. and Letowski, T., "Callsign Acquisition Test (CAT): Speech Intelligibility in Noise," Ear & Hearing, 27, 120-128, 2006.

Revisiting the Meadow Mari vocalic system

Dennis Estill

University of Helsinki, Department of Finnish, Finno-Ugrian and Scandinavian Studies

estill@mappi.helsinki.fi

Abstract

Meadow Mari is a Finnic language spoken in certain parts of Russia, especially in the region of the Volga Bend. There are two literary Mariic languages, Meadow and Hill. The purpose of this article is to define the present system of vowels in Standard Meadow Mari on the basis of empirical research and in the light of traditional concepts. Traditional descriptions of the vocalic system vary considerably both on the close/open and forward/backward scales (compare Alhoniemi, Bereczki, Grigoryev, Pengitov). The specific objectives of this experiment were (1) to ascertain whether the three so-called weak full vowels can be described as such or whether they are better described as reduced vowels, (2) to determine the quality of the central vowel, and (3) to measure the extent of roundedness of all vowels. The paper includes a brief consideration of personal features that may be apparent from analyses of formants 3 and 4. Although formants 3 and 4 provide information about roundedness, their wide variation between individuals offers clues concerning speaker-specific vowel quality. In considering problems (1) and (2) it is necessary to calculate those acoustic parameters that bear on reduction and the integrity of the individual sound of the respective vowels. This was analysed from four perspectives: (a) the positions of the vowels on a chart displaying formants 1 and 2, (b) vowel length or duration, (c) pitch or fundamental frequency and (d) loudness or intensity. In attempting to resolve the problem referred to in (3) above formants 2, 3 and 4 were analysed, although less emphasis was placed on formant 2 for reasons mentioned in the paper. The article presents the results of a comprehensive study of 2,274 Meadow Mari vowel tokens from a corpus of approx. 500 words. The method used is statistical analysis of acoustic measurements using a speech analysis computer program. The findings, which are based on evidence from two informants, suggest that there are only two vowel categories in Meadow Mari, viz. full and reduced, as opposed to views which add a third category, weak. Further, the central vowel is reduced and it is neutral with regard to roundedness, presumably depending on the environment. Roundedness can only be viewed as occurring on a continuum in the case of the vowels /a/, /e/, /u/, /ö/ and /ə/ and not as categorical. Attention is further drawn to the need to determine which symbols should be used to describe the central vowel. Generally speaking IPA *ɣ* and SUT *ə* have been used in this connection. *ɣ* seems unsatisfactory because it implies unroundedness and should probably be replaced by the roundedness-neutral *ə*. In SUT the central vowel could also very well be represented by the symbol *ə*, although in the vowel chart it occupies the position also held by the rounded, near back and mid-close *ǫ*, that is to say, a central vowel positioned between *o* and *ə*. These findings indicate the need to consider adjusting descriptions of at least some aspects of the Meadow Mari vocalic system.

Index Terms: Meadow Mari, reduction, central vowel, rounding, individual features

1. Introduction

The purpose of this article is to define the present system of vowels in Standard Meadow Mari [1] (SMM) on the basis of empirical experimentation and in the light of traditional concepts. This appraisal will include descriptions of the location of SMM vowels in the vowel space along with an analysis of the degree to which they are reduced, and information regarding the extent to which the vowels may or may not be rounded. In this way it should be possible to determine how far previous descriptions, mostly arrived at without any analysis of the acoustic parameters, fit in with the findings. These findings may create a need to adjust our understanding of at least some aspects of the Meadow Mari vowel system.

2. Traditional descriptions of the Meadow Mari vowels

A typical description of the Meadow Mari vocalic system is found in N. T. Pengitov [8] and is shown below:

close	*i y*			*u*
		ə		
mid	*e ö*		*o*	
open		*a*		

According to this arrangement, there are three levels on the close/open scale, viz. close, mid and open. However, /ə/ is given the position of an unreduced close-mid central vowel, while /y/ and /ö/ are shown as front vowels. Pengitov considers reduction to be a general feature, presumably meaning that it can occur in the case of any vowel in specific circumstances. Ya. G. Grigoryev [4], on the other hand, states that in the case of the unstressed vowels /e/, /o/ and /ö/ in word-final position there is reduction, whereas in the case of unstressed /ə/, reduction does not occur and the realisation of the sound is very short. Alho Alhoniemi's [1] configuration of the SMM vowel chart in his description of Mari grammar is the following:

i	*ü*	*u*
e	*ö*	*o*
		a
----------------	----------------	----------------
		ə̂

No reason is given by Alhoniemi for drawing a broken line above /ə̂/, but it may suggest the difficulty of allocating a specific point in the vowel space for what he considers to be a reduced vowel. Alhoniemi also considers /ə̂/ to be a back vowel, as opposed to other descriptions. Otherwise the pattern is very similar to Pengitov and Gábor Bereczki [2], who also separated the central vowel by drawing a line to separate the central vowels from the others and whose chart is shown on the right below. Another problem with the Alhoniemi description is that it does not show the extent of openness and closedness but, rather, presents three equidistant levels. According to Alhoniemi [1] there are three types of vowel in Meadow Mari: strong full, weak full (word final *e*, *o* and *ö*) and the reduced *ə*. One of the most recent descriptions of SMM vowels is that of I. G. Ivanov [5]. In his chart front and back vowels form two separated groups:

	front	back
close	*i, y*	*ə, u*
mid	*e, ö*	o
open		*a*

(I. G. Ivanov 2000)

ü	*i*	*u*
ö	*e*	*o*
		a
ə		

(Bereczki 1990)

The most important difference between Ivanov's layout and the others is that he places *ə* among the close vowels.

3. Measuring the acoustic parameters—background and methodology

The object of the study described below was to determine as far as possible the exact position of SMM vowels in the vowel space, to consider the role, if any, of reduction, to tackle the issue of the proposed strong-weak division and to examine another dimension to the understanding of Meadow Mari vowels, viz. to describe roundedness.

There were two female informants for the experiment, EA and LS, whose ages and places of birth were respectively 39, Zvenigovskij *raion*, and 38, Volzhskij *raion* in the Republic of Mari El. Both are teachers of Mari. For the experiment the standard dialect was considered. The text chosen for recording was composed of several short readings from a teaching manual intended for Russian students [7]. All told the material consisted of 500 words. The recordings took place in an auditorium in the Department of Finno-Ugrian Studies of the University of Helsinki where every attempt was made to keep outside noise down to a minimum. All doors and windows were closed and electrical interference eliminated as far as possible. A Plextalk recorder was used for the recordings and the sound quality later calibrated for fidelity. Although the quality of the sound was considered satisfactory, it did fall a little short of what might be expected in perfect laboratory conditions. The material was read once by each speaker. The microphone used was an AKG D 660S and the acoustic measurements made using the Praat 5.143 program.

Measurements were taken of the values of formants 1–4, and in addition vowel duration, intensity and fundamental frequency (F_0) were recorded. The formant data were means taken from slices of the core of the vowel carefully chosen in order to eliminate as far as possible the possible effects of, for example, co-articulation. The results were then analysed.

4. The findings

The three objectives of this experiment were (1) to ascertain whether the three so-called weak full vowels should be described as such or whether they are better described as reduced vowels, (2) to determine the quality of the central vowel, and (3) to measure the extent of roundedness of all vowels.

For considering problems (1) and (2) it is necessary to calculate and determine those acoustic parameters that bear on reduction and the integrity of the individual sound of the respective vowels. The vowels were analysed from four perspectives: (a) the positions of the vowels on a chart displaying formants 1 and 2, (b) duration, (c) pitch or fundamental frequency and (d) loudness or intensity. In attempting to resolve the problem referred to in (3) above, formants 2, 3 and 4 were analysed.

4.1. Reduced SMM vowels

According to a definition given by John Laver [6], vowel reduction can mean that 'the vowel is pronounced shorter, less loud, lower in pitch and more central in quality'. Therefore, vowel reduction is not only a question of shorter duration. Nor is reduction necessarily correlated with centrality; rather an examination of all relevant factors should be made.

4.1.1. Centrality

To determine the position in the vowel space of the central vowel and the "weak" vowels, measurements were made of formants 1 and 2 and the positions of the vowels were entered into a chart, along with the other seven SMM vowels. These are shown below in figure 2.

The material read by the informants produced 2,274 vowel tokens whose percentage frequency of occurrence is shown in the pie charts below (figure 1).

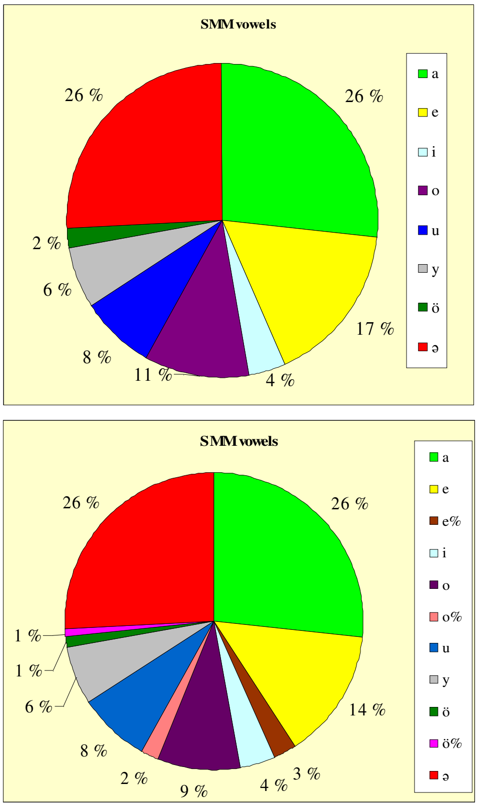


Figure 1. *Percentage of individual SMM vowels in material read by two informants (upper), and percentage of individual SMM vowels with /e/, /o/ and /ö/ divided into categories "full" and "weak" (lower). A % sign is used to indicate "weak" vowels.*

It would be extremely problematic to devise suitable, that is to say fluently readable, material for analysis wherein each vowel would have the same overall frequency, if not least on account of the repetition of the same word(s)—and co-articulation—in the same phonetic environment. Therefore, above all, this analysis took into consideration the need for the reading material to be of the best possible quality, diverse and interesting, in order to produce the right findings. Of course, the readers were given several days to become acquainted with the part to be read, and a number of questions about the text were discussed. There were only 20 occurrences of the "weak" vowel /ö/, but this is unlikely to greatly affect the results as a whole, and as far as the central vowel is concerned approximately 600 tokens were analysed. There were also very slight differences in the sample totals according to which parameter was under consideration.

Figure 2 displays the SMM vocalic system, based on the measurement of approximately 2,300 tokens.

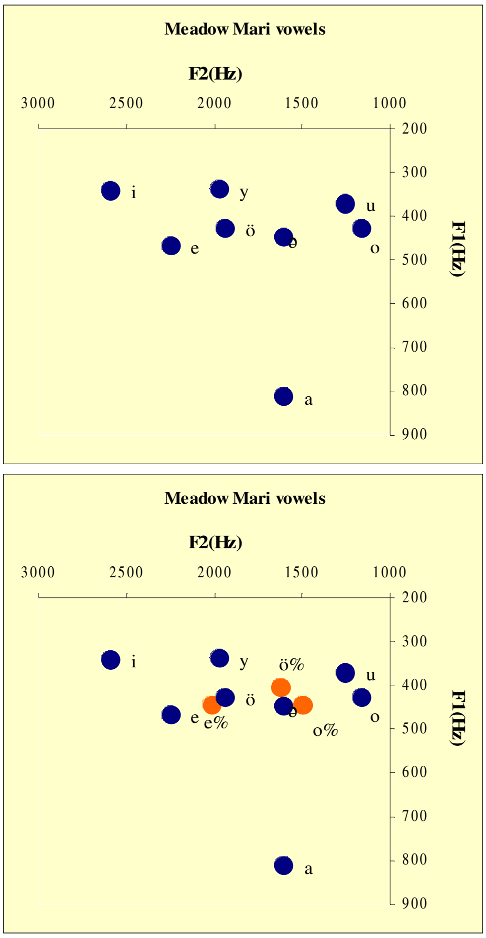


Figure 2. *The eight-vowel SMM vocalic system (upper) and the SMM vocalic system (lower) plus the "weak" vowels (orange). A % sign is used to indicate "weak" vowels.*

The upper vowel chart could thus now be converted into the following form:

	front→			←back
close	*i*	*y*		*u*
mid-close	*e*	*ö*	*ə*	*o*
mid-open				
open			*a*	

With slight alterations these conform quite well to the traditional descriptions shown above. Leaving the mid-open row empty emphasises both the nearness of the top two rows to each other and, accordingly, the distance between /a/ and the other vowels. The chart on the left (b) in figure 2 shows the relative position of the "weak" vowels with regard to the "parent" vowel. It will be observed that all three centralise to the same degree. /e/ moves close to the space occupied by /ö/ and /o/ and /ö/ near that occupied by /ə/. This phenomenon can only be called reduction, and a division into full (= having all the properties of the vowel in question) vowel and "weak" (= having all the properties of the vowel in question in a "weaker" form) adds nothing to a definition of the vowel as reduced. This will be further confirmed when the the features of phonological length, loudness and pitch are considered.

As far as the central vowel /ə/ is concerned figure 2 also shows that it is placed in a mid-close position between /ö/ and /o/. For the moment we can only locate the position of /ə/. Whether it is full or reduced must be resolved by looking at the other properties of the vowel, which are treated below.

4.1.2. *Duration*

The feature most often related to reduction is the phonological length or duration of a vowel. So much so that for some it often seems that this parameter alone is enough to label a vowel as reduced. I have referred to this problem earlier in connection with Moksha [3]. However, the Moksha central vowel is in fact reduced. Comparisons I made between the Moksha /ə/ and the Romanian central vowel clearly indicated this. What about Meadow Mari? The bar chart below (figure 3) compares the duration of the eight SMM vowels and the three "weak" vowels.

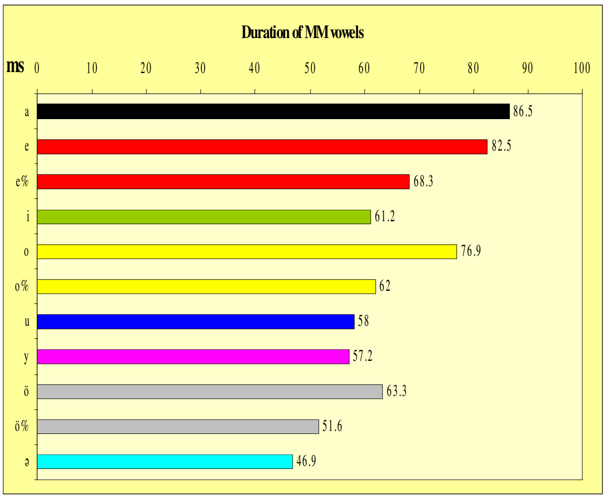


Figure 3. *Duration of eight SMM vowels and three "weak" vowels in milliseconds. A % sign is used to indicate "weak" vowels.*

If duration is the criterion by means of which reduction is determined, then there can be no doubt whatsoever that not only the SMM central vowel, but also the three "weak" vowels are reduced. As percentages of the full vowels the "weak" vowels were 83 (/e/), 78 (/o/) and 82 (/ö/). The colour pairs in figure 3 illustrate this. As far as the central vowel is concerned, it was the shortest of all in duration, with /a/ almost twice as long. Considering the evidence from Romanian, a full central vowel—and the Romanian central vowel is such—is not shorter, at least to any notable degree, than the other full vowels. In fact in Romanian only /a/ was significantly longer than the central vowel. In the case of SMM /ə/ is the shortest vowel of all. The intrinsic relative length of SMM vowels is very close to Romanian: from long to short (as shown in figure 3) a – e – o – i/ö – u/y – ə, and for Romanian a – ă/e – o – â – i – u (ă = ə, â = y). This means that while the Romanian central vowel is long, that of SMM is very short, all other things – almost all – being equal. This seems to be a case for describing the SMM central vowel as reduced. But the matter should be further verified by examining loudness and pitch, which are the next subjects for consideration.

4.1.3. *Intensity*

The acoustic correlation of loudness is intensity. The mean intensity of SMM vowels is shown in figure 4.

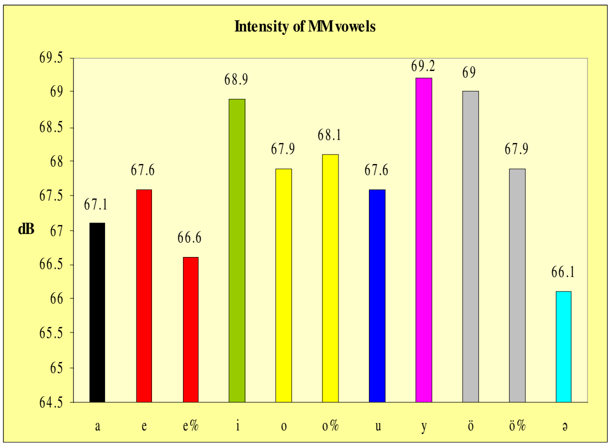


Figure 4. *Intensity of eight SMM vowels and three "weak" vowels in decibels. A % sign is used to indicate "weak" vowels.*

Although there is a significant difference between the intensity of /e/ and /ö/ and their "weak" counterparts, the same cannot be said of /o/. In this case there is no significant difference in intensity between full and "weak". Calculated as totals the difference between these two categories (full and "weak") is 0.5 dB. Therefore, although the degree of intensity seems to be less affected, some reduction does appear to take place. The intensity chart accords with the findings for duration, insofar as the central vowel occupies the last place in the comparison. That is, the level of intensity does not appear to be what should be expected of a full vowel. The case for reduction gathers support from intensity measurements. The final feature for consideration in this section is pitch. Figure 5 depicts the situation in this respect.

4.1.4. *Fundamental frequency*

Figure 5 presents calculations of the fundamental frequency (F_0) of the SMM vowels. As far as the three full vowels are concerned, these figures indicate that there is no difference in this property, be the vowels in question full or "weak". There is, however, enough evidence from the other parameters to show that in this case reduction does take place, since reduction is not defined on the basis of a consensus between the factors causing it. On the other hand, the central vowel still finds itself in the "jumbo" position, although not perhaps to the same extent as earlier. The position of the central vowel in this chart showing F_0 does not conflict with other measurements demonstrating reduction.

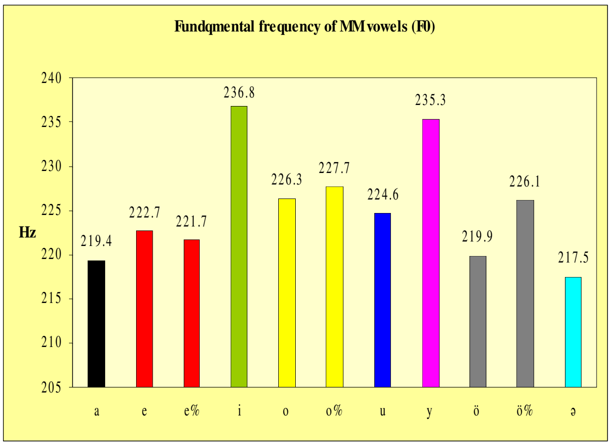


Figure 5. *Fundamental frequency (F_0) of eight SMM vowels and three "weak" vowels in Hertz. A % sign is used to indicate "weak" vowels.*

4.2. Roundedness

Since grammars describe and show Mari vowels as either rounded or unrounded, it is important to establish the criteria for roundedness, while at the same time determining the acoustic parameters related to this feature, in this case in SMM. Using experimental methods, detection of roundedness is generally associated with formants 2 (strongly connected with front~back articulation), 3 and 4. Roundedness is considered to be related to low readings in frequency in these formants. Three formant charts are presented in figure 6 and these display 2D combinations of formants 2, 3 and 4.

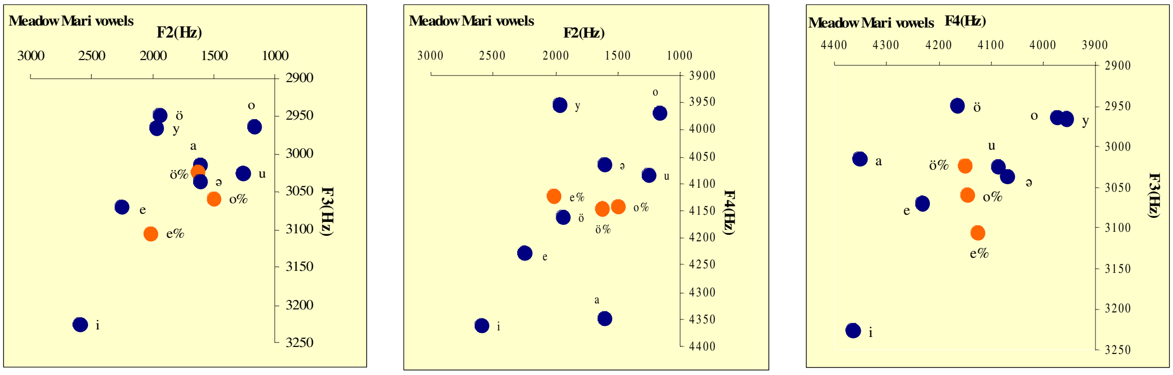


Figure 6. *Formant charts showing 8 SMM vowels and 3 "weak" vowels. Formants 2 and 3 (left), formants 2 and 4 (mid) and formants 3 and 4 (right). A % sign is used to indicate "weak" vowels.*

If the general opinion that low frequency correlates with roundedness is accepted and used as the principal guideline, then the dispersion of points in the three charts suggests that very careful interpretation is demanded. Some rounded vowels such as IPA /ø/ and /y/ are frontal and an interpretation based solely on F_2 would be misleading. F_3 and F_4 are probably more revealing. In theory, if F_3 and F_4 do indicate roundedness this feature should be apparent on a diagonal continuum from top right (roundedness) to bottom left (lip spreading) in charts of the type shown above. In figure 6c the most obviously rounded /o/ and /y/ and the most clearly unrounded /i/ are precisely where they would be expected, at the limits of the continuum. While keeping an eye on the patterns shown in figures 6a and 6b in which the effect of F_2 is evident, most attention should be paid to figure 6c in which F_3 and F_4 are displayed and the drop in frequencies related to backness reduced. In these circumstances /a/, /e/, /u/, /ö/ and /ə/ are located in the area between the two extremes and are either less rounded, /u/, /ö/ /ə/, or less spread, /a/, /e/. The three “weak” vowels are once again centralised, or reduced in terms of roundedness (/o/) varying roundedness (/ö/), and unroundedness (/e/) and their position on the rounded/unrounded continuum should be considered as part of a process of further reduction, thus adding to the evidence already given. /ə/ is so situated that it might well be described as either partially rounded or neutral with respect to the rounded-spread opposition. Conclusion: /ə/ falls in an area between rounded and unrounded and should not be considered an unrounded vowel. It is the mid-central vowel which has no rounded-unrounded variants in the IPA vowel chart. Adding a third dimension to the vowel chart shown in section 4.1.1 using degrees of colouring from green (rounded) to red (lip spread) passing through yellow, the following 3D vowel chart is proposed for SMM:

	front←			→back
close	*i*	*y*		*u*
close-mid	*e*	*ö*	*ə*	*o*
open			*a*	

All of this raises a further question: which IPA and SUT symbols should be used to describe the central vowel? Generally speaking IPA *ɣ* and SUT *ə* have been used in this connection. *ɣ* seems unsatisfactory because it implies unroundedness and should probably be replaced by the roundedness-neutral *ə*. In SUT the central vowel could also very well be represented by the symbol *ə,* although in the vowel chart it occupies the position also held by the rounded, near back and mid-close *ǫ*, that is to say, a central vowel positioned between *o* and ə.

4.3. Individual features

Not only do formants 3 and 4 provide information about roundedness, but also their wide variation between individuals offers clues concerning speaker-specific vowel quality. The extent to which individual features are visible in formants 3 and 4 probably distorts the data relevant to lip roundedness and, thus, these findings should not be treated as absolute relative values. An example of the variation in these formants between speakers, using my informants, is shown in figure 7 in which the values for the two individual informants have been separated from the totals.

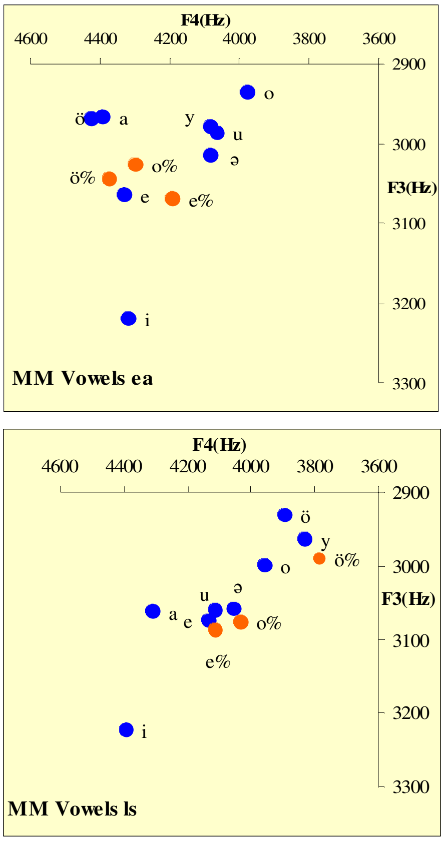


Figure 7. *Charts showing the formant 3 and 4 readings for 8 SMM vowels and 3 “weak” vowels for two informants ea (upper) and ls (lower). A % sign is used to indicate “weak” vowels.*

Informant *ea* has much lower fundamental frequency (F_0), 195,4 Hz compared to 249,6 Hz for *ls,* and F_1 and F_4 are higher in the case of *ea*, 585,9 Hz and 4210 Hz respectively as against 474,5 Hz and 4123 Hz for *ls*. Bearing this in mind and looking at the F_3/F_4 patterns for both informants, a number of differences can be observed. Most strikingly, while the F_3 values for /ö/ and “weak” /ö/ are similar for both speakers, their F_4 values are located at separate ends of the chart. An examination of the placement in the charts of the other vowels also includes many variations. On the other hand, the F_1/F_2 separate charts for *ea* and *ls* show no significant variation in pattern as can been seen in figure 8 below.

It is beyond the object of this article to determine exactly how the complex resonances in the oral cavity actually operate to produce those specific frequencies that reflect personal voice quality and what the particular part played by the third and forth formants (and possibly the second) actually is. What can be ascertained, however, when allowing for these factors, is some general picture of the degree of roundedness that is an important feature of SMM vowels. This being the case it can been said that /i/ is unrounded, /y/ and /o/ rounded and /a/, /e/, /u/, /ö/ and /ə/ either partially rounded or sometimes rounded, for example, depending on the phonetic environment. This diverges from conventional opinion, which regards the SMM as split into two categorical opposites, rounded (/o/, /ö/, /u/, /y/) and spread (/a/, e/,/i/, /ə/).

