ON TEMPORAL ORGANIZATION OF SPONTANEOUS ESTONIAN: PRELIMINARY ANALYSIS RESULTS OF LECTURE SPEECH

Einar Meister[1], Pärtel Lippus[2]

[1] Institute of Cybernetics
at Tallinn University of Technology, Estonia
[2] Department of Estonian and Finno-Ugric Linguistics
University of Tartu, Estonia

Abstract

In the paper some first results of the analysis of the temporal structure of spontaneous lecture speech in Estonian are presented. The analysis is based on the recording of a 45-minutes academic conference presentation given by a male speaker. The speech flow has been divided into major prosodic units – topics, breath groups and prosodic utterances, and the temporal characteristics of the units as well as corresponding pauses have been measured. As a result, the systematic duration patterns at each prosodic level have been found and the relationships between duration of prosodic units and pauses have been discussed.

Keywords: lecture speech, temporal units, breath groups, prosodic groups.

1. Introduction

Progress in speech recognition technology has lead to new challenging applications like recognition and understanding of spontaneous speech. When read speech is recognized with accuracy higher than 95%, the results are much worse in the case of spontaneous speech. This is basically due to substantial acoustic and linguistic differences of read and spontaneous speech styles as well as due to the fact that acoustic and language models are mainly built up using written language and read speech corpora. Two approaches have been used to increase the performance of spontaneous speech recognition: (1) acoustic and language models are build on spontaneous speech corpora, and (2) implementation of knowledge about the structure of spontaneous speech including prosodic and lexical features as well as phonetic-acoustic variability (Furui 2005).

During recent years a great progress has been achieved in Estonian large vocabulary read speech recognition (Alumäe 2005, 2006) and the further research is directed towards extending the existing language-specific methods for spontaneous Estonian. In order to approach this challenging task the recordings of spontaneous speech material and the study of its diverse features have been initiated; the temporal organization of spontaneous monologue (lecture) speech is addressed first. The ultimate goal of the on-going study is to provide the models of different units of spontaneous speech and characterize their prosodic and lexical features that finally should contribute to better recognition of spontaneous speech flow.

As shown in numerous studies the temporal organization of spontaneous speech differs in a great extent from read speech. Hesitations, repeating words, self-repairs, fillers, incomplete utterances, lengthened speech units, different types of pauses, etc. are the most common disfluencies characteristic to spontaneous conversation. Most disfluencies in spontaneous speech reflect the problems in speech planning process (Clark, Wasow 1998).

Temporal features of read speech have been intensively studied in many languages; less attention has been paid to spontaneous speech. In recent years a series of studies have been carried out on temporal structure of Estonian read speech, mainly with a goal to improve the temporal modeling in text-to-speech synthesis (Mihkla 2005, 2006, 2007); pausing in read speech is addressed in (Kerge et al. 2007). Spontaneous discourse in Estonian has been studied from the point of view of conversation analysis (e.g. Hennoste et al. 2005; Koit et al. 2006); a few studies have been addressed to prosodic features in spontaneous Estonian (Krull 1993, 1997). However, there is still very limited knowledge (especially applicable for speech recognition) available about the prosodic characteristics of spontaneous Estonian, including temporal organization.

The current paper introduces the first results of the analysis of temporal structure of spontaneous lecture speech and presents a bird's-eye view of the general structure and characteristics of major temporal units.

2. Method and material

2.1. Methodological issues

The theoretical framework for the analysis of spontaneous speech has been provided and developed mainly for studies of conversation analysis (e.g. Selkirk 1984; Chafe 1994). The widely used concept – intonation unit (IU) – is associated mainly with a characteristic tonal contour and other prosodic cues are rather of secondary importance. A different concept – prosodic unit (PU) – has been introduced for the analysis of spontaneous Mandarin conversation (Liu et al. 2006). PU is defined as "a perceptually coherent prosodic constituent" involving in addition to coherent tonal contour type different perceptually relevant prosodic cues, like pitch reset, final syllable lengthening, breaks, pauses and paralinguistic features. It is expected that prosodic units will be more appropriate for automatic chunking of spontaneous conversation than intonation units.

In our analysis of Estonian material we adopt the term of PU (or prosodic group PG) as it engages more acoustic-phonetic cues that can be involved later in the automatic analysis. Another reason to prefer the term PG relates to the fact that in the case of Estonian tonal contours do not play as significant communicative role as in English, for example.

