Tuomo Saarni

Segmental Durations of Speech

Turku Centre *for* Computer Science

TUCS Dissertations
No 126, January 2010

Segmental Durations of Speech

Tuomo Saarni

University of Turku
Department of Information Technology
FI-20520 Turku, Finland

2010

Supervised by

Professor Tapio Salakoski
Department of Information Technology
University of Turku
Turku, Finland

Professor Jouni Isoaho
Department of Information Technology
University of Turku
Turku, Finland

Reviewed by

Professor Pasi Fränti
Department of Computer Science
University of Joensuu
Joensuu, Finland

Research Professor Markku Turunen
Department of Computer Sciences
University of Tampere
Tampere, Finland

Opponent

Head of Laboratory Einar Meister, PhD
Laboratory of Phonetics and Speech Technology
Institute of Cybernetics
Tallinn University of Technology
Tallinn, Estonia

ISBN 978-952-12-2391-4
ISSN 1239-1883

Abstract

This dissertation considers the segmental durations of speech from the viewpoint of speech technology, especially speech synthesis. The idea is that better models of segmental durations lead to higher naturalness and better intelligibility. These features are the key factors for better usability and generality of synthesized speech technology. Even though the studies are based on a Finnish corpus the approaches apply to all other languages as well. This is possibly due to the fact that most of the studies included in this dissertation are about universal effects taking place on utterance boundaries. Also the methods invented and used here are suitable for any other study of another language.

This study is based on two corpora of news reading speech and sentences read aloud. The other corpus is read aloud by a 39-year-old male, whilst the other consists of several speakers in various situations. The use of two corpora is twofold: it involves a comparison of the corpora and a broader view on the matters of interest.

The dissertation begins with an overview to the phonemes and the quantity system in the Finnish language. Especially, we are covering the intrinsic durations of phonemes and phoneme categories, as well as the difference of duration between short and long phonemes. The phoneme categories are presented to facilitate the problem of variability of speech segments.

In this dissertation we cover the boundary-adjacent effects on segmental durations. In initial positions of utterances we find that there seems to be initial shortening in Finnish, but the result depends on the level of detail and on the individual phoneme. On the phoneme level we find that the shortening or lengthening only affects the very first ones at the beginning of an utterance. However, on average, the effect seems to shorten the whole first word on the word level.

We establish the effect of final lengthening in Finnish. The effect in Finnish has been an open question for a long time, whilst Finnish has been the last missing piece for it to be a universal phenomenon. Final lengthening is studied from various angles and it is also shown that it is not a mere effect of prominence or an effect of speech corpus with high inter- and intra-speaker variation. The effect of final lengthening seems to extend from the final to the penultimate word. On a phoneme level it reaches a much wider area than the initial effect.

We also present a normalization method suitable for corpus studies on segmental durations. The method uses an utterance-level normalization approach to capture the pattern of segmental durations within each utterance. This prevents the impact of various problematic variations within the corpora. The normalization is used in a study on final lengthening to show that the results on the effect are not caused by variation in the material.

The dissertation shows an implementation and prowess of speech synthesis on a mobile platform. We find that the rule-based method of speech synthesis is a real-time software solution, but the signal generation process slows down the system beyond real time. Future aspects of speech synthesis on limited platforms are discussed.

The dissertation considers ethical issues on the development of speech technology. The main focus is on the development of speech synthesis with high naturalness, but the problems and solutions are applicable to any other speech technology approaches.

Acknowledgements

Foremost, I would like to thank my colleague and dear friend Jussi Hakokari for his efforts and ideas during these years. It was not always easy but I could have not done this without you. Sorry for all the pranks I pulled on you.

I would like to thank my supervisors and mentors Prof. Tapio Salakoski, Prof. Olli Aaltonen, and Prof. Jouni Isoaho. Sometimes things just turn out ok.

I am truly grateful for all the support that my colleagues Jyri Paakkulainen, Stina Ojala, Janne Savela, Lotta Alivuotila, Vesa-Petteri Mikkonen, and Mikko Jalonen, have given me. I had the best time with you in- and outside the work.

I would like to thank Prof. Pasi Fränti and Research Prof. Markku Turunen for their diligent reviews of this thesis.

This work was carried out during years 2005-2009 at the Department of Information Technology, University of Turku. I would like to thank the administration and technical staff and colleagues at the department. Further, I would like to thank Maija S. Peltola and her crew at the Phonetics lab. Also, I remember with warmth the two months period at the department of computer science, University of Joensuu.

I would like to express my gratitude to Turku Centre for Computer Science (TUCS) Graduate School for the financial support. This work was also financially supported by the Finnish Funding Agency for Technology and Innovation, the Technology Industries of Finland Centennial Foundation, the Ulla Tuominen Foundation, the Nokia Foundation, and the Henry Ford Foundation.

I would like to thank my father and his spouse Marjo, my mother and my sister and her family, and relatives for their support and encouragement. Hannu, remember those electric gadgets I used to build when I was little? Isn't it funny how little people change when they grow up?

I also would like to thank my all dear friends for their trust that this madness will one day be over and I will get a real job. Didn't see this coming?

And finally, my loving thanks to Annu. You have been the greatest support and encouragement that I have had. I must admire how you still bear my sense of humour and bad singing. Even when it sometimes seemed that I do not need anymore encouragement it was you who just by being there made me realize and remember the most important things in life. Thank you. Can I thank my dog as well? Well, thank you Vinca. It has been a run, hasn't it?

In Turku, January 2010
Tuomo Saarni

List of original publications included in the thesis

I. Tuomo Saarni, Jussi Hakokari, Tapio Salakoski, Jouni Isoaho, Olli Aaltonen: The Role of Duration in Finnish Rule-Based TTS. Speech Analysis, Synthesis and Recognition, Applications of Phonetics, September 19-23, 2005, AGH University of Science and Technology, Kraków, Poland.

II. Jussi Hakokari, Tuomo Saarni, Tapio Salakoski, Jouni Isoaho, Olli Aaltonen: Determining Prepausal Lengthening for Finnish Rule-Based Speech Synthesis. Speech Analysis, Synthesis and Recognition, Applications of Phonetics, September 19-23, 2005, AGH University of Science and Technology, Kraków, Poland.

III. Tuomo Saarni, Jussi Hakokari, Jouni Isoaho, Olli Aaltonen and Tapio Salakoski: Segmental Duration in Utterance-Initial Environment: Evidence from Finnish Speech Corpora. FinTAL 2006, 5th International Conference on Natural Language Processing, 23-25 August 2006.

IV. Tuomo Saarni, Jussi Hakokari, Olli Aaltonen, Jouni Isoaho, Tapio Salakoski: Utterance-initial Duration of Finnish Non-plosive Consonants. NODALIDA 2007, the 16th Nordic Conference of Computational Linguistics, 25-26 May 2007 Tartu, Estonia.

V. Jussi Hakokari, Tuomo Saarni, Tapio Salakoski, Jouni Isoaho, Olli Aaltonen: Measuring Relative Articulation Rate in Finnish Utterances. The International Congress of Phonetic Sciences (ICPhS) 2007, 6-10 August 2007 Saarbrucken, Germany.

VI. Tuomo Saarni, Jussi Hakokari, Jouni Isoaho, Tapio Salakoski: Utterance-level Normalization for Relative Articulation Rate Analysis. Interspeech 2008 incorporating SST'08, 22-26 September 2008, Brisbane, Australia.

VII. Kai K. Kimppa & Tuomo Saarni: Right to one's voice? Ethicomp 2008, the Tenth ETHICOMP International Conference on the Social and Ethical Impacts of Information and Communication Technology, 24-26 September 2008, Mantua, Italy.

VIII. Tuomo Saarni, Jyri Paakkulainen, Tuomas Mäkilä, Jussi Hakokari, Olli Aaltonen, Jouni Isoaho and Tapio Salakoski: Implementing a Rule-Based Speech Synthesizer on a Mobile Platform. FinTAL 2006, 5th International Conference on Natural Language Processing, 23-25 August 2006.

Contents

Part I

Research summary

Chapter 1

Introduction

Speech is a highly evolved way of communication between people. It has advantages making it superior to any other ways to communicate, though it is not free from errors (Ohala 1986). However, speech is error-tolerant and robust in noisy environments, even when the ambient noise is speech or other human generated sounds like coughing (Ohala & Shriberg 1990, Warren 1970, Warren & Warren 1970, Bashford et al. 1992, McDermott & Oxenham 2008).

As the speech is considered to be such a practical way to communicate between people, its usage has slowly shifted to the human-machine interaction. Natural and synthesized speech has become a familiar way to pass information between the user and the system. The Finnish Broadcasting Company (YLE) offers subtitles for television programs in synthesized speech and car navigators guide drivers by voice. However, machines are far behind when it comes to the perception capabilities of humans. Especially, the information needs to be passed somewhat unnaturally compared to human dialogue. It is said that modeling user behavior is the most challenging part in creating spoken dialog systems (e.g. Shin et al. 2002). Still, there are attempts to create a truly conversational human-computer interface (Allen et al. 2001). Hakulinen et al. write about multimodal spoken dialogue systems: "*While some users can learn to use a system just fine with just a static manual or even without any guidance, others have many problems in learning the style of interaction required in human-computer spoken dialogue.*" (Hakulinen et al. 2007). In such multimodal systems that mix spoken language dialogue with a touch screen or a similar modality, the speech is used to support the dialogue with the computer. It has been suspected that speakers may feel hesitant speaking with computers, especially in public, but in their study Lamel et al. (2002) showed opposite results. They found out that 87% of the participants of their study said that they would use speech if the system was located in a train station (Lamel et al. 2002).

Speech recognition is developing at a fast pace, but it has not reached such universality in everyday technologies as synthesized speech has. Some, e.g. the Stopman system (Turunen et al. 2005, Turunen et al. 2006), make use of speech and multimodal interfaces and are publicly available. This system provides timetables for public transport via mobile phone. The Stopman system

has served since 2003 and has received satisfying feedback from the users (Turunen et al. 2005, Turunen et al. 2006).

At the same time, speaker recognition is subject to thorough research to be implemented as a liable biometric identification system. Speaker recognition embodies two different tasks: identification and verification. Speaker identification answers to the question of who is talking, whilst the verification refers to the problem whether this voice belongs to the person he/she claims to be. Recognition is possible even in real time (Kinnunen et al. 2006) and text indepently (Hautamäki 2008). Speaker recognition is described in detail by Campbell and Reynolds (Campbell 1997, Reynolds 2002, Kinnunen & Li 2009).

Research on the segmental durations is a part of speech technology in its very core. Speech technology is increasingly concentrating in modeling human behavior as naturally as possible. The speech recognition has to overcome the problems with natural, free-style speech (Deng & Huang 2004). The speech synthesizer methods need to improve naturalness without losing too much flexibility and the selection-based systems need to become more flexible (see e.g. Kasuya et al. 1999). This development towards naturalness requires high intelligibility while maintaining naturalness and abilities to communicate in a more natural manner with the machine. Therefore, the study of segmental durations is a part of the development towards higher naturalness.

On the other hand, we address intelligibility that must be separated from naturalness. High intelligibility in a speech synthesizer means that the synthetic voice is clear and the speech sounds are easily recognized but it does not have to sound natural or human-like. Intelligibility is important in such cases where the surrounding noise is loud or the situation is otherwise challenging in terms of hearing, e.g. in cars and railway stations. Similarly, the intelligibility is an important factor when creating a speech synthesizer for visually impaired people, who listen to the synthetic speech at a very fast pace, close to the eye-reading speed. In such situations the speech synthesizer should be modified to amplify such effects that are known to enhance intelligibility. Intelligibility in speech synthesizers has been studied e.g. by Venkatagiri (2003) and Logan et al. (1989).

1.1 Speech segments

A speech segment can be defined as any possible well-defined unique member of a group of speech sounds which forms sequential and continuous segments in the speech flow. Common segments are utterances, words and syllables, but others like phones do exist and can be formed. A segment can be or include acoustic silence, e.g. a part of a plosive sound or a pause, but segments must be uninterrupted sequences in time.

Traditionally, speech and written language are segmented into syllables. However, the definition of a syllable in speech is disputable (see e.g. Kenstowicz

& Kisseberth 1979, Trask 1996). R.H. Stetson suggested in the first half of the 20th century that syllables are *chest pulses*, which can be described as "*little ripples on the larger expiratory wave*" (Kelso & Munhall 1988, p. 25). In a frame content/theory in the field of evolution of speech production the syllable is defined as a cycle of continual mouth open-close alternation which can be interpreted as a syllable (MacNeilage et al. 1984, MacNeilage & Davis 1990, MacNeilage 1998). There have been several efforts to define the universal syllable in speech and language by numerous researchers (see e.g. Kenstowicz & Kisseberth 1979). In Finnish, the syllabification rules are quite straightforward, though not completely (Karttunen 1998, Karvonen 2005). Problems arise especially with combined words and loan words. Combined words may be inseparable without the knowledge of the semantics, e.g. '*maanoja*' can be syllabified as '*maan-oja*' ('a ditch of ground') or '*maa-noja*' ('ground support'). In loans words, the single speech sounds should not be taken into different syllables, e.g. '*Por-sche*' in which the '*sch*' is a single sound /ʃ/. This is different in Finnish, because in Finnish each letter mainly represents one speech sound.

In this dissertation, the common speech segment is a phone. A phone is a representative of a phoneme, e.g. in the word '*osa*' ('a part') there are three phonemes /o/, /s/ and /ɑ/, but if the word is pronounced, each speech sound said aloud is a phone. Harris (1951, p. 64) describes phonemes as follows: "*We may try to group segments into phonemes in such a way that all the segments of each phoneme represent sound having some feature in common which is not represented by any segment of any other phoneme: to use articulatory examples, all segments included in /p/ would represent the feature of lip closure plus complete voicelessness (or fortisness) which would not be represented by any other segment.*" In other words, the phoneme is the smallest segmental unit in speech flow that has phones with similar acoustic features within the same phoneme and is distinguishable from other phonemes by these features. Using phones as segments is not uncontroversial by some views since traditionally syllables are considered to be the smallest *reasonable* speech segments. In 1987, Dennis Klatt said *"One of the unsolved problems in the development of rule systems for speech timing is the size of the unit (segment, onset/rhyme, syllable, word) best employed to capture various timing phenomena."* (Klatt 1987), and this dispute is still ongoing. Segmenting speech into phones is criticized to be artificial since acoustic features of phones overlap (see e.g. Browman & Goldstein 1990, Turk et al. 2006). On the other hand, when the annotation into phones is done by a trained phonetician (instead of an automatic annotation algorithm) using appointed segmentation criteria, the result can be coherent. In this dissertation, the corpora are annotated by hand by phonetician Jussi Hakokari, MSc, and the annotation is iterated several times to correct any errors.

In a corpus-based study the corpus size is crucial for statistical analysis. Usually it is not meaningful to study single phonemes since e.g. short vowels are phonologically close to each other. The articulatory movement is similar and the

acoustical features are similar if compared to any other phoneme. Similarly, the statistical analysis is easier when the sample size is kept as large as possible. E.g. foreign phonemes are scarce in the corpora at hand (see e.g. Paper IV).

To overcome these setbacks the phonemes can be categorized into proper sets. This is done by dividing the phonemes into *vowels*, *diphthongs*, *plosives* and non-*plosive consonants*. The categorization is based on the similarity of the articulation movement of phones in the same phoneme category. Also, the phonetically long and short segments are kept in separate categories. This gives us seven categories of phonemes:

Phoneme category	**Abbreviation**
Short Vowels	V
Long Vowels	V:
Diphthongs	D
Short Plosives	P
Long Plosives	P:
Short Non-plosive Consonants	C
Long Non-plosive Consonants	C:

Other categorization methods can be justified, but there is a need to keep the number of categories low, which increases the sample size.

1.2 Segmental durations

The term *segmental duration* refers to the measurable quantity of a speech sound. The quantity of speech sound is therefore measurable in time. Physical quantity differs from the *phonological* length of a speech sound. The phonological length refers to the perception of the speech sound in a language, and usually it cannot be measured in such a simple way as physical quantity.

Finnish is a quantity language, which means that there are long and short speech sounds in the language. Phonologically, short and long phonemes are written in Finnish with one or two letters respectively, see e.g. the Finnish words '*tuuli'*, '*tuli',* and '*tulli'* ('wind', 'fire', and 'customs' respectively). In the first example, there is a long /u:/ sound ('*uu*' in the word '*tuuli*'), which distinguishes it from the second example, which has a short /u/ sound. Similarly, the short /l/ and the long /l:/ distinguish the last two words. To the listener, the difference between a long and a short segment mainly lies in the physical quantity. The double letter is spelled with longer duration, but the duration is *relatively* longer than the short segments in the speech. Therefore, the long and short segments cannot be differentiated merely by time. The duration of a long or a short segment depends on the surroundings. Also, it is known that the long and short segments have several other cues which distinguish them.

Lehtonen (1970), among others, noticed that long segments affect the duration of the surrounding segments. Similarly, in the vowel space of the first two formants the long vowels are more peripheral than the short ones (Lennes 2003). O'Dell (1999) studied the plosive sound /t/ in Finnish and found that "*Evidently there were other factors besides 'pure timing' which affected at least a majority of perceptions for a majority of listeners*". O'Dell, however, does not reveal which the other factors are. Later, O'Dell (2003) studied the matter further in his dissertation and found out that, for example, the fundamental frequency plays a crucial part in the perception of phonological length. However, the view selected in this dissertation focuses on the physical quantity of time. The phonological length is also considered, but only from the categorization point of view.

Here the term segmental *duration* refers to the absolute or relative duration of such segment. Absolute duration means e.g. milliseconds whilst the relative duration is a coefficient that uses surrounding or similar segments to define its value by comparison. E.g. a segment's duration can be 86 milliseconds (absolute) or 120%, which is 20 per cent longer than other segments in the utterance (relative). The relative duration is important when doing normalization. In normalization an unwanted bias and skewness is removed by studying segments in a relative manner.

Information about segmental duration is a key factor in creating models that imitate natural speech prosody. Duration is combined with most of the prosody features. As said earlier, the longer duration affects fundamental frequency, vowel space and the duration of other segments, and possibly vice versa. E.g. the accent is mostly considered to be a compound of fundamental frequency (pitch), intensity (volume), and duration. On the other hand, in this thesis the interest lies mainly in the duration, not in the effects it produces. The segmental duration is chosen to be the center of this study and it is assumed that the durational effects are a result of e.g. phonetic boundaries. This means that we oppose that the boundary is created due to the fact that the durations preceding it have changed.

Better modeling of segmental durations of speech flow increases the level of naturalness of speech synthesis. However, exact models of segmental durations in conversational and spontaneous speech, that is natural speech, may improve the effectiveness and possibilities of other speech technologies, e.g. speaker recognition and speech recognition.

Chapter 2

Overview of the thesis

2.1 Research goals

Science in the field of speech technology is evolving rapidly. New commercial technologies emerge daily and consumers are slowly accepting new interfaces that include speech technology. This dissertation aims at leaving a footprint in the field of speech science and to achieve new viewpoints that will be useful in the vast area of speech technology.

The publications included form three main aspects of research. Firstly, we are interested in the boundary-adjacent effects on segmental durations concerning both final and initial positions in an utterance. To study speech using a corpus approach, we need to define a level of detail on which to operate. This is done, first, by creating phoneme categories which create subgroups of sufficient size. The number of individual phonemes is too low to study them as entities and therefore a categorization is needed. Each category can then be analyzed as an entity on a single phone level. Categorization is also being used to study segmental durations within a word. This study is called the word model approach and it is used for representing durations for a speech synthesizer. In the studies of boundary-adjacent effects we examine how the position in an utterance (especially initial and final) affects the segmental duration, if at all. Similarly, if the position has an effect, how does the alteration vary between phoneme categories? To study boundary-adjacent effects there is a need to normalize data in some cases. The boundary-adjacent areas are prone to be affected in terms of the length of an utterance, which could cause fallacious results. This problem needs to be solved with a normalization method. The normalization method helps to reduce intra- and inter-speaker variation which is a significant problem when using a corpus approach. The study of boundary-adjacent effects is also adding on how to *represent* segmental durations in such a way that modeling and studying speech segments in boundary-adjacent positions is advanced.

Secondly, we are interested in the technological aspects of creating speech synthesis on a mobile platform with limited resources and capabilities. In

this approach we want to measure what the level of efficiency of speech synthesis is on such a limited platform. By efficiency we mean the real-time capability of the speech synthesis procedure and its feasibility. In addition, we want to shed a light on possible enhancements for the future.

Thirdly, there are philosophical considerations on the ethical problems raised by the fast development of speech synthesis. Speech synthesizers are quickly approaching a level where it is hard to tell whether the synthesis is actually a real person speaking. This raises ethical questions such as whose responsibility it is to assure that the synthesis is used in acceptable ways. For example, is it the creator's or the user's responsibility? What measures could and should be taken before the technology is on the market? These questions are pondered from an ethical point of view.

The following list bundles the research goals as questions.

1. How do the segmental durations comport in Finnish?

a) How can we model segmental durations within a word? Can a word model approach be used to enhance the level of naturalness in a speech synthesizer?

b) Is there an effective way to study similar speech sounds as a group and how can this group be justified? Why do we need phoneme categories?

c) How does the vicinity of an utterance boundary affect the segmental durations?

d) How can we reduce variation in the corpus approach when studying segmental durations?

2. How do we implement a speech synthesizer on a platform with limited resources? Can it operate in real time?

3. What ethical considerations should be taken into account when developing a new technology such as a speech synthesizer?

In Table 1, we show how the research goals are bound to the articles included in the dissertation and in which chapters the research questions are answered.

	Paper								
	I	II	III	IV	V	VI	VII	VIII	Goal
3.1 Word models	**X**								**1.a**
3.2 Phoneme Categorization	**X**	**X**	**X**	**X**		**X**			**1.b**
3.3.1 Utterance initial effects				**X**	**X**				**1.c**
3.3.2 Final lengthening		**X**	**X**		**X**	**X**			**1.c**
3.4 Normalization					**X***	**X**			**1.d**
3.5 Synthesis on a limited platform								**X**	**2.**
3.6 Ethical issues							**X**		**3.**

* In this paper we use a prototype of the normalization method

Table 1. In the table we show a linkage between research goals, questions and papers included in the dissertation.

2.2 Research methods

To study segmental durations there are two possible approaches: one is to study durations in a large natural language, e.g. read-aloud sentences and/or a spontaneous speech corpus (*corpus study*), or to use *controlled experiments*, so-called laboratory speech (Turk et al. 2006). It is highly controversial which approach is preferable, since both have disadvantages.

This dissertation is based on corpus studies. In the future, we will use the term *natural speech corpus* to emphasize the difference from *laboratory speech* (controlled experiments). The term corpus refers to a material which combines annotation and the speech signal. Annotation is a marking that points out the start and end position of a speech segment in the speech signal. Our studies base on observations made of natural speech corpora. The advantage of a natural speech corpus study is that the speech is produced in natural situations, especially if compared to controlled studies. A natural speech corpus can consist of monologues, dialogues or sentences read aloud in a non-laboratory environment (e.g. radio news reading). Controlled speech experiments use carefully selected carrier sentences that often place a key word in a selected position in the sentence. The study then compares the acoustic features of the key word in selected positions, e.g. "*I said 'BEEF arm', not 'REEF arm'.*" used by Turk & Sawusch (1997). Sometimes, the key word is a nonsense word used to restrain any content that might affect the reader. A nonsense word does, however, represent the syllable structure of the studied language, e.g. in Finnish "*mipatu*", "*matupi*", and "*mupita*" used by Suomi & Ylitalo (2004). We decided to use a natural speech corpus due to its undeniable correspondence with natural human speech. It does not include nonsense words and it is by nature more improvisational without planning or spontaneous, even though the text is read from a paper, when the reader does not knowingly read it for a speech study and the text is not intended for research purposes. However, some criticism must be expressed of the natural corpus study approach.

Turk et al. (2004) mention that the advantages of speech in natural situations are more or less outweighed by several reasons. First, they claim that the accuracy of the segments is dependent on the accuracy of the automatic segmentation algorithm. Secondly, they point out that some prosodic effects may be obscured when normalization is not successful. Finally, they remark that some prosodic effects may be lost due to the correlating prosodic factors affecting the result (see e.g. van Santen 1994, Campbell 1992). In our approach, the data annotation is not done by an automatic segmentation algorithm but a trained phonetician. Secondly, we have used a normalization method (Paper VI) which resulted in very similar results as seen in our previous studies (Paper V, Paper II). Thirdly, we believe that a statistical analysis can reveal interesting and new information even from a natural speech corpus, if it is done with well-

designed experiments and with sufficient precision. We do agree that some skewness or bias may be embodied in the results due to the correlating prosodic effects, but we doubt if the results themselves are in fact consequences of correlating effects. So far we have received some criticism towards our approach, but it has mainly been based on false predictions or misunderstandings towards our methods. The criticism is discussed in greater detail in the following chapters.

The material used in this research is somewhat two-fold. There is a single-speaker corpus (hereinafter SS) and a multi-speaker corpus (hereinafter MS). The corpora included 2,115 utterances when combined (Paper V). The SS corpus consisted of 967 utterances whilst the MS corpus included 1,148 utterances. The speaking rate in the material was 750 phones per minute (SS) and 800 phones per minute (MS). The size of the multi-speaker corpus varied during the research process due to the fact that more data was annotated and added between the studies. Similarly, the annotation of the single-speaker corpus varied since new levels were added (e.g. stress annotation) and new annotation criteria were implemented (e.g. adding diphthongs instead of two short vowels). Also, the materials were consistently checked and corrected during the research process.

The single-speaker corpus was read aloud by a 39-year-old male from Helsinki. The speaker has a relatively low pitch voice, and a considerable amount of final devoicing (vocal fry, "creak") characteristic of many Finnish speakers. The material consists of sentences from the Finnish periodical Suomen Kuvalehti (Vainio 2001) and represents standard Finnish, the literary language of most media and formal, public events. The multi-speaker corpus is a fragment of FBC-1 (the Finnish Broadcasting Company) speech corpus. It consists of television news broadcasts and two oral presentations on the radio. It featured both studio newsreaders and field reporters. We have little information on the speakers' identity (such as age or residence), but they all speak perfectly fluent standard Finnish and are clearly experienced in speaking in public. Their speaking rate was relatively hasty with the exception of one male reader, who spoke in a calmer manner. All the interviewees were excluded from the material because of their discontinuous speech; as opposed to the readers, they were audibly insecure when interviewed on the national television and tended to pause after every second word or so. Nothing else was rejected; interjections, questions, and lengthy monologues were all accepted. In the multi-speaker corpus there were, in most studies, a total of 15 adult speakers of varying ages; 9 men and 6 women.