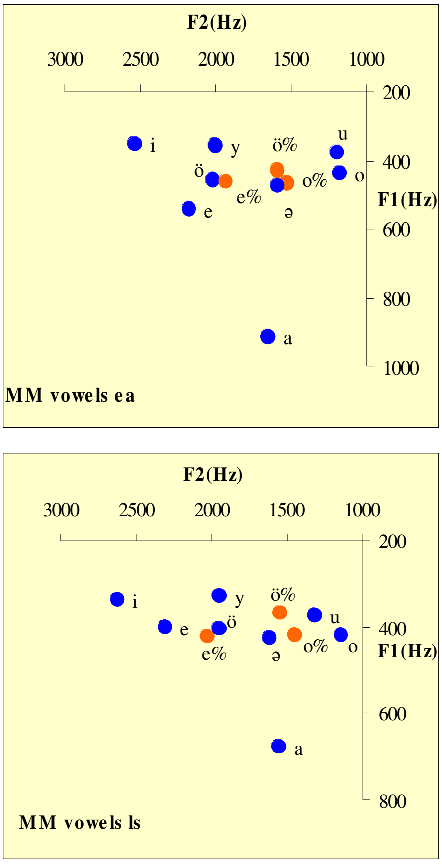


Figure 8. *Charts showing the formant 3 and 4 readings for 8 SMM vowels and 3 "weak" vowels for two informants ea (upper) and ls (lower). A % sign is used to indicate "weak" vowels.*

5. Conclusion

The questions posed at the beginning of this article concerned the "weak" vowels, the nature of the central vowel and the roundedness aspect of SMM vowels. Reduction was shown to be the factor involved in the production of the "weak" vowels and this suggests that SMM vowels should be described as reduced, rather than weak, especially when a clear definition of weak has not been forthcoming. The material analysed showed that the central vowel is reduced even if a "control" or equivalent full vowel does not exist in SMM, since full central vowels such as the Romanian /ă/ display more salient acoustic parameters. The central vowel is just as rounded as it is unrounded, presumably depending on the environment. However, the symbol ə would seem most suitable in IPA and SUT, since it implies reduction, unlike the SUT ọ.

6. References

[1] Alhoniemi, A., "Marin kielioppi", Helsinki: Suomalais-Ugrilainen Seura, 1985.

[2] Bereczki, G., "Chrestomathia ceremissica", Budapest: Tankönyvkiadó, 1990.

[3] Estill, D., "The enigmatic central vowel in Moksha. How central, how reduced?", in S. Werner and T. Kinnunen [Eds], XXVI Fonetiikan päivät 2010, University of Eastern Finland electronic publications: Joensuu, 33–37, 2011.

[4] Grigoryev, Ya. G. = Я. Г. Григорьев, „Марийыкий язык", Марийыкое книжное издательство, 7–9, 1953.

[5] Ivanov, I. G., „Кызытсе марий йылме. Фонетика", Йошкар-Ола: Марий книга савыктыш, 56, 2000.

[6] Laver, J. "Principles of phonetics", Cambridge, 1994.

[7] Зорина З, Г. и др., „Марийский язык для всех", Част 2/3, Йошкар-Ола: Марийское книжное издательство, 2000.

[8] Pengitov, N. T., „Совдоставительная грамматика руссково и марийсеого языков", Часть первая, Йошкар-Ола: Марийское книжное издательство, 20–21, 1958.

Native Finnish and English Speakers’ Fundamental Frequency, Equivalent Sound Level, and Long-Time Average Spectrum Characteristics in Text-reading in Finnish and English

Kati Järvinen & Anne-Maria Laukkanen

Voice Research Laboratory, School of Education, University of Tampere, Finland

kati.jarvinen@uta.fi, anne-maria.laukkanen@uta.fi

Abstract

Fundamental frequency, equivalent sound level and long-time average spectrum characteristics compared between text reading in native and in foreign language. Significant changes in the mean F0 were found for the Finnish subjects, and differences between the Finnish and the English subjects. Significant changes in the relative amplitude and frequency of the spectral peaks were found in both subject groups and between the groups. Differences in the mean frequency of the spectral peaks may reflect differences in formant frequencies, while differences in the spectral slope may be related to differences in F0 and other culture dependent vocal ideals.

Index terms: native language, foreign language, vocal characteristics, F0, Leq, LTAS

1. Introduction

Previous studies suggest that voice changes when speaking a foreign language compared to speaking the native one, and that languages may differ from each other in terms of vocal characteristics, such as F0 and voice quality [1-6]. For example in Finland a relatively low pitch is considered favourable whereas in Britain a higher pitch has traditionally been attributed to a high status in the society [7].

This study investigated native Finnish and English speakers' fundamental frequency (F0), equivalent sound level (Leq) and long-time average spectrum (LTAS) characteristics in foreign language compared to the native language. LTAS shows the average sound energy distribution in a speech or singing sample, and when the sample duration is long enough, the form of LTAS is no longer supposed to be affected by individual speech sounds or prosodic aspects but to reflect overall voice spectral characteristics [8].

1.1. Methods

Sixteen native speakers of Finnish (8 males and 8 females, mean ages 37.9 and 32.4 respectively) and sixteen native speakers of English (8 males and 8 females, mean ages 42.1 and 32.8 respectively) read aloud a text (duration of approximately 1 minute) first in their native language and then in the foreign language. The Finnish subjects had a longer formal education (14 years in average) in the foreign language than the English subjects (0.7 years in average), while the English subjects had a longer residence in Finland (87 % more than 5 years) than the Finnish subjects in an English speaking country (100 % less than 5 years).

The samples were recorded in a well-damped studio using Bruel&Kjaer Mediator, level meter and microphone, placed in front of the subject at a distance of 40 cm from the mouth. The signal was recorded with Sound Forge 7.0 software, using 44.1 kHz sampling frequency, the amplitude resolution was 16 bits. The signal was calibrated for calculation of Leq.

1.2. Acoustical analysis

The samples were analysed for the mean, standard deviation and range of F0 with Praat 5.1.15 signal analysis software. Cross-correlation method was used for F0 detection. In studying F0 analysis in semitones, the zero point was set at 100 Hz. Long-time average spectrum analysis and equivalent sound level measurements were carried out with Intelligent Speech Analyser (ISA) signal analysis system developed by Raimo Toivonen, M.Sc.Eng. Voiceless sounds and pauses were automatically excluded from the LTAS analysis. The frequency range of 10 kHz was used, time window was 50 ms, and Blackman-Harris window weighting was used. The strongest spectral peak in LTAS was set to zero and the relative amplitudes and mean frequencies of the strongest peaks between 0-1 kHz, 1-2 kHz, 2-3 kHz and 3-4 kHz were manually measured.

1.3. Statistical analysis

PASW Statistics 18 software was used for the statistical analysis. Related-samples Wilcoxon signed ranks test was used to study significance of the changes within groups, and Kruskal-Wallis test for significance of the difference in the change between groups. Spearman's correlation coefficient was used for the analysis of correlation between acoustic variables.

2. Results

F0 changed significantly for the Finnish subjects in the target language compared to the native language. F0 was 6.5 Hz (p=.001) and less than 1 semitone (p=.002) higher in English than in Finnish. For the English subjects the change was not statistically significant.

Table 1. *Means, SD and range of F0 in Hz and semitones.*

		F0 (Hz) in native	F0 (Hz) in foreign	F0 (st) in native	F0 (st) in foreign	F0 sd (Hz) in native	F0 sd (Hz) in foreign	F0 range (Hz) in native	F0 range (Hz) in foreign	F0 range (st) in native	F0 range (st) in foreign
Finnish male, n=8	$\bar{x}$	103.87	107.88	0.27	0.88	14.26	15.22	101.88	100.44	16.17	16.03
	sd	19.54	21.01	3.06	3.15	3.81	4.36	19.96	20.30	1.80	2.49
Finnish female, n=8	$\bar{x}$	181.38	190.35	10.08	10.98	24.02	24.48	172.07	169.47	15.22	15.15
	sd	17.32	14.54	1.62	1.28	4.85	4.86	37.56	32.35	3.07	2.47
English male, n=8	$\bar{x}$	105.26	105.87	0.53	0.66	16.64	16.12	117.25	115.84	17.35	17.27
	sd	14.24	13.53	2.36	2.25	6.06	5.81	30.31	32.13	2.61	3.20
English female, n=8	$\bar{x}$	200.01	200.16	11.82	11.85	26.74	25.53	189.55	190.59	15.73	16.03
	sd	11.31	13.45	0.96	1.19	6.44	6.39	34.90	40.23	3.71	3.09

Leq did not change significantly when speaking the foreign language compared to speaking the native one.

Table 2. *Means, SD and range of Leq in dB.*

		Leq (dB) in native	Leq (dB) in foreign	Leq sd (dB) in native	Leq sd (dB) in foreign	Leq range (dB) in native	Leq range (dB) in foreign
Finnish male, n=8	x̄	69.10	69.38	2.97	2.99	17.51	17.29
	sd	4.81	5.37	0.51	0.54	3.97	3.35
Finnish female, n=8	x̄	68.94	69.90	2.55	2.64	13.88	14.90
	sd	1.31	1.96	0.55	0.41	3.17	2.19
English male, n=8	x̄	66.27	66.99	2.57	2.38	14.39	12.58
	sd	4.49	4.30	0.73	0.49	4.70	3.90
English female, n=8	x̄	71.10	71.64	3.23	3.05	18.13	17.69
	sd	4.54	3.82	0.68	0.63	5.74	5.06

In LTAS a significant change in the relative amplitude of the spectral peak was found between 1 and 2 kHz (p=.04) and 2 and 3 kHz (p=.02) for the Finnish subjects when speaking English. The relative amplitude was 1.4 dB lower in English than in Finnish. For the English subjects a significant change in the central frequency of the spectral peak between 1 and 2 kHz (p=.01) was found, frequency being 123 Hz lower in Finnish than in English.

Table 3. *Means and sd of central frequency and relative amplitude of the strongest spectral peaks between 0 and 1 kHz, 1 and 2 kHz, 2 and 3 kHz, 3 and 4 kHz.*

		0-1 kHz (Hz) in native	0-1 kHz (Hz) in foreign	1-2 kHz (Hz) in native	1-2 kHz (Hz) in foreign	2-3 kHz (Hz) in native	2-3 kHz (Hz) in foreign	3-4 kHz (Hz) in native	3-4 kHz (Hz) in foreign
Finnish male, n=8	x̄	336.45	298.77	1157.41	1187.02	2441.32	2484.40	3369.94	3348.41
	sd	140.47	85.70	96.13	83.35	192.58	171.48	163.59	185.95
Finnish female, n=8	x̄	419.89	382.21	1205.85	1213.93	2395.56	2430.56	3488.37	3434.54
	sd	108.58	22.29	94.21	89.85	304.91	379.73	380.00	407.91
English male, n=8	x̄	282.62	253.01	1187.01	1127.80	2398.26	2392.87	3442.62	3493.76
	sd	72.28	56.09	156.74	56.31	152.70	110.66	128.39	98.17
English female, n=8	x̄	403.74	384.90	1491.17	1302.75	2392.82	2349.76	3343.03	3391.48
	sd	98.17	74.09	186.57	127.65	290.53	287.31	385.67	420.08
		0-1 kHz (dB) in native	**0-1 kHz (dB) in foreign**	**1-2 kHz (dB) in native**	**1-2 kHz (dB) in foreign**	**2-3 kHz (dB) in native**	**2-3 kHz (dB) in foreign**	**3-4 kHz (dB) in native**	**3-4 kHz (dB) in foreign**
Finnish male, n=8	x̄	-1.42	-1.34	-11.48	-12.58	-21.49	-23.41	-25.44	-26.47
	sd	1.65	1.81	3.67	3.25	3.45	2.90	4.64	3.89
Finnish female, n=8	x̄	-1.97	-1.75	-11.26	-13.02	-23.29	-24.19	-29.66	-29.12
	sd	0.99	0.89	4.47	3.26	2.54	2.57	3.66	0.96
English male, n=8	x̄	-1.19	-1.19	-12.29	-12.60	-23.11	-21.87	-31.13	-29.01
	sd	1.92	1.59	2.81	1.76	2.93	2.83	3.63	3.11
English female, n=8	x̄	-1.72	-1.97	-13.76	-12.71	-25.49	-25.27	-31.00	-30.26
	sd	2.25	2.20	3.78	3.13	3.64	3.79	2.60	1.28

Table 4. *Means, sd and statistic significance between native and foreign language. ns= non significant.*

Change		F0 (Hz)	F0 (st) (re 100Hz)	F0 sd (Hz)	F0 range (Hz)	F0 range (st) (re 100 Hz)	Leq (dB)	Leq sd (dB)	Leq range (dB)
Finns, n=16	x̄	6.50	0.75	0.71	-2.02	-0.10	0.49	0.07	0.31
	sd	6.81	0.63	2.19	15.78	1.23	1.21	0.27	1.77
	p	0.001	0.002	ns	ns	ns	ns	ns	ns
English, n=16	x̄	0.38	0.08	-0.86	-0.2	0.11	0.62	-0.19	-1.13
	sd	6.06	0.79	2.42	11.10	1.55	2.84	0.37	2.34
	p	ns	ns	ns	ns	ns	ns	ns	ns
Change		**0-1 kHz (Hz)**	**0-1 kHz (dB)**	**1-2 kHz (Hz)**	**1-2 kHz (dB)**	**2-3 kHz (Hz)**	**2-3 kHz (dB)**	**3-4 kHz (Hz)**	**3-4 kHz (dB)**
Finns, n=16	x̄	-37.68	0.15	18.85	-1.43	39.04	-1.41	-37.68	-0.25
	sd	88.44	0.86	114.99	2.31	335.92	1.33	168.18	2.84
	p	ns	ns	ns	0.04	ns	0.02	ns	ns
English, n=16	x̄	-24.23	-0.13	-123.82	0.38	-24.23	0.73	49.80	1.43
	sd	57.71	0.81	140.99	2.20	175.62	1.88	277.52	3.35
	p	ns	ns	0.01	ns	ns	ns	ns	ns

When the changes between the groups (Finnish and English subjects) were studied, a significant difference in the change in F0 was found, in Hertz (p=.007) and in semitones (p=.007). As shown in Figure 1, the English subjects had a larger variation in the change of F0 than the Finnish subjects.

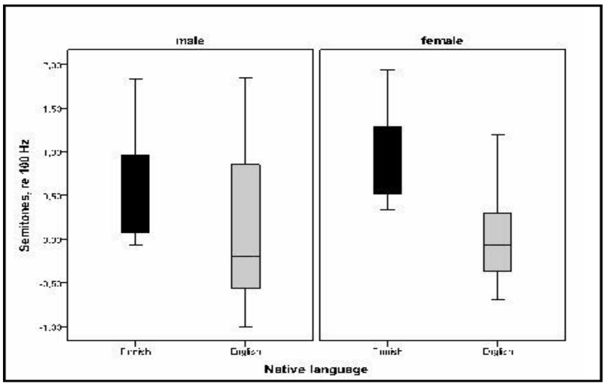


Figure 1. *Change in mean of F0 (in semitones) between native and foreign language for Finnish and English males and females.*

The difference between groups in the LTAS characteristics was significant in the central frequency of the spectral peak between 1-2 kHz (p=.009) and in the relative amplitude between 1 and 2 kHz (p=.04) and 2 and 3 kHz (p=.006). Figure 2 shows the means of the spectral peaks for both groups.

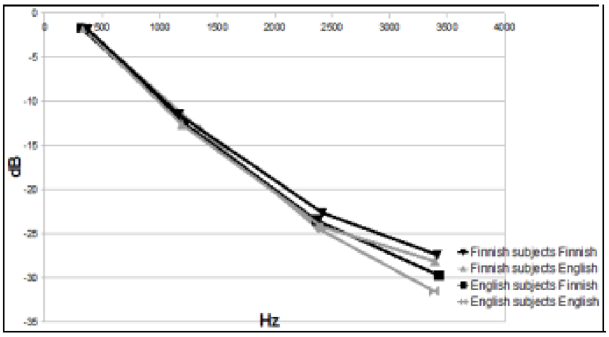


Figure 2. *Means of central frequencies and relative amplitudes of LTAS peaks in Finnish and English subjects' Finnish and English text reading.*

The change in the relative amplitude of the spectral peak between 0 and 1 kHz correlated moderately with the change in the central frequency between 1and 2 kHz (r= .54, p= .03). The change in the central frequency of the spectral peak between 1 and 2 kHz correlated moderately with the change in the relative amplitude between 0 and 1 kHz (r= .59, p= .02) for the Finnish subjects. For the English subjects the change in the frequency between 2 and 3 kHz correlated moderately with the change in the relative amplitude between 1 and 2 kHz (r= .57, p= .02). No correlations with the changes in the spectral peaks and change in F0 or Leq were found.

3. Discussion

The results of the present study suggest a change in vocal characteristics when speaking a foreign language compared to speaking the native one.

Psycho-physiological factors can affect the F0; speaking a foreign language is most likely a more demanding task than speaking the native language and therefore it may cause stress and consequently influence the pitch to rise [9]. Changes in F0 may also be a result of an adaptation to a certain pitch level the speakers aimed to achieve in order to sound more native-like speakers of the target language. The pitch is not merely influenced by physiological factors, but also by cultural ideals [10-12].

The changes in the central frequencies of the peaks in LTAS may be a result of different formant frequencies in the two languages [13]. More frontal articulation can also cause the spectral peaks to move towards higher frequencies [14].

The spectral slope is in general dependent on intensity and type of phonation, getting less steep as intensity rises or type of phonation changes towards more hyperfunctional. A rise in F0 and especially a change from modal into falsetto register is prone to make the spectral slope steeper. In the present material the change in the relative amplitudes of the peaks in LTAS may reflect spectral slope [15-17]. This change did not seem to be related to Leq or F0, but may be due to voice quality differences in languages [4]. Is there perhaps a trend that Finnish then is spoken in a more hyperfunctional way compared to English? On the other hand, the central frequencies and relative amplitudes of LTAS peaks may reflect differences in formant frequencies and their distances from each other [18]. Differences between languages in the amount of voiced/voiceless speech sounds may also affect LTAS. However, in the present material the speech samples were relatively long and for the LTAS analysis all voiceless sounds and pauses were excluded.

Changes in F0 and Leq showed no significant correlations with the changes in frequency and relative amplitude of LTAS which may indicate that there is no evidence that either language is spoken at a lower F0 or louder than the other language. The changes in the relative amplitudes may thus be related to differences in sound qualities in the languages. On the other hand the changes in the relative amplitudes may be due to distance between formant frequencies. Also the amount of voiced/voiceless speech may differ between Finnish and English, which may be actualized in LTAS.

Further study with a larger number of subjects is warranted. Evaluation of perceptual voice quality and naturalness of speech in the target language will be included. Familiarity of the target language may have an effect on voice production [5].

The relative form of F0 distribution will also be studied in addition of means, SD and range of F0. The overall shape of the LTAS [19] should also be addressed. Formant frequencies of individual speech sounds will be taken into account, too.

4. Acknowledgements

This study was supported by the University of Tampere.

5. References

[1] Scherer, K. R. & Giles. H. [Eds], "Social markers in speech", Cambridge, UK: Cambridge University Press, 1979.

[2] Ohara, Y., "Gender-dependent pitch levels: A Comparative study in Japanese and English", in K. Hall, M. Bucholtz, B. Moonwomon [Eds], Locating power. Proceedings of the second Berkeley Women and Language Conference. Berkeley, CA: Berkeley Women and Language Group, 2: 468–477, 1992.

[3] Bruyninckx, M., "Language-induced voice quality variability in bilinguals", Journal of Phonetics, 22: 19–31, 1994.

[4] Wagner, A. & Braun, A., "Is voice quality language-dependent? Acoustic analyses based on speakers of three different languages", 15th ICPhS Barcelona, 651–654, 2003.

[5] Järvinen K, Laukkanen AM, Izdebski K., "Voice Fundamental Frequency Changes as a Function of Foreign Languages Familiarity: An Emotional Effect?", in K. Izdebski [Ed], Emotions in the Human Voice. USA: Plural Publishing, 1: 203–213, 2008.

[6] Järvinen K, Laukkanen AM, Aaltonen O., "Speaking a Foreign Language and its effect on F0", Logopedics Phoniatrics Vocology, (accepted for publication).

[7] Valo, M., "Käsitykset ja vaikutelmat äänestä. Kuuntelijoiden arviointia radiopuheen äänellisistä ominaisuuksista", (Perceptions and appearances of voice. Listeners' evaluation on vocal characteristics in radio speech), Academic dissertation. Studia Philologica Jyväkylaensia 33, 1994.

[8] Cleveland, T. F., Sundberg, J. & Stone, R. E., "Long-Term-Average Spectrum Characteristics of Country Singers during Speaking and Singing", Journal of Voice, 15; 1: 54–60, 2000.

[9] Orlikoff, R. B. & Baken, R. J., "The Effect of the Heartbeat on Vocal Fundamental Frequency Perturbation", Journal of Speech and Hearing Research, 32: 576–582, 1989.

[10] Awan, S. N. & Mueller, P. B., "Speaking Fundamental Frequency Characteristics of White, African American, and Hispanic Kindergartners", Journal of Speech and Hearing Research, 39: 573–577, 1996.

[11] Hudson, A.I. & Hollbrook, A., "Fundamental Frequency Characteristics of Young Black Adults. Spontaneous Speaking and Oral Reading", Journal of Speech and Hearing Research, 25: 25–28, 1982.

[12] Laukkanen, A. M., Mäki, E., Pukander, J. & Anttila, I., "Vertical laryngeal size and the lowest tone in the evaluation of the average fundamental frequency (F0) of Finnish speakers", Logopedics Phoniatrics Vocology 24: 170–177, 1999.

[13] Andrianapoulos, M. V., Darroe, K. & Chen, J., "Multimodal Standardization of Voice Among Four Multicultural Populations Formant Structures", Journal of Voice, 15; 1: 61–77, 2001.

[14] Sundberg J., "Röstlära", (The Science of the Singing Voice), Stockholm: Proprius Förlag, 2001.

[15] Torres da Silva, P., Master, S., Andreoni, S., Pontes, P. & Ramos, L. R., "Acoustic and Long-Term Average Spectrum Measures to Detect Vocal Aging in Women", Journal of Voice, 25; 4: 411–419, 2011.

[16] Kitzing, P., "LTAS Criteria Pertinent to the Measured Voice Quality", Journal of Phonetics, 14: 477–482, 1986.

[17] Nordenberg, M. & Sundberg, J., "Effect on LTAS of vocal loudness variation", TMH-QPSR 45: 1; 93–100, 2003.

[18] Huber, J. E., Stathopoulos, G. M., Ash, T. A. & Johnson, K., "Formants of children, women, and men: The effects of vocal intensity variation", J. Acoust. Soc. Am., 106: 1532–1542, 1999.

[19] Byrne, D. et al., "An international comparison of long-term average speech spectra", J. Acoust. Soc. Am., 96: 4; 2108–2120, 1994.

New speech corpora at IoC

Einar Meister, Lya Meister and Rainer Metsvahi

Institute of Cybernetics at Tallinn University of Technology, Estonia

einar@ioc.ee, lya@phon.ioc.ee, rainer@phon.ioc.ee

Abstract

The paper will give an overview of the new Estonian speech corpora collected in recent years at the Laboratory of Phonetics and Speech Technology, Institute of Cybernetics at Tallinn University of Technology. The development of these corpora has been funded by the National Program for Estonian Language Technology (2006-2010 and 2011-2017).

Index Terms: spoken language resources, speech corpus, multimodal corpus

1. Introduction

Spoken language resources are ultimately necessary in different areas of speech research, ranging from different topics of experimental-phonetic studies to a large spectrum of applications in language technology development and testing. Diversity of research tasks and applications make a demand for different types of speech material varying in many characteristics, e.g. speech style, number of speakers, recording environment and channel, linguistic content, etc. While in some phonetic studies a couple of minutes of speech from 1-2 speakers might be enough, then, for example, in automatic speech recognition (ASR) the amount of speech data is measured in hundreds of hours collected from thousands of speakers. The current state-of-art in ASR exploits statistical models demanding large corpora for training – "there's no data like more data" – still seems to be the best-performing approach. However, in some tasks (e.g. word sense disambiguation) the latter claim is not entirely true [1].

In the past two-three decades large amounts of speech resources have been collected all over the world and new corpora are constantly being produced; the number of speech corpora accessible to researchers and technology developers is growing constantly. The use of available corpora is quite often complicated due to diversity of signal and metadata formats applied in different corpora; there is no unique standard for speech resources available. Yet, there exist a few corpora that have became de facto standards for developing, testing and comparing the performance of speech processing tools, e.g. TIMIT for American English [2], EUROM1 covering many European languages [3], and ANDOSL – the Australian National Database of Spoken Language [4]. An approach to standardize the production and validation of speech corpora has been provided by Bavarian Archive for Speech Signals [5]; know-how on collection, annotation and validation of spoken language resources is provided by SPEX [6], [7].

Systematic collection of Estonian speech resources at the Institute of Cybernetics (IoC) at Tallinn University of Technology begun in the mid 1990s when the Laboratory of Phonetics and Speech Technology at IoC was involved in the BABEL-project (1995-98) [8] funded by the EU Copernicus program. In this project speech databases for five Central and Eastern European languages – Estonian among Bulgarian, Hungarian, Polish and Romanian – were developed. The collection and formatting of the data conforms to the protocols established for EUROM databases. The Estonian Babel Database [9] was designed to serve as a resource for phonetic research on Estonian sound system and prosody as well as for training the acoustic models. The speech material from 70 speakers was recorded while reading the text corpus of CVC-units, numbers, phonetically rich sentences, and short passages. Since the BABEL project several corpora for different purposes have been collected [10], among others the Estonian SpeechDat-like Database [11] involving speech recordings from ca 1300 native speakers.

From 2006 onward the collection of new speech corpora at IoC has been carried out under the National Program for Estonian Language Technology (first phase for the years 2006-2010, second phase for 2011-2017) launched by the Ministry of Science and Education. One of the priority areas of the national program is the development of reusable Estonian language resources; the resources funded by the national program are declared public.

The paper gives an overview of the new corpora collected in recent years and introduces corpora under development.

2. Corpora (almost) completed

2.1. Corpus of radio news

The corpus includes ca 300 hours of news recordings from the Estonian Public Broadcasting collected in the years 2005–2006, and more than 8000 pages of digitized news texts. First the news texts have been used for the training of the language model of the Estonian ASR system and then the raw transcripts have been generated automatically with the ASR. The raw transcriptions have been manually checked and corrected using Transcriber (http://trans.sourceforge.net); in total about 10 % of the news recordings have been transcribed.

2.2. Corpus of talk shows and interviews

The corpus includes ca 20 hours of talk shows from different Estonian broadcast companies typically involving live discussions between the host and 1-2 quests (45 different speakers, 35 male, 10 female). The topics of discussions include economy, politics, etc. The speech of talk shows is prevalently spontaneous involving change of speakers, overlapped speech, characteristics of emotional speech (laugh etc.). All recordings are manually transcribed.

In addition, ca 20 hours of radio interviews have been collected and manually transcribed.

2.3. Corpus of lecture speech

The corpus includes ca 350 hours of recordings of academic university lectures (33 different speakers) and more than 50 hours of conference presentations (more than 70 different speakers). The speech style of academic lectures is variable involving fluent presentation of carefully prepared lecture material as well as chunks of more spontaneous speech. Conference presentations provide even more inter-speaker variability ranging from reading of prepared texts to semi-spontaneous speech. The recordings of 40 speakers (ca 20 hours) have been manually transcribed.