Breathing is inevitable part of speaking. Breath pauses are directly associated with speech planning and are assumed to occur at the boundaries of speech production unit which do not need to coincide with syntactic boundaries – according to Winkworth et al. (1995) about 68% of inhalations in spontaneous speech were located at grammatical boundaries; they conclude that a "breath group" might reflect a single "unit of meaning" rather than a single syntactic structure. Horne et al. (2006) claims that the automatic recognition of inhalations will provide a way of chunking of speech flow into meaningful information units.

2.2. Speech material

The speech material of the current study has been recorded during a public presentation at the academic linguistics conference. Although the talk was based on previously prepared PowerPoint-slides the speech style of the speaker can be still characterized as spontaneous monologue. The speaker is native Estonian male, in the age of 48, with standard Estonian pronunciation.

The speech recording was carried out using a digital recorder Edirol R1 (sampling frequency 44,1 kHz, 16 bits, mono) via close-talking microphone AKG C 444L with preamplifier AKG B29L. The microphone position was constant (ca 5 cm from speaker's lips) during the whole presentation; the level of the input signal was adjusted before the talk; no automatic level control was used. The duration of the presentation was about 45 minutes.

2.3. Segmentation levels

The speech signal has been segmented on different hierarchic levels: phonemes, syllables, words, prosodic units, breath-groups, and topics; pauses at different levels were distinguished: (1) pauses between topics, (2) pauses between breath groups, (3) pauses between prosodic groups.

For phoneme labels Estonian SAMPA symbols (Eek, Meister 1999) were used, on the word level orthographic transcription was implemented (see Figure 1).

The segmentation and labeling has been carried out manually using Praat-program (Boersma, Weenink 2007).

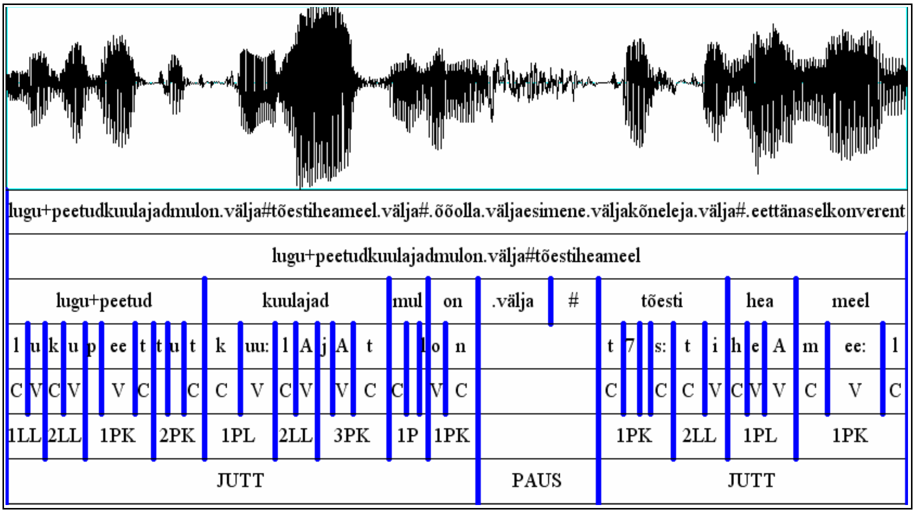
lugu+peetudkuulajadmulon.välja#tõestiheameel.välja#.õõolla.väljaesimene.väljakõneleja.välja#.eettänaselkonverent
lugu+peetudkuulajadmulon.välja#tõestiheameel
lugu+peetud
kuulajad
mul
on
.välja
#
tõesti
hea
meel
l u k u p ee t t u t k uu: l A j A t l o n t 7 s: t i h e A m ee: l
C V C V V C C C V C V V C C V C C C C V C V V C V C
1LL 2LL 1PK 2PK 1PL 2LL 3PK 1P 1PK 1PK 2LL 1PL 1PK
JUTT
PAUS
JUTT

Fig. 1. An example of segmentation

3. Units and patterns of temporal organization

For the analysis of temporal structure three major units – topics, breath groups (BG) and prosodic group (PG) – have been defined. Also the pauses between these units – topic pause, breath group pause (BGP) and prosodic group pause (PGP) were identified and measured.

3.1. Topics

A topic in the context of the current study is defined as the time interval corresponding to the presentation of a single PowerPoint-slide. As this definition is directly related to the specific type of lecture, its temporal characteristics are in great extent subject-specific and may not be actual in the case of different type of spontaneous speech. Nevertheless, bearing in mind the future automatic processing (recognition, summarization, etc) the description and modeling of topic's temporal patterns is relevant, too.