The corpora are sometimes used separately, sometimes combined, and there are studies where solely the single-speaker corpus is used. The reasons are explained in detail in the papers with the exception that the single-speaker corpus was used solely when it was the only corpus at hand. The usage of two separate corpora was based on the idea that it would be interesting to see how

much the results differ between the two. Remarkably little, the study showed (Paper IV).

The studies were done by writing specific scripts that collected the data from the annotated material. Scripting was done by Java and the algorithms were checked several times by sampling. The scripts were small and simple, e.g. collect the durations of all short vowels in the final position of each utterance and save them into a text file. The data was then analyzed, usually by SPSS (Statistical Package for the Social Sciences software, see http://www.spss.com).

To begin with, we must elaborate on the terms accent, prominence, and stress. Prominence is the easiest one to understand, because we know what prominent means: if something is prominent, it is contrastive from its surroundings or it can be exceptional. E.g. one can emphasize words written with capitals '*The APPLE is red.*' or '*The apple is RED.*' in which the capitalized words are prominent. On the other hand, a single word can also be prominent, e.g. imagine a quick warning shout '*STOP!*'.

However, the capitalized words (apple, red) are also accented. The accent is on the word apple or red on the sentence level. There are several different levels of accent, but in our material we have no strongly accented words, and no moderately accented either. This is due to the nature of the speech corpus which is mainly news reading, radio talk, and sentences read aloud. In Finnish, this kind of speech does not include strong prominence. This is typical of standard Finnish which is used in formal situations and is used in official communication. Strong prominence is more likely to be found in radio plays, which are not included in the material.

Stress works on the word level. For example, in the following words the ^ sign marks the stress in the word: '*^photograph*' vs. '*pho^tography*'. The stress can also be used to separate meanings, e.g. '*By ^bus?*' vs. *'^By bus.*', where the first has the raising pitch towards the end to emphasize that it is a question, whilst the latter is just a plain statement. However, it must be reminded that in Finnish the stress is always on the first syllable.

The matter of accent was mainly discarded since we have material that does not include strongly or even moderately accented words. In Finnish the stress follows certain simple rules. The stress is always positioned in the first syllable of a word. The first word has the highest fundamental frequency and it drops towards the end of the utterance. However, the relatively high fundamental frequency is not the only feature of prominence. The relatively long segmental durations and relatively high intensity can be considered to feature with prominence. If we are interested in the first positions of an utterance in Finnish, we can safely assume that the first phones are stressed *always*. On the other hand, if we are interested in the final positions of an utterance, the phones in the end are mainly not stressed. At the same time, the percentage of final lengthened (FL, meaning that the last words or phonemes are spoken slower than the previous segments in the same utterance) utterances was 85-89% in 2-5 word utterances (Paper V), so this would mean that nearly all of the material has a

prominent last word if the FL is considered to be a result of accent. However, subjective listening of the corpora points out clearly that there is a FL with no prominence on the last word. In addition, we carried out a study which separated the prominent words from the non-prominent and showed that the FL is clearly not a result of prominence or stress. The results of this unpublished study and the matter of accent and prominence are discussed later.

2.3 Structure of the thesis

This dissertation is a compilation of eight of my co-authored publications from the past years. The publications were selected to illustrate the research work done from the technological, phonological, and philosophical perspectives. The main focus is on the phonological studies, since they cover the main part of the publication results. Phonological research, however, is concentrated on both segmental duration studies and on the technological part of developing naturalness in the speech synthesis as the main goal. The technological side of the research is, in a sense, the motivation for the dissertation, but it is also visible in the following topics that cover purely technological aspects of this study, such as speech synthesis on a limited platform. The philosophical issues or the ethical viewpoints at the end of the dissertation are there to enhance the technological perspectives.

Although the dissertation at hand is a compilation, there are new unpublished information and research results included in the following chapters: 3.1.1 Implementation, 3.2.2 Vowels and diphthongs, 3.2.3 Consonants, 3.3.1 Utterance initial effects, 3.3.2 Final lengthening. Whenever there is a new or unpublished result in the text, it is mentioned separately.

The dissertation is divided into two parts: Research Summary and Publication Reprints. The research work done for this thesis is written in the first part, Research Summary. In the first part, the introduction chapter gives an overview of what the segmental durations are and why they are interesting. In this chapter the research goals and methods are explained as well as the structure of the thesis. In chapter 3 the research work is examined in detail. In chapter 4 we summarize the dissertation as a conclusion.

The Publication Reprints are in the second part of the thesis. The publications included form three main aspects of research: firstly, the boundary-adjacent effects for segmental durations, concerning both final and initial positions of utterance; secondly, the technological aspects of creating speech synthesis on a mobile platform with limited resources; and, thirdly, the philosophical considerations on the ethical issues raised by the fast development of speech synthesis.

Chapter 3

Overview of the research

This chapter will provide an overview of the research constituting the thesis. The research began in 2005 in a project called "Puheenkäsittelyn uudet menetelmät ja sovellukset" (PUMS, 'New methods and applications in speech processing') mainly funded by the Finnish Funding Agency for Technology and Innovation (TEKES). In the project several Finnish speech technology companies and universities were joined to conduct speech-related research for four years. The research covered the main fields of speech research in Finland, such as multimodal user interfaces, speaker recognition, speech recognition, and speech synthesis. Our team focused on an older method of speech synthesis called rule-based synthesizer, though when our team joined the project there were only three years remaining.

At the beginning of our research work the main focus was to improve the naturalness of the rule-based speech synthesis without losing the intelligibility and controllability of the synthesizer. First, we started out by writing the synthesizer program from scratch. Development was divided into three tasks: high-level synthesis, signal generation, and phonetic phoneme rules. The tasks were given to a computer scientist (author), digital signal processing specialist Jyri Paakkulainen, MSc, and phonetician Jussi Hakokari, MA, respectively.

Once the program was at a stage where it could be used as a test platform for future studies, we established the first studies which aimed at improving naturalness by using better models of segmental durations. This choice of direction was the start kick for the topic of this thesis.

3.1 Word models

The idea of word models was born whilst reading the dissertation of Lehtonen (1970). In the thesis, Lehtonen studied the durational patterns of Finnish. The patterns are presented in the form of vowels (V) and consonants (C), e.g. '*muuli*' ('a mule') is presented as CVVCV. The phonemic variants of long vowels and consonants were studied as one, so the double '*uu*' vowel (phoneme /u:/) in the

word '*muuli'* presents one durational entity or pattern. The form of presentation is not the invention of Lehtonen, but a commonly used way to represent word forms in phonetic literature.

We call the CV structure of a word a *word model*, since it models the segmental durations on the word level. A word model represents all words that fall into the same category of this word form, e.g. '*muuli'*, '*nuori'*, and '*saada'* ('a mule', 'young', 'to receive') all belong to the same word model. In our approach, inflected words were processed as normal word stems, e.g. '*muulin*' goes to the same category (CVVCVC) as '*soitin*' (a genitive form of the word 'a mule' and a stem of the word 'an instrument' respectively). This meant that we did not need a morphological analyzer or competent syllabification methods to dismantle the words being synthesized.

In the research work of Lehtonen (1970), the mean duration of patterns (C or V) varied significantly. For example, Lehtonen showed that in the word form of CVCCV the duration of the first vowel is about one and a half times longer than the second vowel (Lehtonen 1970, p. 141). Later, many others have studied the variation of segmental durations in greater detail (see e.g. Vainio 2001, Suomi et al. 2003, Suomi 2007, Suomi & Ylitalo 2004). The results of these studies for the speech synthesis were interesting since we had so far used fixed segmental durations. The results signalled that by modeling the segmental durations better in the speech synthesis we may be able to achieve higher naturalness.

In addition, it has been showed that in a quantity language, such as Finnish, the vowels are more peripheral in the vowel space when longer in duration (Wiik 1965, Lennes 2003). Though surprisingly, Heikkinen (1979) has found that short vowels in unstressed position have approximately the same quality as stressed long ones. O'Dell (1999) studied the plosive sound /t/ in Finnish and found that "*Evidently there were other factors besides 'pure timing' which affected at least a majority of perceptions for a majority of listeners*". O'Dell, however, does not reveal which the other factors are. Later, O'Dell (2003) studied the matter further in his dissertation and found out that, for example, the fundamental frequency plays a crucial part in the perception of phonological length. The acoustic differences are, however, unperceived by listeners, and therefore the most important factor is the duration. In the end, the duration is the most important difference between short and long vowels.

So, it was obvious that using constant or fixed segmental durations would not be the best choice when enhancing naturalness in the speech synthesis. To improve the naturalness of the synthesizer we have to have a better model for segmental durations. Similarly, the previous studies in Finnish supported the idea that better models for segmental durations would have a positive impact on intelligibility.

3.1.1 Implementation

We borrowed the idea of the durational patterns from Lehtonen (1970) and adapted it to our rule-based speech synthesizer. In the adaptation we used a speech corpus, which was kindly given to us by Dr. Martti Vainio. The corpus is described in detail in his dissertation (Vainio 2001) and briefly in this thesis in chapter 2.2. The words in the corpus were divided into patterns of consonants and vowels as presented earlier, and similar words were allocated into word models, e.g. '*hätä*' ('an emergency') and '*nenä*' ('a nose') fall into the same word model of CVCV words. Lehtonen (1970) and several authors after him have shown that the C and V segments apart greatly from each other by duration. An example of the variation is presented in Figure 1. The material used in this figure is from the single-speaker corpus.

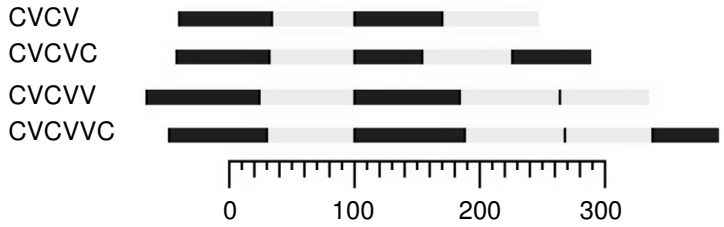


Figure 1. The variation of the mean duration of V and C segments in the word models beginning with CVCV. In the figure, the dark areas represent consonants (C) and the lighter areas vowels (V). The origin of the scale of duration is set to 100ms and for the word models in the end of the first vowel. The figure is taken from Saarni (2005)[1].

Figure 1 shows that the first vowels in the word models have similar durations, and so do the first consonants. On the other hand, the second consonants have very different segmental durations.

The variation of the segmental durations in the single-speaker corpus resembles the variation obtained by Lehtonen (1970). This resemblance can be seen in Table 2. The only remarkable difference can be observed in the last vowels of the word models CVCV and CVCVC. Notice that in the models CVCVV and CVCVVC Lehtonen used only long vowels in the position of VV, whilst the comparison is made with a material that includes diphthongs, vowel pairs and long vowels in the same position. In the latter case, different vowels form two entities, and a single long vowel is divided into two equal halves.

[1] The figure is taken from the author's unpublished master's thesis ("Puhe säännöiksi, säännöt puheeksi – sääntösynteesin kehittäminen", Tuomo Saarni, 2005, University of Turku).

	Saarni 2005	Lehtonen 1970
C	74	72
V	64	65
C	70	78
V	76	102

	Saarni 2005	Lehtonen 1970
C	74	67
V	69	62
C	55	70
V	73	100
C	62	74

	Saarni 2005	Lehtonen 1970
C	90	71
V	76	65
C	84	89
V	80	
V	71	144*

	Saarni 2005	Lehtonen 1970
C	79	73
V	71	67
C	89	83
V	80	
V	71	154*
C	53	59

* Lehtonen used only long vowels here, and therefore had only one segment

Table 2. This table presents the mean duration in milliseconds of each segment in the words beginning with CVCV. The columns on the left represent the data from the single-speaker corpus (Saarni 2005) and, for comparison, on the right there is the data from Lehtonen (1970, pp. 127-128).

In the first study on word models (Hakokari et al. 2005a), we used models which treated plosives and other consonants equally, e.g. such words as "*olla*" ('to be') and "*akka*" ('an old woman') fell into the same word model. This may explain the difference with Lehtonen (1970) in the word models CVCV and CVCVC, since Lehtonen used words that have the same stem, e.g. '*sama*' and '*saman*' ('the same' and 'of the same' respectively, see Lehtonen 1970, p. 106). In our material, very different words fell into the same word model, e.g. '*sama*' and '*kato*' were put into the same model. Also, our method did not pay attention to whether the word was in the first position of the utterance, and therefore if the first consonant happened to be a plosive it hardly had any duration at all due to its acoustic nature. However, these differences between the methods do not explain the resemblance with all the other C and V segments. Similar to the plosive consonants, the diphthongs were not yet annotated in the material and therefore they were treated as sequential two short vowels.

By using the method described above we managed to establish approx. 1,100 word models (Hakokari et al. 2005a). The most frequently used word models occurred more than a hundred times (VCCV, 125 occurrences), whilst many models (628 to be exact) occurred only once. The models were implemented into the speech synthesizer as a part of the durational modeling of speech segments. This means that when a sentence was inputted into the synthesizer it was first divided into single words, and each word was translated into a CV structure, thus into a word model form. Each word could then be

compared with the database of word models created from the corpus. If a matching word model was found the segmental durations were transferred to represent the durations of the current word. If there was no matching word model the system used fixed durations for that word. The user could choose whether to use the fixed segmental durations or the word model durations.

To study the effect of the word models the system was then put to a test where participants assessed the difference between fixed durations and word model durations in a listening procedure. In this test the stimuli were carefully planned so that each word had a word model match.

The stimuli were synthesized using a cascading fundamental frequency contour starting from 100 Hz, rising to 120 Hz within the first syllable, and then lowering linearly towards the end of the utterance reaching 80 Hz. This means that the fundamental frequency rises in the beginning of each word by 20 Hz (120-100=20), but descends towards the end of the word reaching a lower level than it had in the beginning of the word. The decline in the fundamental frequency from the starting level within a word is (100-80) / (*number of the words in the sentence*) Hz. For example, if the sentence had only one word it would reach 80 Hz in the end of the word. However, if the sentence had two words the fundamental frequency would reach 90 Hz in the end of the first word, then rise again to 110 Hz in the first syllable of the latter word and descend to the 80 Hz towards the end of the latter word.

This fundamental frequency contour is typical of a Finnish male speaker without any prominent words in the utterance. The synthesis produced 30-35% longer stimuli with fixed segmental durations, so we used the PSOLA (Pitch-Synchronous Overlap and Add, see Moulines & Charpentier 1990) algorithm to shorten the fixed duration stimuli to match the duration with the word model stimuli. The study included 16 stimuli with durations varying from 0.74 to 4.00 seconds. The participants were 10 women aged 20 to 54, and they reported no deficit in hearing or language. The participants were asked to make a forced choice between either the synthesized sentence with fixed durations or with word model durations. They were asked to identify the version sounding *more natural*. The sentences were given as transcripts to prevent intelligibility issues. The results showed ambiguous results that seemed to favor the word model approach when the stimulus was long. However, the results are discussed in detail later in chapter 3.1.2 Findings.

At the time of the study (Hakokari et al. 2005a), our synthesizer was still under development, and therefore we used a synthesizer which was written at the phonetics laboratory of the University of Turku. The signal generator used was SenSyn version 1.1, by Sensimetrics Corporation. SenSyn is based on the KLSYN synthesizer, which is based on the research of Dennis Klatt (1980). KLSYN is described in detail in Klatt's unpublished paper "KLSYN: A Formant

Synthesis Program"[2]. The synthesis system could only produce a four seconds long stimulus, which prevented the use of longer sentences in the experiment. The short duration of the stimulus was due to the SenSyn program, and unfortunately we had no control over this.

In the latter study on the word models (Paper I), we added a new category of plosives which were separated from the other consonants. The plosives were separated since we noticed that their intrinsic duration pattern diverges from the durations of the non-plosive consonants. For example, if we calculate the mean duration of short non-plosive consonants versus plosive consonants from the unaccented words, we end up at Table 3.

Multi-speaker Corpus			
Final	Penultimate	Other	
170.2	156.9	126.8	Long Plosives
120.3	98.1	81.2	Long Non-plosive Consonants
100.0	84.9	77.5	Short Plosives
76.4	65.5	56.5	Short Non-plosive Consonants
Single-speaker Corpus			
Final	Penultimate	Other	
207.7	154.1	139.8	Long Plosives
144.9	94.0	92.4	Long Non-plosive Consonants
105.1	88.4	85.4	Short Plosives
74.3	61.9	58.5	Short Non-plosive Consonants

Table 3. Intrinsic mean duration in milliseconds of plosives and non-plosive consonants in unaccented words in final, penultimate and other position of the utterances. The results on top are from the multi-speaker corpus. The table is taken from previously unpublished results.

In the latter paper, we also implemented a category of diphthongs instead of interpreting diphthongs as a sequence of two short vowels (Paper I). From the acoustic point of view, it is very hard to mark out the boundary between the two short vowels in a diphthong, and therefore we decided to treat all the diphthongs in the material separately.

By using these phoneme categories we prepared a statistical model of the durational patterns in our single-speaker speech corpus. The model was then implemented to the synthesizer, meaning that whenever a word was synthesized, its word model was first retrieved and the durations of the word model were used in the synthesis process. The new method of creating the word models yielded approx. 2,500 different word models (Paper I) compared to the 1,100 word models in the previous paper (Hakokari et al. 2005a). Again, the material was based on the single-speaker corpus. This time we used a synthesizer of our

[2] see http://www.ling.ohio-state.edu/courses/materials/825/klsyn-dos/klsynman.pdf (visited on third of September 2009)

own which, too, based on the studies of Klatt (1980). The synthesizer was now capable of producing as long sentences as needed, and so the stimuli duration varied from 19 to 44 seconds, which was considerably longer than the stimuli we could produce with the SenSyn program. In this study, the fundamental frequency contour was 100-140-65 (in the previous one it was 100-120-80). The contour was more vivid and lively than the one used before, but still corresponded to a male voice. We also lengthened the final words of the utterances by 10 per cent, and on every linguistic boundary between phrases and sentences we added a short acoustic silence which represented a pause generated by the moment of inhaling. In the study, we had 21 participants, 7 male and 14 female, aged between 23 and 45. The test procedure was again a forced choice between word model durations and fixed durations with preference for better naturalness.

3.1.2 Findings

In the first study with word models, we wanted to find out whether or not the word models enhance naturalness in synthesized speech (Hakokari et al. 2005a). We prepared sixteen stimuli and gathered ten participants. The stimuli were paired so that for each stimulus there was a version with fixed durations and a version with word model durations. The test was a forced choice between the two durational models. The results were ambiguous. Only two of the word model stimuli were preferred by all participants when compared to the stimuli with fixed durations. Six of the word model stimuli were preferred only by one or two listeners. It seemed that the longer stimuli were more often preferred over the ones with fixed durations. The results were unconvincing and the word models needed to be studied more carefully. See the results in Figure 2.

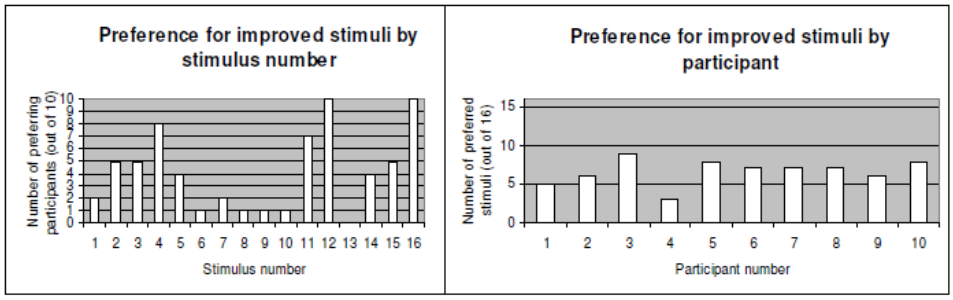


Figure 2. The left figure shows the word model (improved) stimuli preference by stimulus number. The right figure shows participant behavior and their preference for the word model stimuli. The figure is taken from the paper Hakokari et al. (2005a).

In the second study with the word models the models were more detailed (Paper I). We introduced categories of plosive consonants and

diphthongs. In the previous paper, diphthongs were placed in the category of short vowels so that each diphthong was considered as two short vowels. The stimuli were much longer, too, from 19 to 44 seconds, instead of single words or short sentences. The listening test included 21 participants and four stimuli. Again, the results showed that the listeners preferred the word model durations when the stimulus was long. Again, it also meant that listeners preferred fixed durations when the stimulus was short. The overall result was that the word model durations were only slightly preferred by the listeners (53.6%). However, the least experienced listeners of synthetic speech preferred word models over 65 per cent of the time. In Figure 3 we show the preference ratios for each stimulus, where 50 per cent would mean that half of the participants preferred the fixed durations and the second half preferred the word model durations.

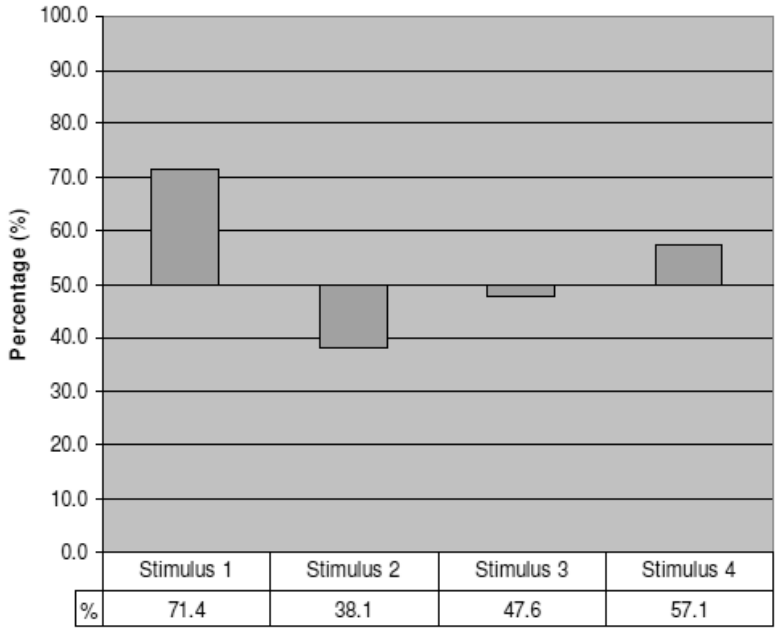


Figure 3. Of the four stimuli, the participants preferred only two stimuli (stimuli 1 & 4) with word model durations. A result of 50% would mean that half of the listeners preferred word model durations. This figure is taken from Paper I.

However, as described in Paper I, it became obvious that using word models as such would lead to exponential growth in the number of word models, and our corpus would not be large enough to produce all possible combinations of the phoneme categories in a highly inflected language such as Finnish.

3.1.3 Considerations

Another possibility would be to use the word stem structure. In this approach the word model is actually only the word stem of similar words, e.g. the words '*susi*', '*haju*', and '*juju*' ('a wolf', 'a smell', and 'a catch' respectively) have the same stem structure when categorized to the C and V structure (CVCV). To construct these models we could use inflected versions of the words if the word stem is still intact, e.g. the inflected version of the word '*susi*' is '*suden*' ('a wolf', 'of a wolf') which is not suitable for usage, but the inflected version '*hajun*' of the word '*haju*' ('of a smell', 'a smell') is usable. The suffixes of the words could be modeled separately from the word stems. For example, the word '*hajun*' ('of a smell') corresponds to two models, CVCV (stem model) and –C (suffix model). The suffix model could be later used to give the duration of the suffix, or the last segment –C, of the word '*jujun*' ('of a catch'). This approach could be suitable even for more complex inflections. For example, in the word '*juju-sta-ni*' ('of my catch') there are two suffixes following the word stem '*juju'*. The word can be used to form a word stem model (CVCV) and two suffix models (CCV, CV). Similar suffixes are found in such words as '*jujuni*' ('my catch') and '*jujusta*' ('from a catch') with the same word stem. Similarly, prefixes can be modeled, e.g. in the word '*epäkohta*' ('disadvantage') the prefix is '*epä-*' ('dis-') and the word stem is '*kohta*' ('a point'). However, there are a couple of unsolved problems in this suggested method. For example, the word stems that are changed in the inflection are problematic. A few examples:

'*susi*' '*suden*' ('a wolf', 'of a wolf')
'*katto*' '*katon*' ('a roof', 'of a roof')
'*satu*' '*sadussa*' ('a fairytale', 'in a fairytale')

This change in the word stem is particularly problematic since it changes the consonant in the middle of the word. The change can even be from a plosive to a non-plosive consonant.

Secondly, the modeling would desperately need an automatic morphological analyzer, as otherwise the syllabification would have to be done by hand, and this is virtually impossible due to the huge amount of data needed in the modeling. However, nowadays this kind of morphological analyzer is available for Finnish. A company called Lingsoft has published a morphological analyzer called FINTWOL. The program is originally developed by Kimmo Koskenniemi (see e.g. Koskenniemi 1983). To further study the approach of word models as described here, the FINTWOL program would be an essential part of it. Unfortunately, this approach has been laid aside and presented here only as a possible direction for future studies of word models.

The word models work on a limited level of detail. One might find this problematic due to the fact that there are several other factors that affect segmental durations. As mentioned before, stress and accent lengthen the segments among other acoustic features. On the utterance level, the boundary-adjacent effects are known to interfere with the segmental durations. However,

these other factors may be implemented instead of revoking the word model paradigm. The simple solution would be to implement durational factors that work successively on different levels of detail. This would cover the main factors of segmental durations. The boundary-adjacent effects are discussed in detail later.