2.4. Foreign accent corpus

The corpus is designed for experimental-phonetic studies of foreign accent phenomena as well as for training of acoustic models for automatic recognition of non-native Estonian speech.

Speaker selection criteria:

- Estonian as a second (or third, etc.) language
- knowledge of Estonian higher than basic
- age > 16

Material:

- self-introduction (name, native language, when and where started to learn Estonian, use of Estonian, etc.)
- 140 short sentences involving main phonological phenomena of Estonian (all vowels and consonants, quantity oppositions, palatalization)
- 2 passages
- 3 pictures

Recording set-up:

- microphones Sennheiser ME3 and Audio-Technica ATM33a
- digital recording (sampling at 44.1 kHz, resolution 16 bit), Adobe Audition 3.0 + Mackie Onyx 1640 or M-Audio Microtrack 24/96

So far about 160 speakers with different language backgrounds have been recorded: Russian (50 speakers), Finnish (30), Latvian (20), German (15), French (12), Italian (5), English (4), Lithuanian (3), Spanish (2), Danish (2), Slovak (2), Japanese (2), Swedish (1), Polish (1), Scottish (1), Irish (1), Azerbaijani (1), Portuguese (1). In addition, a reference group of 20 native speakers of Estonian have been recorded.

3. Corpora in progress

3.1. Multimodal corpora

Recently, facilities for the collection of multimodal corpora including articulatory and acoustic data, hand gestures, head movements, mimics, body posture, etc. have been set up at IoC. In addition to audio & video recorders the facilities include an Electroglottograph (EGG) [12], an Electropalatograph (EPG) [13], an Electromagnetic Articulograph (EMA) [14], and an infra-red 3D Motion Capture System [15].

EGG technique provides the most accurate data on the physical measures of voice fundamental frequency during speech production. Two electrodes are positioned on both sides of the thyroid cartilage and a weak voltage is passed from one electrode to the other. The change in electrical impedance across the throat depends on the contact variations between the vocal folds. These variations are recorded synchronously with the speech signal captured by a microphone.

EPG is a method to study the timing and location of tongue contact with the hard palate during speech production. On the surface of the palate a number of silver contacts are located which register the tongue-palate contact during articulation (Figure 1). Throughout the data capture a speaker has to wear an artificial palate (Figure 2). The data capture process is administered with appropriate software allowing live display of the tongue-palate contact patterns and their time-synchronous recording along with acoustic and/or laryngograph signals.

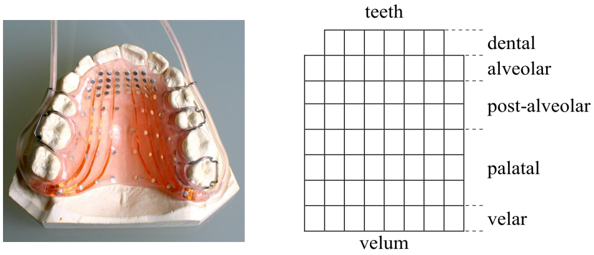


Figure 1. The placement of 62 electrodes on the EPG palate (left) corresponding to six tongue contact areas (right).

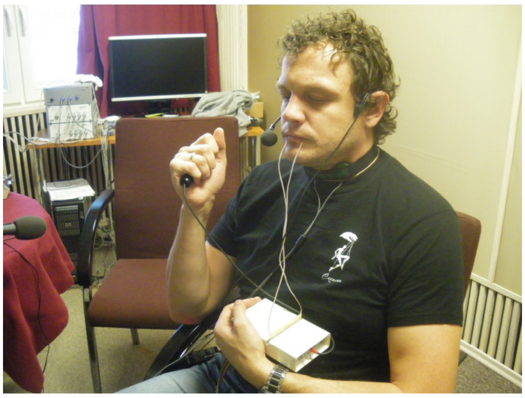


Figure 2. The subject wearing electrodes of the EGG system, the EPG palate, and a close-talking microphone

EMA is a non-line-of-sight motion capture system specifically designed for tracking speech related articulatory movements and articulatory kinematics. It enables real-time 3D data capture from 16 sensors in the measurement volume of 50x50x50 cm, along with synchronized audio. In articulatory studies the sensors are typically glued on the tongue, the lips, the front teeth, and on the jaw, with a sensor glued on the bridge of the nose acting as a reference to compensate for head movement (Figure 3).

The **multi-modal corpus of speech production** is aimed at studying dynamic articulatory patterns and acoustic-articulatory mapping in native Estonian speech; a further goal is to investigate the use of articulatory data in speech technology, e.g. in building of models for audio-visual speech synthesis.

The text corpus compiled for data collection includes CVCV nonsense words and two-syllable real words, and short sentences. Two different recording set-ups have been used:

- EGG + EPG + audio,
- EMA + audio.

In both set-ups the same text corpus has been read by two subjects (1 male, 1 female). The acoustic data is recorded with a close-talk microphone (Sennheiser ME3) at a sampling frequency of 22.1 kHz.

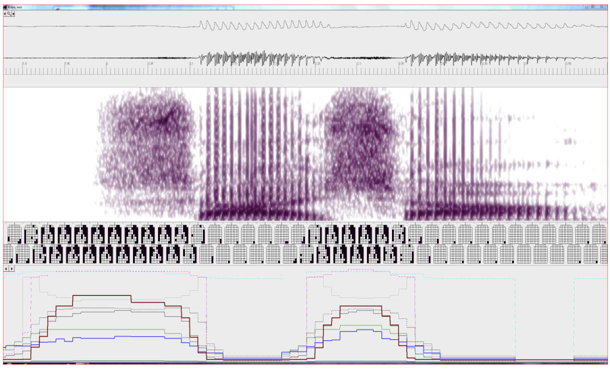


Figure 4. An example of multimodal data for the nonsense word [sɑsʲɑ]. From top to bottom: EGG signal, audio signal, spectrogram, palatogram frames, EPG signals.

In the EMA set-up three sensors are attached to the tongue (tongue tip, tongue blade, tongue dorsum), eight sensors to the lips (three on both upper and lower lips, two on lip corners), one to the jaw, and one sensor to the bridge of the nose.

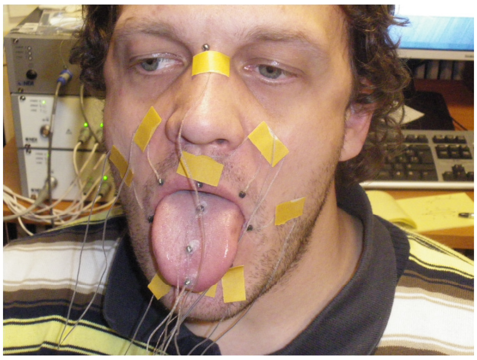


Figure 3. The position of EMA sensors.

For the processing of multi-modal data several software packages (WaveSurfer, MATLAB, R) have been tested, however, the appropriate software environment enabling synchronization, visualization, editing, labelling, animation and analysis of data recorded from different instruments still needs to be developed.

Data collection is in progress, the target amount is one hour of speech from two subjects recorded in two different set-ups.

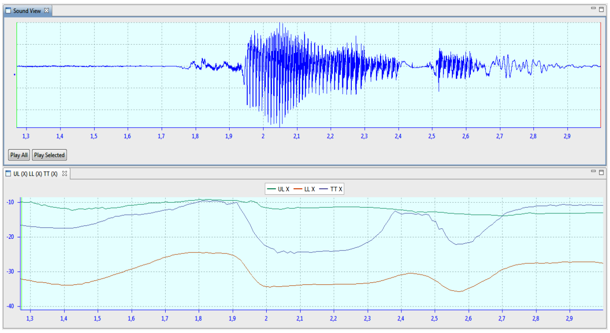


Figure 4. An example of multimodal data for the word [sɑːtɑ] ‘to get’ da-infinitive. Upper window: audio signal; lower window: traces of sensors attached to upper lip (UL, green), tongue tip (blue) and lower lip (yellow).

3.2. Corpus of adolescent speech

The corpus is designed for experimental-phonetic studies of adolescent speech and for training of ASR acoustic models.

Speaker selection criteria:

- native Estonian speakers aged from 8 to 16
- balanced by age and sex

Material:

- *read speech:* phonetically rich sentences, short passages, isolated numbers, PIN-codes, phone numbers, date and time expressions
- *spontaneous speech:* self-introduction, description of pictures, story telling

Recording setup:

- microphones Sennheiser ME3 and Audio-Technica ATM33a
- digital recording (sampling at 44.1 kHz, resolution 16 bit)
- M-Audio MobilePre
- recording software BAS SpeechRecorder

The subjects (up to 200) will be recorded at different schools around Estonia.

3.3. Corpus of spoken named entities

The aim of the corpus is to collect pronunciation variations of different named entities – persons, places, institutions, product brands, foreign names, etc. – uttered in sentence context and in isolation.

A prototype for named entity recognition [16] has been used to collect four types of frequent named entities from Estonian newspapers – names of persons, locations, organizations and facilities. The corpus will be recorded in different acoustic environments via different communication channels (GSM phones, desktop and headset microphones) with up to 200 subjects.

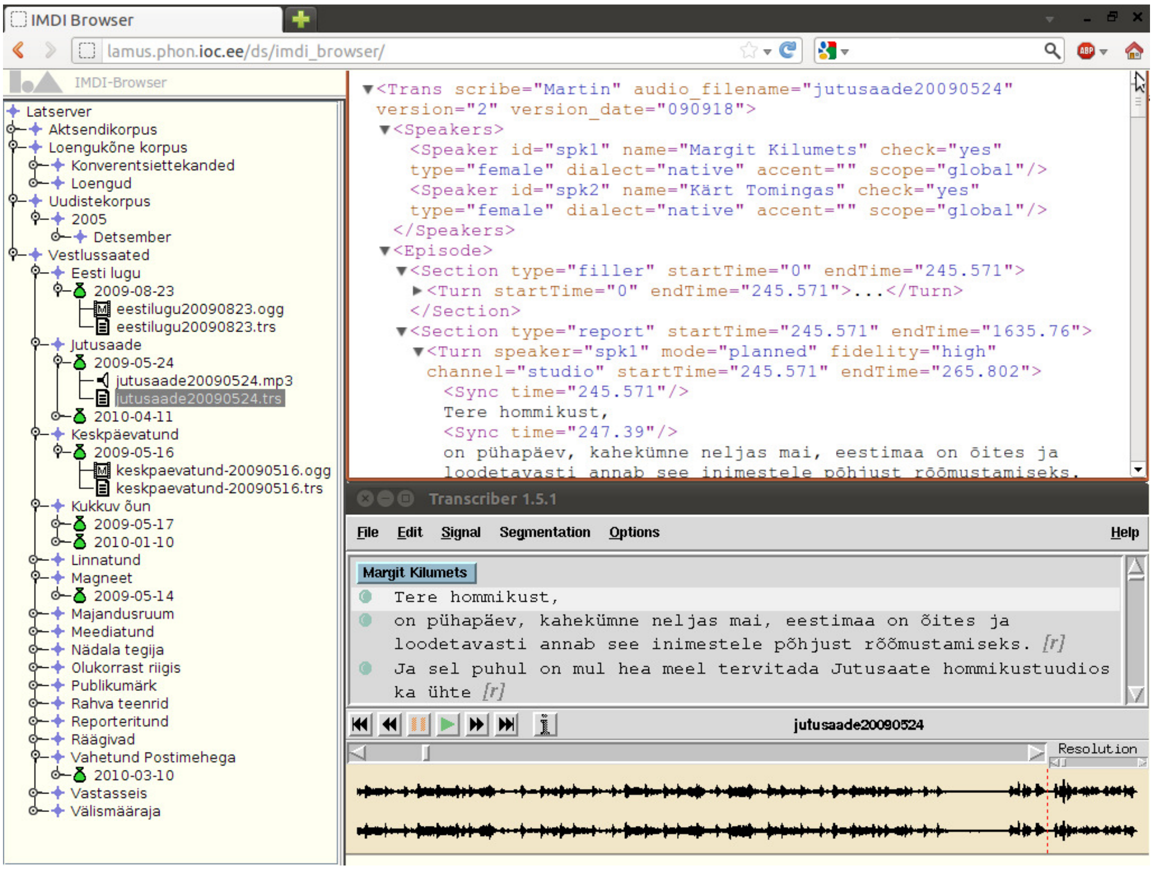
4. Infrastructure

All corpora will be made available via the Center of Estonian Language Resources (CELR). The CELR founding project has been launched in 2012 and it aims to create a new infrastructure for the management, evaluation and sharing of Estonian language resources. The project partners are the University of Tartu, the Institute of Cybernetics, and the Institute of the Estonian Language. CELR will be a part of the pan-European CLARIN network.

For corpus management the LAMUS-system (Language Archive Management and Upload System, http://www.lat-mpi.eu/tools/lamus/) has been adapted (Figure 5).

5. Acknowledgements

This work has been funded by the National Program for Estonian Language Technology (2006-2010 and 2011-2017). We are grateful to all subjects who participated in recordings.



```
▼<Trans scribe="Martin" audio_filename="jutusaade20090524"
 version="2" version_date="090918">
  ▼<Speakers>
     <Speaker id="spk1" name="Margit Kilumets" check="yes"
     type="female" dialect="native" accent="" scope="global"/>
     <Speaker id="spk2" name="Kärt Tomingas" check="yes"
     type="female" dialect="native" accent="" scope="global"/>
   </Speakers>
  ▼<Episode>
    ▼<Section type="filler" startTime="0" endTime="245.571">
      ▶<Turn startTime="0" endTime="245.571">...</Turn>
     </Section>
    ▼<Section type="report" startTime="245.571" endTime="1635.76">
      ▼<Turn speaker="spk1" mode="planned" fidelity="high"
       channel="studio" startTime="245.571" endTime="265.802">
        <Sync time="245.571"/>
        Tere hommikust,
        <Sync time="247.39"/>
        on pühapäev, kahekümne neljas mai, eestimaa on õites ja
        loodetavasti annab see inimestele põhjust rõõmustamiseks.
```

Tere hommikust,

on pühapäev, kahekümne neljas mai, eestimaa on õites ja loodetavasti annab see inimestele põhjust rõõmustamiseks. [r]

Ja sel puhul on mul hea meel tervitada Jutusaate hommikustuudios ka ühte [r]

Figure 5. A screen shot of LAMUS user interface.

6. References

[1] Rehbein, I. and Ruppenhofer, J., “There’s no Data like More Data? Revisiting the Impact of Data Size on a Classification Task”, in Proceedings of LREC'10, Valletta, Malta, pp. 1206–1213, 2010.

[2] Garofolo, J. S., Lamel, L., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., Zue, V., “TIMIT Acoustic-Phonetic Continuous Speech Corpus”, Linguistic Data Consortium, Philadelphia, 1993.

[3] Chan, D., Fourcin, A., Gibbon, D., Granström, B., Huckvale, M., Kokkinakis, G., Kvale, K., Lamel, L., Lindberg, B., Moreno, A., Mouropoulos, J., Senia, F., Trancoso, I., Veld, C., Zeiliger, J., “EUROM – A Spoken Language Resource for the EU”, in Proceedings of the 4th European Conference on Speech Communication and Speech Technology, Madrid, Spain, September 18-21, 1995, Vol 1, pp. 867–870, 1995.

[4] Millar, J.B., Vonwiller, J.P., Harrington, J.M., Dermody, P.J., “The Australian National Database Of Spoken Language”, in Proceedings of ICASSP-94, Adelaide, Vol.1, pp. 97–100, 1994.

[5] Schiel, F., Draxler, Ch., “Production and Validation of Speech Corpora. Bavarian Archive for Speech Signals”, Bastard Verlag, München, 2003.

[6] van den Heuvel, H., “The art of validation”, The ELRA Newsletter, Vol. 5(4), pp. 4–6, 2000.

[7] http://www.spex.nl/

[8] Roach, P., Arnfield, S., Barry, W., Baltova, J., Boldea, M., Fourcin, A., Gonet, W., Gubrynowicz, R., Hallum, E., Lamel, L., Marasek, K., Marchal, A., Meister, E., Vicsi, K., “BABEL: an Eastern European Multi-Language Database”, in ICSLP 96 - Fourth International Conference on Spoken Language Processing, Proceedings, Philadelphia, USA: New York: IEEE, 1892–1893, 1996.

[9] Eek, A., Meister, E., “Estonian speech in the BABEL multi-language database: Phonetic-phonological problems revealed in the text corpus”, in O. Fujimura [Ed], Proceedings of LP'98, Prague: The Karolinum Press, 529–546, 1999.

[10] Meister, E., “Promoting Estonian speech technology: from resources to prototypes”, PhD thesis, Tartu: Tartu University Press, 2003.

[11] Meister, E., Lasn, J., Meister, L., ”SpeechDat-like Estonian database”, in V. Matoušek, P. Mautner [Eds], Text, Speech and Dialogue: 6th International Conference, TSD 2003, Berlin: Springer, Lecture Notes in Artificial Intelligence 2807, 412–417, 2003.

[12] http://www.laryngograph.com

[13] http://www.articulateinstruments.com

[14] http://www.ndigital.com/lifesciences/products-speechresearch.php

[15] http://www.ndigital.com/lifesciences/3dinvestigator-motioncapturesystem.php

[16] Tkatšenko, A., ”Named Entity Recognition for the Estonian Language”, MSc thesis, University of Tartu, 2010.

Modeling turn-taking rhythms with oscillators

Michael L. O'Dell[1], *Tommi Nieminen*[2], *Mietta Lennes*[3]

[1]University of Tampere, [2]University of Eastern Finland, [3]University of Helsinki
michael.odell@uta.fi, tommi.nieminen@uef.fi, mietta.lennes@helsinki.fi

Abstract

Our aim in this paper is to explore ways of modeling the distribution of pause durations in conversation using oscillator models [11], and to consider how these models might be integrated into our Coupled Oscillator Model of speech timing (COM [7, 6, 8]).

1. Overview

Modeling the durations of conversational pauses has recently attracted some attention (cf. the excellent overview in [3]). Wilson & Wilson [11] have modeled conversational turn-taking based on coupled oscillators, and Beňuš [1, 2] tested this model against a database of conversational American English. Beňuš's results provided some support for the model, but the support was weak due to small (although significant) correlations, and a lack of predicted phase patterns.

As Wilson & Wilson pointed out, it is important to gather data on a variety of languages in addition to English. In this paper, we apply Beňuš's analysis to the Finnish Dialogue Corpus [4, 5] and also consider integrating the Wilson & Wilson model into our own speech timing model, which has hitherto lacked an explicit mechanism for dealing with pausing behavior.

2. Wilson & Wilson model

2.1. Motivation for oscillators

There are several facts about turn-taking behavior in spoken dialogue which Wilson & Wilson explain using an oscillator model. According to Wilson & Wilson, "turn transitions with virtually no gap [$<$ 200 ms] are a common occurrence in ordinary conversation." This is testified to in the Finnish corpus as well: Slightly more than a third of the transitional pauses were less than 200 ms in duration (cf. Table 1).

According to Wilson & Wilson and many others, conversational speech also tends to avoid simultaneous starts.[1]

Table 1: Number of transitional pauses for a pair of Finnish speakers (speaker 1, speaker 2)

	$1 \to 2$	$2 \to 1$	Both
Total	145	174	319
$<$ 200 ms	55 (38%)	54 (31%)	109 (34%)

The reason for this is fairly obvious given that conversation has a real, dialogic function. Simultaneous starts after pause (defined as both speakers initiating speech in less than 200 ms of each other) are relatively rare in our Finnish corpus as well: Approximately 6% of pauses ended in simultaneous starts (cf. Table 2).

Table 2: Number of "simultaneous" starts after pause for a pair of Finnish speakers

	$1 \to x$	$2 \to x$	Both
Total	461	409	870
$<$ 200 ms	31 (7%)	22 (5%)	53 (6%)

A fact that is not so obvious is that (according to Wilson & Wilson) pauses "tend to be multiples of some unit length of time," which "ranged from 80 to 180 msec ... with an average of 120 msec." ([11], based on data in [12]).[2] This raises the possibility that turn cycle might be related to some other oscillatory cycle in speech, and Wilson & Wilson suggest possible candidates such as syllable duration, jaw cycles or even the theta rhythm.

2.2. Synchronization and turn cycle

The idea behind synchronization in dialogue is that each participant monitors the speech of the other and tries to keep in synchrony. Arguably such behavior is either a byproduct or a prerequisite of speech perception in general.

[1] It is often assumed that overlapping speech is avoided in general, although this has been questioned along with the assumption that dialogues actually exhibit clear turn-taking structure at all. In the present work we are not directly concerned with overlapping speech, but we hope to return to this question in the future.

[2] Wilson & Wilson refer to this unit length of time, or turn cycle period, as S. Confusingly, the earlier Wilson & Zimmerman article [12] they refer to uses S to mean each speaker's "slot length", which is half of a total turn cycle. Thus Wilson & Wilson's S equals twice Wilson & Zimmerman's S. In what follows we retain S for the slot length and use R for the period of the turn cycle, so that $R = 2S$.

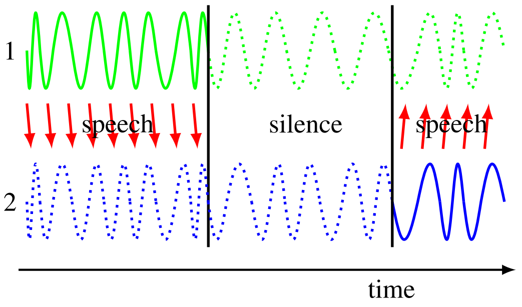


Figure 1: Speakers synchronize during speech, during pause each speaker oscillates between "my turn to start" and "your turn to start".

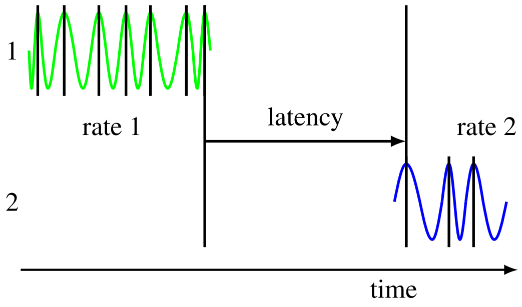


Figure 2: Schematic diagram of Beňuš's measures

During silence, the ability of the speakers to synchronize is considerably weakened. Wilson & Wilson conjecture that the speakers maintain a turn cycle which is also synchronized during speech (possibly related to e.g. syllable rhythm) and then continued during pauses. Such behavior is hypothesized to minimize the "offset" between their conversational turns. Thus, when the current speaker reaches the end of his turn, the current listener may step in with a minimum overlap or gap (when no pause is intended). The participants' oscillators have the same period (when synchronized) but "The listener's cycle is counterphased to that of the speaker." [11] (cf. Fig. 1). Because of this counterphasing, "… the probability of simultaneous starts will be relatively low." [11].

Note that *counterphased* describes the situation from the individual participant's point of view: each one oscillates between phases "my turn to start" and "your turn to start" and these phases are opposed. From a system point of view, however, the two oscillators are actually *in phase*: each one oscillates between phases "1st speaker's turn to start" and "2nd speaker's turn to start" and these phases are in agreement.

2.3. Empirical testing

Beňuš [1, 2] attempted to test the empirical consequences of the Wilson & Wilson model. If a putative turn cycle is a continuation of some rhythm accessible during speech, the question natually arises as to which of the many possible rhythms is the relevant one. Beňuš considered two possibilities in his analysis of a database of conversational American English: syllable rhythm and pitch accent rhythm.

For empirical testing purposes Beňuš compared two measures derived from the database: *latency*, defined as "difference between the end of the chunk [inter-pausal unit] and the beginning of the next chunk" and *rate*, represented by average (syllable or accent) duration within each "chunk" [1]. These measures are illustrated in Fig. 2.

Beňuš also makes an overt terminological distinction between synchronization within speaker vs. between speakers: the first one he calls "isochrony", the second one "entrainment". In both cases, whether "isochrony" proper or "entrainment", the following points hold: a) Rate should be correlated across pause, b) Latency should be correlated with previous rate and c) The latency distribution (normalized by previous rate) should be multimodal, with modes at interval steps.

Results provided some support for the model, but support was weak due to small (although significant) correlations, and a lack of predicted phase patterns.

3. Present study

We set out to apply Beňuš's procedure to Finnish conversational material following in effect Wilson & Wilson's plea for more material from diverse languages. Presently we have studied only one speaker pair, and the results are thus very preliminary, albeit suggestive.

Wilson & Zimmerman estimated S using time series analysis (ARIMA) applied to histograms reinterpreted as a time series. Here we model empirical pause distributions as a mixture of normal distributions (one for each possible turn cycle), imposing various constraints on the means, variances and mixing probabilities. This procedure allows a series of increasingly complex models to be fit to data.

Models of pause duration distributions	
Constant expected duration "no effects model"	$\mathrm{E}(\mathrm{dur}) = \mu$
Cyclical expected duration "Wilson & Zimmerman model"	$\mathrm{E}(\mathrm{dur}) = nR$ or $(n - 1/2)R$
Variable turn cycle "Wilson & Wilson model"	$\mathrm{E}(\mathrm{dur}) = nR(t)$ or $(n - 1/2)R(t)$, $R(t)$ depends on previous speech
Multiple hierarchical cycles "COM model"	$\mathrm{E}(\mathrm{dur}) = c_1 + c_2 n_2 + \cdots + c_k n_k$

A generic graph for these models is shown in Fig. 3. In this figure Z_i is the measured duration, μ_i is the expected duration and σ_i^2 is the duration variance for the ith pause. Expected duration is a function of n_i, the number of silent turn cycles ($\mu_i = R_1 + (n_i - 1)R$, where R is the period of one cycle, and R_1 is the duration of the first cycle). Two parameters, β_0 and β, are included to allow the variance σ_i^2 to increase slightly as n increases ($\ln \sigma_i^2 = \beta_0 + \beta n_i$). The probability of n turn cycles is modeled as a geometric distribution with probability p_0 of success.

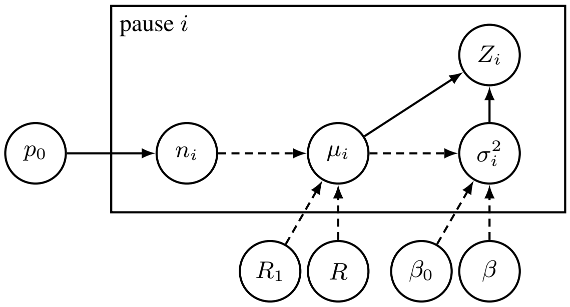


Figure 3: Graph of statistical model

Bayesian inference of periodicity can be based on the ratio of total variance to within cycle variance for the first two cycles (say $\phi = \sigma^2_{\text{total}}/\sigma^2_{\text{within}}$). When this ratio is smaller than two the cyclic structure of the mixture distribution is not apparent, so we use the posterior probability $\Pr(\phi < 2)$ to indicate the significance of periodicity. An almost equivalent alternative which is easier to assess visually is to compare the cycle period (R) with the sum of standard deviations for the first two modes ($\sigma_1 + \sigma_2$, cf. Fig. 4): Periodicity can be considered significant when $R \gg \sigma_1 + \sigma_2$.