The slide show comprised 36 slides, consequently, the same number of topics was identified in the speech flow, as well. Duration of topics ranged from 13 to 210 seconds, median and average durations were ca 55 and 72 seconds, correspondingly.

Pauses between topics involved the slide change simultaneously with one or several inhalation-exhalation cycle(s). The average duration of inter-topic pauses was ca 3 seconds. The number of breath groups in a topic varied from 4 to 33, most of the topics involved 4 to 11 breath groups.

3.2. Breath groups

Each topic involves a diverse number of BGs which are defined as the time intervals between two successive inhalation pauses. In addition to the inhalation interval the BGPs involved typically silent interval(s) as well as filler(s). In total 430 BGs were identified and labeled.

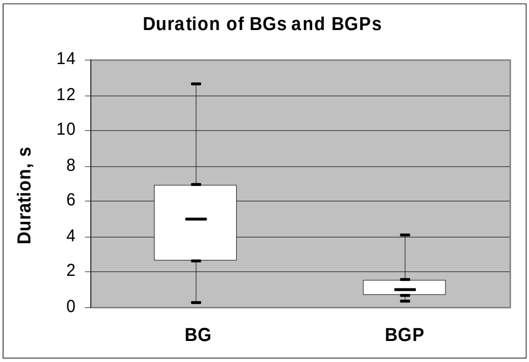

Figure 2. Duration of breath groups (BG) and BG-pauses (BGP)

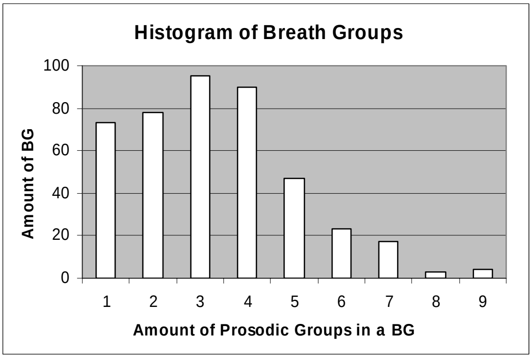

Figure 3. Histogram of breath groups (BG)

The box-plot of temporal characteristics of BGs (Figure 2) shows that 50% of them range from 2.6 s (1st quartile) to 6.9 s (3rd quartile), medium duration of BGs equals to 5 s; the shortest duration is 250 ms and the longest 12.6 s. The duration of BGP ranges from 290 ms to 4.1 s (1st quartile = 0.6 s, median = 1 s, 3rd quartile = 1.5 s).

The number of prosodic groups (PG) in a BG varies from 1 to 9, whereas BGs involving 1 to 4 PGs are dominating; the most frequent BGs have 3 PGs (Figure 3).

3.3. Prosodic groups

The prosodic groups were defined as time intervals within a BG separated by silent or/and filled pause. Often at PG-boundaries other prosodic cues (e.g. final lengthening, F0 reset) were perceptually perceivable; also the co-occurrence of typical fillers and/or particles in PG-initial/final positions was observed. Altogether 1422 prosodic units have been identified.

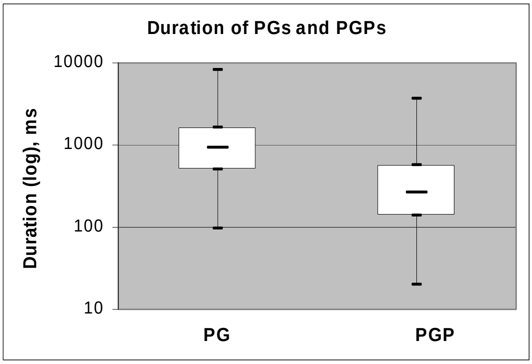

Fig. 4. Duration of prosodic groups (PG) and PG-pauses

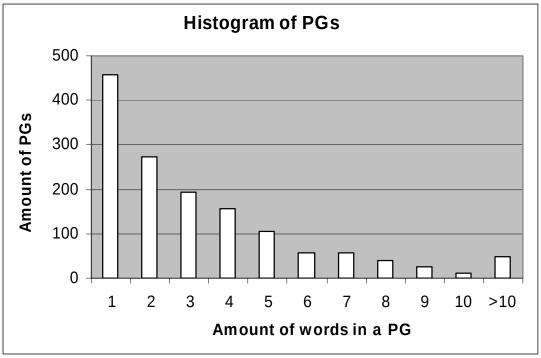

Fig. 5. Histogram of prosodic groups (PG).