3.2 Phoneme categories

In the study of segmental durations we need to address the level of detail. In this dissertation the level of detail is set to phone level. This means that we operate with utterances in which each position is represented by a phoneme. When studying segmental durations it would be easiest to study each phoneme as an entity. This, however, makes it hard to study segmental durations of an utterance since a single utterance rarely includes more than one or two samples of the same phoneme, which makes comparison difficult within the utterance. On the other hand, if the corpus is not exceptionally large we might have difficulties to cover all the positions of utterances on a phone level if we need to find a certain phoneme several times for statistical analysis from the same position. For example, if we want to study the imaginary phoneme X in the 15 first phone-level positions of an utterance in a corpus including 400 utterances with length of 15 phones, this would mean that we have 6,000 phones and each position would have 400 samples (phones). Now, if we imagine that in the language of this study the frequency of the phoneme X is 5 per cent, it would mean that each position only has approximately 20 samples. This sample size is not very large for drawing statistical inference. This problem can be resolved by creating a grouping or a categorization method which is profound and logical. This is the reason for the categorization of the phonemes.

Individual phonemes are mainly based on the categorization in the perception of speech sounds (Kuhl 1991, Repp 1984, Savela 2009). The perception of speech sounds is close to the categorization based on acoustic features. The categorization with acoustic features leads to similar categories with perception with consonants, but with vowels the individual phones may fall into a different category if done only with acoustic analysis. This effect is discussed in detail later in chapter 3.2.2 Vowels and diphthongs. We have to keep in mind that there are numerous possible ways to classify individual phonemes and even numerous opinions on which sounds are graded as phonemes (see e.g. Schneider et al. 2006, Savela 2009). However, in our approach we do not settle for categorization of individual phonemes as mentioned.

One might think of several other ways to categorize phonemes into subgroups, or categories. It might be straightforward, discrete reallocation that is based on the absolute duration of each phone. In this reallocation, we could produce e.g. ten categories which consist of phones divided into ten subgroups

by their durations. This, however, is not very logical, because the reallocation would only be based on the duration which is dependable on factors such as the speaking rate. Another categorization method would involve acoustic features as mentioned with the individual phonemes. This, too, has its problems, such as how to define what the similarity that forms a category is in an acoustic sense. One might ask whether it is sufficient that there are formants or acoustic silences in the phones of the same category.

We have chosen a common phonological method to base the categorization on the motoric movement of the articulatory muscles or on the similarity of the acoustic features of the phones. The phoneme categorization is premised on the similarity of the *articulation movements* and also on the *acoustic similarity* of each phoneme in the same category. Also, the *perception* within each category supports the categorization. For example, Finnish has eight vowels which are recognized with the magnet effect[3] (Raimo et al. 2002, Savela 2009). Of course, our categorization method is controversial, since among the non-plosive consonants there are still very divergent phonemes in this sense. Even in the vowel category one might argue that open and closed vowels should be divided into separate categories. However, this would again decrease the sample size in each category.

The approach we chose is moderate and commonly recognized. The categories we chose are **vowels**, **plosive consonants**, **non-plosive consonants**, and **diphthongs**. We also used their phonemic variants of short and long versions, excluding diphthongs, which are only perceived as long.

Our categorization leads to several positive effects. By keeping the number of categories low we manage to keep the sample size large enough for statistical analysis. On the other hand, by categorization we may use methods that consider segmental durations within individual utterances since it is very likely to have several phonemes from the same category within each utterance, which enables studies that use comparison. Similarly, the study with positions of utterance is likely to have a large sample size on each position. In addition, the categorization provides a commonly accepted way to divide phonemes into subgroups, and it is widely used in phonetic studies in various forms (e.g. Lehtonen 1970, O'Dell 2003). The negative side of the categorization is that there is a possibility that some differences between the phonemes within a category are lost when studying a category as a whole (see chapter 3.2.3 Consonants).

3.2.1 Phonologically long and short

In Finnish there are phonologically short and long vowels and consonants. This means that the three words '*tuli*', '*tuuli*', and '*tulli*' ('fire', 'wind', 'customs') are all different words. These words form two so-called minimum pairs ('*tuli*'

[3] The magnet effect is described in Kuhl 1991.

vs. '*tuuli'* and '*tuli'* vs. '*tulli'*), meaning that the words differ by a single phoneme, and in this case the difference is between long and short phonemes. The quantity system of Finnish is explained in greater detail in Lehiste's study (Lehiste 1965).

In phonetic writing long phonemes have a colon following the phonetic character, for example /l/ is a short phoneme as in the word '*tuli*' and /l:/ is a long phoneme as in the word '*tulli*'. This practice is used in the thesis to separate short and long phonemes.

The difference between long and short is both semantic and physical. In terms of semantics, the choice between a long or short phoneme is significant, because the *meaning* of the word is different. On the other hand, the duration is longer in the absolute sense when spoken in a similar manner. In our study, we calculated the mean duration of phoneme categories as shown in the following table (Paper III). In this study, we omitted the 10 final phones from the data to minimize the effect of final lengthening in the mean duration. The results are shown in the Table 4.

	Mean Duration (SS)	Mean Duration (MS)
Diphthongs	NaN	117.4
Long Vowels	121.6	105.9
Long Non-plosive Consonants	91.9	85.2
Long Plosives	142.4	130.0
Short Vowels	65.1	58.6
Short Non-plosive Consonants	60.3	60.7
Short Plosives	85.3	74.4

Table 4. The table is taken from Paper III. Here we have the mean duration in milliseconds of the phoneme categories in the corpora, both single-speaker and multi-speaker (*SS* and *MS* respectively). In the single-speaker corpora, the diphthongs were not yet separated from the short vowels; thus they were two consecutive short vowels. N.B. The 10 final phones were removed in this study to minimize final lengthening effect.

In our studies, we have noticed that the ratio between the short and long phonemes is between 0.5-0.7. However, Lehtonen (1970, p. 186) recommends using the ratio 0.5 in all positions in the speech synthesizer if the objective is only intelligibility, not naturalness. Lehtonen did find the difference between short and long phonemes to be sometimes even greater than double, e.g. for short vowels between 0.41-0.45 per cent of the long ones (Lehtonen 1970, p. 63). In the Table 5, we show the ratio calculated on the basis of Paper III.

	Category Ratio (Short / Long)	
	SS	MS*
Vowels	0.54	0.50
Non-plosive Consonants	0.66	0.71
Plosives	0.60	0.57

*Diphthongs are included in the long vowels

Table 5. The category ratios in the corpora by categories. The highest ratio is in the non-plosive consonants, whilst the lowest is in the vowels. The data is based on Paper III. N.B. The 10 final phones were removed in this study to minimize the final lengthening effect.

For comparison, in his dissertation Suomi (1980) measured the duration of Finnish plosives /ptk/ in word initial, medial and final positions. The results are shown in Table 6.

	initial	medial	final
/p/	94	96	NaN
/t/	85	85	75
/k/	79	108	NaN

Table 6. Here we have the duration in milliseconds of the short voiceless plosives taken from the word initial, medial, and final positions (Suomi 1980).

	medial	ratio
/p:/	183	0.52
/t:/	184	0.46
/k:/	166	0.65

Table 7. Table presents the duration in milliseconds of long voiceless plosives taken from the word initial, medial, and final position (Suomi 1980). In the right column, we have the short/long ratio of the plosives in the word medial position.

If we calculate the short/long ratio from Suomi's (1980) material, we get 0.54, which is a bit lower than in the SS and MS corpora (0.60 and 0.57 respectively). Suomi's result (as shown in Table 7) may be different due to the fact that he used laboratory speech and very short carrier sentences, e.g. '*Joko ____ luettiin?*' ('*Was ____ read already?*'), in which the measured target word was placed on the line. As will be discussed later, the penultimate position of the target word may have been affected by the final lengthening and therefore the results may differ. In a methodologically similar study by Lehtonen (1970, p. 71), the short/long ratio in plosives is 0.52[4], which is even lower.

[4] Cf. the short/long ratio derived from Lehtonen (1970) was calculated by using a mean value of single phonemes, and therefore it is not an exact mean ratio. Lehtonen did not include sample sizes in the table.

3.2.2 Vowels and diphthongs

In Finnish we have eight vowels /ɑeiouyæØ/ (see e.g. Raimo et al. 2002, Savela 2009). As in other phonemes, the vowels exist in long and short forms. Diphthongs are contour vowels that glide between two vowels. They are considered as long phonemes in our studies. The location of the phonemes is not universal or inter-lingual but depends on the vowel system of the language. The location means the phoneme's location in the vowel space[5] (F1-F2 plane), but it also refers to the position in which the vowel is created, i.e. the tongue position in the mouth. The vowel system means that certain formant pairs of the first two formants of a vowel may be perceived as different vowels in different languages. The difference in perceiving vowels can be seen with the magnet effect (Kuhl 1991, Raimo et al. 2002). A study by Peltola et al. (2003) compares the vowel discrimination between some of the Finnish and English vowels. In Figure 4, we have the Finnish vowel system in the F1-F2 vowel space (Paganus et al. 2006, Savela 2009).

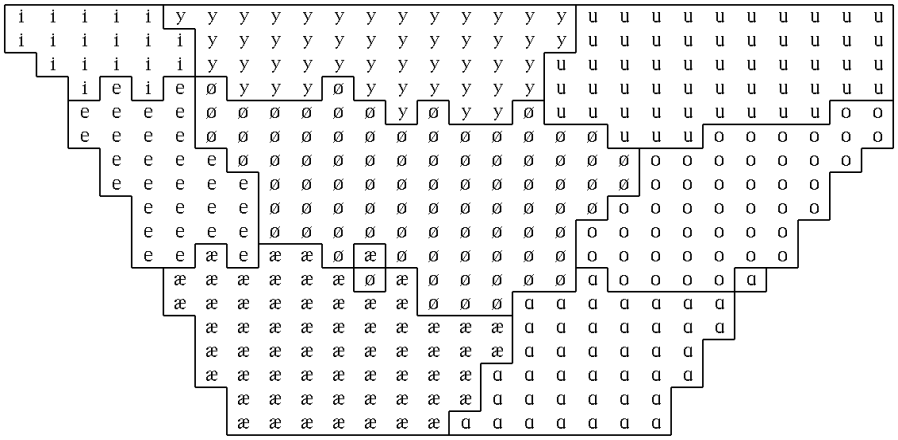


Figure 4. The figure shows the distribution of Finnish vowels in the F1-F2 vowel space. F1 is the Y-axis with origin in the right top corner. The F1 axis goes between 350-860 mels and the F2 axis between 700-1780 mels. The figure is taken from the Turku Vowel Test (Paganus et al. 2006, Savela 2009).

In the Figure 5 we see the Swedish vowel system measured from Finnish Swedish-speaking people. Mainly the vowels are very similar, which can be explained with the close contact of these two languages in Finland.

[5] The vowel space is a planar space induced by the two lowest formants F1 and F2. For more information, see e.g. Paganus et al. 2006. In their study they refer to the vowel space as a vowel chart.

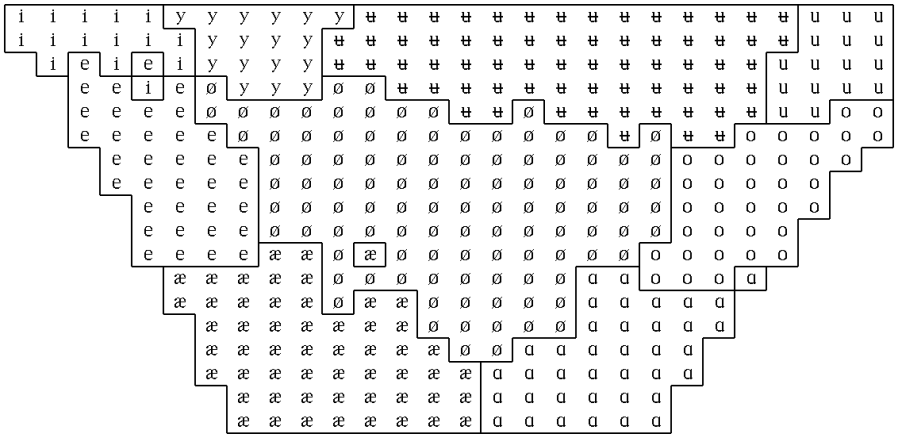


Figure 5. For comparison, here we have the Swedish (Finland) vowel system illustrated in a similar manner. Notice the new vowel /ʉ/ which is partly located on top of the Finnish vowels /y/ and /u/. The figure is taken from the Turku Vowel Test (Paganus et al. 2006, Savela 2009).

In our first studies we did not use diphthongs as an entity. Instead we divided the diphthong into two consecutive short vowels. Making this division was very difficult even for a trained phonetician and we noticed that the task was too subjective. In the end, the diphthong does not actually have two acoustically separated short vowels but a glide between two vowels. So, we decided to use diphthongs as a category of their own.

The mean duration of vowels in the material used is shown in Table 8. The material used contains unaccented vowels and diphthongs from the single-speaker and multi-speaker corpora. The vowels, as well as the other phonemes, are studied in detail later in the section of boundary-adjacent effects.

	MS			SS		
	Duration	Std. Deviation	N	Duration	Std. Deviation	N
Diphthongs	120.4	37.36	1753	137.4	35.50	2235
Long Vowels	114.0	39.07	1451	130.0	35.67	1936
Short Vowels	62.5	24.05	11188	66.7	24.77	14731

Table 8. The table shows the mean segmental durations of unaccented vowels in milliseconds. In the right column there is the single-speaker data. The diphthongs tend to be longer than long vowels and therefore it is justified to treat them as long speech segments. The table contains previously unpublished results.

3.2.3 Consonants

The consonants are divided into categories of voiceless plosives /ptk/ and non-plosive consonants /rsdhjlvnŋm/. The phonemes /fgb/ exist in written Finnish especially in loan words, but it is controversial whether they are a part of Finnish phonology. Finnish also induces the phoneme /ʔ/ (glottal stop), which is not written, but occurs between such word combinations as '*anna olla*' ('let it be'). However, the glottal stop's status as a phoneme is controversial. For more information about the glottal stop in Finnish, see Lennes et al. (2006). The phoneme /ŋ/ is not written in Finnish, but it occurs in certain places such as in the word '*onki*' ('hook and line') where the 'n' is actually pronounced as /ŋ/ because of the following /k/. Similarly, in the word '*Helsingin*' ('Helsinki's') the letter pair 'ng' is pronounced as a long phoneme /ŋ:/.

The voiceless plosives behave similarly among themselves, but very differently when compared against non-plosives. This means that all three plosive sounds include an occlusion stage, where the airflow in the vocal tract is completely stopped, and therefore the sound includes a short period of acoustic silence. In the initial position of an utterance, the plosive has almost no acoustic duration, because it is visible only in the release stage (after the occlusion is cleared) and in the formants of the following speech sound. The occlusion stage is not heard since it is mixed with the respiratory pause (or what ever is the cause for the pause) that precedes the utterance. Table 9 shows the mean durations of unaccented plosives.

	MS			SS		
	Duration	Std. Deviation	N	Duration	Std. Deviation	N
/k:/	142.5	40.7	131	162.1	38.8	188
/p:/	134.8	29.8	57	178.5	43.5	41
/t:/	139.1	34.8	520	150.7	36.4	622
/k/	80.8	24.1	1543	93.0	25.9	1357
/p/	87.0	22.3	484	100.5	22.9	703
/t/	77.5	24.2	2590	88.7	25.6	3349

Table 9. The table contains the intrinsic mean durations of the unaccented voiceless plosives in milliseconds in both corpora (single-speaker in the right columns). In the multi-speaker corpus the phoneme /k:/ is the longest of the long phonemes, whilst the /p/ phoneme is the longest one among the short phonemes. In the single-speaker corpus /p/ is again the longest, but in the long sounds the /p:/ is the longest. However, there are less than a hundred samples in each corpora of the phoneme /p:/. The table contains previously unpublished results.

The non-plosive consonants are a group of acoustically variable speech sounds. All phonemes in this group could be categorized into different

subgroups, e.g. the phonemes /v/ and /j/ are called semi-vowels or glides because they have acoustic characters similar to the vowels (formant patterns), and the phonemes /l/ and /r/ are liquids (/r/ an alveolar trill as well), whilst the phonemes /mnŋ/ are nasals (the sound is formed in the nasal cavity). However, we have mainly studied the category of the non-plosive consonant as a whole. In the tables 10 and 11 we show the intrinsic mean durations of the unaccented non-plosive consonants.

	MS			SS				
	Duration	Std. Deviation	N	Duration	Std. Deviation	N		
/m/	67.0	19.9	1042	67.0	18.7	1260	nasal	labial
/n/	57.0	20.6	2289	52.1	15.9	3515	nasal	dental
/ŋ/	56.1	17.3	190	51.8	17.7	160	nasal	velar
/d/	48.6	15.1	302	51.7	17.2	426	voiced plosive	alveolar tap
/s/	73.8	25.2	2389	78.5	23.7	3047	fricative	alveolar
/v/	55.9	18.6	820	59.8	15.6	1211	approximant	labial
/l/	52.0	15.6	1175	55.8	15.1	1621	approximant	alveolar
/j/	57.9	23.9	766	61.7	19.6	855	approximant	palatal
/r/	62.7	22.5	846	61.5	15.8	999	trill	alveolar
/h/	61.5	19.8	643	68.3	22.4	963	fricative	glottal

Table 10. The table shows the intrinsic mean durations in the category of short non-plosive consonants in milliseconds. In the left column is the multi-speaker corpus. The table contains previously unpublished results.

	MS			SS				
	Duration	Std. Deviation	N	Duration	Std. Deviation	N		
/m:/	87.2	22.6	145	101.4	21.7	140	nasal	labial
/n:/	78.9	23.4	166	93.0	28.2	198	nasal	dental
/ŋ:/	74.5	15.5	9	104.5	33.8	39	nasal	velar
/d:/	84.4	57.7	2	223.1	-	1	voiced plosive	alveolar tap
/s:/	124.7	37.2	272	146.3	38.3	294	fricative	alveolar
/v:/	86.7	16.3	23	-	-	-	approximant	labial
/l:/	76.1	26.8	402	85.1	26.5	563	approximant	alveolar
/j:/	-	-	-	-	-	-	approximant	palatal
/r:/	101.6	50.7	14	103.1	41.2	24	trill	alveolar

Table 11. The table shows the intrinsic mean durations in the category of long non-plosive consonants in milliseconds. In the left column is the multi-speaker corpus. Some long non-plosive consonant phonemes are extremely rare or non-existing (see /d:/ and /j:/). The table contains previously unpublished results.

As we have seen, the non-plosive consonants, too, have long and short variants. However, in certain position in Finnish it is impossible to have a long variant, e.g. in the beginning of a word there can only be a short variant of a consonant. Therefore, an utterance always has a short consonant in the initial position. On the other hand, some long variants are very rare, if not impossible

in the linguistic sense. For example, long non-plosive consonants /j:/ and /v:/ are not known. However, some phonemes may exist in the physiological sense as they are perceived as long phonemes. For example, in the word '*vauva*' ('a baby') one may be fooled to think that there is a double phoneme /v:/ in the latter position ('*vauvva*') due to the previous /u/ that is pronounced in a position close to the following /v/. Therefore, these two phonemes /u/ and /v/ are a form of a similar gliding pair as two vowels in diphthongs.

Current results on the mean durations of phonemes offer an overview to the inter-phoneme duration relationship. On the other hand, the results demonstrate the main durational differences between the multi-speaker and single-speaker corpora. The next chapter will extend the operation of segmental durations to the boundaries of utterances.

3.3 Boundary-adjacent effects

A boundary-adjacent effect refers to a phenomenon that occurs in the vicinity of a phonetic boundary. A phonetic boundary may mean several different edges in the speech flow, e.g. final syllable of an utterance or initial phone of a stressed carrier word, but hereon the phonetic boundary only refers to the edges of an utterance and by edge we mean the beginning or the end of the utterance. As defined before, the utterance here refers to a speech segment that is limited at each end by an acoustic silence, excluding plosive consonants or other voiceless phonemes that include acoustic silence (i.e. glottal stops). Acoustic silence may be caused e.g. by a respiratory pause or a linguistic boundary (end of a sentence) or both, or it can even be a hesitation pause.

The boundary-adjacent effects on segmental durations have been discussed widely. Byrd, Saltzman and colleagues (Byrd, 2000; Byrd, Kaun, Narayanan, & Saltzman, 2000; Byrd & Saltzman, 2003) have described a conceptual approach to boundary-adjacent lengthening (or slowing). In this approach, phrase boundaries are instantiated by so-called prosodic gestures having an extent in time and a dynamic activation level trajectory. The prosodic gesture is centered on the phrase edge and acts to slow down all concurrently active articulatory gestures in proportion to the activation level of the prosodic gesture. The prosodic gesture's maximum level of activation, in turn, is determined by the prosodic boundary strength. Byrd and Saltzman (2003) point out that the prosodic gesture approach has several implications for the patterning of speech. It predicts final lengthening, as the prosodic gesture is centered on the final phrase edge, but it will also predict initial lengthening to the extent that the prosodic gesture overlaps the following initial phrase edge. In addition, it predicts that the degree of lengthening will increase as the boundary is approached, be highest at the boundary, and decrease again as the boundary recedes.

White (2002) studied the boundary-adjacent effects in English in his dissertation. He called the effects near boundaries "domain-edge processes". White studied whether there actually are domain-edge processes, or whether the effects actually result from domain-span processes. He writes: "*Domain-span processes are hypothesized to arise from an inverse relationship between the size of some constituent and the duration of some subconstituent: for example, word-span compression (polysyllabic shortening) and utterance-span compression.*" (White 2002). He does not find evidence of domain-span processes, but suggests that there are domain-edge processes which are responsible for the boundary-adjacent effects. We have also studied this inverse relationship in Finnish and found evidence supporting White (Hakokari et al. 2008) who studied English.

Another interesting study on Brazilian Portuguese made by Barbosa & Madureira (1999) suggests that boundary-adjacent effects exist to separate weaker phrases from each other. This leads to a speaking rhythm that is similar to a hierarchical model of rhythm in Finnish proposed by O'Dell & Nieminen (2001). In an older study Oller (1973) studied the boundary-adjacent effect in English in various intonational patterns and found evidence of both initial and final lengthening. His findings support the "boundary cue" or "juncture cue" theory which suggests that the boundary-adjacent effects are only present to point out the edges. The boundary cue is also supported by Duez (1993), who found that lengthening alone at a boundary can produce a sensation of the speaker temporarily pausing. However, Lindblom has argued that final lengthening exists due to the intensity drop in the end in Swedish (Lindblom 1968), a phenomenon also known in Finnish and English. Lindblom (1968) writes: "*Granted the assumption of the energy per syllable being constant final lengthening of segments becomes a consequence of the intensity being lower in the final part of the basic phrase contour*". This is supported by Öhman (1967) who mentions that the speech gestures tend to relax and therefore slow down towards the final utterances. Öhman's view is supported by Fougeron & Keating (1997) who found that for American English most vowels have less linguopalatal contact in domain-final syllables compared to domain-initial and domain-medial. However, at the same time they report that initial consonants exhibit more linguopalatal contact than domain-medial or domain-final consonants, which might suggest that the initial and final effects operate differently, in opposition to what Barbosa & Madureira (1999) supposed. Similarly, Cho & Keating (2001) found initial lengthening and strengthening to correlate, which is in opposition to the views of Fougeron & Keating (1997).

3.3.1 Utterance initial effects

The initial positions in an utterance have the highest fundamental frequency by default in Finnish. Therefore, it is safe to assume that the first phones of an utterance are accented in this sense. The accent does not automatically lead to

prominence of the first word, since accent is a relative change in the speech flow of acoustic features such as fundamental frequency, intensity, and duration. In our studies, all these three must be higher or longer to establish prominence.

Unlike final lengthening which is found in most studies, domain-initial processes have produced conflicting results. Kaiki et al. (1990) reported initial shortening in Japanese, but the corpus used in their study was criticized for being imbalanced (Campbell 1992). Nagano-Madsen (1992) has found initial shortening in the Eskimo language, while Hansson (2003) has found it in Southern Swedish (syllable duration at word level). Still, initial lengthening is found in several languages and studies. To name a few, Zu & Chen (1998) and Cao (2004) reported initial lengthening in Chinese, Chung et al. (1999) in Korean, and Oller (1973) and White (2002) in English. However, White (2002) found that some speech sounds are actually shortened.

In our study we showed the segmental durations in utterance initial position on a phone level with both corpora (Paper III). In this study, we used the phoneme categories to find out if Finnish has initial shortening or lengthening. The results were intricate; while some categories showed a clear lengthening (vowels and diphthongs), some showed shortening (long plosives). In addition, the non-plosive consonants showed no effect in either direction. But if there was an effect in the initial positions, it only involved one or two of the first positions, making it very short. We can assume that initial effect is only visible in the first syllable, if at all on syllable level. For example, the first phone is a lengthened vowel and the second is a shortened long plosive. However, the effect is remarkably similar in both corpora. In the figures 6 and 7, we have the vowels and plosives in the single-speaker corpora taken from Paper III.

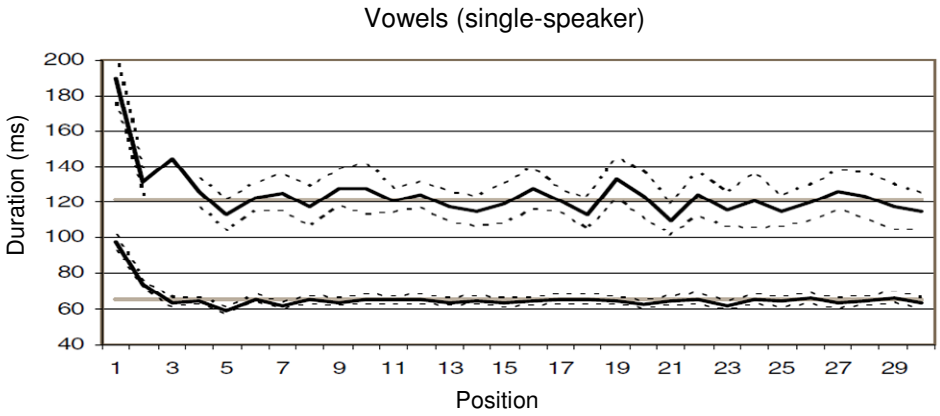


Figure 6. The long (top-most line) and short vowels (bottom-most line) in the single-speaker corpus with confidence limit ($p \leq 0.05$, dashed lines). The figure is derived from Paper III.