3.1. Cyclical expected duration

Beňuš did not look directly at the raw latency distributions in his data for signs of periodicity (and thus did not attempt to estimate S as Wilson & Zimmerman did), but normalized latency duration using syllable (or accent) rate of the previous chunk. Before proceeding to the Wilson & Wilson model, however, we start with the simpler Wilson & Zimmerman model to see whether a clear periodicity in the pause duration distribution can be discerned and whether it agrees with Wilson & Zimmerman's estimate of S with a "range from 40 to 90 ms with a mean of 60.00 ms." [12].

Posterior distributions for R_1, R and $\sigma_1 + \sigma_2$ are shown in Fig. 4 for the four conditions: switches from speaker one to speaker two (1 → 2), switches from speaker two to one (2 → 1), speaker one internal pausing (1 → 1) and speaker two internal pausing (2 → 2). Raw distributions of pause durations in the four conditions are shown in Fig. 5. Also shown in this figure superimposed on the raw distributions are the median posterior fits for the mixture model.

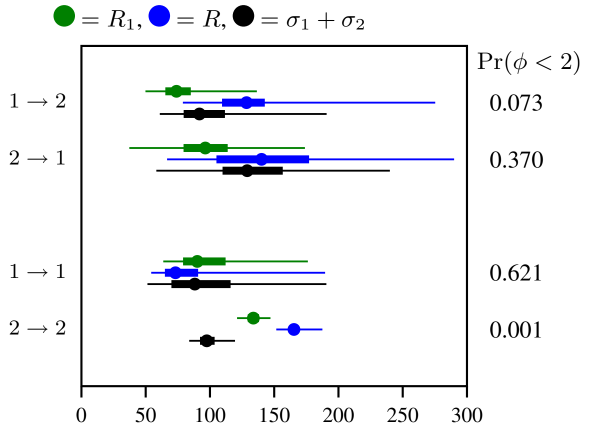


Figure 4: Estimated parameters (ms) for the Cyclical model

The Wilson & Zimmerman model predicts that for between speaker pauses (which contain an even number of slot lengths S), pause duration will be $2kS$, $k = 0, 1, 2, \ldots$, so that $R_1 \approx R$. On the other hand for within speaker pauses (which contain an odd number of slot lengths), the pause duration will be $(2k + 1)S$, so that $R_1 \approx R/2$.

In our data only the within speaker pauses for speaker 2 (2 → 2) showed a significant periodic structure (although condition 1 → 2 was also close to significance; see Figs. 5 and 4). The posterior mean for R for 2 → 2 was 165 ms with a 95 % credible interval of 152–187 ms, which agrees well with Wilson & Zimmerman's estimates, remembering that $R = 2S$. For the within speaker condition the Wilson & Zimmerman model predicts $R_1 \approx R/2$. As shown in Fig. 4, R_1 (posterior median 134 ms) is reliably less than R (posterior median 165 ms), but much greater than $R/2$. This could indicate that the first cycle is slower, or that the two halves of the turn cycle (say S' and S'', so that $R_1{=}S'$, $R = S' + S''$) are not necessarily equal (with $S' > S''$ for speaker 2).

Another interesting feature for the 2 → 2 pauses is that there appears to be a second local maximum in the vicinity of 0.6 to 0.8 s (fourth and fifth bump, cf. Fig. 5). This might indicate the existence of two simultaneous rhythms during pause.

In general, what are the chances of this type of test succeeding? Assuming the turn cycle during pause is a continuation of the syllable cycle during speech, the distribution of durations during speech provides a comparison for judging whether quasiperiodicity could be detected even in an ideal case. To put this another way,

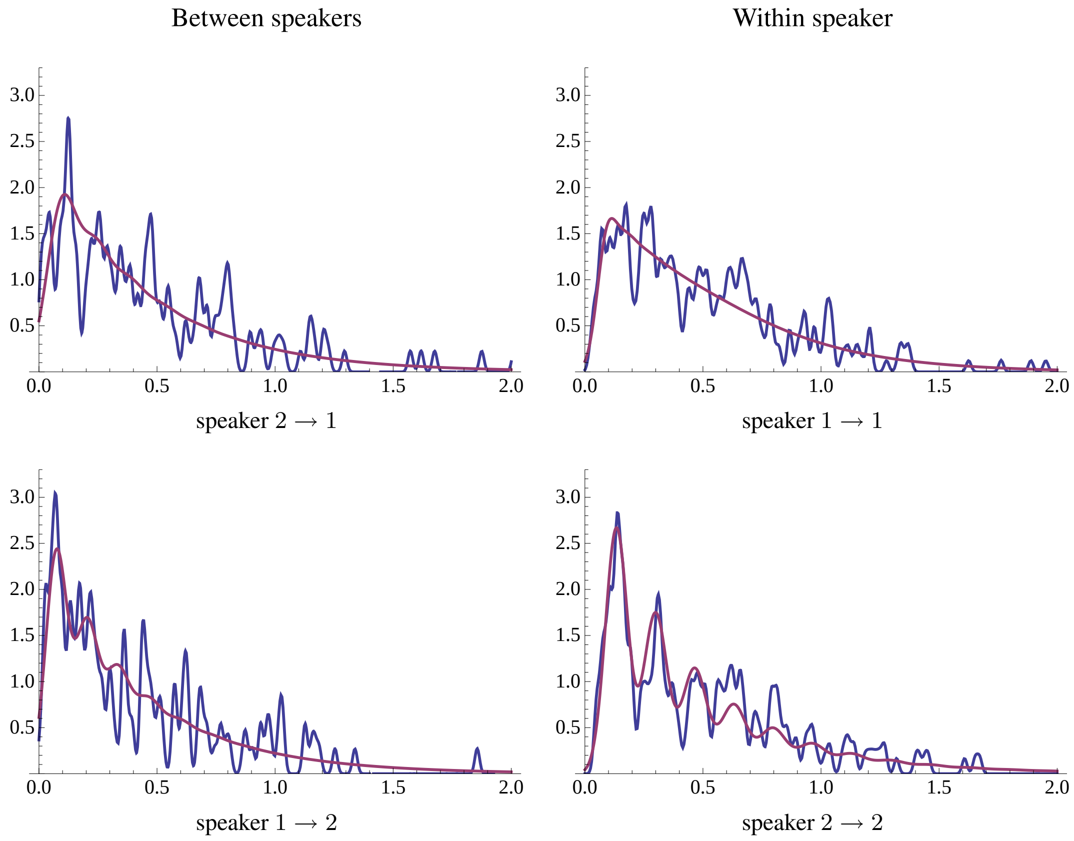


Figure 5: Distributions for pause durations (s)

if we were not sure that speech was composed of syllables, could this be deduced given only the total durations of various units (such as stress groups)? For the present data at least, applying the above statistical procedure to inter-pause groups indicated that periodicity due to recurring syllables during speech is entirely masked by the variability in syllable rate. If pauses are indeed composed of "silent syllables", and if silent syllable rate is as variable as normal syllable rate, then the same may hold for pauses, obscuring the cyclic nature of pausing due to cycle period variation. Of course this cannot be construed as evidence *for* periodicity during pauses, but lack of clear multimodality in the duration distributions does not provide strong evidence against it either. A possible way forward is to look for additional covariates which correlate with the variable turn cycle period.

3.2. Variable turn cycle

Wilson & Wilson [11] hypothesized that turn cycle is a continuation of syllable cycle during speech. If this is the case, we would expect the turn cycle period to vary with syllable rate, rather than being constant (for each speaker or speaker pair). Following Beňuš's lead, we attempt to assess the possible relevance of syllable rate preceding a pause.

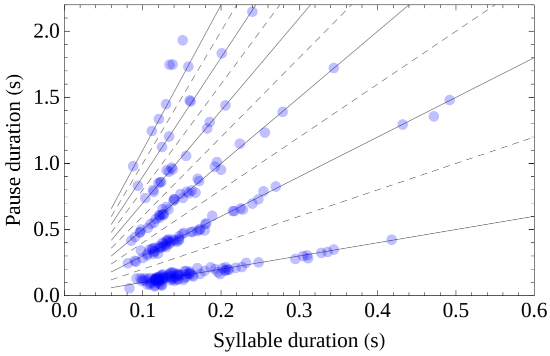


Figure 6: Ideal scattergram of within speaker pauses

In the ideal situation, a scatterplot of pause duration against previous syllable rate would look something like Fig. 6: Pauses with an equal number of "silent syllables" (say k) form slanting stripes because slot length duration ($S(t)$) is tightly clustered around average syllable duration of the previous chunk. A stripe pattern of some kind should be evident even if syllable duration has a non-linear relation to pause duration.

In such a case it is obvious that ignoring syllable rate will radically obscure the periodic pattern. On the other

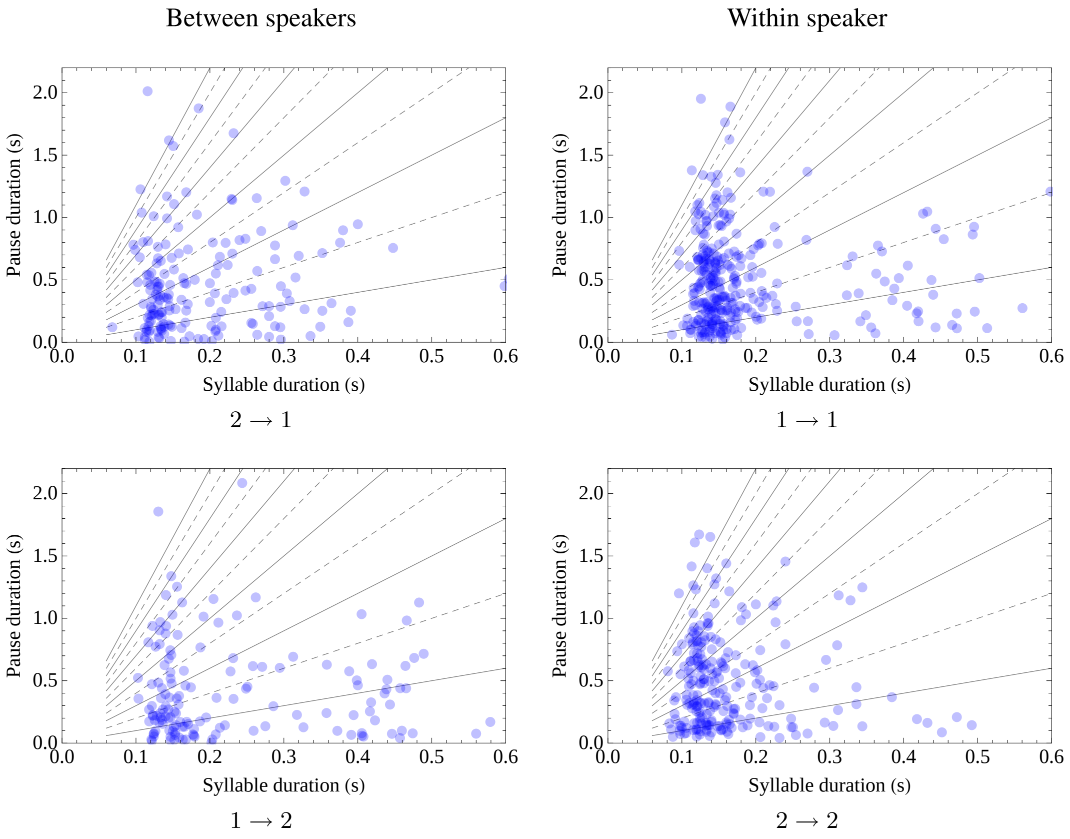


Figure 7: Pause durations by average syllable duration of previous chunk

hand, dividing pause duration by the average syllable duration (say $\hat{S}(t)$) similar to Beňuš's normalization procedure, gives an index ($I = 2kS(t)/\hat{S}(t) \approx 2k$ or $I = (2k+1)S(t)/\hat{S}(t) \approx 2k+1$) which should have an empirical distribution with clear modes at integer values (even for between speaker pauses, odd for within speaker pauses), given that $S(t) \approx \hat{S}(t)$.

For the present data, averaging syllable duration over the entire previous chunk as Beňuš did, produced the scatterplots shown in Fig. 7 for the four conditions. To aid the eye, in both Fig. 6 and Fig. 7 lines have been added indicating where pause duration equals an integer times syllable duration, solid for odd and dashed for even integers.

Fig. 8 shows distributions of pause durations normalized by syllable duration and rescaled to match the unnormalized distributions of Fig. 5 to facilitate comparison. Again, (vertical) lines have been added showing an integer number of syllable durations, solid for odd and dashed for even integers.

Evidence for a possible effect of syllable rate (estimated here by average syllable duration of the preceding chunk) on pause duration is completely lacking in these figures. The scattergrams have no stripes, the normalized duration distributions have no periodic structure. In fact, even the fairly clear periodic structure for the 2 → 2 pauses has been completely obscured in the normalized distribution. Looking at 2 → 2 in Fig. 7 we can see why: The periodic stripes are roughly parallel to the syllable duration axis instead of sloping as in the ideal case (Fig. 6). This suggests that the periodic structure of pauses for 2 → 2 is unrelated to the syllable rate of the preceding chunk.

There are various alternative explanations for the failure to observe a rate effect (apart from the conclusion that pausing is not rhythmic in nature). First of all, shortage of data. Thus far we have studied only one speaker pair, and the effect might be quite weak.

Second, perhaps the syllable rate effect is too short-lived to be observed, Speakers may return to a neutral or preferred turn-taking cycle period fairly rapidly as a pause continues, or natural variation in the period may quickly obscure any initial rate related difference at the beginning of pause. It may also be that during pauses speakers maintain a turn-taking oscillator for a few cycles only. After all, as pause duration increases, the chance of a simultaneous start decreases even without any synchronizing mechanism. In either case the Wilson & Wilson model will be inadequate, because while allowing turn cycle to vary from pause to pause, it assumes a constant

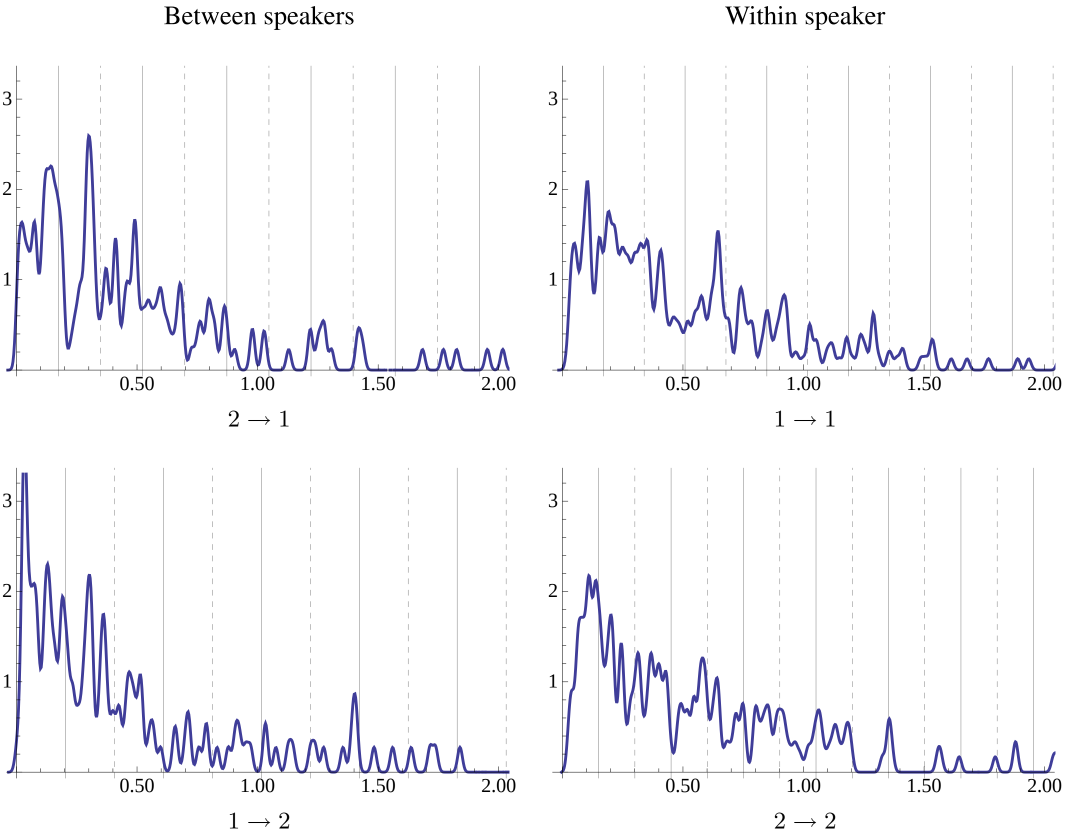


Figure 8: Normalized pause duration distributions (s)

turn cycle during each pause.

A related issue is the adequacy of the rate estimate itself. It may be possible to obtain a better estimate of dynamic rate, for instance by weighting immediately preceding syllables more, rather than using a straight average over the entire previous chunk. In the future we plan to investigate more sophisticated techniques (such as Gaussian Process regression) for estimating various dynamically varying rates during speech and extrapolating those rates during pause.[3]

Finally, given the hierarchical nature of speech rhythm, some other rhythm might prove more relevant to the turn cycle than syllable rhythm. For instance Beňuš considered recurring accents (phrasal stress rhythm), as well as syllables. For Finnish mora rhythm is another candidate worth investigating.

3.3. Coupled Oscillator Model

The next step in our investigation will be to use the Coupled Oscillator Model (COM [8]) to allow multiple, dynamically varying rhythms. This step is important also for our goal of incorporating pausing behavior into the COM.

The Coupled Oscillator Model uses dynamic systems theory to derive a linear regression model for durations (T_1) of various units during speech given the number of synchronized subunits or cycles (n_i) at various levels:

$$T_1 = c_1 + c_2 n_2 + c_3 n_3 + \cdots + c_k n_k, \qquad (1)$$

For instance, our previous analyses of pause group durations in conversational (spontaneous) Finnish speech, allowing for five possible levels, have indicated strong mora and phrasal stress rhythm with possible weaker foot rhythm [7, 6].

Extending the dynamic model to two speakers instead of one is relatively straight forward in principle, since the underlying theory does not require that all oscillators in the system belong to a single speaker. We have, in fact, previously applied the model for analyzing behavior in the so called synchronous speech task, where two speakers read a text out loud together at the same time [9, 10].

A major challenge when modeling the synchronizing behavior of two speakers, however, is how to handle situations such as pauses in which information providing the basis for synchrony is temporarily diminished or absent. One possibility is to introduce stochastic coupling, the

[3] It would also be desirable to include (short) overlap durations as “negative pauses” in the distributions for turn transitions. This idea was also suggested by Heldner & Edlund [3] for a noncyclic (“no effects”) model.

idea being that the synchronizing signal between oscillators (and participants) varies as to its reliability, rather than being modeled as exact. The beginning of silence can be taken to be a strong cue as to the phase of the other participant (explaining why subjects typically pause relatively often in the synchronous speech task), but phase uncertainty grows as silence continues.[4]

Such a characterization leads naturally to a distribution of pause durations with expected value corresponding to the equation (1) above. Following Beňuš we might hypothesize, for instance, that each pause contains an integral number of "silent stress groups" as a continuation of the stress group rhythm of the preceding speech (perhaps with a fixed, preferred number of "silent syllables" per stress group). Since several levels of rhythm are mutually synchronized in the COM, stress group frequency at the beginning of pause should be estimated not merely on the bases of previous stress groups (whether using a raw average or some other technique), but also taking into account all the relevant interacting rhythms on various hierarchical levels such as mora, syllable, etc.

4. Summary

We have begun exploring ways of modeling pause durations in Finnish conversations. Thus far, we have analyzed only one speaker pair but we have developed a general statistical model for testing increasingly complex effects in the gathering material.

The simplest versions of the model do not fit the data (much) better than the "no effects" model, but this may yet change as we look at additional speaker pairs and more sophisticated models.

5. References

[1] Š. Beňuš, "Are we 'in sync': Turn-taking in collaborative dialogues," in *Proceedings of the 10th INTERSPEECH*, 2009, pp. 2167–2170.

[2] Š. Beňuš, A. Gravano, and J. Hirschberg, "Pragmatic aspects of temporal accomodation in turn-taking," *Journal of Pragmatics*, vol. 43, no. 12, pp. 3001–3027, 2011.

[3] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, pp. 555–568, 2010.

[4] M. Lennes and H. Anttila, "Prosodic features associated with the distribution of turns in Finnish informal dialogues," in *Fonetiikan Päivät 2002 / The Phonetics Symposium 2002*, P. Korhonen, Ed. Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, 2002, pp. 149–158.

[5] M. Lennes, "Segmental features in spontaneous and read-aloud Finnish," in *Phonetics of Russian and Finnish*, V. de Silva and R. Ullakonoja, Eds. Peter Lang, 2009, pp. 145–166.

[6] M. O'Dell, M. Lennes, and T. Nieminen, "Hierarchical levels of rhythm in conversational speech," in *Speech Prosody 2008: Fourth Conference on Speech Prosody, Campinas, Brazil*, P. A. Barbosa, S. Madureira, and C. Reis, Eds., 2008, pp. 355–358.

[7] M. O'Dell, M. Lennes, S. Werner, and T. Nieminen, "Looking for rhythms in conversational speech," in *Proceedings of the 16th International Congress of Phonetic Sciences*, J. Trouvain and W. J. Barry, Eds. Universität des Saarlandes, Saarbrücken, Germany, 2007, pp. 1201–1204.

[8] M. O'Dell and T. Nieminen, "Coupled oscillator model for speech timing: Overview and examples," in *Nordic Prosody: Proceedings of the Xth Conference, Helsinki 2008*, M. Vainio, R. Aulanko, and O. Aaltonen, Eds. Peter Lang, 2009.

[9] M. O'Dell, T. Nieminen, and L. Mustanoja, "Assessing rhythmic differences with synchronous speech," in *Speech Prosody 2010 Conference Proceedings*, 2010, pp. 100 141:1–4.

[10] M. O'Dell, T. Nieminen, and L. Mustanoja, "The effect of synchronous reading on speech rhythm," presentation given at *Rhythm Perception & Production Workshop 13*, Leipzig, 2011.

[11] M. Wilson and T. P. Wilson, "An oscillator model of the timing of turn-taking," *Psychonomic Bulletin & Review*, vol. 12, pp. 957–968, 2006.

[12] T. P. Wilson and D. H. Zimmerman, "The structure of silence between turns in two-party conversation," *Discourse Processes*, vol. 9, no. 4, pp. 375–390, 1986.

[4] An interesting finding from our analysis of the synchronous speech task, which may be relevant in the present case, is that while speakers were less synchronized after pause than before, asynchrony did not increase with pause durations greater than approximately 200 ms. This could be taken as further evidence of a silent rhythm during pause.

A Sign Annotation Tool for Phoneticians

Stina Ojala[1,2,*] *and Pertti Palo*[3,4]

[1]Department of Oral Diseases, Turku University Hospital and Department of Oral and Maxillofacial Surgery, University of Turku, Finland
[2]Department of Information Technology, University of Turku, Finland
[3] Institute of Behavioural Sciences, Faculty of Behavioural Sciences, University of Helsinki, Finland
[4]Institute of Mathematics, School of Science, Aalto University, Finland
[*]Corresponding author: `stiroj@utu.fi`

Abstract

We report progress in designing and implementing a sign annotation tool for sign language researchers. We employ a user-centred design approach to develop the tool. The prototype will be made available in open-source form to the research community. In this paper we identify tasks that the sign annotation tool should support and describe a prototype and its user interface. Furthermore, this paper is a call for participation from fellow researchers to contribute their results to this project.

1. Introduction

When phoneticians want to analyse acoustic speech – i.e., speech – there is a standard system to use: Praat [1]. But when phoneticians want to analyse sign language – or signed speech – there are no adequate systems to choose from (see [2] (Chp. 1.4), [3]). Without such a system phonetic research of sign language is very slow if not outright impossible. The current sign analysis tools, such as SignStream© [4], are aimed at sign or utterance-level analysis, that is linguistic analysis, not at phonetic-level precision or articulatory event analysis.

Research in sign articulation studies is fairly recent and for the most part it has been conducted via analogue and digital video material. The difficulty we see is that the research techniques and tools to date have been mostly isolated from each other. Thus, all research groups have their own tools suited for their specific tasks; there is no common platform for each group to test and validate their contribution in the context of a fully functioning sign annotation tool for sign production and articulation research.

We are working towards creating a complete sign annotation platform for researchers. We employ a user-centred design approach to build an open-source based system in the spirit of [5], [6]. We hope that sign researchers will participate by sending us their recommendations and design criteria and/or participating in the implementation of the system.

In this paper, we list a hierarchical dispersion of the tasks necessary for a useful sign annotation tool. This list is meant to be a starting point for colleagues to change and expand to fit to their own needs. We have derived an initial task list by looking at a particular researcher's task sequence in the field. Our first attempt to validate the tool design is to exemplify a scenario based on a stereotypical sequence of steps a researcher may go through. We anticipate developing a number of typical scenarios to validate the flexibility of the tool design to accommodate different researchers' needs.

We have begun to develop an articulatory sign annotation tool architecture to support the tasks we see. The architecture is based on object oriented design. As such it is structured to provide modularity and flexibility so researchers can easily add and subsitute parts as relevant to their own work.