Duration range of PGs (Figure 4) is rather large – from 0.1 s to 8 s, whereas half of the PGs fit into time interval 0.5 s (1st quartile) to 1.6 s (3rd quartile); median and mean durations of PGs are 0.9 s and 1.2 s, correspondingly. Pauses between PGs are typically filled or silent pauses varying from 20 ms to 3.6 s (1st quartile = 140 ms; 3rd quartile = 560 ms); median duration of PGPs is 265 ms and mean = 415 ms. Relatively large difference between median and mean values of PGP durations points to the skewed distribution.

Histogram of PGs (Figure 5) shows that one-word prosodic groups are most frequent; about 75% of PGs contain 1 to 4 words (median = 2 words, mean = 3 words).

3.4. A model of temporal structure

Based on the current analysis results we propose a hierarchical model of temporal organization of lecture speech. The temporal patterns of different units are based on the median values of measured unit durations.

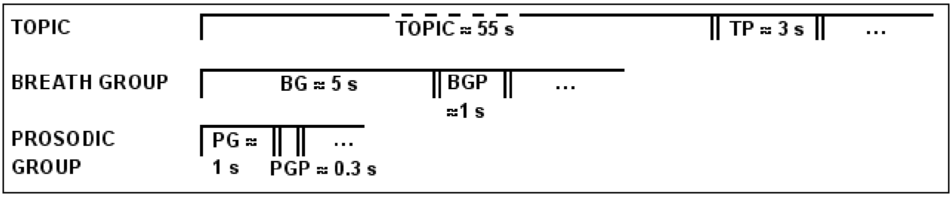

Fig. 6. A model of temporal structure of lecture speech

4. Discussion and further work

The preliminary analysis results of temporal organization of a spontaneous monologue speech have been introduced and the temporal patterns of three major units have been presented. Our findings are in line with the results of other similar studies (e.g. Swedish GROG-project) where speech chunks with 2-4 words have been found to be dominant in radio interviews (Strangert 2004) and duration of speech production units of 2-2.5 s has been observed (Horne et al. 2006) (it should be mentioned that differently from our analysis Horne's speech production units can contain internal pauses); timing constraints of speech processing around 2 s are found also in memory research (Baddeley 1997).

Although the current results give just a bird's-eye view to the sophisticated structure of spontaneous speech, they already provide valuable information for building models to be implemented in spontaneous speech processing. Automatic recognition of inhalations and PG-pauses will help to divide speech flow into different meaningful chunks which can be more effectively processed at the later stages of speech recognition.

The future work will focus on further analysis of acoustic characteristics of BGs and PGs and should involve diverse spontaneous speech data (different speakers, speaking environments, etc). Prosodic and lexical/semantic cues on different temporal units as well as characteristics of boundary-cues will be studied.

References

Alumäe, T. 2005. Using adaptive stochastic morphosyntactic language model for two-pass large vocabulary Estonian speech recognition. In: Kokkinakis,G. et al. (eds.) *Proc. of 10th International Conference Speech and Computer,* Patras: University of Patras. 515–518.

Alumäe, T. 2006. Sentence-adapted factored language model for transcribing Estonian speech. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing.* Piscataway, N.J.: IEEE, (1). 429–432.

Baddeley, A. 1997. Human Memory: Theory and Practice. Hove: Psychology Press.

Boersma, P.; Weenink, D. 2007. Praat: doing phonetics by computer (Version 4.6.36) [Computer program]. Retrieved September 2, 2007, from http://www.praat.org/

Chafe, W.L. 1994. Discourse, consciousness, and time: the flow and displacement of conscious experience in speaking and writing. Chicago: University of Chicago Press.

Clark, H.; Wasow, T. 1998. Repeating words in spontaneous speech. *Cognitive Psychology* 37. 201–242.

Eek, A.; Meister, E. 1999. Estonian speech in the BABEL multi-language database: Phonetic-phonological problems revealed in the text corpus. In: Fujimura,O. (eds.) *Proceedings of LP'98*. Prague: The Karolinum Press, (II). 529–546.

Furui, S. 2005. Spontaneous speech recognition and summarization. In: *Proceedings of the Second Baltic Conference on Human Language Technologies,* Tallinn, Estonia. 39–50.