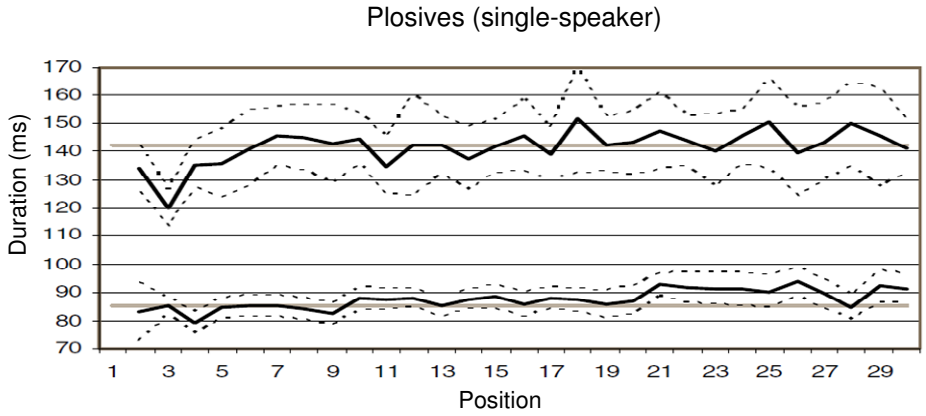


Figure 7. The long (top-most line) and short voiceless plosives (bottom-most line) in the single-speaker corpus with confidence limit ($p \leq 0.05$, dashed lines). Please note that plosives were not measured in the first position. The figure is derived from Paper III.

The study of initial positions left the non-plosive consonants without a clear lengthening or shortening (Paper III). Therefore we needed an elaborated study of this phoneme category. In our next study on initial positions we took a closer look at the category of non-plosive consonants (Paper IV). In this study, instead of phoneme categories, we used phonemes as the level of detail. The results were, again, threefold. The phonemes /r/ and /s/ were lengthened in the first position in both corpora. The phoneme /m/ was shortened in the first position in both corpora, but the phoneme /n/ was only shortened in the multi-speaker corpora (in the first position). The phoneme /j/ was the only approximant which was clearly shortened in the first position, or actually in the two first positions in the single-speaker corpus. The other approximant /v/ was shortened in the first position only in the MS corpus, whilst the phoneme /l/ showed no effect at all. The phoneme /h/ was shortened in the first position in the SS corpus and in the two first positions of the MS corpus. Again, the effective distance of the boundary remained very short; only one or two positions (see Paper IV). Yet again, this effect might have been lost if the measurements would have been done using a syllable-level.

In our next study on final lengthening, we also looked into the matter of initial shortening (Paper V). In this study we conducted two experiments. In the first one we counted or rank-ordered the relative articulation rate of each word in an utterance (in 2-5 word utterances). Rank-ordering means here that the words in the same utterance are put in order by their articulation rate, meaning that if in a 2-word utterance has the latter word articulated in slower manner than the first then this utterance adds one to the last position into the column of two word utterances. This does not only give the number of final lengthened utterances, but also the initial shortened/lengthened utterances on a word level. The results

showed that 85-49 per cent of the utterances had the first word spoken faster than any other word in the utterance (initial shortening). On the other hand, 0-15 per cent of the 2-5 word utterances had the initial word spoken slower than any other word in the utterance (initial lengthening). The experiment used a combined corpus of both single- and multi-speaker corpora. The results are shown in the Table 12.

	2-word (n=240)	3-word (n=255)	4-word (n=281)	5-word (n=244)
1st word	36	17	0	2
2nd word	204	26	23	5
3rd word		212	42	18
4th word			216	36
5th word				183
IS %	85.00	63.13	55.87	48.77

Table 12. The table shows the results of the articulation rate of 2-5 word utterances. Of the two-word utterances (n=240) 204 had the last word spoken slower than the first word, or similarly, in the three-word utterances 26 utterances had the medial word spoken the most slowly. Initial shortening (IS) was found in 85-49 per cent of the utterances, whilst the chance level would be between 50-20 per cent. The table is taken from Paper V.

For the next experiment presented in Paper V we calculated the actual articulation rates and sketched them for comparison. This experiment showed that almost all 2-9 word utterances had initial shortening with statistical confidence. The Figure 8 shows an example from Paper V.

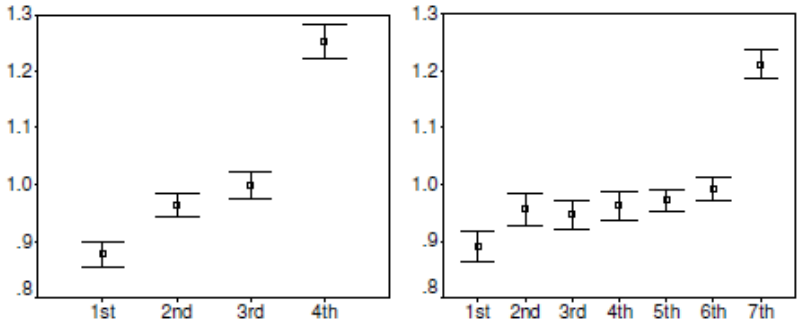


Figure 8. The table shows the articulation rates of four (left) and seven-word utterances. In the Y-axis we have the relative articulation rate (1.0 being the mean articulation rate for all of the utterances), and the slower the rate, the higher the number. The X-axis shows the words from the utterances in respective order. Both figures show that the first word of the utterance is articulated faster than the other words in the utterance with confidence limit ($p \leq 0.05$). This figure was taken from Paper V.

In our unpublished study of accent and final lengthening we studied the effect of prominence on final lengthening. The method used was similar to the previous study (Paper V) with the exception that the prominent words were annotated. At the same time, the MS corpus was extended. The results show that the first word was very rarely prominent. There were 1,203 prominent words of which only 5 were in the initial position. In addition, the results show remarkably similar evidence that support the presence of initial shortening. The unpublished results are shown and discussed under the next topic of final lengthening.

3.3.2 Final lengthening

In our previous study we showed that Finnish may exhibit final lengthening (FL) since speech segments show lengthening in the final and prepausal words (Paper II). The study was focused on mean durations of phoneme categories in prepausal (final) words, penultimate words, and words positioned elsewhere, and it used the single-speaker corpus. The results show that in all phoneme categories there was a clear FL which extended to the penultimate words as well. There were also differences between phoneme categories: the phonemes were 23-51 per cent longer in final words than in medial words. In addition, the penultimate words had 3-6 per cent longer phonemes than medial words.

In 2007, we studied the occurrence tendency of FL in Finnish on a word level (Paper V), and found that it is, surprisingly, on the same level as in the Swedish language as studied by Hansson (2003). Swedish is Nordic Germanic language whilst Finnish is Finno-Ugric language and therefore the languages are not related. In our study we used a combined corpus of single-speaker and multi-speaker corpora. The experiment showed that the material had final lengthening in 75-85 per cent of the 2-5 word utterances, meaning that the last word was spoken more slowly than any other word in the utterance in most of the utterances. Using Hansson's method, in which the last word is only compared against the penultimate word, the result was 85-89 per cent. Hansson's result was 71-82 per cent of FL in Swedish (Hansson 2003). In comparison, Mihkla (2005) reported 60 per cent of FL in Estonian, a language closely related to Finnish, which is a surprisingly low figure. However, we did not find the details of the method Mihkla used so we were unable to compare the results in greater detail.

In another experiment in the same paper, we showed that FL is a statistically relevant phenomenon on the word level in Finnish (Paper V). The experiment in question showed the distribution of mean durations on the word level in 2-9 word utterances with confidence limits ($p \leq 0.05$). The results show that the final word has significantly longer mean duration than any other word in the utterances. This study, as far as we know, was the first paper to reveal FL in

Finnish. This was a remarkable study, since it has been widely debated whether the quantity languages have FL. Previously, it has been shown that other quantity languages have FL, but Finnish remained an unclear case. However, there are some misunderstandings about FL in Finnish, e.g. Vaissière (1983) claims that Finnish has "little (if any) final lengthening" by referring to Lehiste (1965) who does not study segmental durations at all, but syllabic structure and the function of phonological length. Therefore, Finnish is usually named as a language with no FL and being the last outpost against the universality of FL effect. Our study shows for the first time that FL is statistically significant in Finnish, and the universality of the effect becomes possible.

In our unpublished study, we used similar methodology as in the previously discussed papers (Paper V, Hansson 2003), with the exception that in this study we had annotated the prominent words. This made it possible to separate FL in unaccented and accented words. In addition, the multi-speaker corpus was extended with new material. The results show that the prominence is more common in the final position than in any other position. However, the non-prominent final words are still very likely to embody FL. Surprisingly, the prominent words in utterance final position do not embody FL as often as the non-prominent, and similarly, the prominence in a word did not mean that the word would directly be the most slowly articulated in the utterance. Only 10-26 per cent of the prominent words which were articulated in the slowest manner were in final position in 2-9 word utterances, whilst among the non-prominent words the result was over 70 per cent. The results are shown in Tables 13 and 14.

	2-word	3-word	4-word	5-word
N	251 (62)	274 (104)	307 (157)	264 (173)
1	37 (0)	24 (1)	20 (0)	12 (0)
2	214 (55)	32 (5)	22 (2)	17 (4)
3		218 (35)	33 (4)	18 (2)
4			232 (43)	19 (1)
5				198 (35)
FL %	85.3	79.6	75.6	75.0
PFL %	25.7	16.1	18.5	17.7

Table 13. Here N is the number of utterances, the number in parentheses is the number of prominent words in these utterances, and the first column on the left shows the position of a word in respective utterance. Similarly, e.g. on the first line in the column for three-word utterances we see the number (24) of the three-word utterances that had the first word articulated in the slowest way. The number in parentheses (1) means that only one of these was prominent. The results in this table have not been previously published.

	6-word	7-word	8-word	9-word
N	254 (215)	213 (207)	143 (152)	107 (131)
1	13 (0)	14 (1)	3 (0)	1 (0)
2	21 (1)	16 (3)	7 (0)	2 (0)
3	24 (6)	16 (3)	7 (1)	9 (0)
4	20 (5)	12 (4)	15 (5)	9 (1)
5	18 (3)	9 (2)	14 (1)	4 (2)
6	158 (24)	10 (1)	8 (0)	7 (2)
7		136 (16)	12 (1)	6 (0)
8			77 (11)	8 (3)
9				61 (6)
FL %	62.2	63.8	53.8	57.0
PFL %	15.2	10.1	14.3	9.8

Table 14. Here N is the number of utterances, the number in parentheses is the number of prominent words in these utterances, and the first column on the left shows the position of a word in respective utterance. The results in this table have not been previously published.

In Table 13 the FL percentage is between 75 and 85, meaning that the last word is articulated in the slowest manner in most of the utterances. The prominent word was found in the final lengthened word in 16-26 per cent of the cases (prominent word with Final Lengthening, PFL %). In Table 14 we notice once again that there is a remarkable amount of FL found in the material. In more than half of the long utterances the final word is the most slowly articulated word in the utterance. The prominence was found in the final lengthened word in 10-15 per cent of the cases (prominent word with Final Lengthening, PFL %).

In 2008 we studied a normalization method for segmental durations and used the method to demonstrate FL in Finnish, with the purpose of clearing out any possible misunderstandings that Finnish does not have FL (Paper VI). In this paper, we use a simple normalization method that reduces inter- and intra-speaker variation which can cause skewness and bias to boundary-adjacent effects. The study utilized a combined corpus of both single-speaker and multi-speaker corpora. In addition, as before, the study involved phoneme categories, but this time we used a phone-level approach to describe the FL in each category. The phone-level approach gives detailed information about how an articulation rate behaves in the final positions of utterances. The results support the previously published papers. An example of the results is shown in Figure 9. In the figure, the Y-axis shows the mean duration of each position using a relative duration, in which 1.0 would be the mean articulation rate in the same utterance. The values over one were articulated in a slower manner, e.g. a value of 1.1 for some position Y in a phoneme category X means that in this position the phones are 110 per cent of the mean duration of the same phonemes in the same utterance. Please note that the FL effect is included in the mean duration,

making it higher than the mean duration would be without the FL. The X-axis shows the position of phones in reverse order, e.g. value 1 means that it is the final position of utterances.

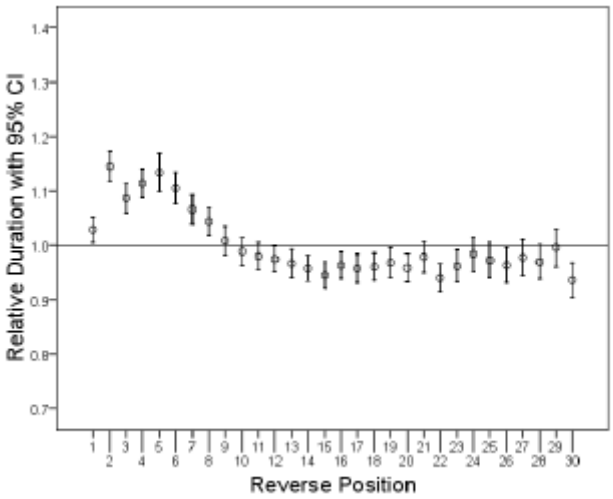


Figure 9. The figure shows the mean durations in final positions of utterances in the category of short vowels. The final lengthening extends from the final position to the 7th reverse position. The figure is taken from Paper VI.

Recently Nakai et al. (2009) claimed that they had *established* FL in Finnish by using laboratory speech in their study of Northern Finnish. They only refer to our earlier Paper II after which we have made several studies on this matter as discussed earlier. However, we have used a totally different methodology and approach, as shown.

In our unpublished study, we ran a complete statistical analysis of the data (ANOVA). The purpose was to find out whether segmental duration is affected by the corpus, the phone's environment in utterance (final, penultimate, or other), and/or the phonemic category. We treated segmental duration as the dependent variable and environment, phonemic category, and the corpus as independent variables. The effect of environment was studied post hoc with Tukey's Honestly Significant Difference (abbreviated *Tukey's HSD*). The idea of the HSD is to ensure that the groups we created have significantly different means. Before the test we had ensured that the data was normally distributed, independent, with equal variations. The results showed that position, category, and corpus had a significant effect. There was also a significant effect in the interaction of position and corpus, category and corpus, and in the interaction of position, category and corpus. These results are part of a journal paper in preparation; therefore some results are left unpublished here. The results from Tukey's HSD post-hoc test on word position are listed in Table 15.

Tukey's HSD	(I) Position in Utterance	(J) Position in Utterance	Mean Difference (I-J)	Std. Error
	Penultimate	Other	5.5*	0.2
	Final	Other	22.9*	0.3
	Final	Penultimate	17.3*	0.3

Based on observed means.
The error term is Mean Square (Error) =.001
*The mean difference is significant at the .05 level.

Table 15. Tukey's HSD post-hoc test on position in milliseconds (Mean Difference and Std. Error), both corpora included.

The results show that there is a gap of almost 23 ms in the mean segmental durations between the final and the medial/initial words. Again, there is even a significant difference of more than 5 ms between the penultimate words and the words positioned otherwise. The gap of 23 ms is quite long, keeping in mind that the mean durations in the categories varied from 60.3 (short plosives in single-speaker corpus) to 142.4 (long plosives in the single-speaker corpus), see Table 4.

In our other study in preparation, therefore unpublished, we studied FL on word level using a similar approach as in Paper II, but used a word level instead of a phone level (cf. the previous Figure 9). In this study we calculated the mean articulation rate for each word in 2-5 word utterances, once again using the combined corpus of both single-speaker and extended multi-speaker corpora. In addition, the corpus had prominence annotated, which helps to separate FL from the prominent words.

Tables 16 and 17 present the results. To separate the effect of accent we exclude the prominent words from the mean. The prominent words are used to calculate a separate mean of prominence for each position, but the confidence interval was left out due to the small number of prominent words. An average articulation rate of the utterances has the coefficient 1.0, and a value higher than that means slower articulation rate.

	2-word		3-word		
Word position	1	2	1	2	3
CI, high	0.98	1.25	0.94	1.02	1.28
Mean	**0.95**	**1.22**	**0.91**	**1.00**	**1.25**
CI, low	0.93	1.18	0.89	0.97	1.21
Prominent Mean		**1.21**		**0.96**	**1.19**

Table 16. In the table we see the mean articulation rate of words in all two- and three-word utterances ("Mean" row). The confidence interval was calculated with $p \leq 0.05$, which is shown on the line "CI, high" that is the upper interval and "CI, low", the lower interval. The words are positioned in an increasing order from left to right, the final word being the one on the far right. The table contains previously unpublished results.

From Table 16 we notice that in the two-word utterances there were no prominent words in the first position. The final word (coefficient 1.22) was clearly articulated more slowly than in the initial position (coefficient 0.95). The prominent words have a faster articulation rate than the average words in the position, but the coefficient is within the confidence intervals ($p \leq 0.05$). Again, in the three-word utterances there were only one prominent word in the first position, but it was not included since it is not a proper mean value. Please note that the mean articulation rate of the non-prominent words is actually slower than that of the prominent ones. Moreover, in both positions of the prominent words the articulation rate coefficient is below the confidence intervals ($p \leq 0.05$). The mean articulation rates show a clear FL in the final word.

	4-word				5-word				
Word position	1	2	3	4	1	2	3	4	5
CI, high	0.92	0.99	1.03	1.29	0.89	0.97	0.97	1.01	1.25
Mean	**0.90**	**0.97**	**1.00**	**1.25**	**0.87**	**0.94**	**0.95**	**0.99**	**1.22**
CI, low	0.88	0.94	0.98	1.22	0.84	0.92	0.93	0.97	1.19
Prominent Mean		**1.01**	**0.98**	**1.22**		**0.94**	**0.96**	**0.94**	**1.25**

Table 17. In the table we see the mean articulation rate of words in all four- and five-word utterances ("Mean" row). The confidence interval was calculated with $p \leq 0.05$, which is shown on the line "CI, high" that is the upper interval and "CI, low", the lower interval. The words are positioned in an increasing order from left to right, the final word being the one on the far right. The table contains previously unpublished results.

Again, in Table 17 there were no prominent words in the first position of the four-word utterances. This time, the prominent words in the second position have a slower articulation rate than the average rate in that position, though the prominent words have a faster articulation rate in other positions. The mean articulation rates show a clear FL in the final word. In the five-word utterances, there were no prominent words in the first position. The prominent words have similar articulation rates with the mean articulation rates in respective position, but in the final position the articulation rate is slower with prominent words. The mean articulation rates show a clear FL in the final word.

The previous unpublished Tables 16 and 17 show that the final word of the utterances between 2-5 words was articulated more slowly than the medial and initial words. The figures also show that the prominent words, too, are under the FL effect. In addition, the results suggest that Finnish has initial shortening even though it is not as strong as FL.

Current results confirm that Finnish has final lengthening that is by no means an effect of prominence. As discussed earlier, Vaissierè (1983) mentioned Finnish, Estonian, and Japanese as having little or no final lengthening. Since all three languages are quantity languages, it was thought that FL would somehow disturb the quantities. Therefore the quantity languages should resist final

lengthening. However, the presence of FL has been shown in Estonian by e.g. Krull (1997) and Mihkla (2005) even though the language has three opposing quantities (cf. Finnish has two; short and long phonemes). Similarly, FL has been shown in Japanese by Takeda et al. (1989). This has left Finnish as the last quantity language without final lengthening, and therefore FL could not previously be considered a universal phenomenon in languages. We have now shown that the universality of FL is still possible.

The tables 13-17 (excl. 15) show an interesting difference between the prominent and non-prominent words, especially in the last position. In the utterances of 2-4 words the final words that are prominent are actually articulated faster than the non-prominent words. Similarly, the figures show no evidence that prominence would lengthen or shorten the segmental durations in other positions either. White and Mády (2008) found that Hungarian, which is a language related to Finnish, has no lengthening effect that is caused by the prominence. In addition, they found that Hungarian has final lengthening in prominent words just as we have shown here. Suomi et al. (2002) studied Finnish with laboratory speech and found that strong accent produced higher lengthening than other degrees of accent. They also found that contrastive focus causes lengthening. However, we must keep in mind that there was no strong accent in our material. We must also point out that their material did not consider the effect of penultimate word position in their carrier sentences. Here we have an example of their carrier sentence with a strong accent present in the target word '*kato*': '*Sanoin että KATO pelotti, en sanonut että KATTO pelotti*' ('I said that DEARTH frightened (me), I didn't say that the CEILING frightened (me)'), (Suomi et al. 2002). It is possible that the target word is lengthened because of the penultimate position (comma usually creates an acoustic pause and breaks the sentence into two utterances, just as strong prominence may do), not merely due to the contrastive focus. In addition, if we calculate the distance from the end of the utterance we notice that the last phoneme /o/ of the word '*kato*' is in the position 7 (starting from the end) which is very close to the FL influence boundary. FL ends for example in the 8th position for short vowels (see Figure 9 or Figure 3 in Paper VI). Similarly, our results show that the lengthening extends even to the 8th position of an utterance in the category of short plosives, and in this position is the phoneme /t/ of the word '*kato*' in the carrier sentence (cf. Figure 6 in Paper VI). Therefore it seems that the results of Suomi et al. (2002) may be affected by FL.

Let us discuss the possible origin of final lengthening. As said previously, Lindblom (1968) suggested that FL origins from the intensity drop in the end of utterances in Swedish, but a similar effect is also present in Finnish. His idea is supported by Öhman (1967) who talks about relaxation of the speech gestures towards the end of utterances. We find both ideas to be close to our own. In our view, FL is due to what we call *the effective breathing rhythm*. Breathing must be in rhythm with speaking; otherwise the speech would be constantly interrupted by inhaling, and probably by exhaling. Therefore the

speech is divided into utterances that are in rhythm with breathing. In between every utterance the speaker inhales and fills his/her lungs with air. While articulating the next utterance the speaker exhales, which creates an airflow needed for the speech. Now, if the speaker runs out of air in the middle of the utterance the speech is interrupted, and again the speaking style becomes unfavorable and inefficient. This can be avoided by using as little air as possible until the speaker is sure that the utterance is near its end. This time, he/she has plenty of air left in the lungs to be used in the last parts of the utterance. This creates FL as the speaker empties the lungs by lengthening the speech segments.

One could argue that it would be easier to exhale, but this is not the case. This would mean that the speaker has to exhale the rest of the air after the utterance and then inhale. Exhaling and inhaling within the same pause would cause lengthening in the pause duration. A lengthened pause is a cue of turn-taking, and therefore the speaking companion might take the speaking turn. This is one of the reasons why the pause between the utterances must be short.

On the other hand, our approach for the plausible origin of FL includes a hypothesis. The hypothesis is that the utterance must not be planned totally. This means that if the utterance is completely planned the speaker should be able to predict when the utterance ends and plan the usage of the air in the lungs so that it runs out exactly when the utterance ends. This does not mean that if the speaker knows the utterance it should not have any FL since he/she knows what is to be said. The planning in this sense is a much more detailed process of articulatory movements and other bodily air usage that is taking place. We argue that the speaker does not have the possibility to foresee this process in such detail that is required for this kind of planning, especially in conversational speech on which we have focused here.

3.4 Normalization method

Next we will discuss normalization in the use of a natural speech corpus (cf. laboratory speech with designed carrier sentences). In this context normalization means control over such effects that cause variation that is not studied. We know that there are various sources of variability such as the effect of preceding and following phonetic segments (cf. Liberman et al. 1967), stress (e.g. Klatt 1975, Peterson & Lehiste 1960), and syntactic structure (e.g. Klatt 1975, Oller 1973). However, these effects are not what interest us, whilst we are more interested in the effects of utterance length (e.g. Klatt 1973, Oller 1973) and speaking rate (e.g. Miller 1981), since these are the problematic effects in the natural speech corpus. There are several normalization methods especially in speaker recognition (e.g. Goodman et al. 1986, Li et al. 2002), but we have not found any method suitable for our approach.

The natural speech corpus may have high variation in the articulation rate. This can be a problem when studying segmental durations due to the fact

that some utterances may have higher impact on the results. This is possible especially if we only use mean durations. The research aims at capturing the essence of the segmental durations in the corpus and this can be lost with unbalanced and variable articulation rates. Next we will discuss this problem.

Boundaries: To study the boundaries of utterances of various lengths we need to set an imaginary origin. If the study is about segmental durations we set the origin to be the initial phone of an utterance, because every utterance in the corpus has one. On the other hand, if we study the final parts of the utterances we can shift the origin to the final phone of the utterance. Every utterance has a final position, and therefore we can dissect the segmental durations in reverse order starting from the final position.

Various lengths of utterances: Now that we have set the origin we can consider the problems of variable utterance length. Let's imagine we have three utterances (A, B, and C): A has 10 phones, B has 20 phones, and C has 30. Now A influences only 10 positions, whilst C influences 30.

Variable speaking rates: If A is articulated in a very slow manner it will have a large impact on the mean durations in the 10 final positions. This means that even though B and C were articulated with a steady articulation rate it would seem that in the 10 positions where A is influencing the average articulation rate is slower. This could lead to e.g. a false final lengthening effect.

Intra-speaker variation: This means that a single speaker can articulate utterances with varying articulation rate. One utterance may be spoken slowly whilst the next is articulated at a faster pace. Similarly, the different utterances may vary in length.

Inter-speaker variation: Different speakers may have very different articulation rates, which is a similar problem to the intra-speaker variation. In addition, speakers may have a habit to use very long or very short utterances on average. Intra- and inter-speaker variations are similar problems, and in fact from the viewpoint of normalization they are the same.

These problems must be solved so that the study finds the behavior of segmental durations without skewness or bias resulting from variation in the corpus. One solution is to use a normalization method, which eliminates or diminishes the variation. Normalization means that each utterance is scaled to have a similar impact on mean durations.

In our study (Paper VI), we approach the problem on utterance level. In this approach each utterance forms an entity that has its own pattern of segmental durations, and this pattern can be compared to any other utterance. Since every utterance has its own pattern, by using a scale that is uniform between the utterances, the patterns are comparable. The pattern is created on the phone level of the utterance. Each phone is classified into its respective category of long/short plosive, non-plosive consonant, vowel or diphthong. Then the mean duration of each category is calculated, e.g. the mean duration of short vowels in all positions in that utterance is 65 ms. Then the duration of each

phone is divided by the mean value of the phone's respective category, e.g. the phone /a/ in a certain position with the duration of 72 ms is divided by 65 ms, leading to the coefficient of 1.11. After this, the coefficients are used to replace the original durations in the utterance. The same routine is repeated with every utterance. When the coefficients have been calculated the material can be treated as segmental durations, but now the coefficients between utterances are comparable, since each vector of coefficients in an utterance represents the pattern of segmental durations. The method is illustrated in Figure 10.