2. Tasks in Phonetic Sign Research

The tasks for an articulatory oriented sign language research tool would be similar to those available in

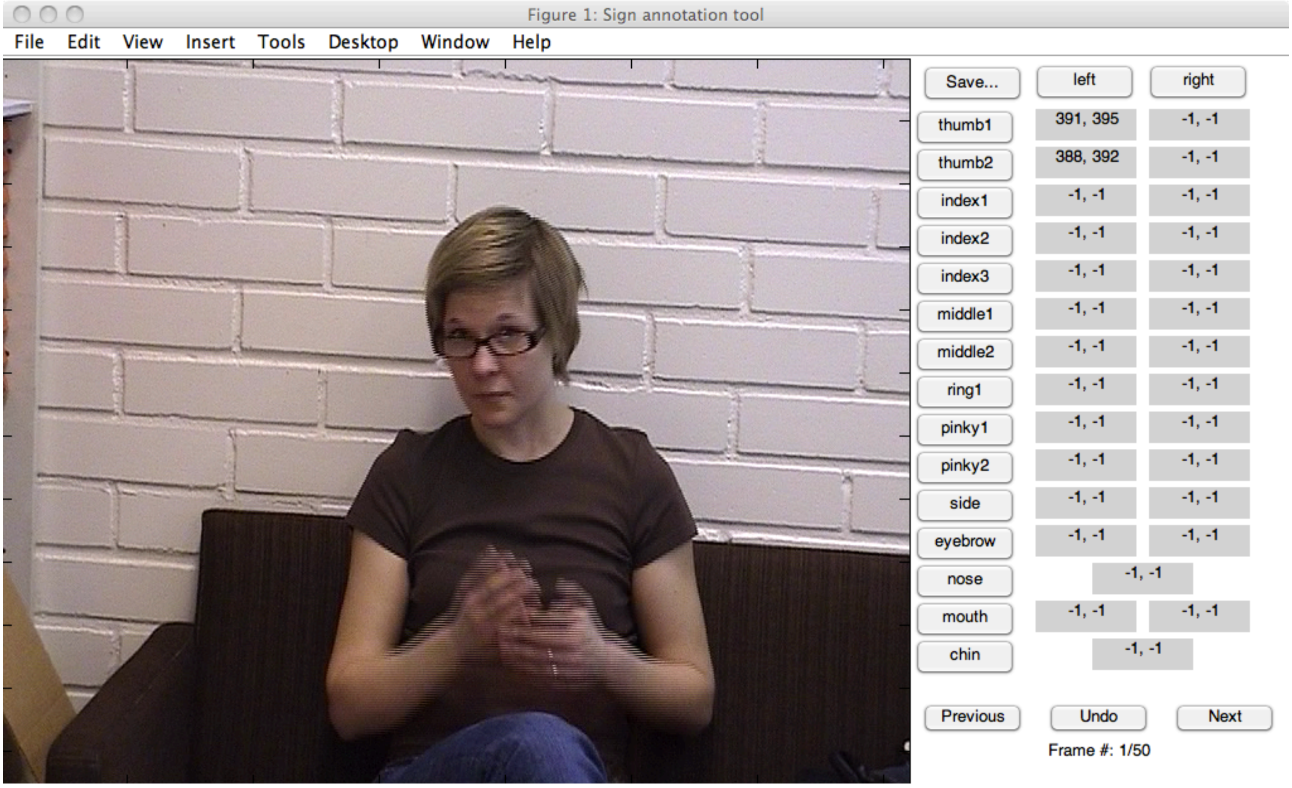


Figure 1: Sign annotation tool
File Edit View Insert Tools Desktop Window Help
Save...
Previous
Undo
Next

	left	right
thumb1	391, 395	-1, -1
thumb2	388, 392	-1, -1
index1	-1, -1	-1, -1
index2	-1, -1	-1, -1
index3	-1, -1	-1, -1
middle1	-1, -1	-1, -1
middle2	-1, -1	-1, -1
ring1	-1, -1	-1, -1
pinky1	-1, -1	-1, -1
pinky2	-1, -1	-1, -1
side	-1, -1	-1, -1
eyebrow	-1, -1	-1, -1
nose	-1, -1	
mouth	-1, -1	-1, -1
chin	-1, -1	

Frame #: 1/50

Figure 1: GUI of the sign annotation tool.

various speech analysis tools. The most basic need is to track a specified set of articulators in time. To date it has been possible only with time-consuming and error-prone manual tracking from a video recording. While not exhaustive at this point, our survey provides a picture of some of the tasks that will need to be supported. We intend to use the tool in our own sign coarticulation research, sign prosody studies and comparative linguistic studies among others. Our own research goals also help us in task definitions.

We have identified several tasks that are performed within a sign articulatory study setting. The tasks include:

- videorecording
- definition and eliciting the desired utterances from the corpus for the specific study design
- time-tagging coarticulatory trajectories into a file
- definition of desired articulator set for the specific study design
- loading an avi-format videoclip
- tagging each articulator in a particular video frame
   1. articulators of head and face
   2. articulators of right hand
   3. articulators of left hand
- defining possible non-visible articulator(s)
- etc . . .

3. Scenarios

Some of the possible research scenarios that would benefit this kind of an annotation tool for sign research include:

- investigating same sign in different contexts
- coarticulatory patterns from one handshape to another
- coarticulatory patterns from one place of articulation to another
- investigating same sign in different individuals
- averaging between coarticulatory patterns to establish a sign space within a sign language
- investigating sign acquisition
- investigating clinical sign phonetics and linguistics

- building sign corpora with phonetic level precision
- ...

4. System Design

4.1. Prototype

The prototype sign annotation tool is a fairly limited matlab script. It is started by loading an avi-format videoclip and provides a rudimentary way of tagging articulator positions in each frame. The set of articulators is currently fixed. Figure 1 shows the GUI of the prototype system.

Figure 2 a) shows the prototype's block structure. The class structure is based on the View-Control-Model design pattern. However, it is not fully implemented as the Matlab class SignAnnotationTool has a dual role as the View and the Control while SignVideoData is the Model class.

The prototype is started by creating an instance of SignAnnotationTool, which creates a SignVideoData object to read the video clip and contain the articulator data. During tagging SignAnnotationTool handles the display of the video frames and processing of user input. Saving the articulator data is handled by a Writer class called ArticulatorDataWriter.

4.2. Expanding and modifying the system

Figure 2 b) shows possible expansions and modifications of the system architecture. The first thing to consider is whether Matlab is the optimal implementation platform for this type of a tool. After all, quite a bit of the potential and required functionality needs advanced GUI features and this is not what Matlab was originally designed for.

The GUI of the tool needs improvements as well as the system in general. Desirable features for the GUI include:

- Praat-like annotation tiers
- time series analysis tools for articulator movement
- articulator tag visualization

The system should also be able to handle the following:

- Undo in articulator tagging (only a very limited feature has been implemented).
- Session saving and retrieval for continuing tagging on a file and checking previously tagged articulators.
- Reading different video codecs/formats.
- Changing the articulators set used for tagging.
- Editing of articulator sets used for tagging and analysis.
- Exporting data to different formats.

Implementing these features calls for the addition of several classes in the program structure. In the back end of the system VideoReader classes should be added to handle different video formats and ArticulatorDataReader class or classes to handle reading of a video clip and articulator data concerning the particular video clip.

Closer to the GUI, articulator and video metadata needs a class structure to represent it and enable the functioning of Visualizer, Analyzer and Annotator classes.

5. Conclusions

This type of multidisciplinary co-operation in designing a user interface makes use of different perspectives of individual researchers into this project. Ultimately this type of collaboration brings end-user feedback nearer, or more precisely yet, into the core of the developer team itself. Thus, it gives opportunities for problem-solving and debugging in a tighter cycle within the process without having to wait for end-user comments on the finished software or programme.

The tool itself is mainly aimed for diminishing the error-prone and laborous manual tracking of coarticulatory patterns. Although in its initial stages it still is quite limited in the tasks it performs, it is aimed at being open for all researchers to develop and add expansions according to their needs. Similarly to Praat, this open source application could prove to be as versatile and flexible a tool in the future.

We would like to kindly invite all fellow researchers to contribute to this project. We welcome all proposals for future dimensions for the application as well as other modes of support, such as participating in the programming itself.

6. References

[1] P. Boersma and D. Weenink, “Praat: doing phonetics by computer (version 5.3.11),” Retrieved March 29, 2011, from http://www.praat.org/, 2011, [Computer program].

[2] S. Ojala, “Towards an integrative information society: Studies on individuality in speech and

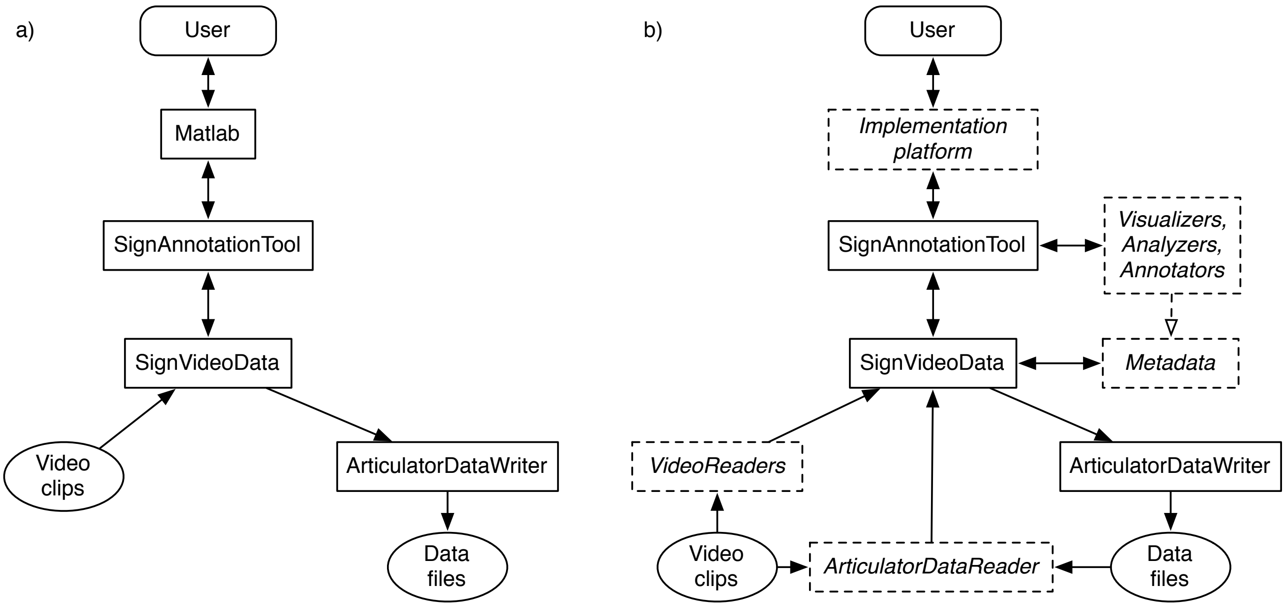


Figure 2: Block diagrams of a) the prototype annotation tool and b) some possible expansions and modifications to the annotation tool.

sign," Ph.D. dissertation, University of Turku, May 2011, tUCS Dissertations No 135.

[3] M. E. Tyrone, *Gesture and Sign Language in Human-Computer Interaction Lecture Notes in Computer Science*, 2002, ch. Overview of Capture Techniques for Studying Sign Language Phonetics.

[4] C. Neidle, S. Sclaroff, and V. Athitsos, "Signstream^{TM}: A tool for linguistic and computer vision research on visual-gestural language data," *Behavior Research Methods, Instruments, and Computers*, vol. 33, pp. 311 – 320, 2001.

[5] S. S. Fels, F. Vogt, B. Gick, C. Jaeger, and I. Wilson, "User-centered design for an open source 3D articulatory synthesizer," in *ICPhS 2003*, 2003.

[6] F. Vogt, S. S. Fels, B. Gick, C. Jaeger, and I. Wilson, "Extensible infrastructure for a 3D face and vocal-tract model," in *ICPhS 2003*, 2003.

Articulating Finnish Vowels: Results from MRI and sound data

Pertti Palo[1,2], *Daniel Aalto*[1,3], *Olli Aaltonen*[1], *Risto-Pekka Happonen*[4],
Jarmo Malinen[2], *Jani Saunavaara*[5], *Martti Vainio*[1]

[1] Institute of Behavioural Sciences, Faculty of Behavioural Sciences,
University of Helsinki, Finland
[2]Institute of Mathematics, School of Science, Aalto University, Finland
[3]Department of Signal Processing and Acoustics, Aalto University, Finland
[4]Department of Oral Diseases, Turku University Hospital and Department of Oral
and Maxillofacial Surgery, University of Turku, Finland
[5]Medical Imaging Centre of Southwest Finland, Turku University Hospital , Finland
*Corresponding author: pertti.palo@aalto.fi

Abstract

We present anatomic and acoustic data from a pilot study on the Finnish vowels [ɑ, e, i, o, u, y, æ, ø]. The data were acquired simultaneously with 3D magnetic resonance imaging (MRI) and a custom built sound recording system. The data consist of a single static repetition of each vowel with constant f_0. The imaging sequence was 7.6 s long and had an isotropic voxel size of 1.8 mm. We report results of listening tests and acoustic analysis of audio data as well as manual analysis of MR images.

Index Terms: Formant analysis, spectral LPC, MRI

1. Introduction

Vowel production has been studied with several imaging methods. The earliest such studies used X-ray imaging [1, 2, 3, 4]. Nowadays, MRI is preferred because no known health hazards are associated to it [5, 6]. Here we report simultaneous MRI and audio data from one test subject pronouncing Finnish vowels. In addition to the images, we assess the quality of the vowels based on a listening experiment of the audio data.

The data examined in this study was acquired for developing a mathematical and computational model of speech production (for a detailed report and further references, see [7] and references therein). We aim at maximal spatial resolution with minimal movement artifacts. The simultaneous audio recording provides an indirect measure of the stability of the vocal tract and a reference point for model validation.

Magnetic resonance imaging (MRI) is a widely used tool to acquire three dimensional (3D) anatomic data of the vocal tract (VT) for speech production studies, simulation and articulatory synthesis [8, 9, 10]. Bones, teeth and most of the small details under the voxel size (1.8 mm in our case) are not visible in MRI. On the other hand tissues containing water and lipids are clearly visible along with mucus which is can be indistinguishable from the actual tissues.

3D MR imaging sequences provide a poor time resolution. We employed a 7.6 s long version of a sequence called VIBE as detailed in [7, 11]. As the conditions are less than ideal for the test subject – requiring a supine position, an extremely long production and being subjected to intense acoustic noise – extra care needs to be taken in evaluating and validating the data.

We use three separate methods to evaluate the same vowel production event. Hence, a single empirical data point is connected to anatomic, acoustic and linguistic contexts.

2. Materials and methods

2.1. The data set

In this study we evaluate a data set consisting of the Finnish vowels [ɑ, e, i, o, u, y, æ, ø]. The set consists of a single production of each of the vowels uttered by a native male speaker. For each production we acquired a simultaneous 3D MRI scan and an audio recording. A detailed report on the data acquisition is available in [7, 11]. For perceptual and acoustic evaluation clear speech samples were extracted from the recording before and after the MRI sequence in the same manner as in [7].

2.2. Perceptual evaluation of audio data

Two samples of clear speech were extracted manually from the MRI recordings for each of the eight vowels. The first sample – the begin sample – was a 200 ms sample directly before the onset of the MRI noise. The second sample – the end sample – was a 200 ms sample located 100 ms after the end of the MRI noise.

These samples were listened to by 20 female students of phonetics with no known hearing defects and whose ages ranged between 20 and 39 years (mean 26 years, s.d. 5 years). Two listeners were bilingual speakers of Finnish and Swedish and all the rest were native speakers of Finnish. The first three listeners used Sennheiser HD 250 linear II earphones during the test and the rest used Sony MDR-7510 earphones. In both cases the listening experiment was run with Max/MSP software (version 6.0.3) running on a MacBook Pro laptop with Mac OS X (version 10.6.8).

In the experiment, the listeners were asked to categorise the vowels samples they heard and rate the sample's prototypicality and nasality. The test was a forced choice test and the listeners could listen repeatedly to the sample they were rating.

2.3. Acoustic evaluation of audio data

The samples used in the perceptual assessment were analysed with LPC. As the recording system does not have a flat frequency response [7], we employed the measured power spectral response of the system in compensating the FFT spectrums of the samples. The spectral linear prediction algorithm [12] was then used to obtain formant estimates for these samples. The fundamental frequency f_0 of each of the samples was estimated with the autocorrelation method. All of the acoustic analyses were carried out with Matlab release 2010b running on a MacBook Pro laptop with Mac OS X (version 10.6.8).

2.4. Evaluation of MRI data

We measured the cross sectional area of the smallest opening within the vocal tract and the opening distance of the jaw for each vowel articulation. The jaw opening was measured as the distance between the maxilla and the mandible as shown in Figure 1. Also, we measured the cross sectional area of the lip opening for those articulations where it was possible to define a cutting plane limited by the lips. All articulatory measurements were done with OsiriX (version 3.9) on a MacBook Pro laptop with Mac OS X (version 10.6.8).

3. Results

The prototypicality and nasality scoring proved to be inconclusive. In contrast, the categorisation part of the experiment yielded clear result as seen in Table 1. The confusion matrices displayed there show that [æ] and [u] in this data are not very representative at the end of the productions. It should be noted that many of the listeners reported that the productions in general were not very prototypical, but that they were nonetheless clearly categorisable in most cases. Two other frequently reported observations were the machine like quality of the speech and the fact that some of the listeners felt that some of the samples were shorter than others.

Table 2 lists the results of the acoustic analysis of the samples. As can be seen the subject was able to sustain a fairly stable f_0 and in most cases the formants provided by the analysis show only a small drift. However, there is a relatively large difference in the formants of [e], [i], [u], and [æ]. In the cases of [e] and [i], the formant extraction algorithm has produced one or more artifactual formants. In the cases of [u] and [æ], the articulation has changed considerably. These views are supported by the confusion matrices in Table 1.

Figure 1 shows the position of the narrowest constriction for each vowel in the vocal tract (between the lips and the epiglottis). Table 3 lists the articulatory measures: Jaw opening (distance of the maxilla and the mandible), lip opening (inner distance between the lip surfaces), and the smallest area (size of the narrowest constriction in the vocal tract) for each vowel. Lip opening area is also listed for rounded vowels.

In the present data the three dimensional features of the articulations are readily visible. In all of the current vowel productions the tongue is grooved and asymmetric with respect to the mid-sagittal plane. In the vowels [y, i, e, ø, u] the tongue is in contact with the palate and in [u, a, o, æ] with the pharyngeal wall.

4. Discussion

Our observations of the position of the tongue and its groovedness are well in line with the observations of earlier studies [3, 4]. It should be noted that this is the first 3D data set on Finnish and as such is potentially richer in detail than previously collected data. An X-ray image produced in the traditional way (rather than with computed tomography) is an average of the tissues in one direction. In contrast, MRI produces images as slices through the tissues. The difference is demonstrated by comparing Figures 1 and 2. However, as can be seen from our results, the MRI data will provide additional detail, while the original understanding of vowel articulation remains well founded.

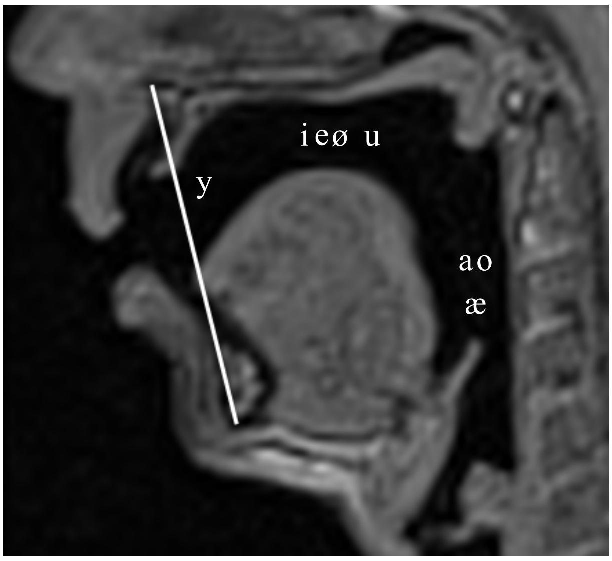


Figure 1: Position of the narrowest constriction of each vowel shown on the midsagittal cut of [ø]. Also shown is a line demonstrating measuring of the jaw opening distance.

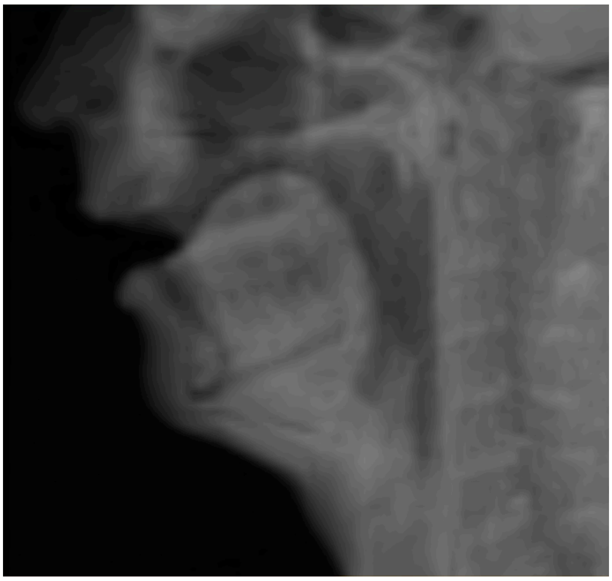


Figure 2: An average image of the 3D MR image stack of [e] produced by averaging in the direction perpendicular to the sagittal plane.

It is difficult to produce good vowels in the conditions required by MRI. As our data on [u] and [æ] show, the articulatory position is liable to change during the long productions as well as being different from that employed in spontaneous speech [13]. This likely to be due to several different effects acting simultaneously. The supine position is likely to affect the po-

Table 1: Confusion matrices for the listening experiments. a) Beginning samples and b) end samples.

a)	Categorised as							
Target	[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[ø]
[ɑ]	20	0	0	0	0	0	0	0
[e]	0	20	0	0	0	0	0	0
[i]	0	1	17	0	0	2	0	0
[o]	0	0	0	19	1	0	0	0
[u]	0	0	0	1	19	0	0	0
[y]	0	0	0	0	0	19	0	1
[æ]	0	0	0	0	0	0	20	0
[ø]	0	0	0	0	0	0	0	20

b)	Categorised as							
Target	[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[ø]
[ɑ]	20	0	0	0	0	0	0	0
[e]	0	18	0	0	0	0	0	2
[i]	0	1	19	0	0	0	0	0
[o]	0	0	0	17	3	0	0	0
[u]	0	0	0	8	12	0	0	0
[y]	0	0	0	0	0	17	0	3
[æ]	12	0	0	0	0	0	8	0
[ø]	0	0	0	0	0	0	0	20

Table 2: f_0s and formants F1-F4 for vowel samples with target $f_0 = 110$Hz.

Sound		[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[ø]
f_0 (Hz)	begin	107.6	108.1	108.4	107.3	107.3	108.4	105.3	106.8
	end	110.8	109.2	109.4	110.2	111.4	110.5	107.6	109.7
F1 (Hz)	begin	658	272	255	403	269	294	748	419
	end	644	524	238	392	342	303	764	452
F2 (Hz)	begin	1059	560	987	753	636	1577	1532	1488
	end	989	1993	722	717	714	1539	1245	1360
F3 (Hz)	begin	2763	1898	3039	2298	2186	2057	2278	2008
	end	2530	2436	2183	2181	2160	2012	2373	2088
F4 (Hz)	begin	3643	2625	3473	3487	3381	3281	3511	3321
	end	3715	3504	3078	3252	3103	3148	3531	3253

Table 3: Articulatory measures from the MR images.

Phoneme	[ɑ]	[e]	[i]	[o]	[u]	[y]	[æ]	[ø]
Jaw opening (cm)	7.4	7.3	6.8	8.2	8.2	7.0	8.6	8.2
Lip opening (cm)	1.3	1.7	1.2	0.5	0.6	0.6	3.1	1.0
Smallest area (cm^2)	1.3	1.6	0.3	0.5	0.9	1.8	2.5	3.9
Lip opening area (cm^2)	na	na	na	0.7	0.3	0.3	na	1.9

sition of the tongue. The noise of the MRI machine will cause a Lombard effect on the subject's speech. The emptying of the lungs will affect the position of the articulatory organs via the movement of the thorax. Furthermore, the long productions are more likely to produce more extreme articulation as can be seen in e.g. the very narrow lip opening of [u] and [y] in Table 3. Taking into account these considerations, this data can be used in modeling speech production not only at the given data points but also by extrapolating from them.

5. Acknowledgements

We wish to express our gratitude to Mr Olli Santala for helping with the MAX/MSP implementation of the listening experiments. The study has been supported by the Academy of Finland (proj. numbers 128204, 125940, and Lastu 135005).

6. References

[1] S. Jones, "Radiography and pronunciation," *The British Journal of Radiology*, vol. 2, no. 2, pp. 149 – 150, 1929.

[2] A. Sovijärvi, "Röntgenogrammeja suomen yleiskielen vokaalien ääntymäasennoista," *Virittäjä*, 1938.

[3] T. Chiba and M. Kajiyama, *The Vowel, Its Nature and Structure*. Phonetic Society of Japan, 1941.

[4] A. Sovijärvi, *Suomen kielen äännekuvasto*. Gummerus, 1963.

[5] T. Baer, J. C. Gore, S. Boyce, and P. W. Nye, "Application of MRI to the analysis of speech production," *Magnetic Resonance Imaging*, vol. 5, pp. 1 – 7, 1987.

[6] O. Engwall and P. Badin, "Collecting and analysing two- and three-dimensional MRI data for Swedish," *TMH-QPSR*, no. 3-4/1999, pp. 11–38, 1999.

[7] P. Palo, "A wave equation model for vowels: Measurements for validation," Licentiate Thesis, Aalto University, Institute of Mathematics, 2011.

[8] A. Hannukainen, T. Lukkari, J. Malinen, and P. Palo, "Vowel formants from the wave equation," *Journal of the Acoustical Society of America Express Letters*, vol. 122, no. 1, pp. EL1–EL7, 2007.

[9] M. Stone, G. Stock, K. Bunin, K. Kumar, M. Epstein, C. Kambhamettu, M. Li, V. Parthasarathy, and J. Prince, "Comparison of speech production in upright and supine position," *Journal of the Acoustical Society of America*, vol. 122, no. 1, pp. 532 – 541, 2007.

[10] P. Švancara and J. Horáček, "Numerical modelling of effect of tonsillectomy on production of Czech vowels," *Acta Acustica united with Acustica*, vol. 92, pp. 681 – 688, 2006.

[11] D. Aalto, J. Malinen, P. Palo, O. Aaltonen, M. Vainio, R.-P. Happonen, R. Parkkola, and J. Saunavaara, "Recording speech sound and articulation in MRI," in *Biodevices 2011*, Rome, Italy, 2011, pp. 168 – 173.

[12] J. Makhoul, "Spectral linear prediction: Properties and applications," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 3, pp. 283 – 296, 1975.

[13] O. Engwall, "Are statical MRI data representative of dynamic speech? Results from a comparative study using MRI, EMA and EPG," in *In Proceedings of International Conference on Spoken Language Processing 2000 (ICSLP 2000)*, 2000, pp. I: 17–20.

Factors affecting the categorizing and rating of the goodness of L2 vowels

Terhi Peltola

Institute of Behavioural Sciences, University of Helsinki, Finland

terhi.peltola@helsinki.fi

Abstract

A foreign accent can cause difficulties for the listener to understand a language learners' speech, especially when the correct pronunciation of the foreign speech sounds is problematic for the learner due to category goodness correspondence between speech sounds (Best 1991). For Hungarian Finnish learners the most problematic Finnish vowels are /æ/ and /e/, due to phonemic and orthographic differences. This can sometimes create confusions and amusing sentences, such as *Hän lehti takaisin* instead of *Hän lähti takaisin* ('she leaf back' instead of 'she went back'). The current paper is an on-going quantitative investigation on which factors affect the categorization and goodness rating of foreign pronounced vowels. The stimuli were extracted from recordings of a previous study (Peltola 2011). The different ways of production were reading and imitating. In the present study Finnish university students rated the goodness of these problematic vowels pronounced by Hungarian students separately and in simple CV-syllables /kV, pV, tV/ on the Likert scale (from 1–7). Three hypotheses were tested in the current paper. Firstly, the effect of the ways of production of the speech is investigated: are the L2 read and imitated vowels categorized and rated differently by native speakers? Secondly, the effect of musicality of both the speaker and the rater are investigated. Thirdly, the effect of context is investigated: were the vowels in syllables rated better than single vowels? Read and imitated stimuli were rated differently, musicality was found to affect ratings in certain ways and syllables were rated better than single vowels.