Hennoste, T.; Gerassimenko, O.; Kasterpalu, R.; Koit, M.; Rääbis, A.; Strandson, K.; Valdisoo, M. 2006. Cue-based Interpretation of Customer's Requests: Analysis of Estonian Dialogue Corpus. In: *Advances in Natural Language Processing/ Lecture Notes in Artificial Intelligence*. Springer. 206–213.

Horne, M.; Frid, J.; Roll, M. 2006. Timing restrictions on prosodic phrasing. In: *Nordic Prosody IX*. Frankfurt am Main: P.Lang. 117–126.

Kerge, K.; Pajupuu, H.; Tamuri, K. 2007. Where should TTS-synthesizer pause and breathe? In: *The Third Baltic Conference on Human Language Technologies,* Kaunas, Lithuania.

Koit, M.; Valdisoo, M.; Gerassimenko, O.; Hennoste, T.; Kasterpalu, R.; Rääbis, A.; Strandson, K. 2006. Processing of Requests in Estonian Institutional Dialogues: Corpus Analysis. In: Sojka, P.; Kopecek, I.; Pala. K. (eds.) *Text, Speech and Dialogue: 9th International Conference, TSD 2006*. Springer (Lecture Notes in Artificial Intelligence). 621–628.

Krull D. 1993. Temporal and tonal correlates to quantity in Estonian spontaneous speech. In: *Papers from the Seventh Swedish Phonetics Conference.* Reports from Uppsala University Linguistics (RUUL) 23.

Krull D. 1997. Prepausal lengthening in Estonian: Evidence from Conversational speech. In: Lehiste, I. and Ross, J. (eds.) *Estonian Prosody: Papers from a Symposium.* Tallinn. 108–121.

Liu, Y-F.; Tseng, S-C.; He, Y-F.; Huang, Y-J.; Lee, T-L. 2006. A preliminary study for spontaneous speech understanding and processing: prosodic units in Mandarin Chinese. In: *Proceedings of LPSS'2006*. 97–112.

Mihkla, M. 2005. Modelling pauses and boundary lengthening in synthetic speech.. In: *Proceedings of the Second Baltic Conference on Human Language Technologies.* Tallinn, Estonia. 305–310.

Mihkla, M. 2006. Comparison of statistical methods used to predict segmental durations. In: Aulanko, R,; Wahlberg, L.; Vainio, M. (eds.) *The Phonetics Symposium 2006: Fonetiikan Päivät 2006, Helsingi, 30.-31.08.2006.* Helsinki: University of Helsinki. 120–124.

Mihkla, M. 2007. Modelling Speech Temporal Structure for Estonian Text-to-Speech Synthesis: Feature Selection. *TRAMES*, 11(61/56), 2. 284-298.

Selkirk, E. 1984. The Phonology and Syntax: the relations between sound and structure. Cambridge: MIT Press.

Strangert, E. 2004. On modeling of conversational speech. In: *Proceedings of FONETIK 2004*. Department of Linguistics, Stockholm University.

Winkworth, A.; Davis, P.; Adams, R.; Ellis, E. 1995. Breathing patterns during spontaneous speech. *Journal of Speech and Hearing Research*, 38(1). 124–144.

EINAR MEISTER is a head of the Laboratory of Phonetics and Speech Technology, Institute of Cybernetics at Tallinn University of Technology since 1990. He graduated at Tallinn Technical University in 1982 in the field of electronics, in 1998 he received his M.Sc. in computer science and system engineering and in 2003 he defended his Ph.D. in general linguistics at Tartu University. His research interest include experimental phonetics, speech analysis and synthesis, speech and speaker recognition, speech databases, speech technology applications. Einar Meister is the author of more than 70 scientific papers and has been a project manager of numerous national and international projects. Currently he is involved in NordForsk network "Variation in speech production and speech perception" and responsible for two national research projects. He has taught different courses on speech technology at Tartu University and Tallinn University of Technology. He is the member of the board of the National Estonian Language Technology Programme 2006-2010. E-mail: einar@ioc.ee

PÄRTEL LIPPUS received his MA degree in Estonian and Finno-Ugric linguistics (thesis “Affricate sound in Võru dialect: acoustic analysis”) at the University of Tartu in 2005. From 2005 he is working on his Ph.D. at the Institute of Estonian and General Linguistics at the University of Tartu with thesis on Estonian word prosody. His main research interest is Estonian prosody. E-mail: partel@murre.ut.ee