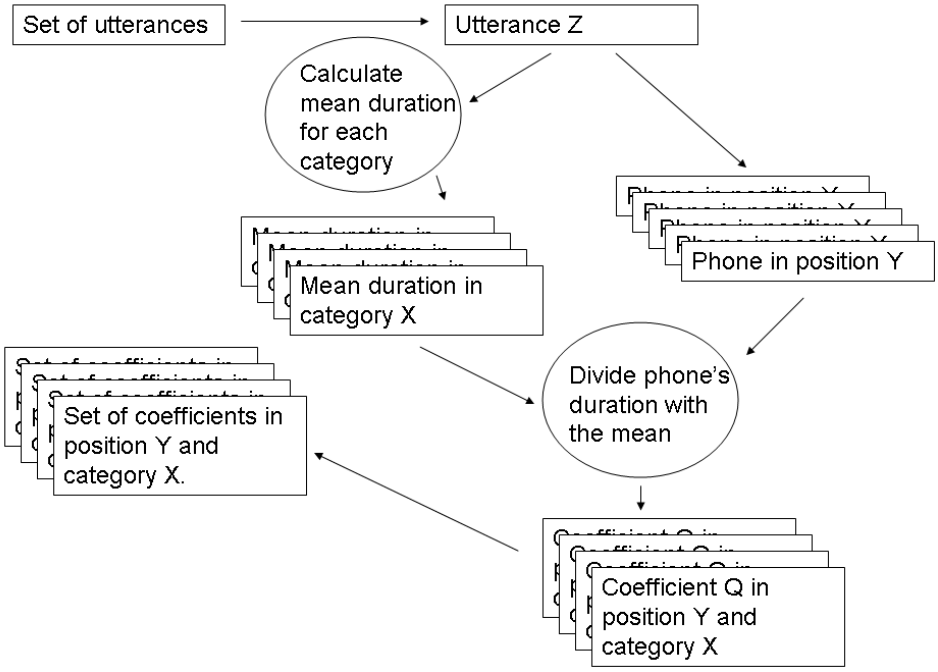


Figure 10. Steps in calculating normalized duration coefficients for each sound category and position. The figure is taken from Paper VI.

The method requires explication. The phonemes are normalized not only within each utterance but also within each phoneme category. The normalization within each utterance is done due to the fact that we want to catch the articulation rate within each utterance. If an utterance has final lengthening it is only visible when comparing the segmental durations within the same utterance. However, we cannot use the duration of each segment to capture FL due to the quantity system, since short and long phonemes would confuse the comparison. So we need a better approach to capture the segmental durations in the utterance. We could use single phonemes, but it is highly possible that a certain phoneme is present in the utterance only once, and there is no point for comparison. On the other hand, we could compare the short and long phonemes against themselves,

but this would lead to high durational variation within the comparison group. Therefore, we use the phoneme categories which were presented in chapter 3.2.

The normalization method helps to reduce variability by capturing the pattern of segmental durations within the utterance. If the utterance has no final lengthening its vector of coefficients has only values close to 1.0 towards the end (if not even less than one). Again, if we have utterances A, B, and C (10, 20, and 30 phones respectively), A will still affect only 10 positions, but if it is spoken very slowly its effect is no greater than the effect of B or C in the positions it represents. This means that the variable length of utterances with variable speaking rates will not cause false final lengthening or initial shortening effects. If one of these three utterances has a great effect of FL and the other two do not, it is visible as one third of its magnitude in the means of coefficients.

The beauty of this method is that it is intuitive and easy to implement, and still it can separate the pattern of segmental durations without complex formulations. However, there is still variability that even this method cannot resolve. For example, the variable length, especially with very short utterances, will only affect the positions they represent. This can only be solved by cutting the material so that the very short utterances are left out. However, the very short utterances can only embody a lengthening or shortening effect within the positions they represent, and it has to take place within the same utterance.

As we have now discussed the matter of studying segmental durations from the viewpoint of speech technology, it is time to take a closer look at the development of speech synthesis.

3.5 Synthesis on a limited platform

Nowadays, a speech synthesizer on a mobile platform is common. Synthetic speech is used to read text messages, addresses, and e-mails. However, the naturalness and intelligibility of synthesized speech do not by any means meet high standards at the moment. The problem seems to be twofold. The loudspeaker output does not achieve the acoustic dynamics of speech and the synthesizers used are not trimmed at the expected levels.

Speech synthesis on a mobile platform can be built in two ways. Computation can be done on the platform or by using a distributed system in which the computation is done on a server in a network. The distributed system is not by definition synthesis on a limited platform, and therefore it is not in our interest. When doing synthesis on a limited platform we encounter several problems. The main problem has previously been the limitations in the computational capacity. Other problems include limited memory and loudspeaker quality. We will skip the loudspeaker issues in this review. The limitations in the computational capacity, as well as with the memory issues, are slowly being removed. The development of the capabilities of mobile platforms

is advancing at a tremendous pace (e.g. Rasmusson et al. 2004, Bertozzi & Benini 2008).

In 2005, mobile phones were not commonly implementing speech synthesizers. At the time, we were developing our rule-based speech synthesizer on a normal PC, but we noticed that it only used limited computational and memory footprint. Soon we began to implement the system on a mobile platform, Nokia 6680 mobile phone (Paper VIII). The system was modified to meet the requirements of Java™ using CLDC 1.1, Connected Limited Device Configuration (Sun Microsystems 2003) and MIDP 2.0, Mobile Information Device Profile (Motorola 2002). At the time, the Nokia 6680 had an average MIDP 2.0 performance according to the result database of the JBenchmark J2ME benchmarking tool (see JBenchmark Home Page 2006). The synthesizer needed some adjustments before it could be implemented. For example, the Java™ used in MIDP 2.0 only supported integers.

The structure of the synthesis was common for rule-based synthesizers (Klatt 1980). The structure is shown in Figure 11. The horizontal line in the figure presents the division between high-level synthesis and signal generation which is the actual formation of the sound signal.

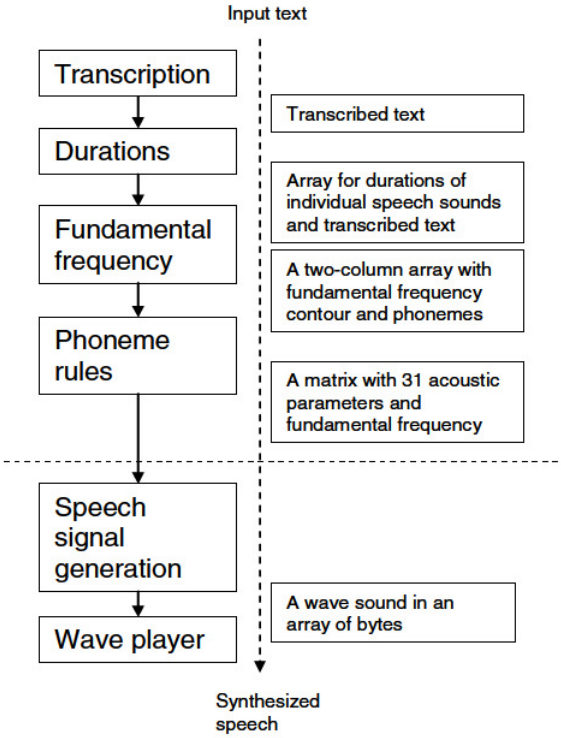


Figure 11. The figure shows the structure of the speech synthesizer and the synthesis phases on the mobile platform. The horizontal dotted line separates the high level synthesizer from the signal generator. The figure is taken from Paper VIII.

The implementation was reviewed by measuring its capabilities for a real-time speech synthesis (Paper VIII). The performance of the synthesizer was tested with six sentences with various lengths (approx. 3-11 seconds). The synthesis procedure was divided into two stages; high level synthesis and signal generation which could be measured in time. Our main concern was the high-level synthesis, whilst the signal generator was only used to produce the speech signal. To measure the performance we needed a suitable approach to measure the time used in comparison to the duration of the resulting synthesized sentence, and thus *a time cost ratio* was developed. This means that the time consumed to synthesize the sentence was divided by the duration of the synthesized sentence, e.g. if it took one second to synthesize a 2-second sentence, the time cost ratio would be 0.5. If the time cost ratio was less than one, it could be said that the synthesizer works in real time.

The results showed that the system did not work in real time as a whole. The time cost ratio varied from 1.59 to 4.44, worsening along with the length of the sentence. The performance, however, was clearly due to the signal generation process, in which the time cost ratio varied from 1.30 to 4.18. In the

high-level synthesis, the time cost ratio varied from 0.29 to 0.45, and it was not directly dependent on the length of the sentence.

Since the publication of Paper VIII in the summer 2006, we have made some improvements to the signal generation process. First, we lowered the sampling frequency from 16 kHz to 8 kHz, which had no significant effect on the voice quality. Also, we added a simple threading to the system. Now the high-level synthesizer, the signal generator, and the wave player (see Figure 11) are all separated into different threads, which make it possible to overlap the processes, e.g. simultaneously playing the synthesis while creating a new one in the background. We also added new and improved phonemes to the system and enhanced the letter-to-phoneme translation. However, these modifications were not reviewed or published.

At the time of the study there were not many, if any, speech synthesizers on mobile platforms available on the market. However, they started to emerge at a fast pace in the following years. Nowadays, the speech synthesizer is almost a standard accessory in a mobile phone. However, if we compare the quality of synthetic speech to the system available now we find little or no development in the naturalness or intelligibility. The synthesizers that use other methods than rule-based synthesis may have slightly better naturalness, but still they seem to make errors with, or even lack, the basic phonetic rules of prosody, e.g. segmental durations and fundamental frequency. One might suspect that the approach developers have chosen is too much focused on engineering speech than understanding the basics of phonetics. As far as intelligibility is concerned, our system is still competing even with the best systems on the market. The controllability of the rule-based system makes it easy to enhance such prosodic features that make it easier to understand synthetic speech. For example, we can modify the fundamental frequency to make the words more isolated or lengthen the segmental durations to slow down the articulation rate without comprising the intelligibility, or we can strengthen the boundary-adjacent effects to isolate the utterances better.

When developing technologies such as speech synthesis or any other (speech) technology, it should be considered what kind of an impact it may have on the society in which it is used. In the next chapter, we consider the ethical impact a new technology may have.

3.6 Ethical issues

When developing a speech synthesizer or any other speech technology, its impact on ethical issues should always be considered. During our project with speech synthesis we decided to take a closer look at the issues that might come up as synthesizers are improving at a fast pace.

In our study, we considered the possibility that synthesized voice cannot be distinguished from the person it imitates (Paper VII). This view was

considered from three perspectives: 1) Is there a right to one's voice, and if so, what kind of right? 2) What are the borderline cases, if any, and how should they be solved? 3) What uses are clearly permissible and what are clearly not permissible? (Paper VII).

The answers found varied from technical solutions to legal remedies. Also, we compared the situation with similar problems found in other fields of technological development (see e.g. Weckert & Adeney 1994, Evans & Mahoney 2004).

The overall conclusion of this study (Paper VII) was that the technological development is likely to cause misuses which can be either borderline cases, which need to be considered from an ethical point of view or cases that are clearly misuses and need to be addressed in legislation. In addition, the design and implementation of such technology has to cater the possibility of such misuse, whether it is about the natural rights of the person it imitates or the duties of the user and the producer, and, of course, the consequences of such software which are often very hard to predict. In the next figure we present the ethical issues in the development of speech synthesis with capabilities to imitate a person's voice.

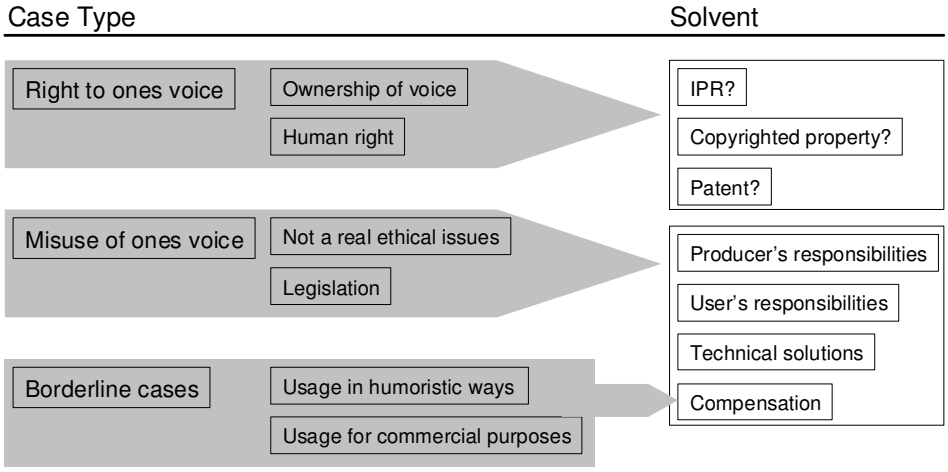


Figure 12. In this figure we see the threefold problem of synthesis development.

Chapter 4

Conclusion

In this thesis we set three main goals: a better knowledge of segmental durations in Finnish, a feasibility study of the speech synthesis technology on limited platforms, and philosophical considerations of ethical issues in speech synthesis development. The following chapter 4.1 answers the research questions set in the beginning.

4.1 Summary of the publications

In this chapter we compile the publications included with short summaries. The summaries do not include any new information not mentioned in the previous chapters of this dissertation. Any additions to the results added in the thesis are attached after publications with similar topic.

Paper I
The Role of Duration in Finnish Rule-Based TTS
Tuomo Saarni, Jussi Hakokari, Tapio Salakoski, Jouni Isoaho, Olli Aaltonen. Speech Analysis, Synthesis and Recognition, Applications of Phonetics, September 19-23, 2005, AGH University of Science and Technology, Kraków, Poland.

In this paper we studied the segmental durations for the speech synthesizer to better model the Finnish prosody. We applied simple scripting procedures to a speech corpus to find statistical patterns, or word models as we called them, and implemented them into the speech synthesizer.

The word models were built by dividing letters into seven categories: long and short plosives, non-plosive consonants, vowels and diphthongs. Each word in the corpus was then assigned a word model by placing the letters into these seven categories. Similar words were then in the same word model, e.g. the words '*kissa*' and '*pannu*' ('a cat' and 'a pan' respectively). Then the mean duration of each phoneme in a word model was calculated. When a sentence was synthesized the durations were taken from the similar word model.

It had been shown that the durations of individual phones are highly sensitive to their position within the word (e.g. Lehtonen 1970). In our previous study we discovered that word models may improve naturalness in long sentences (Hakokari et al. 2005a). The synthesizer was now used in a listening test to find out whether more detailed word models improve the naturalness of the synthesis when compared to fixed segmental durations. Once again, the word models appeared to improve naturalness in longer speech stimuli, but not in the shorter ones. 15 out of the 21 participants preferred the word model stimulus when the stimulus was considered to be long, but the result was only 8/21 participants when the stimulus was short.

Paper II

Determining Prepausal Lengthening for Finnish Rule-Based Speech Synthesis

Jussi Hakokari, Tuomo Saarni, Tapio Salakoski, Jouni Isoaho, Olli Aaltonen. Speech Analysis, Synthesis and Recognition, Applications of Phonetics, September 19-23, 2005, AGH University of Science and Technology, Kraków, Poland.

This study investigated prepausal lengthening (or final lengthening) in Finnish. The motivation of the study was to implement the possible effects into the speech synthesizer to increase the naturalness of the synthesized speech. We used a single-speaker corpus to estimate the level of lengthening in final, penultimate, and other words of the utterances. The data was divided into categories of vowels and consonants (separating the long and the short), but the amount of prepausal lengthening effect was also calculated for the individual phonemes in the final word position.

The procedure collected phonemes from the final, penultimate and other words. Then the mean duration of the phoneme category was calculated within the word (e.g. short vowels in the final words). The phonemes were compared against the positions (e.g. short vowels in final, penultimate and other positions). For the individual phonemes, the phones in the final words were compared against all other phones of the same phoneme (e.g. /i/ phonemes in final and in other positions of the utterances in the corpus).

The results show a clear lengthening in the final word and a slight effect remaining for the penultimate words as well. The lengthening in the final word varied from 23.4 per cent (short vowels) to 50.9 per cent (long consonants). The penultimate words were only slightly lengthened, 2.5-6.0%. The lengthening of individual phonemes varied greatly. In short vowels the scale of lengthening varied from 17.3-34.8 per cent, whilst for the long vowels the scale was 17.8-52.7 per cent. In the consonants, the greatest lengthening was shown in the sibilant /ʃ/, 132.2% (N.B. N=4 in final word position) and in the long consonant tremulant /r:/, 91.9% (N.B. N=9 in final word position). For reference, the most

common consonant /t/ was lengthened by 27.0 per cent (N=669 in the final word position). However, there was a consonant that was actually shortened in the final word position, /g/ by -7.3 per cent (N=7), but the /g/ and /ʃ/ sounds both only occur in non-Finnish loan words.

The study was the first to show that Finnish does have lengthening effect in the final word position, even though there were misconceptions in previous literature that Finnish along with some other quantity languages would somehow resist it (e.g. Vaissiere 1983).

Paper III

Segmental Duration in Utterance-Initial Environment: Evidence from Finnish Speech Corpora

Tuomo Saarni, Jussi Hakokari, Jouni Isoaho, Olli Aaltonen and Tapio Salakoski. FinTAL 2006, 5th International Conference on Natural Language Processing, 23-25 August 2006.

This study reported the utterance-initial segmental durations in Finnish. The research was based on the comparison of two qualitatively different speech corpora (multi-speaker vs. single-speaker) and it was carried out on a phone level.

The phonemes were divided into categories of vowels, diphthongs, plosives and non-plosive consonants with their phonologically short and long counterparts. Information was gathered on each phone, including the phone's position in the utterance (e.g. first, second, third etc. phone in the utterance), the phoneme category (e.g. short vowel), and duration in milliseconds. Then each phone in the category was arranged in descending order by position, and mean duration was calculated for each position (e.g. mean duration of all short vowels in the first position). The category was presented position-by-position with the confidence limit ($p \leq 0.05$).

The results show initial lengthening in vowels and diphthongs, and initial shortening in long plosives. The vowels and diphthongs were significantly lengthened in the first position. The long plosives were shortened in the third position, but not significantly in the second position. The results were ambivalent for other categories. Surprisingly, the results show no significant difference between the corpora.

Paper IV

Utterance-initial Duration of Finnish Non-plosive Consonants

Tuomo Saarni, Jussi Hakokari, Olli Aaltonen, Jouni Isoaho, Tapio Salakoski. NODALIDA 2007, the 16th Nordic Conference of Computational Linguistics, 25-26 May 2007 Tartu, Estonia.

In this study we investigated the segmental durations in the utterance-initial position using two qualitatively different speech corpora. The study concentrated on non-plosive consonants due to the fact that our previous study showed no effects for non-plosive consonants (Paper III) when studied as a category.

The use of two corpora was based on the comparability of qualitatively different speech corpora. The aim was to find lengthening or shortening effects on the initial positions of an utterance in a boundary-adjacent area. The segments were selected to be phones to achieve a detailed level of the phenomena of the articulation rate. Each phoneme was treated similarly to the phoneme categories in Paper III which provided positional behavior on a phoneme-by-phoneme level of each individual non-plosive consonant.

The results show all possible combinations; lengthening, shortening, and no effect depending on the phoneme in question. Phonemes /s/ and /r/ were lengthened significantly in the first position and slightly in the second position. The phonemes /m/, /j/, /v/ (in multi-speaker corpus only), /h/, and /n/ (in MS only) were shortened in the initial area. The lengthening or shortening only occurred in the very first positions, with the exception of the phoneme /h/ in multi-speaker corpus which was shortened all the way till the fourth position. The comparison between the corpora showed minor differences. Never in the initial positions was there a situation where another corpus would have showed shortening whilst the other showed lengthening.

In addition to paper III and IV, we studied articulation rate in utterances and discovered that the initial words were articulated faster than the other words in the utterance. The results showed also, that the prominence was extremely rare in the initial word.

Paper V
Measuring Relative Articulation Rate in Finnish Utterances
Jussi Hakokari, Tuomo Saarni, Tapio Salakoski, Jouni Isoaho, Olli Aaltonen. ICPhS The International Congress of Phonetic Sciences 2007, 6-10 August 2007 Saarbrucken, Germany

In this paper we studied final lengthening in Finnish using two methods to show *how often* final lengthening is present in utterances of 2-5 words and *how much* there is final lengthening in utterances of 2-9 words. At the same time we studied how often and how much there is initial shortening. The material used was combined from the single- and multi-speaker corpora.

The study used a relative measuring method that compared the phone lengths within the same utterance to eliminate the effect of variable speaking rates between utterances and speakers. In this methodology the duration of a phone was divided by the mean duration of the phones in the same category in the material. The entire word was then given a coefficient which was the mean value of the phones in the word. Finally, each word was rank-ordered by their coefficients. The rank order shows which word in the utterance was uttered most

slowly etc. The rank-ordering was carried out for all utterances of 2-5 words. In the other experiment the amount of final lengthening was studied by tracing the coefficients as a sign of speech rate for each word in respective order. This time the coefficient was used to calculate the mean speaking rate for each word in the same position (e.g. final position of every 5-word sentence). The result showed how the speaking rate progressed in all utterances of 2-9 words.

The results were compared against a similar study carried out on Swedish (Hansson 2003). Results show that 85-75 per cent of the utterances are affected by final lengthening, and 85-49 per cent are initially shortened. The results were remarkably similar to those reached in the study on Swedish, even though the languages in question are not related. The articulation rate seems to be considerably faster in the beginning of the utterance and slow down towards the end of the utterance, whilst the final word is articulated significantly more slowly than the penultimate one.

Paper VI
Utterance-level Normalization for Relative Articulation Rate Analysis
Tuomo Saarni, Jussi Hakokari, Jouni Isoaho, Tapio Salakoski.
Interspeech 2008 incorporating SST'08, 22-26 September 2008, Brisbane, Australia.

In this paper we described a normalization method for analyzing speech corpus with several speakers and articulation rates. The method was developed to decrease the inter- and intra-speaker variation effect in the study of segmental durations. Normalization is needed when the corpus has utterances of variable lengths and articulation rates, and the various speaking rates of different speakers affect the duration of segments. The method was then used to study final lengthening in Finnish using speech corpora of several speakers and speaking styles.

In the study we described a digestible normalization method suitable for speech material with several trends of variation. Firstly, intra-speaker variability means that a speaker uses variable speaking rates between utterances. Secondly, inter-speaker variability means that different speakers may use different speaking rates. Problems arise e.g. if we are interested in studying the speaking rate in final positions of utterance and a slow-speaking participant uses only short sentences and produces more utterances than others, as the segments in the final position are greatly affected by this slower speaker. Our solution was to handle each utterance as an entity with its own speaking rate behavior. The approach divides the phones in the utterance into phoneme categories and calculates the mean duration of each category by using the representative phone durations. If there is only one representative, the category is discarded for this utterance. Then each position is given a coefficient which is calculated by dividing the duration of the phone in position by the mean duration of its phoneme category. This is done with all of the utterances in the material. After

this, we can calculate the mean coefficient for each position of each utterance, e.g. the final position of the whole utterance. Each utterance is then equal and a slower speaking rate does not affect it?? as much as with the absolute durations.

To present the use of the normalization methodology, we used the approach to demonstrate final lengthening in the multi-speaker corpus. The results show the pattern of final lengthening in Finnish on a phone level with statistical significance ($p \leq 0.05$). The results substantiate the views of our previous studies (Paper VI, Paper II). The results also support White (2002), who claims that there is no support for the domain span effect, but for the domain edge effects (Paper VI).

In addition to papers II, V, and VI, we studied the effect of prominence in final lengthening. The results showed that the prominence has a little or no effect at all on the occurrence of final lengthening. Thus, Finnish has final lengthening that is not dependent on prominence. Also, our additional studies showed that final lengthening is distinct even after the material is both normalized from inter- and intra-speaker variation and the prominence is extracted from the material. It was also showed, that final lengthening is statistically significant even after the prominent words were excluded.

Paper VII
Right to one's voice?
Kai K. Kimppa & Tuomo Saarni. Ethicomp 2008, The Tenth ETHICOMP International Conference on the Social and Ethical Impacts of Information and Communication Technology, 24-26 September 2008, Mantua, Italy.

In this paper we considered the ethical issues raised by the fast development of speech synthesis systems. This has become important since synthesizers are achieving a level where it is hard to distinguish the person whose voice is copied from the computer. We were interested in the ownership of a person's voice as well as in the ethical questions related to copying this voice. This paper provided an overlook to the technology of speech synthesis and other related speech applications.

The ethical and social questions were divided into three categories: right to one's voice, misuse of one's voice, and borderline cases. Firstly, we had to consider whether we have ownership of our voice, whether it is a natural human right, or neither. Ownership could be compared to IPRs, copyrighted properties or patents, all of which are problematic. The misuse of such a high-level speech synthesis is quite clear in the sense that it could be easily used in morally unacceptable ways, but these problems cannot be considered as "real" ethical issues. The borderline cases are the hardest to grasp. Such cases would include the use of personated voice for comedy purposes or in TV commercials, or even in commercial products. These cases should be considered individually.

On the technical side of the issue, what are the responsibilities of the producer? Should the program contain a solution that restrains misuse and how

far should the restrains extend? We suggested that a study of the programs' effects should be carried out. Another possibility would be a watermark that gives out the user in a misuse situation. On the legal side, compensation for the person whose voice is used should be considered.

Paper VII
Implementing a Rule-Based Speech Synthesizer on a Mobile Platform
Tuomo Saarni, Jyri Paakkulainen, Tuomas Mäkilä, Jussi Hakokari, Olli Aaltonen, Jouni Isoaho and Tapio Salakoski. FinTAL 2006, 5th International Conference on Natural Language Processing, 23-25 August 2006.

In this paper we presented an approach to implement a Finnish speech synthesizer on a mobile platform. In this case, the mobile platform was a Java™-supported mobile phone, but the base could have been any similar platform with limited resources. This paper showed the structure of the synthesizer and evaluated its performance.