Index Terms: foreign accent, imitating, goodness rating

1. Introduction

In the current paper I describe a study in which native Finnish speakers categorized and rated the goodness of vowels which the Hungarian Finnish learners had produced in Finnish sentences. The vowels in focus were Finnish /e, æ/, which are difficult for Hungarians to learn to categorize (Winkler et al. 1999) due to the category-goodness equivalence to Hungarian /æ, ɛ/ and to production difficulties arising from different orthographic rules.

I begin by considering the theoretical background from the learners' and the listeners' perspective, and then go on to describe the vowel systems in more detail. After this I introduce the research questions in more detail and describe the current study, and in turn this is followed by a thorough discussion of the results.

Kuhl (1995) describes adults as language bound listeners, whereas children are citizens of the world, who are capable of learning any language from their ambient world. The threshold age seems to be between six and ten months, after which children are only able to categorize the speech sounds that exist in their native language. Supporting this view, Näätänen et al. (1997) found the native-language categories corresponding to differences in MMN in their brain response studies.

During the first year of their lives children gather statistical information from the ambient world and the languages around them, be they one, two or more. The adult human being's perceptual world is already based on native language models, and while not every bit of information available in the new language will be perceived, when it is, it will always be with a native bias. (Kuhl 1995, Best 1995.) According to some more recent studies adults are collecting information from the ambient language environment as well and adjusting their speech accordingly, although perhaps not with the same proficiency as children. According to Strange's (2012) Automatic Selective Perception (ASP) working model, late language learners must employ more attentive energy in order to gather enough information to form and adjust to new categories. He is comparing children acquiring their first language, early second language learners and late second language learners. According to Strange's model, information is gathered from the foreign-language with selection perception routines which use acoustic information selecting the most reliable acoustic parameters. Whereas Chang (2012) postulates that phonetic shift also applies to native language categories from the very beginning of second language learning. He studied phonetic shift native vowels of American English natives studying Korean in a Korean environment.

Liberman and Mattingly (1995) as well as Best (1995) describe the connection between speech sounds and their perception directly and the cues for speech are speech gestures or distal events. Thus, the speech sounds are not converted to acoustical signal nor analysed acoustically. Rather they are analysed directly as objects of information. According to Liberman and Mattingly (1995) speech is processed in a special speech module. Mattingly (1990) sees the connection between perception and production very simply, even as if it were not there at all; both are based on gestures.

Best (1991) developed the PAM (Perceptual Assimilation Model) to describe how adults' perceptual world is defined by the speech categories that are formed thorough the active perception of the native language during childhood by perceiving the distal events of speech. This model's patterns describing sound assimilations are good tools to describe two languages contrastively, but do not consider how perception is changing thorough language learning or only as an effect of the ambient environment. In the subsequent PAM-2 model the language learners' view is also considered in detail (Best et al. 2007). Best claimed already earlier that phonetic training helps in perceiving foreign sounds, although she considers that perceiving is based on phonological models.

Flege (1987, 1995) explains in his SLM (Speech Learning Model) how native language speech categories govern the learning of new speech categories. According to him, near native proficiency is possible for most speech sounds. Later Flege et al. (2003) found, in their study of Native Korean immigrants, that the more a foreign accent was perceived, through L2 listening and speaking, the more it contributed to better ratings. The Critical Period Hypothesis (Lenneberg 1967) was not found to reflect the perceived accent.

Kuhl (1995) explains the perception of speech sounds thorough the psychoacoustics of the language in NLM Theory

(Native Language Magnet). This theory describes the distortion of acoustic perception, caused by native prototypes. According to their studies, neighbouring sounds of prototypes are heard as more similar than the less-prototypical ones in the category periphery and the perceived differences did not reflect the acoustic differences.

Kuhl's NLM theory is supported by Weber's findings. In Weber et al.'s (2011) study on accented speech, participants showed better results recognizing foreign accented speech spoken by a person with the same accent. Their findings suggest that linguistically experienced L2 learners are better at recognizing words with the same accent. This indicates that the native perceptual world is much more bound to native categories than the language learners' perceptual world.

Another effect which is distorting the heard speech is the McGurk effect (McGurk and McDonald 1976). In studies where subjects heard syllables while seeing different syllables pronounced, they tended to report hearing the actual visual input. This means that the visual stimulus is stronger than the auditory stimulus and it is distorting what the person hears. What is the case in foreign-language situations? It can be interpreted that language learners rely more on what they see than on what they hear. This is supported by Kuhl et al.'s (2003) study on children learning foreign sounds. They found that video stimuli were not enough for children to learn new speech sounds, children needed a stronger motoric model and a human interface. It seems, though, that children learn new speech sounds only in interaction with adults.

Kuhl's (1995) findings of the NLM effect suggests that human beings are bias perceivers and raters of speech sounds, especially the ones closer to the native prototype acoustic area, because native category magnets distort perception. Winkler et al. (1999) suggest that native Hungarians with little experience in Finnish have difficulties with some /e, æ/ categories vowels because they are in the native Hungarian /ɛ/ prototype area. Weber et al.'s (2011), Strange's (2012) and Chang's (2012) findings give an insight into the perceptual flexibility of the language learner. Nevertheless, native speakers are good raters of the goodness of foreign speech sounds, even though a little biased and magneted towards the native prototypes.

Imitation is the basis for learning. In this study, imitation is considered as phonetic, somewhat free of native categories and allophones (Mitterer and Ernestus 2008). Imitation is also considered to rely on the imitation of the gestures seen or distal events heard. However, the stimulus imitated is perceived thorough the native pattern, which might distract the learner to concentrate on features irrelevant in the target language.

Finnish and Hungarian vowel systems are not identical, as described earlier. The Finnish vowel system consists of 8 vowels, all of which have short-long counterparts. Hungarian also has phonemic length, but the Hungarian vowel system is not symmetric regarding the quantities and qualities of vowels: all of the short vowels do not have a longer counterpart. In Hungarian there is traditionally considered to be 7 short and 7 long vowels, that is altogether 14 vowel phonemes. There is some dialectal variation in the vowel inventory. In some dialects the vowel inventories are more symmetrical. Nevertheless this is discarded in this study, because the informants were studying in Budapest, which is a dialectally neutral area.

The vowels in focus in this experiment are the Finnish vowels /e, æ/. This opposition is problematic for a Hungarian learning Finnish as mentioned earlier. In standard Hungarian there are only 2 short front vowels /i, ɛ/ and two long front vowels /i:, e:/, whereas there is 3 front vowels in Finnish /i, e, æ/ all of which have longer phonemic pairs. However, the short vowel /ɛ/ has an orthographical longer pair /e:/, which differs from its shorter pair with its quality. These two Hungarian vowels seem to correspond to the Finnish /e, æ/ vowels categories, but make it problematic for Hungarians to learn their Finnish counterparts.

2. Research questions

Three questions were formed to test the collected data, the goodness ratings. Questions were formed to investigate which factors are relevant in categorizing foreign accented speech sounds.

As mentioned, the stimuli consist of four different types of vowels: categories /æ/ and /e/, both imitated and read. First question tested is if read and imitated stimuli was categorized and rated differently by the native Finns. Further the author considers which stimuli group do the Finnish categorize best (read vs. imitated, /æ/ vs. /e/)? In other words does the vowel quality or method of production affect the categorization?

In pre-test questions in both experiments the participants were asked if they had any musical hobbies and evaluate their musical abilities. This way we can study the effect of musicality: Are musical Hungarian informants' vowels categorized and rated better? Are musical Finnish goodness raters better categorizers than their non-musical peers?

Half of the stimuli were vowels, half syllables. However, they were cut from the very same sentences; therefore they were the very same vowels. The third question to answer is, does context, in this case unvoiced plosives /p, t, k/, affect the categorizing and rating of the vowels?

3. Method and procedure

The stimuli were collected in a previous experiment during autumn 2010 at ELTE University Budapest, Hungary. Participants in this earlier experiment were university students studying Finnish language and culture as a minor subject for the first semester at ELTE University in Budapest. The participants had studied Finnish for a period of 12 weeks, four lessons a week with a native Hungarian, but bilingual proficient Finnish language university lecturer. Four informants out of this group were chosen for the second experiment, two women and two men, mean age of 18.5 years, and they were studying various majors at the university. One male and one female were chosen because they reported musical hobbies and talents; the other two were selected because they did not report such hobbies or talents.

The task given in this previous experiment was first to read out loud sentences in Finnish, and after this, to imitate the very same sentences pronounced by a native Finnish speaker. The model for the imitated stimuli was heard through earphones. Gathered /e/ and /æ/ vowels and syllables were extracted from the whole sentences, lengthened, and their intensity was strengthened with Praat software tools (Boersma et al. 2005). Because of the sounds manipulation, they were simplified into mono sounds. The duration of a vowel stimulus was approximately 350 ms, whereas the duration of a syllable stimulus was slightly shorter, 300 ms. Lengths were decided objectively by the author. Altogether the stimuli consisted of 188 sounds (94 vowels and 94 syllables) half of these imitated and half read by the Hungarian students. The vowels and syllables were extracted from the very same words, the only difference was the preceding consonant, or the lack of it, and the length of the stimulus.

The participants of the current experiment were 26 Finnish students at the University of Helsinki, 2 male, 24 female. Their average age was 23.23 and their majors varied from English philology to phonetics. The experiment was conducted as a part of an introduction to experimental methods course for undergraduate students in the Faculty of Behavioural Sciences. 7 raters' results were discarded, either because of too many times misses or uniform answers. The Likert scale is commonly used as a goodness rating classificator. The scale from 1-7 seemed most usable, considering the purpose and raters in this study. Scales 1-5 and 1-9 are also often used in phonetic goodness ratings.

The participants listened to the stimuli in a normal lecture auditorium without any hearing enhancements from the auditorium loud speakers. The acoustics in this auditorium can be described as good. The experiment took less than 14 minutes at the end of a normal lecture. The participants were asked to listen to the stimuli carefully, categorize and rate the goodness of the vowels and syllables heard according to the Likert scale using the following values: 1 'ä', 2 'almost ä', 3'poor ä', 4'neither', 5'poor e', 6'almost e' and 7'e'. While categorizing, the participants knew that the vowels were produced by Hungarians studying Finnish, and they were warned of the sounds quality, which was very non-human like after the sound manipulations.

For data analysis the group of raters was divided into two subgroups depending on any musical hobbies and skills they reported, either to the 'musical' or the 'non-musical' group.

4. Results

The goodness ratings were typed into the computer, saved as a text file and these data were processed with the statistics program R. First, the results were compared in bar plots, after which statistical Mann-Whitney statistical significance tests were conducted.

4.1. Barplots

An overall look at the data reveals that imitated stimuli receive less best ratings, either '1' or '7'. This is surprising since imitating should have a positive effect on the quality of L2 speech sounds and improve it. However, since the imitation task was conducted only once, lower overall goodness ratings for the imitated stimuli seem more reasonable.

Comparing the musical and the non-musical raters' goodness ratings reveals that musical raters tend to give more 'poor' categorizations to the stimuli, whereas the non-musical raters tend to give more 'almost' categorizations. This indicates that musical raters rate the stimuli slightly differently, more strictly, when they classify it not to be a good example of the category. It seems that the musical raters are more strict goodness raters than the non-musical.

Comparing the same group of informants, musical and non-musical, reveals that the non-musical tend to receive more 'ä' and 'almost ä' ratings to their /æ/ productions, than the musical. It is very surprising that the musical informants read and imitated /æ/ vowels ratings' do not differ much.

/e/ and /æ/ syllables seem to receive the same amount of correct categorizations from the Finnish raters. Whereas, it is especially the imitated /e/ vowels which receive more 'neither' categorizations. The same tendency does not apply for the /æ/ vowels, neither, read or imitated. The /e/ and /æ/ in syllables seem to receive more '1' or '7' goodness ratings from the Finns than the single vowels. This indicates that the preceding consonants carry relevant information for the vowel categorization and goodness rating, even though the goodness rating of the vowels and syllables was not found to differ statistically.

4.2. Statistical testing

The goodness rating of the read and imitated vowels differed statistically with a p-value of 1.317e-14. This indicates that the different experiment settings, reading and imitating, change the perceived sound quality.

The musical and non-musical raters' goodness ratings somewhat differed statistically with a p-value of 0.04918, which indicates that the musical skills and hobbies affect the goodness rating of these vowels. However, the p-value leaves some room for questioning the strength of the effect.

However, a more detailed look reveals that, both the non-musical and the musical raters goodness rating of read and imitated /e/-sounds differed statistically with a p-values of <0.001 (non-m) and <0.001 (m), whereas the goodness rating of the /æ/-sounds did not differ statistically in either of the groups (non-m p=0.89, m p=0.32).This indicates that the raters' musical skills and hobbies do affect the goodness ratings of the imitated /e/-sounds, but not the imitated /æ/-sounds.

The goodness rating of the musical and non-musical informants sounds differed statistically with a p-value of <0.001. A closer look at the same groups reveals that the goodness rating of their imitated sounds differs with a p-value of <0.001, but the goodness rating of the read sounds does not differ statistically (p=0.13). These p-values indicate that musical hobbies and skills affect the imitation skills, but not the reading skills.

The goodness rating of the read and imitated /æ/-sounds did not differ statistically (p=0.5298), but the ratings of the read and imitated /e/-sounds did differ with a p-value of <0.001. The same can be seen in more detailed inspection of the musical (p-values /æ/=0.9, e<0.001) and non-musical (p-values /æ/=0.14, /e/<0.001) informants' groups. This indicates that the /e/-sound qualities differ whether imitated or read, whereas the /æ/-sound qualities do not differ in different experiment settings.

The goodness rating of the vowels and the syllables did not differ statistically (p=0.03) in the whole study group.

5. Discussion

The goodness rating of the read and imitated stimuli were found to differ statistically. Bar plots revealed further that imitated stimuli were rated less often to be good examples of the category they represented. Imitated stimuli received half or two thirds of the best ratings, '1' or '7', compared to the read stimuli in the whole data. This indicates that the imitated stimuli qualities were poorer than the read stimuli. As mentioned earlier, the imitation task was conducted only once, whereas multiple repetitions most likely had resulted in better ratings for the imitated stimuli.

Musicality was found to be a factor affecting the production of the /e/s, as well as the goodness rating of the foreign accented /e/s but not the production or rating or the /æ/s. The musical informants got more 'neither' ratings for their /e/s as the non-musical got slightly more /æ/ ratings for their /e/s. This indicates that the musical informants were able to produce the category difference between /e-æ/ slightly better than their non-musical peers.

The musical raters gave worse ratings than their peers without musical hobbies. It seems that as raters, the ones with musical hobbies were stricter. An interesting question is why the musical are stricter raters than the non-musical? It is very likely that their strictness is learned through their musical hobbies. Learning to play an instrument involves learning to be critical and to handle criticism as well as training to hear differences in auditory signals not heard before.

Syllable stimuli received more correct and better ('ä/e' and 'almost') ratings from the Finns than the vowel stimuli, even though they were not found to differ statistically. It seems that the preceding consonants contribute more information about the preceding vowels and thus make it easier for the raters to rate it as a good or an almost good example of the native category, but it does not affect the categorization much.

Another interesting future pursuit is the effect of repetitions on vowel qualities. Does a higher amount of repetitions lead to better acoustic qualities and goodness ratings from native speakers? Yet another interesting detail for further studies is if the musical raters would rate differently the speech sounds located in the acoustic prototype-periphery-area compared to their non-musical peers. In other words, how does musicality affect the native vowel category inside hierarchies?

Did the Hungarian informants imitate features in the Finnish sounds, which made it very foreign for the Finnish ear but which was not considered in this study and what was it? Another factor to consider in the future is that the L2 learners might imitate incorrect and redundant features in the native stimuli speech sounds. As mentioned, adults need to concentrate more on the L2 speech they are learning than children learning speech sounds. Better learning could be reached with interactive speaker stimuli than recorded sounds, and of course with more repetitions.

6. Acknowledgements

The author would like to thank Professor Olli Aaltonen for guidance and Pertti Palo for statistical advices.

7. References

[1] Best, C. T.,"The Emergence of Native-Language Phonological Influences in Infants: A Perceptual Assimilation Model", Haskins Laboratories status report July-December 1991, SR 107/108,Haskins Labs, New Haven, CT., 1-30, 1991.

[2] Best, C. T., "A direct realist view of cross-language speech perception", in W. Strange [Ed.], Speech perception and linguistic experience: Issues in Cross-language research, Timonium, MD. York Press, 171-204, 1995.

[3] Best, C.T. and Tyler, M., "Nonnative and second-language speech perception: Commonalities and complementaries", in O.-S. Bohn, M. Munro [Eds.], Language Experience in Second language Speech Learning. In honor of James Emil Flege, 13-34, 2007.

[4] Boersma, P. and Weenink, D. 2005, Praat: doing phonetics by computer (Version 4.3.01) [Computer program]. Retrieved from http://www.praat.org/.

[5] Chang, C. B., "Rapid and multifaceted effects of second-language learning on first-language speech production", in Journal of Phonetics 40, 249-268, 2012.

[6] Flege, J., "The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification", in Journal of Phonetics,15, 47-65, 1987.

[7] Flege, J., "Second Language Speech earning, Theory, Findings and Problems"in W. Strange [Ed.], Speech Perception and Linguistic Experience: Issues in Cross-language research, Timonium, MD. York Press, 229-273, 1995.

[8] Flege, J.E., Birdsong, D., Bialystok, E., Mack, M., Sung, H. andTsukada, K., "Degree of Foreign Accent in English sentences produced by Korean children and adults",in Journal of Phonetics 34, 153-175, 2003.

[9] Kuhl, P.K. and Iverson, P., "Linguistic Experience and 'The Perceptual Magnet Effect'", in W. Strange [Ed.], Speech Perception and Linguistic Experience: Issues in Cross-language research, Timonium, MD. York Press, 121-154, 1995.

[10] Kuhl, P. K.Tsao, F.-M. and Liu, H.-M., "Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning",in Proceedings of the National Academy of Sciences 100, 9096-9101, 2003.

[11] Lenneberg, E.H., "On Explaining Language",in Science, New Series, Vol. 164, No. 3880, 635-643, 1967.

[12] Liberman, A. M. and Mattingly, I. G., "The motor theory of speech perception revised",in Cognition 21, 1-36, 1985.

[13] Mattingly, I. G., "The global character of phonetic gestures",in Journal of Phonetics 18, 445-452, 1990.

[14] McGurk, H., MacDonald, J., "Hearing lips and seeing voices", in Nature 264, 746-748, 1976.

[15] Mitterer, H. and Ernestus, M., "The link between speech perception and production is phonological and abstract: Evidence from a shadowing task", in Cognition 109, 1-36, 2008.

[16] Näätänen, R. and Tiitinen, H., "Auditory information processing as indexed by the mismatch negativity", in M. Sabourin, F. Craik, M. Robert [Eds.], Advances in psychological science: Biological and cognitive aspects, Hove, U.K. Psychology Press, 145-170, 1997.

[17] Peltola, T., "Some remarks on the effect of imitation in novel vowel qualities", in Folia Uralica Debreceniensia 17, Debreceni Egyetem finnugornyelvtudomány tanszék, Debrecen. 46-53, 2011.

[18] Raimo, I., Savela, J. andAaltonen, O., "Turku Vowel Test", inFonetiikanpäivät, Akustikan ja äänenkäsittelytekniikan laboratorio, TKK. Otaniemi, 45-52, 2003.

[19] Strange, W., "Automatic selective perception (ASP) of first and second language speech: A working model", in Journal of Phonetics 39, 546-466.

[20] Weber, A., Boersma, M. and Aoyagi, M., "Spoken word recognition in foreign accented speech",in Journal of Phonetics 39, 479-491, 2011.

[21] Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., Czigler, I., Csépe, V., Ilmoniemi, R. and Näätänen, R., "Brain responses reveal the learning of foreign language phonemes", in Psychophysiology 36, 638-642, 1999.

A preliminary comparison of Estonian and Finnish plosives

Kari Suomi [1], *Einar Meister* [2]

[1] Phonetics, Faculty of Humanities, University of Oulu, Finland
[2] Institute of Cybernetics at Tallinn University of Technology, Estonia
kari.suomi@oulu.fi, einar@ioc.ee

Abstract

The paper reports the results of a pilot study on the comparison of Estonian and Finnish plosives. The occlusion and burst durations of word-initial and short/single word-medial plosives was measured in a set of segmentally comparable target words in the two languages read in carrier sentences. The results showed that Estonian short plosives in both word positions have a shorter occlusion duration than Finnish single plosives, and that in the word-medial position, Estonian short /k/ has a shorter burst duration than Finnish single /k/. Cross-linguistic comparisons suggest that it is Estonian medial /k/ that has exceptionally short burst duration.

Index Terms: Estonian, Finnish, plosives, place of articulation

1. Introduction

The starting motivation for this comparison was Leho Võrk's impressionistic description of some differences between Estonian and Finnish short/single plosives [1]. According to Võrk (p. 14), "a word-initial [invariably short] plosive in native Estonian words and in old loanwords is regularly written using the graphemes **p, t, k**. They are preferably pronounced in a somewhat weaker manner than in Finnish"[1]. As for medial short/single plosives, Võrk wrote that "in Estonian, [the graphemes] **b, d** and **g** denote short, voiceless lenis plosives [as in the word *luba, kade* and *lugu*]. They are voiceless like Finnish **p, t** and **k** [as in the words *lupa, kate* and *luku*], but they are pronounced very loosely and with a weak pressure of air, so that also their explosion burst is weak" (p. 15). Notice that, despite the potentially confusing spellings, both Estonian and Finnish traditionally lack a voicing contrast in plosives, although both languages show signs of acquiring one, under pressure from foreign languages. In this experiment, however, no oppositions based on voicing alone are involved.

Võrk thus claims that in Estonian, word-initial plosives are "preferably" pronounced as weaker than in Finnish, and that there is, between e.g. the Estonian words *luba, kade* and *lugu* and the Finnish words *lupa, kate* and *luku*, a difference such that in Estonian the medial plosive is pronounced more loosely and with a weaker explosion burst than in Finnish. Võrk's impressionistic descriptions are in agreement with our similar impressionistic intuitions, and we decided to investigate whether such differences can be observed experimentally.

This can be considered a pilot study. The materials come from another experiment designed to investigate the durational realisation of quantity in the two languages. Target word selection was determined by the existence of word triplets in Estonian which differ from each only in terms of quantity but not in terms of segment quality, and the existence of sufficiently similar word pairs in Finnish (see below). Therefore, the materials were not specifically selected to enable the comparison here undertaken, one consequence of which is that there were no target words with word-initial /p/. Even so, we believe that the results are suggestive.

2. Methods

The materials come from [2]. From among the target words of that larger study all those words were chosen for this experiment that contained initial and/or medial plosives. In the larger experiment, two types of target word sets (which are only a small selection of the different quantity patterns in either language) were chosen, one set in which the quantity oppositions are mainly signalled by consonant duration (the C set) and another set in which the quantity oppositions are mainly signalled by vowel duration (the V set). Target words, all with stress on the first syllable, were selected in such a way as to minimise segmental differences between the target words within a given set. Example word series from both sets are:

	Estonian			Finnish	
	Q1	Q2	Q3	C/V	CC/VV
C set	*kade*	*kate*	*katte*	*katu*	*katto*
V set	*keda*	*keeda*	*keeda*	*kita*	*kiito*

Notice that in e.g. *keeda* (Q2) and *keeda* (Q3) the quantity difference is not indicated in orthography. Although the focus of the present study is on short plosives, we looked at (over)long/double plosives to determine whether patterns observable in short/single plosives also apply to (over)long/double plosives. Meaningful carrier sentences for the target words were constructed, e.g. (with target word underlined):

Q1 *Ütlesin, et olen kade maja pärast.*
Q2 *Ütlesin, et tuleb kate paigaldada.*
Q3 *Ütlesin, et tahan katte paigaldada.*
C *Sanoin, että vilkas katu suljettiin.*
CC *Sanoin, että vanha katto korjattiin.*

Using context manipulations, the sentences were elicited in such a way that the target words occurred under three degrees of prominence: unaccented, thematically accented and contrastively accented. Degree of prominence had no statistical effect on the parameters measured in this

[1] Translation of this and the next quotation by KS. In the Finnish original: "Sananalkuista [aina lyhyttä] klusiilia merkitään viron omissa sanoissa sekä vanhoissa lainasanoissa säännöllisesti kirjaimilla **p, t, k**. Ne äännetään mieluimmin jonkin verran heikompina kuin suomessa", and "Virossa [kirjaimet] **b, d** ja **g** tarkoittavat lyhyitä, soinnittomia leenisklusiileja [kuten sanoissa *luba, kade* ja *lugu*]. Ne ovat soinnittomia, kuten suomen **p, t** ja **k** [kuten sanoissa *lupa, kate* ja *luku*], mutta ne äännetään hyvin löyhästi ja heikolla ilmanpaineella, jolloin niiden eksploosio-paukahduskin on heikko".

experiment, and therefore prominence will not be referred to in the text below.

Nine female speakers were recorded in both languages, the Estonian speakers in Tallinn and the Finnish speakers in Oulu, using high quality digital equipment and highly similar instructions. The numbers of short/single plosive tokens studied were as follows:

	Estonian	Finnish
Word-initial (only /t/ and /k/)	648	430
Word-medial (/p/, /t/, /k/)	232	241

The durational measurements were made using Praat [3] and standard segmentation criteria.

3. Results

When segment durations are compared across languages, using different speaker groups, it is important to control that any differences observed are not due to potential differences in mean speaking rate across the speaker groups. Evidently, and luckily, there was no such rate difference in [2], and hence in this experiment. Thus it was observed in [2] that (i) the grand mean duration of V1 across the quantities was statistically the same in both languages, (ii) the grand mean duration of C2 across the quantities was the same in both languages, and (iii) the absolute amount of accentual lengthening was the same in both languages. However, the mean total target word duration was longer in Estonian (377 ms, s.d = 50.0) than in Finnish (358 ms, s.d = 30.7), a difference that was statistically significant [F(1, 88) = 4.09, p< 0.05]. If anything at all could be predicted from this difference, the only feasible prediction would be that durations of segments (and their parts) would be longer in Estonian than in Finnish. Below, however, we present results to the opposite effect (i.e. shorter durations in Estonian). The obvious conclusion then is that the results represent real differences between the two languages; they are not due to differences in speaking rate.

3.1. Word-initial plosives (C1)

Notice that in both languages C1 is outside the quantity system (i.e., there is no quantity opposition in C1).

3.1.1. Closure duration

Closure duration was shorter in Estonian (70 ms, s.d. = 11.8) than in Finnish (83 ms, s.d. = 13.8), [F(1,34) = 9.24, p = 0.005]. The difference was systematic in both places of articulation studied (/t/ and /k/). Obviously this difference constitutes part of the more general pattern according to which the duration of C1 is shorter in Estonian than in Finnish. As can be seen in Table 1, based on the results in [2], i.e. including initial consonants other than plosives, C1 had a systematically shorter duration in Estonian than in Finnish, both in absolute and in proportional terms.

3.1.2. Burst duration

There was no cross-language difference in mean burst duration which was 22 ms (s.d. = 5.8) in Estonian and 24 ms (s.d. = 8.2) in Finnish, [F< 1]. In Table 1 above, C1 duration includes burst duration: since there was no difference in this parameter, the Estonian – Finnish differences in total C1 duration must be due to differences in closure durations. In both languages, /k/ had a longer burst (Estonian: 27 ms; Finnish: 30 ms) than /t/ (Estonian: 18 ms; Finnish: 17 ms) [F(3, 32) = 17.28, p < 0.001].

Table 1. *Mean absolute and proportional duration of C1 in the three degrees of prominence ("una" = unaccented, "acc" = accented, "con" = contrastively accented). Standard deviations in parentheses. Data from [2].*

	Absolute (ms)		Proportional (% of total word duration)	
	Estonian	Finnish	Estonian	Finnish
una	67 (9)	81 (14)	20.1 (3)	25.9 (3)
acc	68 (11)	82 (12)	20.9 (3)	25.4 (2)
con	104 (22)	111 (19)	22.4 (4)	25.4 (2)

3.2. Word-medial short/single plosives (C2)

We looked at all medial plosives, but there were cross-language differences in burst durations only among the short/single plosives (and the differences in closure duration due to quantity are beyond this paper). Example words with medial short/single plosives are Estonian *kade* (Q1), *keda* (Q1), *keeda* (Q2), *keeda* (Q3) and Finnish *katu* (CVCV), *kita* (CVCV), *kiito* (CVVCV). That is, we did not distinguish between short/single plosives according to word structure, something that might be profitably done in a future, larger-scale study.

3.2.1. Closure duration

Closure duration of was again shorter in Estonian (62 ms, s.d. = 12.5) than in Finnish (71 ms, s.d. = 10.9) [F(1, 52) = 9.02, p = 0.004]. The difference was systematic in each of /p/, /t/ and /k/.

3.2.2. Burst duration

It turned out that the places of articulation behaved differently with respect to burst duration, see Table 2 and Figure 1.

Table 2. *Mean (and median)burst duration (ms) of medial short/single stops.*

	Estonian	Finnish	Statistical significance
/p/	13.0 (14.3)	17.2 (14.4)	n.s. (p = 0.578)
/t/	16.1 (15.0)	15.6 (14.7)	n.s. (p = 1.000)
/k/	19.3 (22.5)	33.4 (31.5)	p < 0.001

That is, burst duration of medial plosives was shorter in Estonian than in Finnish only in the velar place of articulation, but in the velar place the difference was very clear. It is noteworthy, as can be seen in Table 2, that in Estonian the difference between the velar plosives and the other ones is very small, in contrast to Finnish.

There was also a difference between Estonian and Finnish in the number of burstless (i.e. completely voiced) medial plosive tokens: 19 (5.9%) of the Estonian tokens were completely voiced, while this was the case in only 2 (0.8%) of the Finnish tokens. In Estonian, the percentages of completely voiced tokens were 1.1% for /p/, 3.3% for /t/ and no less than 25.9% for /k/. What is exceptional in this pattern is that, usually it is the case that the velar plosives are the least prone to becoming voiced (for more on this see below).

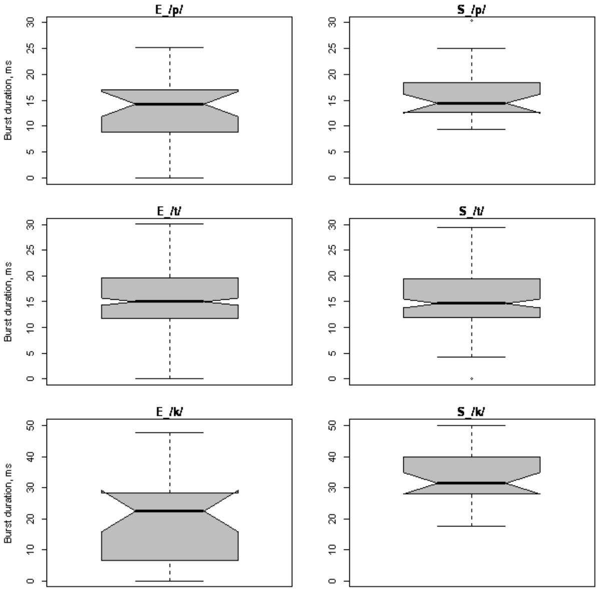


Figure 1. *Box-plots of burst durations of medial short/single stops (the left column: Estonian, the right column: Finnish).*

Our Finnish results on medial plosive burst duration are in agreement with previous ones [4]; for Estonian, corresponding results are not available. In both [4] and in the present experiment, /k/ had clearly longer burst duration (VOT) than /p/ and /t/. And this is the usual pattern across languages, see the small selection of languages in Table 3. (Notice that although Central Swedish voiceless plosives are aspirated at the onset of stressed syllables, at the onset of unstressed syllables they are unaspirated. The data in [5] referred to in Table 3 concern such unaspirated medial /p/, /t/ and /k/.)

Table 3. *Mean burst durations (ms) of medial plosives as reported in the present study and in some earlier studies.*

	Present study (Estonian)	Present study (Finnish)	Suomi, 1980 (Finnish)	Ringen & Suomi, 2012 (Fenno-Swedish)	Helgason & Ringen, 2008 (Central Swedish)
/p/	13	17	11	10	13
/t/	16	16	16	18	23
/k/	19	33	25	25	31

In light of the information in Table 3, it is obviously a special characteristic of Estonian that the burst duration of medial /k/ (<g>) is exceptionally short; it is not the case that the burst duration of Finnish /k/ is unusually long. It thus seems that, in word-medial position (at least intervocalically), Estonian /k/ differs from universal regularities according to which the burst duration of velar plosives is clearly longer than that of labial and coronal (dental or alveolar) plosives. Our results show that with respect to non-short velar plosives, Estonian behaves according to universal tendencies (i.e., velar stops had longer burst duration than labial and dental plosives).

4. Discussion

It appears clear that Estonian and Finnish short/single plosives differ from each other with respect to occlusion duration both in the word-initial position and in the word-medial position. According to our results, a difference in burst duration in the medial position concerns the velar place of articulation only. In Estonian there were also many more instances of completely voiced tokens, a circumstance that may contribute to the subjective impression of "weakness". It is very much possible that these durational differences are sufficient to explain the impressionistic differences between Estonian and Finnish short/single plosives mentioned by Võrk.

Võrk also mentioned the relative weakness of the burst of Estonian medial short plosives. To test the existence of such an intensity difference between the two languages, a more controlled experiment is needed in which also Sound Pressure Level (SPL) is measured.

The Estonian short medial /k/ thus differs with respect to its burst duration from the corresponding plosives in e.g. Finnish, Fenno-Swedish and Central Swedish. Moreover, the Estonian medial short /k/ had completely voiced tokens much more often than did /p/ and /t/, although aerodynamic considerations would suggest the opposite. For example, if a member of the voiced plosive set is missing in a language, it is usually /g/ that is missing since, for aerodynamic reasons (the existence of a small, unexpandable cavity above the glottis), maintaining the transglottal pressure difference required for phonation is relatively difficult in velar plosives [6]. In principle, the exceptional behaviour of the Estonian medial short /k/ may have its source in either glottal or in supraglottal manoeuvres, or in a combination thereof. It is clear that this Estonian consonant deserves closer examination.

5. Acknowledgement

We wish to thank Catherine Ringen for useful comments on earlier versions of this paper.

6. References

[1] Võrk, L., "Viron kielen ääntämys", Tekijä, 1972.

[2] Suomi, K., Meister, E., Ylitalo, R. & Meister, L., "Durational patterns in Northern Estonian and Northern Finnish", Journal of Phonetics (in print).

[3] Boersma, P. & Weenink, D., "Praat: doing phonetics by computer" (Version 5.2.09) [Computer program]. Retrieved January 9, 2011, from http://www.praat.org/.

[4] Suomi, K., "Voicing in English and Finnish stops", Publications of the Department of Finnish and General Linguistics of the University of Turku 10, 1980.

[5] Helgason, P. & Ringen, C., "Voicing and aspiration in Swedish stops", Journal of Phonetics, 36, 607–628, 2008.

[6] Maddieson, I., "Voicing and gaps in plosive systems", in M. Dryer and M. Haspelmath [Eds], The World Atlas of Language Structures Online. Munich: Max Planck Digital Library, chapter 5, 2011.

[7] Ringen, C. & Suomi, K., "The voicing contrast in Fenno-Swedish stops", Journal of Phonetics, 40, 419–429, 2012.

Emotions and speech temporal structure

Kairi Tamuri, Meelis Mihkla

The Institute of the Estonian Language, Estonia
kairi.tamuri@eki.ee, meelis.mihkla@eki.ee

Abstract

The focus of the article is on whether emotions could be traced in the temporal structure of Estonian speech. There are two research questions, namely, (a) Do emotions affect speech rate? and (b) What detectable traces, if any, might emotions generate in speech prosody? To answer question (a), the articulation rate of emotional utterances was measured and the results were compared with those on neutral speech. The difference revealed was statistically significant. To answer question (b), the relations between emotions and the temporal characteristics of words with a vowel-centered structure were investigated. Sound durations were measured, various durational relations were computed and various combinations of the characteristics where subjected to statistical analysis. The results revealed a certain difference between the temporal characteristics of Q2 and Q3 feet, and a loss of the difference between the second and third quantity degrees in sad speech.

Index Terms: emotional speech, speech rate, word prosody, quantity degrees, Estonian

1. Introduction

The Estonian Emotional Speech Corpus is currently used to study the acoustic characteristics of *basic emotions* – anger, sadness and joy – and of neutral speech. The aim is to ascertain the definitive and distinctive parameters enabling recognition of emotions in Estonian speech as well as their identifiable synthesis.

According to literature, every emotion has its own acoustic parameters, which distinguish the emotion from the others of its kind and from neutral speech. In many parts of the world, emotion acoustics have yielded commendable results, including the main acoustic characteristics of emotions in speech. These are: fundamental frequency and intonation, intensity, formants and articulation precision, pauses and speech rate (the list is not closed) [1], [2], [3], [4], [5], [6]. The results of the acoustic analysis of Estonian emotional speech, concerning pauses [7], as well as formants and articulation precision [8], also confirm that emotions do affect those parameters.

Our studies have shown, for example, that variation in speech rate may, inter alia, be eloquent of the speaker's state of emotion. Thus, a very slow rate may indicate that the speaker is sad or depressed. It should be kept in mind, however, that speech rate is a rather subjective characteristic, which depends not only on the state of emotion but also on the speaker's gender, age, speech style, language, cultural space, communicative situation etc. [9].

Speech rate is measured in speech segments per unit time (e.g. speech sounds per second), either with pauses included or excluded [10]. In the present study a speaker's rate of articulation is measured on speech material where pauses have been deleted. According to literature, certain emotions (e.g. joy, anger, fear) are articulated more rapidly than neutral speech, whereas some others (e.g. sadness, disgust) are associated with slower than neutral articulation [1], [10], [3], [5], [6]. Although the available studies of emotional speech have used different speech material (acted speech vs. speech with induced emotion vs. spontaneous speech; sound vs. sound + image) from different languages and cultures (as is known, emotional expression differs across cultures and emotions can only be identified from sound within a culture, see, e.g. [11], [12], [13], there has been practically no variation in speech rate results, notably, in sad utterances the speech rate is lower than neutral, whereas in angry and joyous utterances it is typically higher than neutral [12], [14], [15], [13], [16]. Presumably, for these three basic emotions, the speech rate tends to follow a universal pattern.

Estonian is a word-central language. It is in word prosody that quantity degrees are manifested, which certainly belong to the pivotal phenomena of Estonian phonetics. Notably, in Estonian the phonologically significant three-way opposition between the Q1, Q2, and Q3 quantity degrees is manifested in the foot [17]. On the acoustic level, the distinctive features of those three quantity degrees include the duration ratio of the rhyme of the stressed syllable and the nucleus of the unstressed syllable[1], and the F0 contour [18], which together form a mutually complementary system. The temporal parameter is the duration ratio (V1:V2) of the stressed and unstressed syllables in the word; this ratio is hitherto the most stable parameter distinguishing between the quantity degrees. The numerous different experiments from more than a half century have established the following general V1:V2 ratios for the Estonian quantity degrees: 2:3 for the first degree (Q1), 3:2 for the second (Q2), and 2:1 for the third quantity degree (Q3) [19], [20], [21], [22], [23], [24]. As, according to previous research [25], the duration ratio of the stressed and unstressed syllables covers three fourths of data variation this ratio was pklaced in the focus of the present study as well. Besides the parameters mentioned earlier it has also been suggested that quantity degrees could be determined from the durational ratios of adjacent speech sounds [26], using perception and weighting of durational differences. Up to now, emotional speech has never been studied from the point of view of temporal characteristics of word prosody.

The present study investigates the possible connections between the temporal characteristics of uttered words with a CV[V]CV structure and emotions, considering the words stressedness, phrase position and part of speech. Statistical methods (logistic regression and CART) are used to study combinations of different parameters, possibly enabling to detect small, covert, yet essential connections between the input parameters and emotions [27].

[1] The quantity degree of a foot is defined as follows: $\sigma_{\text{stressede}}$(nucleus+[coda]) / $\sigma_{\text{unstressed}}$(nucleus).

2. Hypotheses and research material

Hypothesis One states that emotions affect Estonian speech rate. Estonian being a word-central language hypothesis Two concerns word prosody: emotions affect word temporal structure, while the influence can be detected in temporal parameters and in the duration ratio V1:V2 of the stressed and unstressed syllables, which is the main distinctive feature of quantity degrees. In addition it is investigated in precisely which temporal parameters of word prosody the possible rate specifics of emotional speech is manifested.

The acoustic base of the study consists of the Estonian Emotional Speech Corpus (EESC[2]) of the Institute of the Estonian Language. The corpus has been generated on the principle that emotions can sufficiently well be identified from natural, non-acted speech and that natural speech synthesis should be based on non-acted speech [5]. The corpus contains read sentences of anger, joy and sadness, and neutral speech. Those basic emotions also cover the following emotions: anger = displeasure, irony, dislike, contempt, schadenfreude, rage; joy = gratitude, happiness, pleasure, enthusiasm; sadness = lonelyness, disconsolateness, uneasiness, hopelessness. 'Neutral' means 'without particular emotions'. The corpus items (text paragraphs) have been selected so that their content is likely to excite a state of emotion in the reader. Therefore the reader has not been prompted as to with what emotion the paragraph should be read. The corpus contains paragraphs of journalistic texts read by a female voice, which have been segmented into sentences, words and speech sounds. The emotional colouring (anger, joy, sadness) or neutrality of the corpus sentences has been found by using perception tests, see [28].

To study the speech rate, at least three-word emotional (joyous, sad, angry) or neutral sentences were used, the emotionality or neutrality of which had been confirmed by more than 50% of perception test listeners. To study speech prosody, words of all three quantity degrees with a CV[V]CV structure were picked from the corpus (see Table 1).

Table 1. *Material for speech rate and speech prosody investigation*

Emotion	No. of sentences	CV[V]CV words		
		Q1	Q2	Q3
joy	55	45	28	20
sadness	84	62	42	31
anger	77	83	40	24
neutral	98	111	35	37
TOTAL	314	301	145	112

3. Method

To find out whether emotions actually affect speech rate, the articulation rate of emotional vs. neutral speech was measured. EESC contains journalistic texts, where all emotional sentences are different, therefore speech sounds per second was considered the most adequate unit of measurement. As from the phonological point of view, long speech sounds represent sequences of two short phonemes [29] long sounds were counted as two sounds for speech rate calculations. Emotion results were compared pairwise and with neutral speech. In addition it was investigated whether variation of emotionality is accompanied by speech rate variation within an utterance. For that purpose separate speech rate measurements were conducted on phrase-final word and non-phrase-final words[3].

To find out whether emotions cause changes in prosody, correlations between the temporal parameters of the vowel-centered word structure CV[V]CV and emotions were modelled, considering the stressedness, phrase position and part of speech of the word. Spech sound durations were measured and the duration ratios of V1/V2, V1/C1 and V2/C2 were computed. Logistic regression and the CART method were used to prove the significance of the effect caused by emotion characteristics in word temporal structure. The corpus has been tagged in the Praat environment. The measurements were analysed and modelled using the SYSTAT12 package.

Klaus R. Scherer [30] has elaborated a model that predicts the effect of emotion on voice expression. Scherer's *component process model* (CPM) considers the psychological and physiological factors associated with emotional expression and demonstrates that there exist certain emotion-specific acoustic patterns. Scherer describes emotion as a series of *adaptive changes*, which are mutually related. Having received an emotional impulse the nervous system affects the speaker's breathing as well as the muscular tension of the speech organs, which all causes changes in the acoustics of the speech signal. The CPM also predicts what difference emotions will probably cause in the speech rate as compared to normal speech. For a comparison of the results of the present study with those of the CPM and some earlier studies see Chapter 5.

4. Results

4.1. Speech rate

The results reveal that emotions do affect the overall speech rate and that of the non-phrase-final words, while the differences across emotions as well as between emotional and neutral speech are statistically significant (see Tables 2 and 3). The rate of pronunciation of the final word of the phrase has no differentiating power, neither between emotions nor between emotional and neutral speech.

Table 2 shows that the overall speech rate is the highest in anger utterances and the lowest in sadness utterances: anger (17.5 sound/s) > joy (17.1 sound/s) > neutral (16.9 sound/s) > sadness (16.6 sound/s). Table 3 indicates that for emotion pairs differences in overall speech rate are statistically significant. As for neutral speech its overall rate only differed significantly from anger utterances, whereas no significant difference was observed between the rates of neutral vs. joy or neutral vs. sadness.

To find out whether speech rate may, depending on changing emotions or emotionality, also vary within a single utterance, separate rate measurements were conducted on phrase-final word and non-phrase-final words.

As is demonstrated in Table 2 speech rate differences are the most salient in non-phrase-final words. Here, too, the rate is the highest in anger utterances and the lowest in sadness utterances: anger (18.44 sound/s) > joy (17.62 sound/s) > neutral (17.54 sound/s) > sadness (17.04 sound/s). The

[2] See http://peeter.eki.ee:5000/ (19.04.2012).

[3] As phrase-final lengthening also has an inevitable effect on word temporal structure the final word of a phrase were considered separately from the rest.

differences in the pronunciation rate of non-phrase-final words were statistically significant in emotion pairs as well as between emotional and neutral speech (see Table 3). Here the olny exception is joy, in which case the articulation rate of non-phrase-final words is not significantly different from that of neutral speech.

The average articulation rates of phrase-final word do not differ much (see Table 2), let alone significantly (see Table 3), across emotions.

Table 2. *Average pronunciation rates in sounds per second.*

Emotions	Overall	Non-phrase-final words	Phrase-final word
joy	17.1	17.6	14.0
sadness	16.6	17.0	14.4
anger	17.5	18.4	14.4
neutral	16.9	17.5	14.1

Table 3. *Results of ANOVA pairwise analysis of speech rate differences (statistically significant differences, p < 0,05, are highlighted by gray background).*

Emotions	p-value for speech rate differences		
	overall	non-phrase-final words	phrase-final word
joy vs. anger	0.010	0.043	0.452
joy vs. sadness	0.039	0.036	0.317
joy vs. neutral	0.557	0.810	0.985
sadness vs. anger	0.001	0.001	0.837
sadness vs. neutral	0.107	0.033	0.237
anger vs. neutral	0.031	0.008	0.370

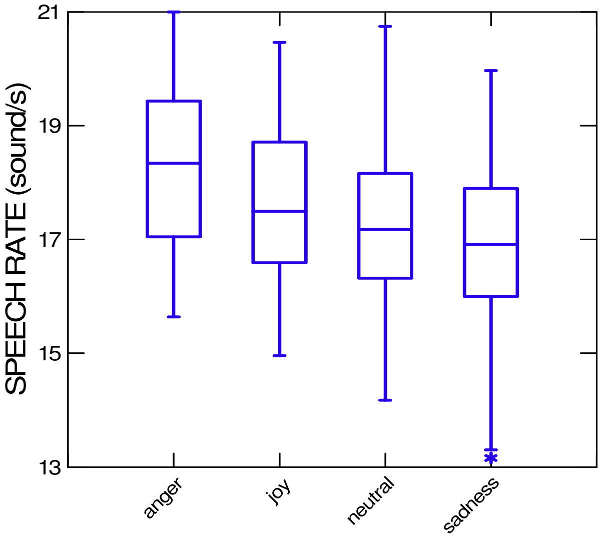


Figure 1: *Speech rate of non-phrase-final words.*

A closer look at the variation of emotional and neutral speech rates reveals that in non-phrase-final words the rate of neutral speech varies considerably less than that of emotional speech (see Figure 1).

In phrase-final word speech rate variation does not differ much across emotions (see Figure 2), which means that phrase-final lengthening is realized similarly for all emotions.

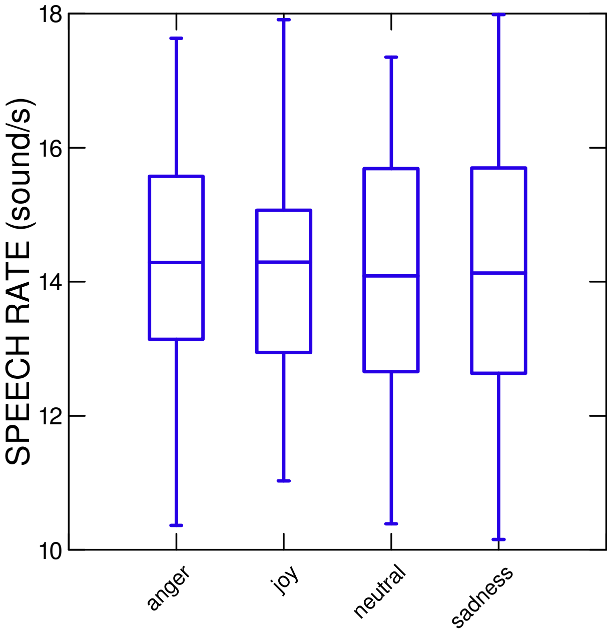


Figure 2: *Speech rate of phrase-final word.*

4.2. Word prosody

Table 4 contains the mean values of the temporal parameters of nearly six hundred words with a vowel-centered structure C1V1[:]C2V2 as distributed across emotions and quantity degrees. This table also contains the most relevant parameters of the duration model: V1, C2, V2 and V1:V2. Experimental attempts of modelling emotional speech have shown that like in neutral speech [25], the duration ratios of adjacent speech sounds, V1:C1 and V2:C2, are considerably less important than the classical ratio of V1:V2. Table 4 uses boldface to highlight those parameters whose averages displayed significant statistical differences across emotions (p < .05).

For the first degree, Q1, the temporal parameters were quite similar, with no significant differences between emotions or emotional vs. neutral speech. More salient differences between emotional and neutral speech could be observed in the Q2 foot. In emotional speech, the duration of the vowel of the stressed syllable, V1, was significantly shorter than in neutral speech (see Figure 3).

Table 4. *Temporal parameters of emotional and neutral speech in words with a CV[:]CV structure across three quantity degrees (Q1, Q2, Q3).*

Emotion	Q1				Q2				Q3			
	V1	C2	V2	V1:V2	**V1**	C2	V2	V1:V2	V1	C2	**V2**	V1:V2
joy	66	53	90	0.82	**126**	47	78	1.73	141	53	**67**	2.38
sadness	63	53	81	0.83	**132**	51	70	2.11	154	63	**80**	2.20
anger	69	48	91	0.82	**124**	49	71	1.80	153	61	**64**	2.76
neutral	65	52	87	0.80	**145**	54	78	1.94	150	58	**63**	2.64

However, there were no considerable differences in the durations of the consonant, C2, and vowel, V2, of the unstressed syllable.

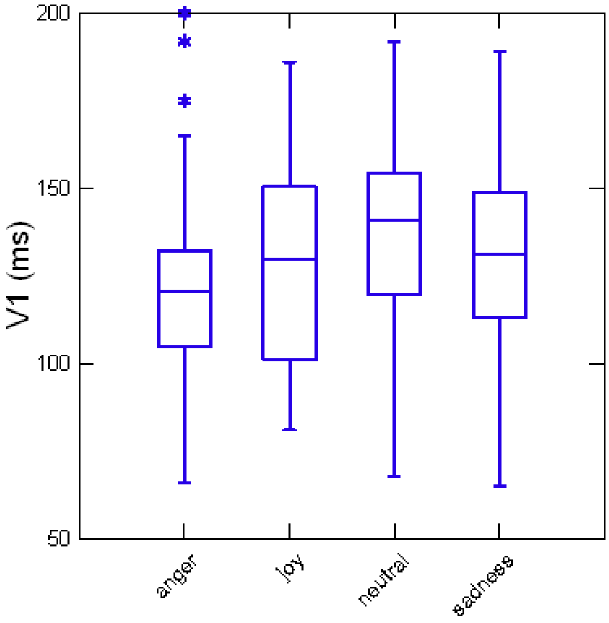


Figure 3: *Distributions of the durations (in ms) of the stressed syllable vowel V1 in words with a Q2 foot for different emotions and neutral speech.*

Although there is a noticeable difference between the average duration ratios (V1:V2) of the vowels of the stressed and unstressed syllables in the words with a Q2 foot when uttered with different emotions, those differences between the mean values are not significant statistically (p>.05). Evidently this is due to the behaviour of the unstressed syllable vowel V2, which differs from that of V1. The highest value of V1/V2 is observed in Q2 words of sad speech. This ratio (2.11) is rather more like the third quantity degree Q3, which in sad speech equals 2.20. Thus, in case the vowel-centered words of the given material are pronounced sadly, their second and third quantity degrees converge, so that the three-way opposition is replaced by a dual one. In the words with a Q3 foot statistically significant differences can be observed between the mean values of the unstressed syllable vowels, V2, if articulated in emotional speech (see Figure 4). For joy, anger and neutral speech, the durations of the unstressed syllable vowel V2 are relatively similar (67, 64 and 63 ms, respectively). A notable V2 lengthening is observed in the Q3 foot in the case of sad speech.

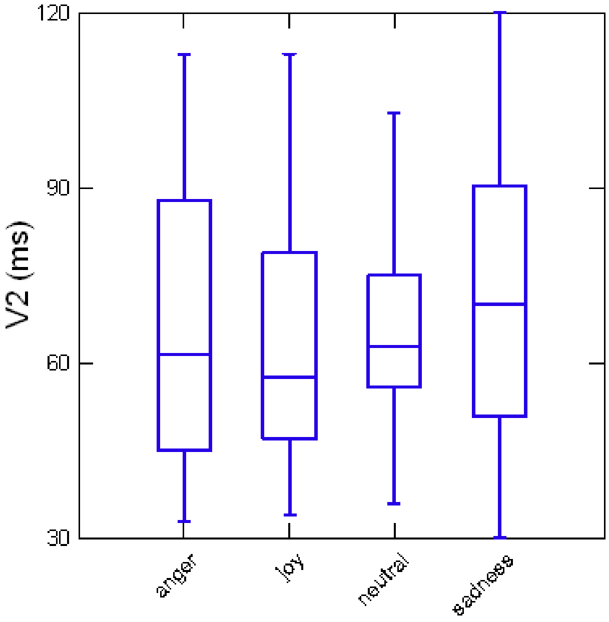


Figure 4: *Distributions of the durations (in ms) of the unstressed syllable vowel V2 in words with a Q3 foot for different emotions and neutral speech.*

According to the speech rate study previously discussed, sad speech was the slowest. What exactly is the role of V2 stretchingin the speech rate decrease of sad speech is pending future studies focused on words with a consonant-centered structure. An analogous situation occurs in the case of anger, where the speech rate is the highest and, in a Q2 foot in words of a vowel-centered structure, the unstressed syllable vowel is the shortest. Again, the significance of that local change cannot be judged from the vowel-centered structure.