The mobile phone (Nokia 6680) that was used supported Java™MIDP 2.0 (Motorola 2002) and CLDC 1.1 (Sun Microsystems 2003). Originally, the synthesis program was made for personal computers, but the rule-based method used only required relatively limited storage space and computing capabilities. Therefore it was suitable for a limited device platform. The paper showed in detail the steps of the synthesis methods used in the program. The methods were divided into high level and signal generation phases. The synthesis itself was done on the high level and the acoustic signal was generated in the signal generator, based on Klatt (1980).

The performance was tested with six sentences of various lengths. The duration of the sentences varied from 3.06 to 10.68 seconds. The synthesis procedure was divided on the program level into two stages (high level and signal generation) which could be measured in time. To measure the performance we needed a suitable approach to measure the time used in comparison to the duration of the resulting synthesized sentence, and thus *a time cost ratio* was developed. This means that the time consumed to synthesize the sentence was divided by the duration of the synthesized sentence, e.g. if it took 1 second to synthesize a 2-second sentence, the time cost ratio would be 0.5. If the time cost ratio was less than one, it could be said that the synthesizer works in real time.

The results showed that the system did not work in real time. The time cost ratio varied from 1.59 to 4.44, worsening along with the length of the sentence. The performance, however, was clearly due to the signal generation, in which the time cost ratio varied from 1.30 to 4.18. In the high level synthesis, the time cost ratio varied from 0.29 to 0.45, and it was not directly dependent on the length of the sentence. The results hinted that to achieve a real-time synthesis the signal generator must be developed further. However, at the time of the study, a speech synthesizer on a mobile phone was still an oddity.

4.2 Conclusion and future work

Whilst speech provides such an effective way to communicate, we will be faced with speech technology increasingly in the future. Speech technology will be implemented as biometric identifiers, speech synthesizers, and speech recognizers, to name a few. The technology *will* evolve, and it is only a matter of time before all of these technologies are commonly available.

This dissertation is a part of this future trend. Although we only scratched the surface of this wide field of speech technology, we touched the essence of speech by approaching the implementation of technology from the perspectives of naturalness, intelligibility, and controllability. Naturalness, the human-like aspect and factor, is the key in making the speech synthesizers accepted by the masses. Intelligibility can be seen as the capability to go beyond human speech by producing synthesized speech that is even more error-tolerant than natural human speech. Controllability is the capability to glide between naturalness, intelligibility, and any other aspects of speech that we can create, and thereby master.

The future remains open for some aspects covered in this dissertation. Will the discourse on the universality of final lengthening be opened? Would the study of universality be the direction in which to proceed? Especially, it would be interesting to study the cause of FL from the biological point of view, as suggested earlier to be the origin of the phenomenon.

Also, there are interesting technological roads left to uncover for future studies. One of these could be other platforms with limited resources than mobile phones, e.g. any household items with computing capabilities. In addition, the intelligibility capabilities of a rule-based speech synthesizer should be studied separately on the basis of naturalness. One possible application could be the announcement systems in public places with a lot of noise.

Whatever the direction of the study, it should be kept in mind that whenever studying an interesting phenomenon, whether it is an effect in segmental durations or a new application of speech synthesis, the purpose of the study should not be limited to getting to the bottom of the phenomenon, but to decrease the confusion in the matter. As Hawkins & Blakeslee (2004) so exquisitely summed up, too often science is attracted by the trivial details when the entirety is lost. Similarly, we have to remind ourselves to sometimes take a step back and consider whether we are approaching the matter wisely. With this in mind, it is good to end this dissertation.

References

Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., & Stent, A. (2001): Towards conversational human-computer interaction. AI Magazine, 22, 27/37.

Barbosa, P. & Madureira, S. (1999): Toward a hierarchical model of rhythm production: Evidence from phrase stress domains in Brazilian Portuguese. In J. Ohala, Y. Hasegawa, M. Ohala, D. Granville & A. Bailey (eds.), Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, 1–7 August 1999. Linguistics Department, University of California, Berkeley, pp. 297–300.

Bashford, JR J.A., Riener K.R., & Warren, R.M. (1992): Increasing the intelligibility of speech through multiple phonemic restorations. Perception & Psychophysics 1992, 51 (3), pp. 211-217.

Bertozzi, D., & Benini, L. (2008): Hardware Platforms for Third-Generation Mobile Terminals. In Memories in Wireless Systems. Signals and Communication Technology. Micheloni, Campardo, Olivo (Eds.). Springer Berlin Heidelberg. July 24, 2008. pp. 1-28.

Browman, C. P., & L. Goldstein, (1990): Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (Eds.), Papers in Laboratory Phonology I: Between the grammar and the physics of speech (pp. 341–338), Cambridge, UK: Cambridge University Press.

Byrd, D., & Saltzman, E. (2003): The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. Journal of Phonetics, Volume 31, Issue 2, April 2003, pp. 149-180.

Byrd, D. (2000): Articulatory vowel lengthening and coordination at phrasal junctures. Phonetica, 57(1), 3-16.

Byrd, D., Kaun, A., Narayanan, S., & Saltzman, E. (2000): Phrasal signatures in articulation. In M. B. Broe & J. B. Pierrehumbert (Eds.), Papers in laboratory phonology V, pp. 70-87. Cambridge: Cambridge University Press.

Campbell, J. P. (1997): Speaker recognition: A tutorial, Proceedings of the IEEE, vol. 85, pp. 1437-1462, September 1997.

Campbell, N. (1992): Segmental elasticity and timing in Japanese speech. In. Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (eds.): Speech Perception, Production and Linguistic Structure. Amsterdam: IOS Press, pp. 403-418.

Cao, J. (2004): Restudy of segmental lengthening in Mandarin Chinese. Proceedings of Speech Prosody 2004 (SP-2004), Nara, Japan, pp. 231-234.

Chung, H., Gim, G. & Huckvale, M.(1999): Consonantal and Prosodic Influences on Korean Vowel Duration. Proceedings of Eurospeech, Vol.2. Budapest, Hungary (1999), pp. 707-710.

Deng, B.L. & Huang, X.(2004): Challenges in adopting speech recognition. Communications of the ACM, Vol. 47, No. 1 (2004) pp. 69-75.

Duez, D. (1993): Acoustic Correlates of Subjective Pauses. Journal of Psycholinguistic Research, Vol. 22(1). (1993), pp. 21-39.

Evans, J., & Mahoney, J. (2004): Ethical and Legal Aspects of Using Digital Images of People: Impact on Learning and Teaching. Ethicomp 2004, University of the Aegean, Syros, Greece, 14 to16 April 2004, pp. 289-297.

Fougeron, C., & Keating, P. A. (1997): Articulatory strengthening at edges of prosodic domains. Journal of the Acoustical Society of America 101 (6), pp. 3728-3740.

Goodman, D.J., Lockhart, G.B., Ondria, J.W., & Wai-Choong, W. (1986): Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications, IEEE Transactions on acoustics, speech, and signal processing. Vol. ASSP-34, 6.

Hansson, P. (2003): Prosodic Phrasing in Spontaneous Swedish. Academic Dissertation. Travaux de l'institut de linguistique de Lund 43. Lund: Lund University (2003).

Heikkinen, H. (1979): Vowel reduction in the English of Finnish learners. In J. Lehtonen & K.Sajavaara (Eds.) Papers in contrastive phonetics, Jyväskylä. Cross-Language Studies 7, Jyväskylä, University of Jyväskylä.

Hakokari, J., Saarni, T., Jalonen, M., Aaltonen, O., Isoaho, J., & Salakoski, T. (2005a): Word Model-Determined Segmental Duration in Finnish Speech Synthesis and its Effect on Naturalness. Second Baltic Conference on Human Language Technologies, April 4-5, 2005, Tallinn, Estonia.

Hakokari, J., Saarni, T., Salakoski, T. Isoaho, J., & Aaltonen, O. (2005b): Determining Prepausal Lengthening for Finnish Rule-Based Speech Synthesis. Speech Analysis, Synthesis and Recognition, Applications of Phonetics, September 19-23, 2005, AGH University of Science and Technology.

Proceedings of Speech Analysis, Synthesis, and Recognition: Applications of Phonetics (SASR 2005), Kraków, Poland.

Hakokari, J., Saarni, T., Salakoski, T. Isoaho, J., & Aaltonen, O. (2007): Measuring Relative Articulation Rate in Finnish Utterances. ICPhS The International Congress of Phonetic Sciences 2007, 6-10 August 2007 Saarbrucken, Germany. Proceedings of the 16th International Congress of Phonetic Sciences, Universität Des Saarlandes, paper ID-1401, pp. 1105-1108.

Hakokari, J., Saarni, T., Isoaho, J., & Salakoski, T. (2008): Correlation of Utterance Length and Segmental Duration in Finnish Is Questionable. Interspeech 2008 incorporating SST'08, 22-26 September 2008, Brisbane, Australia.

Hakulinen, J., Turunen, M., & Räihä, K-J. (2007): Tutoring in a Spoken Dialogue System. In Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue: 239-242, 2007.

Harris, Z. S. (1951): Methods in Structural Linguistics. The University of Chicago Press, Chicago.

Hautamäki, V., Kinnunen, T., & Fränti, P. (2008): Text-independent speaker recognition using graph matching, Pattern Recognition Letters, 29 (9), 1427-1432, 2008.

Hawkins, J., & Blakeslee, S. (2004). On Intelligence. New York: Holt.

JBenchmark Home Page (2006): http://www.jbenchmark.com/phonedetails.jsp?D=Nokia%206680&benchmark=v2, Kishonti Informatics LP. Accessed on 27th of August 2009.

Kaiki, N., Takeda, K., & Sakisaga, Y. (1990): Statistical Analysis for Segmental Duration Rules in Japanese Speech Synthesis. In proceedings of the 1990 International Conference on Spoken Language Processing. Kobe, Japan (1990), pp. 17-20

Karttunen, L. (1998): The proper treatment of Optimality Theory in computational phonology. Finite-state methods in natural language processing. pp. 1–12.

Karvonen, D. (2005): Word Prosody in Finnish. Ph.D. thesis, University of California at Santa Cruz.

Kasuya, H., Maekawa, K., & Kiritani, S. (1999): Joint Estimation of Voice Source and Vocal Tract Parameters as Applied to the Study of Voice Source Dynamics, ICPhS 99, pp. 2505-2512.

Kelso, J.A.S., & Munhall, K.G. (1988, eds): R.H.Stetson's Motor Phonetics - A Retrospective Edition. Little, Brown and Company (Inc.). 1988.

Kenstowicz, M., & Kisseberth, C. (1979): Generative Phonology: Description and Theory. Academic Press.

Kinnunen, T., Karpov, E., & Fränti, P. (2006): Real-time speaker identification and verification, IEEE Transactions on Audio, Speech and Language Processing, 14 (1), 277-288, January 2006.

Kinnunen, T. & Li, H. (2009): An overview of text-independent speaker recognition: From features to supervectors. Journal of Speech Communication, Volume 52, Issue 1, January 2010, pp. 12-40.

Klatt, D. H. (1973): Interaction between two factors that influence vowel duration. Journal of the Acoustic Society of America, 54, pp. 1102-1104.

Klatt, D. H. (1975): Vowel lengthening is syntactically determined in a connected discourse. Journal of the Acoustical Society of America, 3, pp. 129-140.

Klatt, D., (1980): Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America 67. 971-995.

Klatt, D.H. (1987) Review of text-to-speech conversion for English. Journal of the Acoustical Society of America, 82, pp. 737-793.

Koskenniemi, K. (1983): Two-level morphology: A general computational model for word-form recognition and production. Publication no. 11, Department of General Linguistics, University of Helsinki, Helsinki.

Kuhl, P. K. (1991): Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. Percept. Psychophys. 50:93–107.

Krull, D. (1997): Prepausal lengthening in Estonian: evidence from conversational speech. In Lehiste, I. & Ross, J. (Eds.), *Estonian Prosody: Papers from a Symposium, Proceedings of the International Symposium on*

Estonian Prosody, Tallinn, Estonia. Tallinn: Institute of Estonian Language, pp. 136-148.

Lamel, L., Bennacef, S., Gauvain, J. L., Dartigues, H., & Temem, J. N. (2002): User evaluation of the M kiosk, Speech Communication, Volume 38, Issues 1-2, September 2002, pp. 131-139.

Lehiste, I. (1965): The function of quantity in Finnish and Estonian. Language 41, 447-456.

Lehtonen, J. (1970): Aspects of quantity in standard Finnish. Jyväskylä: University of Jyväskylä.

Lennes, M. (2003): On the expected variability of vowel quality in Finnish informal dialogue. In: Solé, M., Recasens, D., Romero, J., (eds.) Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS). pp. 2985–2988.

Lennes, M., Aho, E., Toivola, M. & Wahlberg, L. (2006): On the use of the glottal stop in Finnish conversational speech. *The Phonetics Symposium 2006.* Publications of the Department of Speech Sciences, University of Helsinki, 53, pp. 93-102.

Li, Q., Zheng, J., Tsai, A., & Zhou, Q. (2002): Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition. - IEEE Transactions on Speech and Audio Processing, 2002 vol. 10, no. 3, pp. 146–157.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967): Perception of the speech code, Psychological Review, 74, pp. 431-461.

Lindblom, B. (1968): Temporal Organization of Syllable Production, STLQ Progr. Status Rep., Stockholm, Sweden, Roy. Inst. Technol. 2(3).

Logan, J. S., Greene, B. G., & Pisoni, D. B. (1989): Segmental intelligibility of synthetic speech produced by rule, J. Acoust. Soc. Am. 86, pp. 566–581.

MacNeilage, P. F., Studdert-Kennedy, M. G., & Lindblom B. (1984): Functional precursors to language and its lateralization, Am J Physiol Regulatory Integrative Comp Physiol 246:912-914, 1984.

MacNeilage, P. F., & Davis, B. L. (1990): Acquisition of speech production: Achievement of segmental independence. In W. I. Hardcastle, & A. Marchal,

(Eds.), Speech production and speech modeling, 55–68. The Netherlands: Dordrecht.

MacNeilage PF. (1998): The frame/content theory of evolution of speech production. Behav. Brain Sci. 21:499–511

McDermott, J. H., & Oxenham, A. J. (2008): Spectral completion of partially masked sounds. PNAS 105, 5939-5944

Mihkla, M. (2005): Modelling pauses and boundary lengthenings in synthetic speech. Proceedings of the Second Baltic Conference on Human Language Technologies, Tallinn.

Motorola (2002): Mobile Information Device Profile for Java™ 2 Micro Edition – Version 2.0. Motorola and Sun Microsystems JSR-118. 2002. James Warden (Maintenance). See http://jcp.org/aboutJava/communityprocess/mrel/jsr118/index.html. Accessed on August 19th 2009.

Moulines, E. & Charpentier, F. (1990): Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis using Diphones. Speech Communication, (9):453–467.

Nagano-Madsen, Y. (1992): Temporal characteristics in Eskimo and Yoruba: a typological consideration. In Huber (ed.): Papers from the Sixth Swedish Phonetics Conference held in Gothenburg. Technical Report No. 10, Department of Information Theory, School of Electrical and Computer Engineering, Chalmers University of Technology, Gothenburg. pp. 47-50.

Nakai, S., Kunnari, S., Turk, A., Suomi, K., & Ylitalo, R. (2009): Utterance-final lengthening and quantity in Northern Finnish. Journal of Phonetics, Volume 37, Issue 1, January 2009, pp. 29-45.

O'Dell, M. (1999): Some factors affecting perception of stop quantity in Finnish. In: Out Loud: Papers from the 19th Meeting of Finnish Phoneticians (J. Järvikivi & J. Heikkinen, eds.), pp. 76-85. University of Joensuu, Faculty of Humanities, Studies in Languages 33.

O'Dell, M. & Nieminen, T. (2001): Speech rhythms as cyclical activity. In: 21. Fonetiikan päivät Turku 4.-5.1.2001 (S. Ojala & J. Tuomainen, eds.), Publications of the Department of Finnish and General Linguistics of the University of Turku, 67, pp. 159-168.

O'Dell, M. (2003): Intrinsic Timing and Quantity in Finnish. Acta Universitatis Tamperensis 979, University of Tampere Press.

Ohala, J.J. (1986): Against the direct-realist view of speech perception. Journal of Phonetics, 14, pp. 75-82.

Ohala, J. J., & Shriberg, E. E. (1990): "Hypercorrection in speech perception", In ICSLP-1990, pp. 405-408.

Oller, K. (1973): The effect of position in utterance on speech segment duration in English. The Journal of the Acoustical Society of America, 51, pp. 1235-1247.

Paganus, A. Mikkonen, V-P., Mäntylä, T., Nuuttila, S., Isoaho, J., Aaltonen, O., & Salakoski, T. (2006): The Vowel Game: Continuous Real-Time Visualization for Pronunciation Learning with Vowel Charts. In Advances in Natural Language Processing, 5th International Conference, FinTAL 2006 Turku, Finland, August 23-25, 2006 Proceedings, Aug 2006.

Peltola, M.S., Kujala, T., Tuomainen, J., Ek, M., Aaltonen, O., & Näätänen, R. (2003): Native and foreign vowel discrimination as indexed by the mismatch negativity (MMN) response. Neuroscience Letters, Volume 352, Issue 1, 27 November 2003, pp. 25-28.

Peterson, G.E. & Lehiste, I. (1960): Duration of syllable nuclei in English. Journal of the Acoustical Society of America, 32, pp. 175-184.

Raimo, I., Savela, J. & Aaltonen, O. (2002): The Turku Vowel Test. Fonetiikan päivien paperit 2002. Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, Report 67, 45-52.

Rasmusson, J., Dahlgren, F., Gustafsson, H., & Nilsson, T. (2004): "Multimedia in mobile phones - the ongoing revolution", Ericsson review, vol. 02, 2004.

Repp, B. (1984): Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), Speech and language: Advances in research and practice, Vol. 10. New York: Academic Press.

Reynolds, D. A. (2002): An overview of automatic speaker recognition technology,
in Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002), Orlando, FL, 2002, pp. 4072–4075.

Saarni, T. (2005): Puhe säännöiksi, säännöt puheeksi – sääntösynteesin kehittäminen [In Finnish]. Unpublished master's thesis, University of Turku.

Saarni, T., Hakokari, J., Salakoski, T,. Isoaho, J., & Aaltonen, O. (2005): The Role of Duration in Finnish Rule-Based TTS Speech Analysis, Synthesis and Recognition, Applications of Phonetics, September 19-23, 2005, AGH University of Science and Technology, Kraków, Poland.

Saarni, T., Hakokari, J., Isoaho, J., Aaltonen, O., Salakoski, T. (2006a): Segmental Duration in Utterance-Initial Environment: Evidence from Finnish Speech Corpora. Advances in Natural Language Processing: 5th International Conference on NLP, FinTAL 2006, Turku. Published as volume 4139 in Springer series "Lecture notes in Artificial Intelligence". pp. 576-584.

Saarni, T., Paakkulainen, J., Mäkilä, T., Hakokari, J., Aaltonen, O., Isoaho, J., and Salakoski, T. (2006b): Implementing a Rule-Based Speech Synthesizer on a Mobile Platform. FinTAL 2006, 5th International Conference on Natural Language Processing, 23-25 August 2006.

Saarni, T., Hakokari, J., Aaltonen, O., Isoaho, & J., Salakoski, T. (2007): Utterance-initial Duration of Finnish Non-plosive Consonants. NODALIDA 2007, the 16th Nordic Conference of Computational Linguistics, 25-26 May 2007 Tartu, Estonia. Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007. University of Tartu, Tartu, 2007. pp. 160-166.

Saarni, T., Hakokari, J., Isoaho, J., Salakoski, T. (2008): Utterance-level Normalization for Relative Articulation Rate Analysis. Interspeech 2008 incorporating SST'08, 22-26 September 2008, Brisbane, Australia. Proceedings of the 9th Annual Conference of the International Speech Communication Association incorporating the 12th Australasian International Conference on Speech Science and Technology (SST 2008), p. 538-541.

Savela, J. (2009): Role of Selected Spectral Attributes in the Perception of Synthetic Vowels. PhD Dissertation. TUCS Dissertations, no. 119, June 2009.

Schneider, K., Lintfert, B., Dogil, G., & Möbius, B. (2006): Phonetic Grounding of Prosodic Categories. In: Sudhoff, S. et al. (eds), Methods in Empirical Prosody Research. Berlin: de Gruyter, 1-27.

Shin, J., Narayanan, S., Gerber, L., Kazemzadeh, A., & Byrd. D. (2002): Analysis of user behaviour under error conditions in spoken dialogue. In Proceedings of ICSLP, 2002.

Sun Microsystems (2003): Connected Limited Device Configuration Specification – Version 1.1. Sun Microsystems. JSR-139. 2003. Jonathan Courtney (Maintenance). See http://jcp.org/aboutJava/communityprocess/final/jsr139/index.html. Accessed on August 19th 2009.

Suomi, K., Toivanen, J., Ylitalo, R. (2003): Durational and tonal correlates of accent in Finnish, Journal of Phonetics, Volume 31, Issue 1, January 2003, pp. 113-138.

Suomi, K., Ylitalo, R. (2004): On durational correlates of word stress in Finnish, Journal of Phonetics, Volume 32, Issue 1, January 2004, pp. 35-63.

Suomi, K. (2007): On the tonal and temporal domains of accent in Finnish, Journal of Phonetics, Volume 35, Issue 1, January 2007, pp. 40-55.

Takeda, K., Sagisaka, Y., and Kuwabara, H. (1989): On sentence-level factors governing segmental duration in Japanese, J. Acoust. Soc. Am. 86, pp. 2081-2087.

Trask, R. L. (1996): A Dictionary Of Phonetics And Phonology, Routledge, London, UK.

Turk, A. E., Sawusch, J. R. (1997): The domain of accentual lengthening in American English, Journal of Phonetics, Volume 25, Issue 1, January 1997, pp. 25-41.

Turk, A., Nakai, S., & Sugahara, M. (2006): Acoustic Segment Durations in Prosodic Research: A Practical Guide. In: Sudhoff, S. et al. (eds), Methods in Empirical Prosody Research. Berlin: de Gruyter, 1-27.

The Turku Vowel Test. Dept. of Phonetics, University of Turku. http://fon.utu.fi/ Accessed on August 20th 2009.

Turunen, M., Hakulinen, J., Salonen, E-P., Kainulainen, A. & Helin, L. (2005): Spoken and Multimodal Bus Timetable Systems: Design, Development and Evaluation. Proc. of 10th International Conference on Speech and Computer (SPECOM 2005): 389-392, 2005.

Turunen, M., Hurtig, T., Hakulinen, J., Virtanen, A., & Koskinen, S.(2006): Mobile Speech-based and Multimodal Public Transport Information Services. In Proceedings of MobileHCI 2006 Workshop on Speech in Mobile and Pervasive Environments, 2006

van Santen, J.P.H. (1994): Assignment of segmental duration in text-to-speech synthesis. Computer Speech and Language 8, pp. 95-128. 1994.

Vainio, M. (2001): Artificial neural networks based prosody models for Finnish text-to-speech synthesis. PhD Dissertation, University of Helsinki, Dept. of Phonetics.

Vaissière, J. (1983): Language independent prosodic features. A. Cutler & R. Ladd (Eds.) Prosody: Models and Measurements, pp. 53-65. Springer Verlag.

Venkatagiri, H.S. (2003): Segmental intelligibility of four currently used text-to-speech synthesis methods, J. Acoust. Soc. Am. 113, 2095.

Warren, R. M., & Warren, R. P. (1970): Auditory illusions and confusions. Scientific American, 223, pp. 30-36.

Warren, R. M. (1970): Perceptual Restoration of Missing Speech Sounds. Science 23 January 1970:Vol. 167. no. 3917, pp. 392–393

Weckert, J., & Adeney, D. (1994): Ethics in Electronic Image Manipulation. Ethics in the computer age, Galtinburg, Tennessee, United States, pp. 113-114.

White, L.S. (2002): English speech timing: a domain and locus approach, University of Edinburgh PhD dissertation, 2002.

White, L., & M´ady, K. (2008): The long and the short and the final: phonological vowel length and prosodic timing in Hungarian. In 4th Speech Prosody Conference, Campinas, Brasil, pp. 363–366, 2008.

Wiik, K. (1965): Finnish and English Vowels. Annales Universitatis Turkuensis, Series B, University of Turku.

Zu, Y., & Chen, X.(1998): Segmental Durations of a Labelled Speech Database and its Relation to Prosodic Boundaries. In Proceedings of the 1st International Symposium on Chinese Spoken Language Processing (ISCSLP 1998)

Öhman, S. (1967): Word and sentence intonation: a quantitative model, STLQ Progr. Status Rep. Stockholm 2/3, pp. 20–54.

Part II

Publication reprints

Paper I

The Role of Duration in Finnish Rule-Based TTS

Tuomo Saarni, Jussi Hakokari, Tapio Salakoski, Jouni Isoaho, Olli Aaltonen. Speech Analysis, Synthesis and Recognition, Applications of Phonetics, September 19-23, 2005, AGH University of Science and Technology, Kraków, Poland.

The Role of Duration in Finnish Rule-Based TTS

Tuomo Saarni[1] & Jussi Hakokari[2]

Tapio Salakoski[1], Jouni Isoaho[1] & Olli Aaltonen[2]

[1]Department of Information Technology

[2]Phonetics Laboratory

University of Turku, Turku, Finland

University of Turku

FIN-20014, TURKU

tuomo.saarni@utu.fi & jussi.hakokari@utu.fi

ABSTRACT

We are developing a rule-based Finnish-language TTS system. Our primary concern is to find ways to increase naturalness in the synthesis. Our approach is to observe tendencies in natural language through acoustic analysis and data mining, and to implement our findings into the synthesizer. We have concentrated on modeling duration, which is an essential part of Finnish prosody. The language exhibits contrasting phonemic lengths and the durations of individual phones are highly sensitive to their position within a word. We have developed a duration model ("word models") based on how the syllabic structure of a word correlates with segmental durations in a natural speech corpus. We have implemented and automatized the word models, and studied through listening tests whether they improve naturalness in the synthesis. We compared the word model–determined segmental durations with with fixed ones. The result was ambiguous: the word models appear to improve naturalness in longer speech stimuli, but not in the shorter ones.