5. Discussion

According to the prediction of CPM, anger makes the speech rate rise as compared to neutral speech. This is confirmed by our results, where the speech rate in anger utterances is indeed higher than in the neutral ones (17.5 sound/s *vs.* 16.9 sound/s), while the difference is statistically significant (p = 0.008). In the case of anger, rise in the speech rate has also been observed in some earlier studies, e.g. [1], [5], [31], [6].

For sadness the CPM predicts a fall in the speech rate as compared with neutral speech. According to our results sadness alos brings about a lower than neutral speech rate (16.6 sound/s *vs.* 16.9 sound/s), but the difference is not statistically significant (p = 0.107). Again, a lower speech rate has been observed in sad speech by many other researchers of emotional speech acoustics, e.g. [1], [3], [6].

For happiness, the CPM prediction reads that the speech rate should be lower than neutral, whereas elation should raise it higher than neutral. The present study, however, does not treat happiness, joy and elation separately and the category of 'joy' covers happiness and elation as well (see Ch. 2). According to our results the joy utterances are produced more rapidly than the neutral ones (17.1 sound/s *vs.* 16.9 sound/s), but the diffrence is not significant statistically (p = 0.557). Again, a higher speech rate has also been observed in joy utterances by several earlier researches, e.g. [1], [5], [6].

As Estonian is unique for its three-way opposition of quantity degrees (Q1, Q2, Q3) it is complicated to achieve a word prosodic comparison with other languages and with universal language models. The classical duration ratios [21], turned out to hold for emotional speech as well, except that in sad speech the duration ratio V1:V2 of the stressed and unstressed syllables was almost similar for the second and third quantity degrees, due to which the three-way opposition gave way to a dual one. There was no statistically significant correlation between emotions and V1:V2, which is the main distinctive feature of Estonian phonetic quantity degrees. The general durational model of the Estonian speech sounds [32], does not include emotion as an argument feature. In the words with a vowel-centered structure emotion was only significant for the stressed syllable vowel V1 in Q2 foot words and for the unstressed syllable vowel V2 in Q3 foot words. Future experiments should deal with the possible effect of emotions on words with a consonant-centered structure CVC[C]V and on monosyllables, with a view of an overall durational model of emotional speech.

The results of the present study once again confirm that speech rate tends to follow a universal pattern, notably, it is higher in utterances of anger and joy, whereas sadness makes it fall. Of the three emotions analysed, anger is the farthest from neutral speech; the difference is statistically significant.

6. Conclusion

The present analysis of read Estonian emotional utterances has proved that emotions do affect the speech rate. According to our measurements, the overall speech rate was higher than neutral in utterances of joy and anger, whereas sadness made it drop lower than neutral. However, the difference was statistically significant only between anger and neutrality. As for emotion pairs the speech rates were always different, and the difference was statistically significant.

Separate analyses were carried out on phrase-final word and non-phrase-final words. The differences were more salient in the latter, showing the highest rate for angry utterances and the lowest for sad ones. Almost all speech rate differences in emotion pairs as well as between emotions and neutral speech proved statistically significant, except for joy vs. neutral, which difference was not significant in non-phrase-final words. In phrase-final word the articulation rates did not substantially differ. Consequently the influence of emotions on speech rate is confined to non-phrase-final words.

The working hypothesis of the possible influence of emotions on the temporal characteristics of the words with a vowel-centered structure CV[V]CV and on the duration ratio of the stressed and unstressed syllables was but partly confirmed. Emotions did not have a significant influence on the main distinctive feature, V1:V2, of Estonian quantity degrees, however, the average durations of V1 in words with a Q2 foot differed across emotions, as well as the duration of the vowel V2 of the unstressed syllable in words with a Q3 foot. In sad speech, the parameters of the second and third quantity degrees converged, so that the typical three-way quantitative opposition was reduced to a dual one. But this was proved just for words with a vowel-centered structure. To reach a final conclusion about the influence of emotions on Estonian word prosody and the emotion-induced local changes in the speech rate our research should also cover the correlation of emotions with the temporal parameters of words with a consonant-centered CVC[C]V structure and of monosyllabic words.

7. Acknowledgements

The study was financially supported by the specifically funded research programme SF0050023s09 and the Estonin Science Foundation Grant No. ETF7998.

8. References

[1] Murray, I. R., Arnott, J. L., “Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech”, in Computer Speech and Language, 22 (2), 107-129, 2008.

[2] Toivanen, J., Waaramaa, T., Alku, P., Laukkanen, A. M., Seppänen, T., Väyrynen, E., Airas, M., “Emotions in [a]: a perceptual and acoustic study”, in Logopedics Phoniatrics Vocology, 31, 43-48, 2006.

[3] Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Deng, Z., Lee, S., Narayanan, S., Busso, C., “An acoustic study of emotions expressed in speech”, in INTERSPEECH-2004, 2193-2196, 2004.

[4] ten Bosch, L., “Emotion, speech and the ASR framework”, in Speech Communication, vol. 40, 213-225, 2003.

[5] Iida, A., Campbell, N., Higuchi, F., Yasumura, M., “A corpus-based speech synthesis system with emotion”, in Speech Communication, 40 (1-2), 161-187, 2003.

[6] Banse, R., Scherer, K. R., “Acoustic profiles in vocal emotion expression”, in Journal of Personality and Social Psychology, 70 (3), 614-636, 1996.

[7] Tamuri, K., “Kas pausid kannavad emotsiooni?”, in Eesti Rakenduslingvistika Ühingu Aastaraamat, 6, 297-306, 2010.

[8] Tamuri, K., “Kas formandid peegeldavad emotsiooni?”, in Eesti Rakenduslingvistika Ühingu Aastaraamat, 8, 231-243, 2012.

[9] Laver, J., “Principles of Phonetics”, Cambridge University Press, Cambridge, pp 534, 1994.

[10] Braun, A., Oba, R., “Speaking Tempo in Emotional Speech – a Cross-Cultural Study Using Dubbed Speech”, in ParaLing’07, 77-82, 2007.

[11] Altrov, R., Pajupuu, H., “Estonian Emotional Speech Corpus: Culture and Age in Selecting Corpus Testers”, in I. Skadiņa, A. Vasiļjevs [Eds], Human Language Technologies – The Baltic Perspective – Proceedings of the Fourth International Conference Baltic HLT 2010, Amsterdam: IOS Press, 25-32, 2010.

[12] Burkhardt, F., Audibert, N., Malatesta, L., Türk, O., Arslan, L. M., Auberge, V., “Emotional prosody – does culture make a difference?”, in Proceedings of Speech Prosody. Dresden, Germany, May 2-5, 2006.

[13] Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P., “Emotional speech: Towards a new generation of databases”, in Speech Communication, 40 (1-2), 33-60, 2003.

[14] McIntyre, G., Roland, G., “Researching emotions in speech”, in Proceedings of the 11th Australian International Conference on Speech Sciences & Technology, 264-269, 2006.

[15] Wilting, J., Krahmer, E., Swerts, M., “Real vs. acted emotional speech”, in Proceedings of Interspeech 2006 ICSLP, Pittsburgh, PA, USA, 805-808, 2006.

[16] Scherer, K. R,. “Vocal communication of emotion: A review of research paradigms”, in Speech Communication, 40 (1-2), 227-256, 2003.

[17] Lehiste, I. “Search for phonetic correlates in Estonian Prosody”, in Estonian Prosody: Papers from a Symposium, Proceedings of the International Symposium on Estonian Prosody, 11-35, 1997.

[18] Ross, J., Lehiste, I., “The temporal Structure of Estonian Runic Songs”, in Phonology and Phonetics **1**, 2001.

[19] Lehiste, I., “Segmental and Syllabic Quantity in Estonian”, in American Studies in Uralic Linguistics, vol 1. Bloomington: Indiana University, 21-28, 1960.

[20] Liiv, G., ”Eesti keele kolme vältusastme kestus ja meloodiatüübid”, in Keel ja Kirjandus, nr 7-8, 412-424, 480-490, 1961.

[21] Eek, A., Meister, E., “Simple Perception Experiments on Estonian Word Prosody: Foot Structure vs. Segmental Quantity”, in Estonian Prosody: Papers from a Symposium, Proceedings of the International Symposium on Estonian Prosody, 77-99, 1997.

[22] Krull, D. “Word-prosodic features in Estonian conversational speech: some preliminary results”, in PERILUS (Phonetic Experimental Research, Institute of Linguistics, University of Stockholm) XVII, 45-54, 1993.

[23] Lippus, P., Pajusalu, K., Allik, J., “The tonal component in perception of the Estonian Quantity”, in Proceedings of the 16th International Congress of Phonetic Sciences, 1049-1052, 2007.

[24] Kalvik, M-L., Mihkla, M., Kiissel, I., Hein, I., ”Estonian: Some findings for modelling speech rhythmicity and perception of speech rate”, in P. Sojka, A. Horak, I. Kopecek, K.Pala [Eds], Text, Speech and Dialogue. Berlin/Heidelberg: Springer Verlag, 314-321, 2010.

[25] Kalvik, M-L., Mihkla, M., ”Modelling the temporal structure of Estonian speech”, in I. Skadina, A. Vasiljevs [Eds], Human Language Technologies. The Baltic Perspective: Proc. 4th IC. Amsterdam: IOS Press, 53-60, 2010.

[26] Eek, A., Meister, E., “Foneetilisi katseid kvantiteedi alalt”, in Keel ja Kirjandus*,* nr 11-12, 815-837, 902-916, 2003.

[27] Sagisaka, Y., “Modeling and perception of temporal characteristics in speech”, in Proc. 15th ICPhS Barcelona, 1-6, 2003.

[28] Altrov, R., Pajupuu, H., “Estonian Emotional Speech Corpus: Content and options”, in G. Diani, J. Bamford, S. Cavalieri [Eds], Variation and Change in Spoken and Written Discourse : Perspectives from Corpus Linguistis, Amsterdam: John Benjamins, 2012. [forthcoming]

[29] Eek, A., “Eesti keele foneetika I”, TTÜ Kirjastus, 2008.

[30] Scherer, K. R., “Vocal affect expression: A review and a model for future research”, in Psychological Bulletin, 99 (2), 143-165, 1986.

[31] Juslin, P. N., Laukka, P., “Communication of emotions in vocal expression and music performance: Different channels, same code?”, in Psychological Bulletin, 129 (5), 770-814, 2003.

[32] Mihkla, M., “Modelling speech temporal structure for Estonian text-to-speech synthesis: feature selection”, in Trames : Journal of the Humanities and Social Sciences, 11(3), 284-298, 2007.

Microduration in Finnish and Estonian vowels revisited: methodological musings

Stefan Werner [1], *Einar Meister*[2]

[1] Department of Linguistics, University of Eastern Finland, Finland
[2] Institute of Cybernetics at Tallinn University of Technology, Estonia
stefan.werner@uef.fi, einar@ioc.ee

Abstract

The influence of vowel duration on the perception of different vowel qualities in Finnish and Estonian has been the topic of several of our recent studies. For the present paper, we reconsidered some of our methodological choices, analyzed result data in different ways and tried to establish the reliability of our test design.

Index Terms: Estonian, Finnish, vowels, intrinsic duration

1. Introduction

The studies on microprosody in several languages have established systematic differences in the intrinsic features of vowels – open vowels tend to have lower F0, higher intensity and longer duration than close vowels (e.g. [1], [2], [3], [4], [5]). Our recent studies address the intrinsic vowel duration in quantity languages like Estonian and Finnish, mainly focusing on the role of intrinsic duration on the perception of different phonological categories, i.e. vowel contrasts in close-open dimension and short vs. long durational oppositions. We have shown experimentally that in boundary conditions when spectral as primary features do not provide sufficient information for category discrimination in close-open vowel pairs, the intrinsic duration of vowels acts as a secondary feature facilitating the perceptual decision [6]. In a subsequent study we have found further evidence for the impact of intrinsic vowel duration by examining the categorical short vs. long distinction – the vowel quality (hence intrinsic duration of a vowel) plays a significant role in the discrimination of Estonian short vs. long phonological category [7]. The latter result is rather surprising since in quantity language like Estonian duration has to be intentionally controlled by a speaker to signal quantity contrasts and this “higher order” control can “override” the intrinsic features.

The aim of our current paper is to verify our previous findings on short vs. long category discrimination by different groups of subjects involving Estonian and Finnish subjects, and to address a number of methodological issues like different test setups, intra-subject variations in repeated experiments, different methods applied in the statistical analysis of the results.

2. Methods and data

2.1. Stimulus corpus

For the perception experiments a stimulus corpus involving short vs. long category oppositions in close vowel /i/ and open vowel /a/ in CV(:)CV carrier words was designed. The stimuli were created from the nonsense words /kaka/, /kiki/, /papa/, /pipi/, /tata/, and /titi/ pronounced in isolation by a native Estonian male speaker. In all words the duration of the stressed vowel (V1) was manipulated from 100 ms to 190 ms in 10 ms steps which consequently resulted in six stimulus sets from CVCV to CV:CV – /kaka/ vs. /ka:ka/, /papa/ vs. /pa:pa/, /tata/ vs. /ta:ta/, /kiki/ vs. /ki:ki/, /pipi/ vs. /pi:pi/, /titi/ vs. /ti:ti/. The durations of the other segments were kept constant (C1(burst) = 25 ms for /k/, 15 ms for /p/ and /t/; C2 = 75 ms; V2 = 240 ms); the F0 was set to a constant value of 100 Hz in both vowels. The number of different stimuli in all sets was 10. The manipulation of stimuli was done with Praat [8].

In Estonian, the stimulus sets constitute a continuum from a word in quantity one (Q1) to a word in quantity two (Q2) achieved by changing the duration of the first-syllable vowel.

2.2. Test variations

Factors whose potential influence we wanted to assess were:

- test setup: self-paced (individual) vs. timed (group) test
- test-retest intra-subject variation
- test evaluation: reaction times as additional support
- test evaluation: statistical modelling

To investigate the possible effect of imposing a time limit on the subjects we designed two slightly different group versions of our quantity perception tests, one of which was administered to Estonian subjects, the other to Finnish subjects. The Estonian version contained the full set of different stimuli from the individual tests – two vowel qualities and three consonant articulation places – with each stimulus played three times and an inter-stimulus interval of five seconds, whereas the Finnish version only used three of the six stimuli (/kaka/, /kiki/, /papa/) but played every stimulus five times using an inter-stimulus interval of three seconds. The Estonian group involved 40 subjects whereas 30 subjects where native speakers of Estonian (EST-L1) and 10 non-native subjects with Russian-language background (EST-L2); the Finnish group involved 17 native speakers (FIN-L1).

For both native groups short vs. long category discrimination is natural since both Estonian and Finnish exploit the duration cue contrastively; also L2 subjects are able to discriminate Estonian short and long contrasts despite non-categorical role of duration in Russian [9].

In order to check for test-retest variation, one native Finnish and two native Estonian subjects underwent the same test several times; for the Finnish subject, reaction times were now also recorded. Instead of a linear regression analysis of response frequency in terms of duration (as in [6]), we fitted more complete binomial logistic regression models, with and without random effects.

3. Results

3.1. Short vs. long boundaries

Overall group test results are in line with our previous studies: vowel openness correlates (non-significantly) positively with stimulus duration in all subjects' groups (Figure 1). In EST-L1 group the boundary mean in the case of high vowel lies at 146.8 ms and in the case of low vowel 151.7 ms, in EST-L2 the categorical boundary values are slightly lower – 142 ms and 147.3 ms for high and low vowel, correspondingly. The boundary values for the Finnish group lie even at shorter vowel durations – at 135.9 ms in the case of high vowel and at 144.5 ms in the case of low vowel. The difference between the two means is significant in two native groups, in EST-L1 group at the 0.001 level (Welch two-sided t-test, t = 3.9; df = 174; p < 0.001) and in FIN-L1 group at the 0.01 level (t = 2.7; df = 35; p < 0.01); in the EST-L2 group the difference in category mean values between low and high vowel turned out to be insignificant (t = 1.4; df = 35.9; p = 0.17).

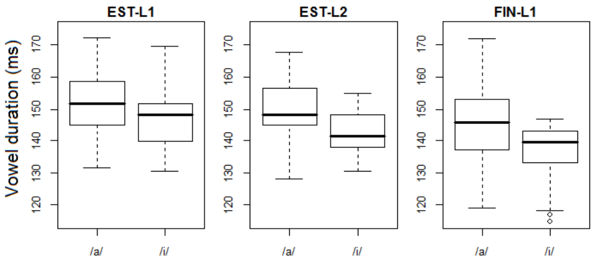


Figure 1. *Boxplots of the distributions of Estonian and Finnish speakers' category boundaries between "long" and "short" in low and high vowels.*

The variation of responses broken down by stimulus duration is shown in Figure 2 (the difference in frequency range is due to the higher number of observations per duration and subject in the Estonian test). The area of indecision around the category boundary seems to spread out slightly more in the Estonian data. But this can be due to, at least partly, the greater variation of segmental contexts in the Estonian stimulus material and the larger number of Estonian test subjects (here EST-L1 and EST-L2 groups are pooled together).

3.2. Test setup

Test setup does not seem to have a systematic influence on the perception test results in our one-subject Finnish case study. The subject's category boundary ranged from 148.4 ms to 133.7 ms in the four identical individual tests which were self-paced, and was 142.3 ms in the time-controlled group test.

3.3. Test-retest variation

As illustrated in the mosaic plots of Figures 3 and 4 intra-subject test-retest variation turns out to be moderately high for our Finnish case who went through the test six time (see previous section), but minimal for the two Estonian two-test cases: category boundaries are at 151.7 and 150.0 for subject AK and at 144.7 in both tests for subject MK. All in all, the category boundaries between long and short are not affected in a way that would challenge our overall results for both language and intra-subject variation between tests does not exceed within-test intra-subject variation for repeated stimuli. Even in the Finnish case, median durations for long vs. short only fluctuate between adjacent conditions: 170 vs. 160 ms and 120 vs. 110 ms for long and short responses, respectively.

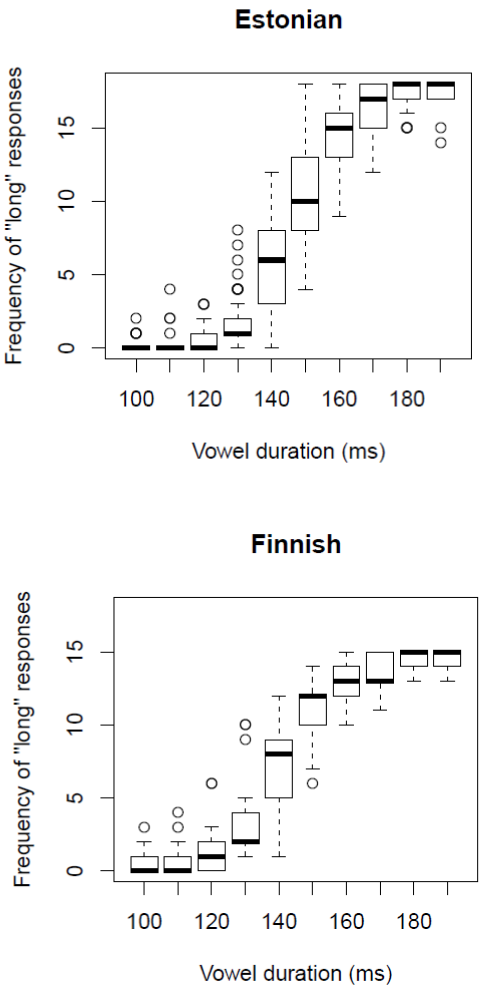


Figure 2*: Boxplots of the proportions of Estonian and Finnish speakers' "long" responses across stimulus durations. The whiskers extend to 1.5 times the interquartile range, indicating a 95% confidence interval for the difference in medians.*

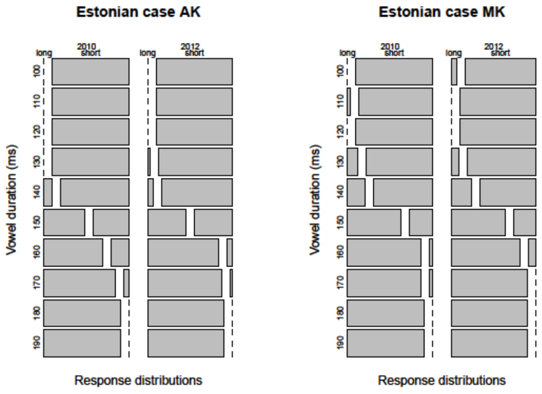


Figure 3. *Test-retest comparison of response distributions for two Estonian subjects.*

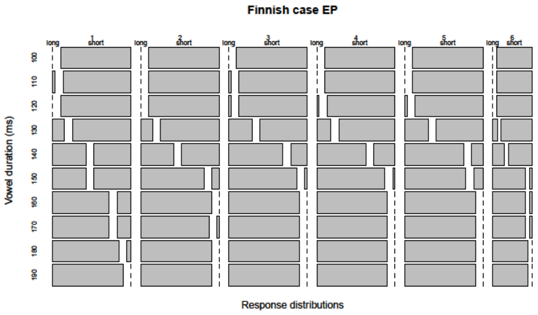


Figure 4. *Test-retest comparison of response distributions for Finnish subject.*

3.4. Reaction time

In two of the four self-paced tests of Finnish subject EP reaction times were measured. As can be seen from Figure 5, there seems to be a slight trend for reaction time to increase towards the category boundary which lies at 135.5 ms for these two tests. There is a weak but significant negative correlation (r=-0.16, p<0.001) between the squared distance of stimuli's' duration from the category boundary and reaction time. If a similar trend could be observed in other subjects as well it would lend additional support to our estimation of the category boundaries.

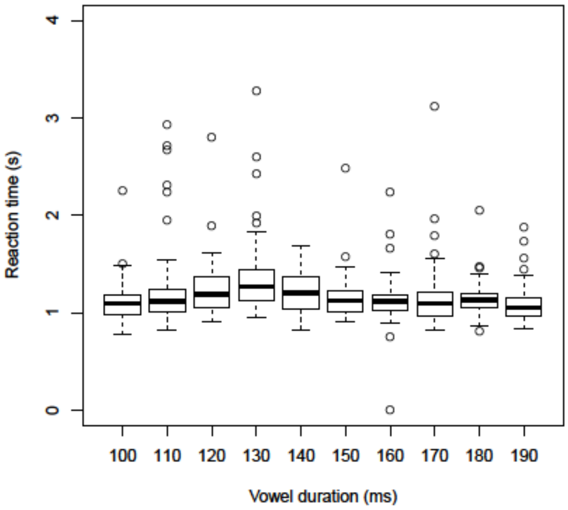


Figure 5. *Boxplots of Finnish subject EP's reaction times (from two tests). The whiskers extend to 1.5 times the interquartile range, indicating a 95% confidence interval for the difference in medians. (The zero value at 160 ms must be due to an inadvertent keypress.)*

3.5. Statistical models

We fitted binomial logistic regression models using R's glm() and glmer() functions. The mixed model analyses adding subject, stimulus, and/or presentation order as random effects did not produce results that significantly differed from the fixed-effects-only model: the only relevant factor, in addition to stimulus duration, is the consonantal context. Vowel openness, although affecting the categorical boundary in a minimal model with duration as the only factor (146 ms vs. 151 ms and 136 ms vs. 144 ms for high vs. low in Estonian and Finnish speakers, respectively), does not improve the model fit significantly. Subjects' sex and age reduced model deviance even less. Table 1 shows as an example the deviance analysis of a three-factor model for the Finnish group data.

Table 1. *Analysis of deviance table for a binomial logistic regression model (logit link function) of duration perception in the Finnish group test with factors duration, consonant place of articulation and vowel openness.*

	Df	Deviance	Resid.	Resid.	Pr(Chi)
			Df	Dev.	
NULL			2549	3523.4	
dur	1	1733.08	2548	1790.4	2e-16 ***
cons	1	108.41	2547	1682.0	2e-16 ***
vow	1	1.47	2546	1680.5	0.2247

4. Discussion

Our new tests with Estonian and Finnish subjects lend further support to our previous findings on the connection between intrinsic vowel duration and perceptual vowel categorization in quantity languages. Our case study of reaction time measurements in addition to perceptual ratings also shows the same trend.

On the basis of our one-subject case study it seems that the influence of variations in the test set-up can be neglected but more data will be needed to prove this point. Finally, more sophisticated statistical analyses with mixed models instead of fixed-factors-only models do not introduce new insights into our data.

All in all, the collection of results from new data and reconsiderations of methodological solutions presented here consolidates the concept of micro- and macroduration interplay developed already in our earlier studies.

Acknowledgement

This work has been partly supported by the target-financed theme No. 0140007s12 of the Estonian Ministry of Education and Research. We are grateful to Lya Meister for conducting the perception experiments with Estonian listeners and to all subjects who participated in the study.

References

[1] Peterson, G. E., Lehiste, I., "Duration of syllable nuclei in English", Journal of the Acoustical Society of America, 32(6):693–703, 1960.

[2] Solé, M.J., "Controlled and mechanical properties in speech: a review of the literature", in M.J. Solé, P. Beddor, M. Ohala [Eds], Experimental Approaches to Phonology, Oxford: Oxford University Press, 302–321, 2007.

[3] Di Cristo, A., "De la microprosodie à l'intonosyntaxe", Thèse d'Etat, Université de Provence, Aix-en-Provence, 1978.

[4] Wahlen, D. H., Levitt, A. G., "The universality of intrinsic F_0 of vowels", Journal of Phonetics 23, 349–366, 1995.

[5] Meister, E., Werner, S., “Intrinsic microprosodic variations in Estonian and Finnish: acoustic analysis”, in R. Aulanko, L. Wahlberg, M. Vainio [Eds], Fonetiikan Päivät 2006 = The Phonetics Symposium 2006, Publications of the Department of Speech Sciences, University of Helsinki, 53:103–112, 2006.

[6] Meister, E., Werner, S., “Duration affects vowel perception in Estonian and Finnish”, Linguistica Uralica, 45(3), 161–177, 2009.

[7] Meister, E., Werner, S., Meister, L., “Short vs. long category perception affected by vowel quality”, in W.-S. Lee and E. Zee [Eds], ICPhS XVII: [Proceedings of] the 17th International Congress of Phonetic Sciences, August 17-21, 2011, Hong Kong: City University of Hong Kong, 1362–1365, 2011.

[8] Boersma, P., Weenink, D., “Praat: doing phonetics by computer” (Version 5.3) [Computer program], Retrieved October 15, 2011, http://www.praat.org/

[9] Meister, L., Meister, E., “Perception of the short vs. long phonological category in Estonian by native and non-native listeners”, Journal of Phonetics, 39(2), 212–224, 2011.