1. Introduction

While rule-based speech synthesis is versatile and, when correctly configured, intelligible, its weakness lies in naturalness; it is very difficult to produce synthetic speech that sounds humanlike without resorting to samples of recorded speech and the concatenative methods. Our aim is to investigate just how far one can go with rule-based synthesis by carefully modeling

characteristics of natural speech. At the moment we are mostly concerned with duration and its effect on naturalness. Duration is a part of speech prosody and important to a natural rhythm of speech. Duration is also a delicate matter in Finnish speech synthesis; the language is cited as a "quantity language" [3].

The Finnish language has contrasting phonemic length; all the vowels and the majority of consonants may occur either short or long and thus form minimal pairs. The short vowels tend to be slightly more central in the vowel space than the long ones [7, 4], but the decisive factor is duration [7]. Finns are generally unaware of any qualitative differences between the two phonemic lengths. The duration is not absolute even in the widest sense, but relative to the segment's position within a word and the word's position within a sentence.

This is a continuation to an earlier study in which we compared "word model"–determined segmental durations to fixed durations [1]. Word models, inspired by Lehtonen's work [3], are mean durations based on consonant-vowel sequences data mined from natural speech corpora. For instance, in our training corpus the word form VC (a vowel followed by a consonant) has mean durations of 69 ms (V) and 52 ms (C). The synthesizer retrieves the data from the word model bank and makes the vowel ~70 ms and the consonant ~50 ms long in all words of the form VC. Our word model bank had ~1100 entries. The results were encouraging; long sentences with long words in them were deemed more natural than the sentences synthesized with fixed durations [1]. In this study we made a more complex set of word models; plosives are now distinguished from the rest of the consonants and diphthongs are separated from long vowels. Now the speech corpus yielded ~2500 entries. Furthermore, we used much longer samples of synthetic speech than in the previous study. The speaking rate was also slower; a faster speaking rate used in the first experiment proved confusing to the naïve participants. In this study we investigated whether the improved word models enhance naturalness in synthetic speech.

2. Methods

2.1. Data analysis

The speech corpus from which the word models were extracted consisted of 692 declarative Finnish sentences containing ~6500 words. The corpus, described in more detail in [6], is read aloud by a 39-year-old male from Helsinki and adds up to 69 minutes of recording. The corpus was consistently segmented and annotated at word and phone levels to make data mining possible. The previous word model bank which did not differentiate diphtongs and long vowels or plosives and the rest of the consonants was problematic. Consequently, all the words occurring in the corpus were this time established as sequences of plosives (P), other consonants (C), vowels

(V), and diphthongs (VV). Durations for each segment within a model are included into the bank; multiple occurrences of a single model involve calculating a mean duration for each segment. The number of established word models was ~2500, reflecting the long words typical of a highly inflected language with a small phoneme inventory such as Finnish. In fact, the ~2500 word models are not nearly enough for synthesizing free input. For TTS purposes, the user can choose either fixed durations or a generic word model to determine the segmental durations in case the system encounters a word that falls outside the bank. Both the wavelength and the 10 ms time resolution of the signal generator dictate that the durations prescribed by the word models cannot be reproduced with utmost accuracy; the system uses rounding and waveform interpolation in producing a continuous speech signal.

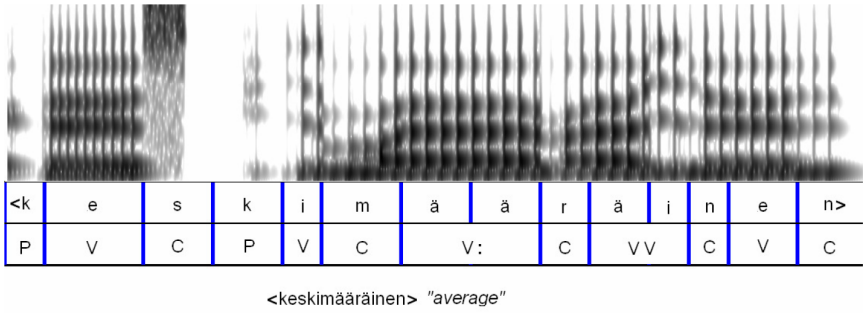
2.2. Stimuli

The eight stimuli were four paragraphs of text synthesized into speech in two different ways. We made sure that all the words in the stimuli had a representation in the word model bank. In other words, the generic models or fixed durations were not used. The stimuli were all Standard Finnish, but one of them (stimulus A) contained two foreign proper names ("Vladimir" and "Visentini"). The first set of stimuli used the word model bank to determine segmental durations. The second set of stimuli used fixed durations based on mean values found in the same speech corpus. The speech rate (= overall duration) of the stimuli was thus practically equal. All the phonemically short segments were ~70 ms in duration, while the long ones were ~140 ms.

Table 1. *Stimulus information.*

	Words	Characters	Duration fixed	Duration word-model
Stimulus A	85	606	44.24 s	42.61 s
Stimulus B	40	261	19.20 s	19.14 s
Stimulus C	75	464	34.13 s	32.62 s
Stimulus D	65	426	31.59 s	31.57 s

The stimuli were sound files (.wav) with a sample rate of 10 kHz. The signal generator is still under development and produces occasional disturbances (pops and clicks) similar to those produced by the Klatt synthesizers [2]. The disturbances were not manually edited out of the signal, but the participants were asked not to pay attention to them.

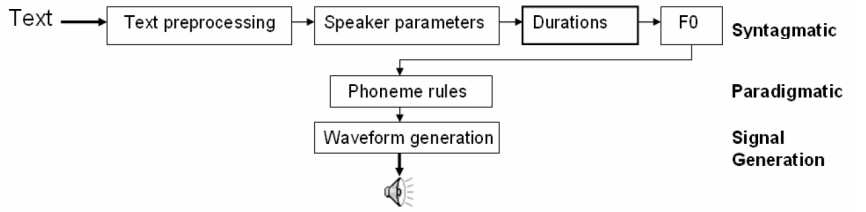


Picture 1. *An example of word model –determined segmental durations in a synthesized word.*

A simple cascading model for fundamental frequency was used in generating the stimuli to highlight the effect of segmental durations. F0 starts at 100 Hz in the beginning of a sentence and climbs up to 140 Hz by the end of the first syllable. Gradually, F0 falls down to 65 Hz by the end of sentence, going up 40 Hz intermittently at each word boundary. Consequently, a drawn F0 contour looks like a gently sloping saw tooth pattern that is tilted towards the left hand side. F0 does not go all the way down to 65 Hz at a phrase boundary within a sentence (i.e. a comma or a semicolon in the input text), but rises 5 Hz in addition to the ordinary 40 Hz rise at word boundaries. There was a 150 ms silence interval at phrase boundaries, and a 350 ms interval at sentence boundaries.

There was also a 46 s test file the participants heard before they began. It was generated using the older word models presented in [1]. There was also a prepausal lengthening module switched on rendering all phrase- and sentence-final words 10 % longer than the rest. Otherwise the configuration was identical to that of the actual test stimuli.

The rule-based synthesizer used for stimulus generation consists of three levels of processing. The syntagmatic level contains preprocessing and a number of modules for prosody and speaker parameters to choose from. The paradigmatic level holds the phoneme and allophone inventories. The third level is signal generation, now handled by JPSyn, a Klatt –type software of our own design.



Picture 2. *The structure of the TTS system.*

2.3. Participants

There were 21 participants, 7 women and 14 men. One of the participants was left-handed. Their average age was 28 years, the eldest being 45 and the youngest 23. The participants were asked about their primary and secondary dialect background, since there is considerable dialectal variation in Finnish speech prosody, segmental durations included. The majority of the participants were speakers of the South-Western dialects of Finnish. One of the South-Westerners was bilingual in Swedish and Finnish. There were six primarily Southern (includes the capital city Helsinki) speakers; an additional three listed a secondary background in Southern dialects.

The participants had to evaluate their experience with synthetic speech in general. The scale extended from 1 (hears synthetic speech several times a week) to 5 (has never been exposed to synthetic speech); the average for the group was 3.5, roughly corresponding to a few times a month.

2.4. Listening procedure

We preferred the test to be taken in a comfortable environment and through a likely medium for speech synthesis use. The participants got to access the entire test material over the internet and carry out the evaluation in their homes. They were instructed to exclude any disturbing noise or movement from their vicinities before beginning and to set the volume in their loudspeakers to a loud but comfortable level. There was also a synthetic 46 s test file (a greeting of a sort) the participants heard first; the test file utilized neither of the duration models under examination.

The test itself was a forced choice paradigm. The participants were instructed to listen to each of the stimulus pairs over as many times as they wanted. They were asked to mark which one of the sentences (A1 or A2, B1 or B2, etc.) they thought sounded more natural and better corresponded to human speech rhythm; the order of the stimuli in a stimulus pair was scrambled, and nothing about the alternative duration models was disclosed to the participants. They were specifically instructed not to let intelligibility issues affect their judgment.

Finally, they were asked to submit their personal information, including age, sex and handedness, and to identify their primary and secondary dialect backgrounds. They would also estimate their amount of personal experience with synthetic speech. The participants submitted their results and information by e-mail using an electronic answer sheet.

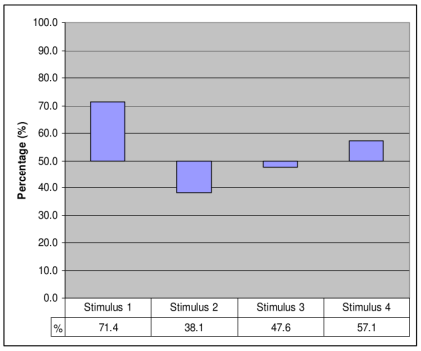
3. Results

The results show that the participants preferred the stimuli with word modeled segmental durations only slightly (53.6 % of the stimuli). The first and the longest one of the stimuli was preferred the most (71.4 % for the

word models). The second and shortest one of the stimuli was the least preferred (38.1 % for the word models).

Men preferred the word models more than women (58.9 % vs. 42.8% of the stimuli). Four of the participants preferred all the instances of the word models, while two of them preferred all of the fixed duration stimuli; no common denominer was found for them. Dialect background had no significant effect, but those more experienced with synthetic speech (reported an experience level of 2 or 3) were likely to prefer fixed durations (40 % for the word models). The least experienced ones (experience level of 4 or 5) were likely to prefer the word models (65.9 %); they rated the first stimulus better 90.9 % of the time. Three of the participants had preferred the stimulus they heard last in all four cases; this may or may not represent an unconscious bias, but it does not affect the outcome significantly.

Table 2. *Preference for the word-modeled stimuli and partial raw data (1=preference for the word models).*



	Stimulus 1	Stimulus 2	Stimulus 3	Stimulus 4
Participant 1	1	1	1	1
Participant 2	0	1	0	0
Participant 3	1	0	0	1
Participant 4	1	1	0	1
Participant 5	0	0	0	1
Participant 6	1	0	1	0
Participant 7	1	0	1	0
Participant 8	1	1	1	0
Participant 9	1	1	1	1
Participant 10	1	0	0	1
Participant 11	1	1	0	1
Participant 12	1	1	1	1
Participant 13	0	0	0	1
Participant 14	1	0	1	0
Participant 15	1	0	1	1
Participant 16	1	1	1	1
Participant 17	1	0	0	1
Participant 18	0	0	1	0
Participant 19	1	0	0	0
Participant 20	0	0	0	0
Participant 21	0	0	0	0

4. Discussion

The results show that the word models either enhance naturalness (stimulus A), hinder it (stimulus B), or have no effect at all (stimuli C and D). It appears that the longer the synthesized sample is, the more the word models enhace naturalness. That is in line with the preliminary findings in the previous study [1]. Several weaknesses can be identified in the word model approach. First, the word models require a large database. The ~6500 words in the corpus produced ~2500 models. The syllabic structure of Finnish, a highly inflected language, is so complex, that establishing an adequate database would require a much greater corpus. A complete set of word models would require an astronomical amount of entries, since there is no theoretical upper limit for the length of word forms in written Finnish. It may be of interest to examine how the word models perform in a language

which requires a limited set of word models (shorter words and less inflection).

Second, the fact that rule-based synthesis is computatively non-expensive and requires little memory capacity is one of the greatest advantages of the method. To implement a vast database would be a compromise in the latter respect. Third, word models represent a single speaker's speaking style. That may be seen as a disadvantage, if one wants to create a generalized, impersonal speaker. On the other hand, a TTS system might contain several speaker profiles with corresponding individual or dialect-specific segmental durations. Fourth, word models based on a large sample size tend to lose some of their shape due to averaging. Conversely, a word model that is based on a single token may reproduce effects of syntactic environment or information structure that are ill-fitted to other contexts.

The overall preference for the word model–determined durations is so weak, that their implementation is not necessarily justified considering the weaknesses. Fixed durations appear to do well in comparison even though they are counterintuitive; it is unlikely a natural language would operate with fixed durations. In fact, there is a chance that duration is not that important from the vantage point of speech perception. People are generally unaware that the acoustic correlate of phonemic length, duration, is relative. They are surprised to hear that within just one word, a phonemically long vowel may be shorter in duration than another short vowel. During speech perception, the brain apparently registers each speech sound as either long or short. The only thing that catches one's ear in the synthesis is when a segment is abnormally short or long; that happens occasionally with the current word models. Carefully modeled phonemes, transitions, F0, and the oft-neglected intensity may prove to contribute more to naturalness than duration. Sakamoto and Saito [5] studied synthetic speech modeled after donor speakers (VoiceFonts), and found that duration has a relatively small effect on speaker recognizability compared to other variables.

We have developed another model that combines the phonemes' intrinsic durations (some are longer than others on the average) with a generic word model. The generic word model makes the syllables grow shorter towards the end of the word, a tendency observed in the corpus as well as in Lehtonen's material [3]. In addition, there is a prepausal lengthening effect. If avoiding large databases is no question of principle, one could of course bring in more variables and create a very extensive word model bank. For one thing, the word models could be sensitive to syntactic roles (necessitates a syntactic parser for data mining purposes). Alternatively, the word model could cover only a limited sequence of phones, for instance the first eight in any word. The remaining phones would be dealt with using a generic model. Nevertheless, we are discouraged to continue the current line of investigation into word models. We are inclined

to find alternative methods to model segmental durations in rule-based text-to-speech synthesis.

5. Conclusion

We compared fixed segmental durations to those measured from a natural speech corpus. The results show that word modeled durations improved naturalness only partially according to the listeners' judgment. Therefore, we suggest that the syntactic structure of the sentences should be taken into consideration if the word model approach was developed further. The statistical analysis of duration in various syllabic structures alone is inadequate at least for a language such as Finnish.

References

[1] Hakokari, J., Saarni, T., Jalonen, M., Aaltonen, O., Isoaho, J. & Salakoski, T., 2005. Word-model determined segmental duration in Finnish speech synthesis and its effect on naturalness. M. Langemets & P. Penjam (Eds.) *Proceedings of the 2nd International Conference on Human Language Technologies.* Tallinn: Raamatutrükikoda. 137-142. Available online at http://users.utu.fi/tuiisa/pubs/

[2] Klatt, D., 1980. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America 67*. 971-995.

[3] Lehtonen, J., 1970. *Aspects of quantity in standard Finnish*. University of Jyväskylä.

[4] Lennes, M., 2003. On the expected variability of vowel quality in Finnish informal dialogue. M. Sóle, D. Recasens & J. Romero (Eds.) *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, pp. 2985-2988.

[5] Sakamoto, M., Saito, T. 2002. Speaker recognition evaluation of a VoiceFont-based text-to-speech system. *Proceedings of 7th International Conference on Spoken Language Processing,* pp. 2529-2532.

[6] Vainio, M., 2001. *Artificial neural network based prosody models for Finnish text-to-speech synthesis*. Helsinki: Yliopistopaino.

[7] Wiik, K., 1965. *Finnish and English vowels*. Turku: University of Turku.

Paper II

Determining Prepausal Lengthening for Finnish Rule-Based Speech Synthesis

Jussi Hakokari, Tuomo Saarni, Tapio Salakoski, Jouni Isoaho, Olli Aaltonen. Speech Analysis, Synthesis and Recognition, Applications of Phonetics, September 19-23, 2005, AGH University of Science and Technology, Kraków, Poland.

Determining Prepausal Lengthening for Finnish Rule-Based Speech Synthesis

Jussi Hakokari[1] & Tuomo Saarni[2]

Tapio Salakoski[2], Jouni Isoaho[2] & Olli Aaltonen[1]

[1]Phonetics Laboratory

[2]Department of Information Technology

University of Turku, Turku, Finland

jussi.hakokari@utu.fi & tuomo.saarni@utu.fi

ABSTRACT

We are developing a Finnish rule-based TTS system. Our primary concern is to enhance naturalness in the synthesis by observing tendencies in natural language and implementing the findings into the synthesis. We have concentrated on modeling duration, which is essential to the Finnish language due to contrasting phonemic length and the fact that the durations of individual phones are highly sensitive to their position within a word. One of the many durational considerations is prepausal lengthening. Most languages exhibit lengthening of speech sounds at the ends of phrases and sentences, but there is no applicable data available on the phenomenon with regard to Finnish. Data mining a speech corpus, we have set out to investigate what is the best justified way to implement prepausal lengthening to a Finnish rule-based TTS system. The results show that there is considerable lengthening of the entire prepausal word. Moreover, some of the effects appear to extend to the penultimate word. The data acquired in the study provides us adequate information for modeling prepausal lengthening in TTS.

1. Introduction

We are in the process of developing a Finnish rule-based TTS system. Our approach is to tackle the problem of naturalness associated with the rule-based methods by data mining natural speech corpora and trying to model the set of rules accordingly to best correspond to natural speech. For now, we are using a single-speaker corpus and aim to model that speaker instead of making compromises between a host of voices and speaking styles. Furthermore, we are developing a signal generator software that would perform better than the previous synthesizers such as Klatt. While a fully synthetic rule-based synthesis has its advantages, tolerable naturalness is rightly associated with the concatenative methods that rely on samplings

of recorded human speech. Much of our research has so far concentrated on quantity, which is a central issue in Finnish, a language with contrasting phonemic length and long, multisyllabic words [4].

We set out to model prepausal (final, preboundary) lengthening, a common phenomenon in speech, in our synthesis and came to an unexpected discovery. While there is virtually no literature on the subject specifically with regard to Finnish, several published and unpublished sources claim that namely the Finnish, Estonian (a language closely related to Finnish), and Japanese languages exhibit either little or no prepausal lengthening. Vaissière [11] presents the claim, referring to Lehiste [6]. Lehiste's work, however, deals with phonological length (the phonemic contrast between short and long speech sounds) and the syllabic structures of Finnish and Estonian. She does not discuss any acoustic phonetic duration, prepausal lengthening or other. Curiously, D'Odorico & Carubbi [3] report final syllable lengthening (FSL), a more specific instance of prepausal lengthening, to be actively suppressed in languages such as Finnish, Estonian, and Japanese. They cite Oller [9] as the source; Oller's study again mentions Finnish in no way. Conversely, Bishop [1] refers to a study by Johnson & Martin [5] claiming that final lengthening effects have been reported in Finnish among other European languages. Johnson and Martin, on the other hand, refer to Lehtonen [7], whose study deals with segmental durations in a number of syllable and phoneme environments. Lehtonen does not, however, present any conclusive data namely about prepausal lengthening. As a conclusion, it has to be admitted there is little explicit evidence that Finnish speakers would lack prepausal lengthening, a phenomenon thought by many to be a universal tendency and physiological in nature. A prepausal lengthening effect has been observed in Estonian, but it appears weaker than in English and not as pervasive [8].

Vainio's [10] investigations on the same corpus used in this study revealed a prepausal lengthening effect. Vainio's study, however, involved prosody in general and he did not describe the effect in much detail. We set out to investigate the effect more closely in order to obtain sufficient data to implement in Finnish TTS. We used data mining to compare mean durations of segments in prepausal words to those in other positions. While shedding some light on the issue of prepausal lengthening in Finnish and thus contributing to anyone interested in its potential status as a universal feature is a worthy goal, our primary concern is to find a phonetically justified and computationally inexpensive manner of implementing the effect to a TTS system.

2. Methods

2.1. Speech Material

The speech material was a single-speaker corpus of 692 Finnish declarative sentences read aloud by a 39-year-old male from Helsinki. The corpus contained ~6500 words. The sentence durations ranged from ~2 to ~20 seconds adding up to 69 minutes of recorded speech. The sentences are randomly picked from a Finnish periodical (Suomen Kuvalehti), with the exception that foreign words and foreign proper names have been avoided. The material is clearly articulated and represents Standard Finnish. The speaker has a relatively low pitch voice, and has a distinctive creak (final devoicing) in the ends of sentences, a common trait among Finnish speakers.

The corpus was segmented and annotated at word and phone levels. Pauses (silence) were separately marked in the phone level annotation. The annotation appears accurate and consistent. All the data mined information presented in the paper ultimately relies on the annotator's judgement on phone boundaries. A thorough description of the corpus and the annotation criteria are found in Vainio [10].

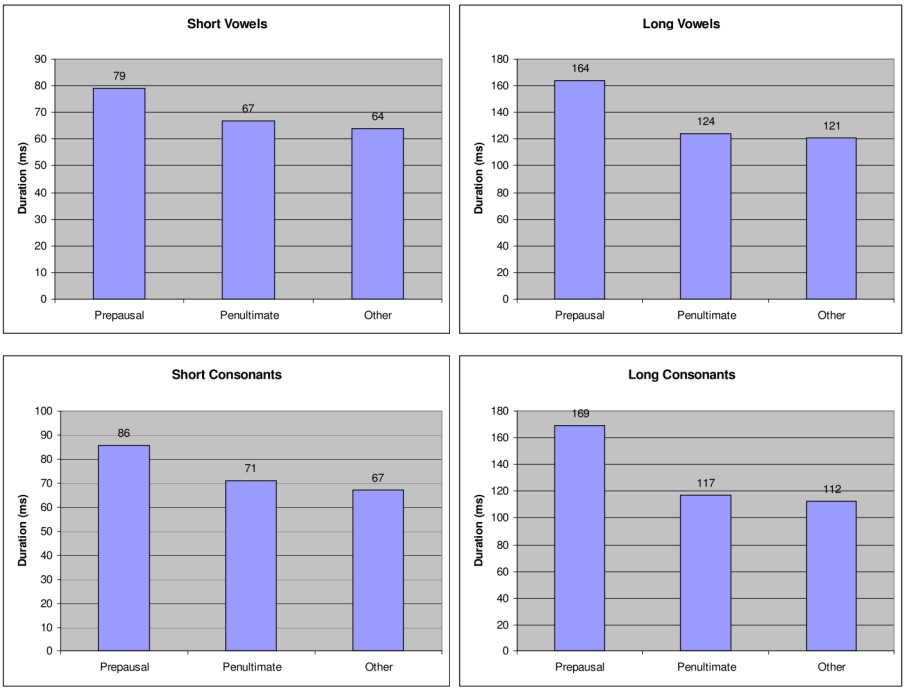
2.2. Data Mining and Procedure

Calculating mean segmental durations was done automatically using the word and the phone level annotation in the corpus. The mean segmental durations of the entire phrase-final (prepausal) words are contrasted with the rest of the speech material. Thus, syllables, final or penultimate, are not examined alone. Multisyllabic wordforms are more the rule than the exception in Finnish; the words in a prepausal position may consist of a number of syllables. Furthermore, the ends of phrases and the ends of sentences were not separated; the method treats all the words that precede a silence or the end of an utterance in the recording equally (with the exception of silences during plosive consonants, which, of course, do not trigger a pause). The penultimate words, the ones that precede the prepausal word, are also examined to determine whether the effect is extended beyond the last word before a pause.

Both the prepausal and the penultimate words were compared against the rest that consisted of phrase-initial and medial words exclusively. The rest, in other words, did not include the prepausal or the penultimate words themselves. There were two levels of examination. First, we examined the data by grouping individual phones into four categories: phonemically short consonants (C), long consonants (C:), short vowels (V), and long vowels (V:). The idea was to find out whether there is a lengthening effect and whether it applies to all the four categories. Second, we examined the lengthening effect phoneme-by-phoneme to find out if there were differences in the behaviour of different speech sounds.

3. Results

The present data suggests there is a clear prepausal lengthening effect even when the entire prepausal word is examined. That applies to all the speech sounds. The phonemically short consonants were 28.4 % longer in a prepausal position than in other positions (penultimate positions excluded). The corresponding figures for the long consonants are 50.9 %, 23.4 % for the short vowels, and 35.5 % for the long vowels. Contrary to what was expected, the lengthening selectively affected even the penultimate word. All the four categories were longer in the penultimate words compared to the overall duration of words in the medial and the phrase-initial positions (C:= 4.5 % longer, V= 4.7 %, V:= 2.5 %, and C= 6.0 %).



Picture 1. *The amount of lengthening in prepausal and penultimate positions compared to other positions.*

Table 1. *Sample sizes of all the prepausal, penultimate, and initial/ medial phones in the study.*

	Sample Size N
Prepausal Short Vowels	3422
Prepausal Long Vowels	397
Prepausal Short Consonants	3700
Prepausal Long Consonants	405
Penultimate Short Vowels	2915
Penultimate Long Vowels	276
Penultimate Short Consonants	3152
Penultimate Long Consonants	293
Other Short Vowels	12836
Other Long Vowels	1289
Other Short Consonants	13694
Other Long Consonants	1407

The prepausal lengthening effect was also examined on the phoneme level. All the vowels and the consonants were affected, including the phonemically long and short sounds, but to varying degrees. The short consonants with a sample size less than 10 are all non-native and only occur in loan words. Of these the voiced velar plosive /g/ (N=7) was an exception; it was actually shorter in prepausal positions. The lengthening of the penultimate words was not examined at the phoneme level.

Table 2. *Phoneme-to-phoneme results. The sample size refers to occurrences in a prepausal position.*

Phoneme	Percentage of prepausal lengthening	Sample size N
ø	25.4	52
æ	18.1	273
ɑ	17.3	861
e	24.6	569
i	28.4	854
o	27.8	366
u	27.1	309
y	34.8	138
b	15.1	8
d	32.9	78
f	36.4	6
g	-7.3	7
h	34.6	187
j	24.2	93
k	27.2	400
l	22.0	317
m	21.6	224
n	14.1	499
p	23.2	141
r	22.0	202
s	32.9	618
t	27.0	669
ʋ	17.4	223
ŋ	49.9	24
ʃ	132.5	4

Phoneme	Percentage of prepausal lengthening	Sample size N
ø:	52.7	2
æ:	42.0	64
ɑ:	35.5	133
e:	44.5	56
i:	32.7	52
o:	41.8	19
u:	25.6	55
y:	17.8	16
k:	42.2	36
l:	53.5	106
m:	37.3	19
n:	48.3	31
p:	25.2	14
r:	91.9	9
s:	60.3	78
t:	52.0	98
ŋ:	39.5	13

4. Discussion

Our data clearly suggests a prepausal lengthening effect of the phrase-final word. The current data is based on a single speaker. While the sample size of individual phones is adequate, we cannot generalize the effect to the entire Finnish-speaking community; a future study must include several speakers, preferably representatives of different dialect backgrounds. Our data contains declarative sentences exclusively; a future investigation should establish whether the phenomenon is present in questions and commands as well. The current data represents newspaper speech. Our study does not answer the question whether prepausal lengthening takes place in (spontaneous) conversational speech. However, a recent study of prepausal lengthening in Russian by Volskaya & Stepanova [12] suggests that the effect is not style specific. This study has only addressed segmental duration on the word level. The role of syllabic structure is yet to be determined. Furthermore, it is necessary to produce a more detailed description of how the effect is distributed towards the phrase boundary.

The phoneme-to-phoneme examination revealed that there are differences between individual phonemes, but we failed to find any systematics in the variation other than that the short voiceless consonants were lengthened somewhat more than the voiced ones (~29.1 % vs. ~20.0 %). That is in line with the tendency first observed by Cooper & Danly [2] The intrinsic mean durations of phonemes are easy to implement as such in a TTS system, and more naturalness may be achieved by using multipliers for phones in penultimate and prepausal positions. That presents a very simple and straightforward solution for introducing prepausal lengthening into a Finnish rule-based TTS system and warrants experimentation.

5. Conclusion

The categorical notion that the Finnish language lacks prepausal lengthening does not seem to hold in the light of our current data. Also the claim that Finnish speakers somehow resist prepausal lengthening effects lacks evidence. We have shown a significant lengthening effect that is observable even when the entire prepausal word is under examination. We have also extended the lengthening effect to the penultimate word; the penultimate words in their entirety show a weak effect especially in the phonemically long consonants. Its significance is dubious, however, and we are inclined to conduct a more detailed study on the distribution of the lengthening effect. Even though the data is based on a single-speaker corpus and should not be overly generalized, it serves as a basis for designing a preliminary module that introduces prepausal lengthening into our Finnish rule-based TTS system.

References

[1] Bishop, J. B. 2003. *Aspects of intonation and prosody in Bininj Gun-wok: an autosegmental-metrical analysis*. PhD thesis, School of Graduate Studies, University of Melbourne.

[2] Cooper, W. E., Danly, M. 1981. Segmental and Temporal Aspects of Utterance-Final Lengthening. *Phonetica*, 38, pp. 106-115.

[3] D'Odorico, L., Carubbi, S. Prosodic characteristics of early multi-word utterances in Italian children. *First Language*, 23(1), pp. 97-116.

[4] Hakokari, J., Saarni, T. Jalonen, M., Aaltonen, O., Isoaho, J., Salakoski, T. 2005. Word model-determined segmental duration in Finnish speech synthesis and its effect on naturalness. *Proceedings of The Second Baltic Conference on Human Language Technologies*, Tallinn.

[5] Johnson, K., Martin, J. 2001. Acoustic vowel reduction in Creek: effects of distinctive length and position in the word. *Phonetica*, 58(1-2), pp. 81-102.

[6] Lehiste, I. 1965. The function of quantity in Finnish and Estonian. *Language 41*, 447-456.

[7] Lehtonen, J. 1970. *Aspects of quantity in Standard Finnish*. Jyväskylä: K.J. Gummerus Oy.

[8] Mihkla, M. 2005. Modelling pauses and boundary lengthenings in synthetic speech. *Proceedings of The Second Baltic Conference on Human Language Technologies*, Tallinn.

[9] Oller, K. 1973. The effect of position in utterance on speech segment duration in English. *The Journal of the Acoustical Society of America,* 51, pp. 1235-1247.

[10] Vainio, M. 2001. Artificial Neural Network Based Prosody Models for Finnish Text-to-Speech Synthesis. Academic dissertation, University of Helsinki.

[11] Vaissière, J. 1983. Language independent prosodic features. A. Cutler & R. Ladd (Eds.) *Prosody: Models and Measurements*, (53-65). Springer Verlag.

[12] Volskaya, N., Stepanova, S. 2004. On the temporal component of intonational phrasing. *Proceedings of the 9th Conference "Speech and Computer"*.

Paper III

Segmental Duration in Utterance-Initial Environment: Evidence from Finnish Speech Corpora

Tuomo Saarni, Jussi Hakokari, Jouni Isoaho, Olli Aaltonen and Tapio Salakoski. FinTAL 2006, 5th International Conference on Natural Language Processing, 23-25 August 2006.

Segmental Duration in Utterance-Initial Environment: Evidence from Finnish Speech Corpora

Tuomo Saarni[1], Jussi Hakokari[2], Jouni Isoaho[1], Olli Aaltonen[2], and Tapio Salakoski[1]

[1] Turku Centre for Computer Science, Finland
{tuomo.saarni, jouni.isoaho, tapio.salakoski}@it.utu.fi
[2] Phonetics Laboratory, University of Turku, Finland
{jussi.hakokari, olli.aaltonen}@utu.fi

Abstract. This study examines segmental durations produced by Finnish speakers in utterance-initial environments. We have established a method to statistically examine segmental duration on the phone level in speech corpora. The two corpora represented in this study consist mainly of television news broadcasts and short texts read aloud by professional speakers. Previous studies conducted have been contradictory; there are reports of initial shortening in certain languages and lengthening in others. Our results are conclusive in neither way, but suggest a qualitatively differentiated behavior. We have observed lengthening of all utterance-initial vowels, diphthongs included, and shortening of phonologically long plosive (stop) consonants. No other speech sounds are significantly affected. These findings hold in both corpora, in despite of different speakers and annotators.

1 Introduction

Final, prepausal, and pre-boundary lengthening have been reported in a great number of languages, leading us to believe it is in fact, depending on the point of view, either a phonetic or a linguistic universal. Final lengthening refers to the human tendency to slow down articulatory movements in the ends of utterances or syntactic units, effectively increasing the duration of individual speech sounds. There is debate over what are the origin and the possible function of the phenomenon. We do know lengthening is a powerful boundary signal, especially since lengthening alone at a boundary can produce a sensation of the speaker temporarily pausing [3]. White [12] calls such modulations of speech in both initial and final environments domain-edge processes.

There are reports of domain-edge processes at the other end of the utterance, as well. Unlike final lengthening, on which the studies mostly agree, domain-initial processes have produced conflicting results. Some languages are mentioned to display shortening of initial speech sounds in the literature, while perhaps the majority is reputed to lengthen utterance-initial speech sounds. Kaiki et al. [8] report having found shortening in Japanese, although Campbell [1] has criticized their corpus of imbalance. Nagano-Madsen [10] reports initial shortening in Eskimo. Hansson [6] makes a very strong case for initial shortening in Southern Swedish. Initial lengthening has been

T. Salakoski et al. (Eds.): FinTAL 2006, LNAI 4139, pp. 576 – 584, 2006.

reported, among others, in Korean [2], Standard Chinese [13], and English [12]. The locus and extent of lengthening or shortening varies from language and method to another. White [12], for instance, has found certain speech sounds to occur shorter while the others are lengthened. Furthermore, initial lengthening is sometimes credited as a consequence of initial strengthening, a phenomenon causing stronger contact between the tongue and the palate in domain-initial consonant articulations. The study at hand will only address segmental duration in Finnish. We will move phoneme by phoneme from phrase-initial towards medial positions and observe how long different kinds of phonemes are realized.

2 Methods and Materials

Many studies into domain-edge processes and segmental duration have relied on behavioral experimental designs, such as reading aloud nonsense words embedded into carefully designed sentence structures. Our methodology is different; we wish to observe segmental duration statistically in vivo; in ordinary, unrestricted speech flow. Furthermore, while the studies generally operate on word or syllable level, we have chosen a phone-level approach. The method appears novel in speech timing research. We designed search scripts to read the annotation files which contained the speech sounds' identities along with their position and duration information. The output of the scripts was manually inspected to verify the resulting data.

2.1 Speech Corpora

We studied the effect of utterance-initial position on segmental durations using two corpora. Both are in Standard Finnish, the literary language of both spoken and print media. The first one is a single-speaker corpus consisting of 964 utterances, read aloud by a 39-year-old male. The corpus is described in more detail in Vainio [11]. The other one is a multi-speaker corpus featuring 9 men and 4 women, all professional speakers. It originates with the Finnish Broadcasting Company and consists of news reading, interviews, and oral presentations. There are 802 utterances altogether. Both corpora were annotated manually by trained phoneticians, although the annotation strategies were slightly different due to independent annotators.

2.2 Procedure

Segmental duration was examined by two criteria: position by position and by the phoneme category. We mapped all vowels, consonants, etc. into a chart by their position in their respective utterances. The phoneme categories were vowels, nonplosive consonants, and voiceless plosives. Their phonologically contrasting short and long counterparts were further separated. The annotation of the multi-speaker corpus also recognized diphthongs, which were marked as two consecutive short vowels in the single-speaker corpus. Therefore short vowels are not entirely

comparable between the two corpora, as the single-speaker material contains short vowels that are in fact diphthongs cut in two halves. To clarify the method established, let us consider the following example from the single-speaker corpus:

/suui mene:/ (transl. 'the summer passes')

In the short non-plosive consonant chart (below in the results section), the first value (position 1) is the mean duration of all short non-plosive consonants that are the very first phone in an utterance (/s͡uui/). The phonemically long vowel /e:/ in /mene:/ is therefore in the eight position of the long vowel chart. /ʊ/ in /suui/ is in the third position of the short non-plosive consonant chart. By organizing the phones separately by the category, we are not tied to higher-level units such as syllable or word duration, as the previous studies have. We can get accurate information on as how long different kinds of phones are realized in the utterance-initial position, and how their duration evolves as we move further towards the end. For reference, the mean duration of all phones of a category (regardless of their position) is represented as horizontal lines in the charts.

Our studies have revealed a significant prepausal lengthening effect in both the single-speaker [5] and the multi-speaker corpora (unpublished). The lengthening takes place as early as the 10th last phone in the utterance. To study duration in initial position, it was necessary to eliminate any prepausal lengthening from the speech material. It was done by removing the last ten phones from each utterance in the corpora, effectively excluding all utterances with 10 or less phones in them. For instance, we have found very short utterances entirely affected by lengthening, and they will not give accurate information on specifically utterance-initial phenomena. The procedure left 934 utterances of the original 964 (all remaining ones 10 phones shorter than before) for the single-speaker corpus, and 679 of the 802 utterances for the multi-speaker one. Utterance is used in a purely acoustic sense here; any single, continuous chunk of speech, limited by silence (pauses) at both ends, qualified as an utterance. While those pauses usually co-occur with syntactic boundaries, the annotation was not syntactically motivated.

Table 1. The sample sizes and mean durations in milliseconds of the phoneme categories in both corpora (having removed the 10 last phones of each utterance)

	Single-speaker corpus		Multi-speaker corpus	
	Sample size	Mean duration	Sample size	Mean duration
Diphthongs			817	117,4
Long vowels	1456	121,6	649	105,9
Short vowels	15014	65,1	5484	58,6
Long non-plosive consonants	934	91,9	456	85,2
Short non-plosive consonants	11157	60,3	5359	60,7
Long plosives	667	142,4	284	130,0
Short plosives	5043	85,3	2452	74,4

3 Results

The data is shown in the following charts. The positions are on the horizontal axis and the mean durations (absolute values) on the vertical axis. The bold line represents the mean duration of the given phoneme category in, and only in its respective position. The dotted lines represent the confidence limit (level of confidence $p \leq 0.05$). The straight grey line is the mean duration of all phones of the given category, regardless of their position. We considered statistically significant a case in which the entire confidence limit is situated above or below the mean.

The phonologically long sounds, presented as the upper line in the graphs below, have more variation because they are relatively infrequent in Finnish, even when compared to other quantity languages [4]. In our data, the phonologically long speech sounds make up less than 10 % of all sounds [table 1]. The variation occasionally

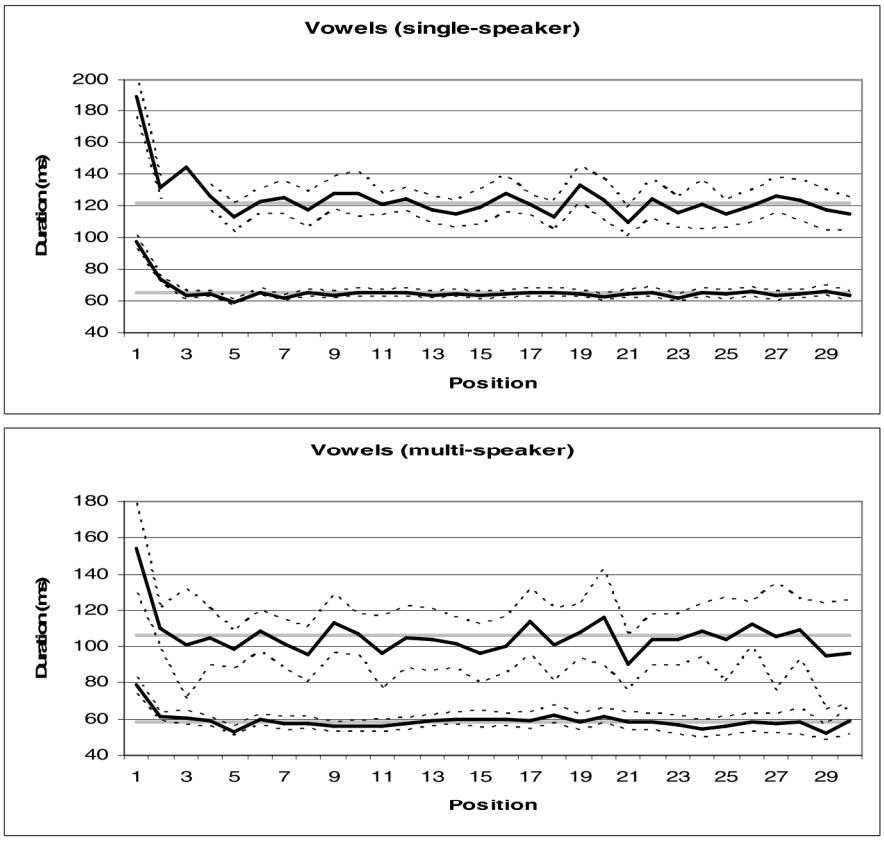


Fig. 1. Short and long vowels. The upper solid lines represent phonologically long sounds.

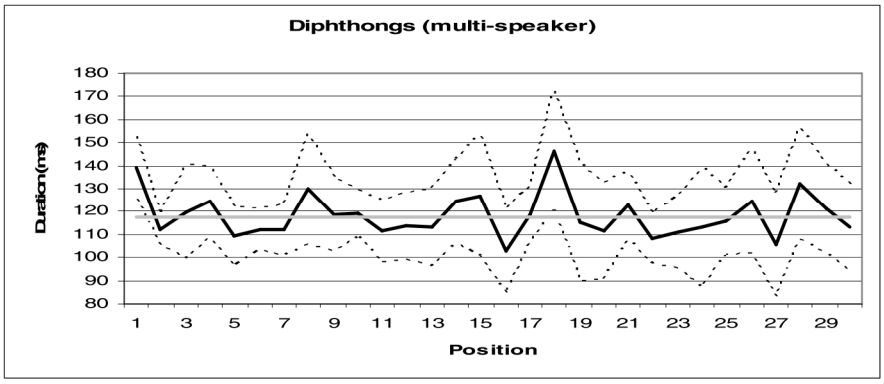


Fig. 2. Diphthongs (not available in the single-speaker corpus)

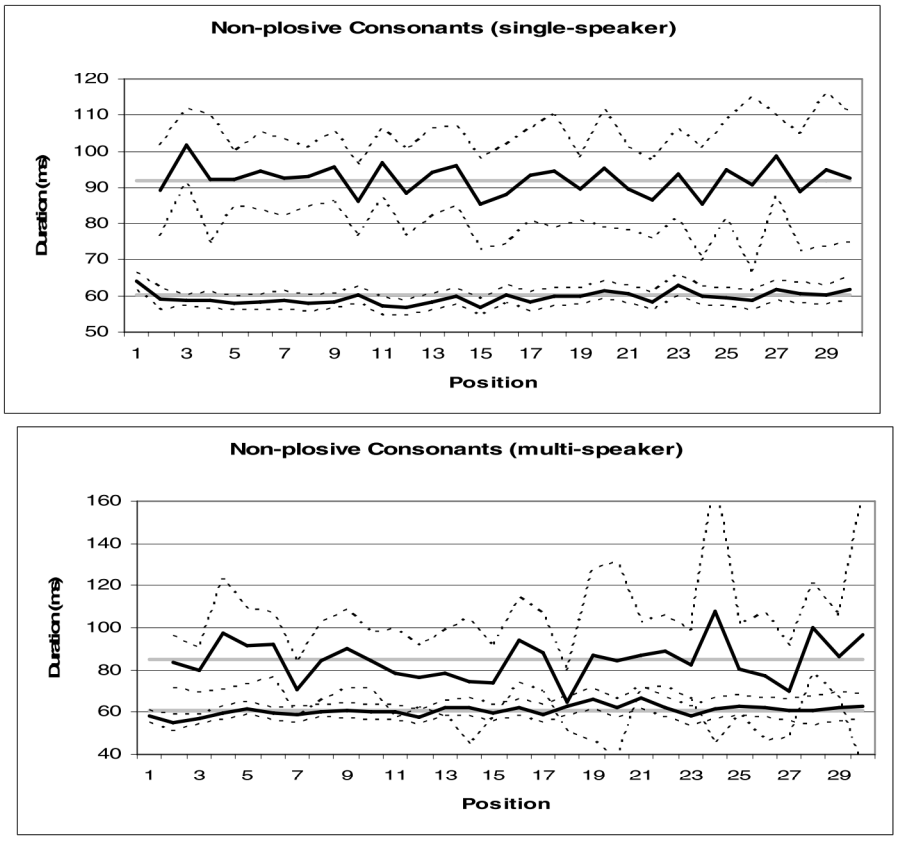


Fig. 3. Non-plosive consonants. The upper solid lines represent phonologically long sounds.

introduces statistically significant points in seemingly haphazard positions, and we will only discuss issues we feel may carry weight. Generally, the figures are more reliable towards the initial position, since the sample size decreases towards the end of the scale. Every utterance in the study is at least one phone long, of course, and therefore contributes to the first position.

Figure 1 shows that vowels, both short and long, are lengthened in the first position in both materials. There is a trace amount of lengthening in the second position, as well. The third position appears longer with long vowels in the single-speaker data, but that is insignificant as there is only one sample of the position (hence the missing level of confidence). The difference between the two materials is most likely due to individual variation.

According to figure 2, diphthongs in the multi-speaker corpus behave in the same fashion as the rest of the vowels. The lengthening is clear in the first position. The diphthongs were annotated only in the multi-speaker corpus, hence the lacking single-speaker figure. A statistically significant lengthening effect also appears in position 18, which we suspect is merely an artifact.

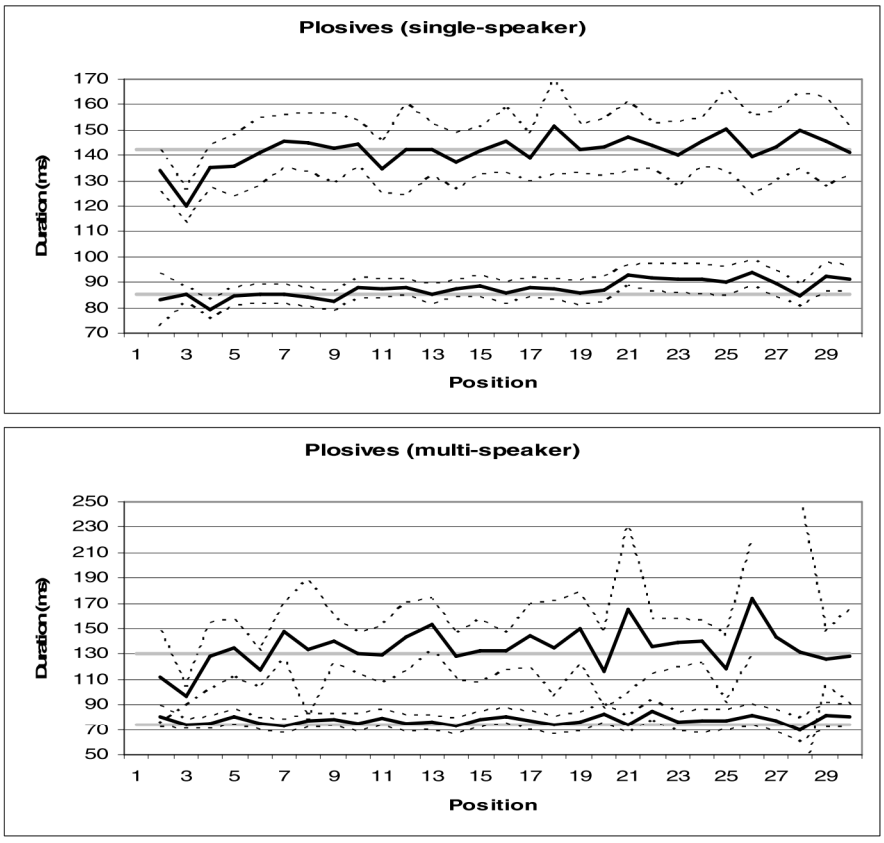


Fig. 4. Plosive consonants. The upper solid lines represent phonologically long sounds.

Figure 3 suggests the short non-plosive consonants are slightly longer in the 1st position in the single-speaker corpus, but not in the multi-speaker corpus. Instead, the short ones are slightly shorter until the 2nd position. We are uncertain whether those minimal deviations are meaningful. The long consonants did not deviate from the average duration significantly. Long consonants, plosives included, have no value in the 1st position; they are invariably geminates occurring in medial positions.

Figure 4 indicates the phonologically long plosives (geminates) were shorter in initial environment. The shortening was strongest in the 3rd position in both materials. The second position is similarly shorter than the mean in both, but lacks statistical significance in the multi-speaker data. The 3rd position corresponds to the boundary between first and the second syllable. The short plosives were not affected significantly.

There are no initial long consonants, and the short plosives were excluded since the majority of them have no reliably discernible duration. We can usually only spot the explosion phase, while the actual onset of articulation remains invisible and inaudible. The 27th position in the multi-speaker data has only one sample and therefore no confidence limit.

4 Discussion

The Finnish language is characterized by a very small phoneme inventory, restrictive phonotactics, and a fixed first-syllable lexical stress. However, a pervasive quantity system (almost doubling the inventory) and long, multi-syllabic words compensate in preventing homophony. The phonological structure makes Finnish different from many other European languages, and has in fact led some to question whether domain-edge processes, such as final lengthening, can be observed in the language. Our work [5] with both corpora clearly indicates Finnish is not exempt. Final lengthening has also been confirmed in Hungarian [7] and Estonian [9], both genetically related quantity languages previously thought not to display lengthening on similar grounds as Finnish.

The stressed syllable (always the first one in Finnish) is generally thought to prescribe longer duration. The lengthening of the vowels in the first position cannot be explained away as a simple stress issue, however. Namely, the second position of a vowel, showing little if any lengthening, corresponds to a consonant-initial first syllable, such as /su.ʊi/ (‘summer’). That points toward a boundary process. The applicable structures are either a single-vowel syllable, such as /e.si.ne/ (‘an object’) or vowel-initial closed syllable /is.ku/ (‘an impact’). Furthermore, the lengthening applies to long vowels as well as diphthongs, such as /ei.len/ (‘yesterday’) and /uː.ti.nen/ (‘a piece of news’). We would also expect lengthening of consonants until the 3rd position if we were dealing with lengthening of the stressed syllable. There is nothing of such nature in our data.

Phonologically long plosive consonants are affected in reverse fashion. While they cannot exist word-initially, they occur significantly until the second and the third position. The phonologically long consonants are all voiceless and geminate

(/pː tː kː/); they do not occur within syllables but at syllable boundaries. Expected cases of shortening thus include structures such as /kuk.ka/ ('a flower'), /ilk.ka/, (a given name), /uː.ta/ ('something new'), and /ap.pi/ ('a father-in-law'). We have not conducted behavioral listening tests, but the acoustic results suggest the shortening does not obscure the phonemic contrast between short and long plosives in these positions. The ratio is narrowed from ~1:1.7 to ~1:1.4 in the single-speaker data, and from ~1:1.7 to ~1:1.3 in the multi-speaker data.

A future investigation would benefit from at least three expansions. The inclusion of a spontaneous or conversational speech corpus is needed to determine whether the effects are style-specific. Examining not categories (vowels, plosives, etc.), but phonemes separately, would be useful, since any lengthening or shortening may have to do with specific articulatory dynamics of certain speech sounds. A simple division into categories may be too crude. Nevertheless, increasing sample sizes would yield more reliable and more easily interpreted results.

It is notable that the two corpora used in this study are very different. The first one has only one speaker reading aloud short passages of text, while the other is a mixture of speakers in various environments with considerably varying speaking rates. The annotation is also conducted independently by different annotators. Still the somewhat unexpected results are congruent between the corpora.

5 Conclusion

Various studies suggest speakers tend to articulate either faster or slower when they begin to speak or carry on speaking having paused. The phenomenon, although not studied to a great extent, has become known as initial lengthening or shortening. We have studied the effect of utterance-initial environment on segmental duration in Finnish. We have established a method of studying segmental duration on the level of individual phones instead of syllables or words. While it does forfeit the advantage of other domains, especially the syllabic structure, it allows us to examine the speakers' behavior in greater detail.

In the light of our current data, initial environment does affect segmental duration in Finnish, but it cannot be described as either initial lengthening or initial shortening. Vowels and phonologically long plosive consonants were significantly affected, but in the opposite manner. Utterance-initial vowels (as opposed to all first-syllable vowels) were lengthened, while long plosives were shortened up until the third syllable. The rest were affected minimally if at all. If the initial environment does shorten certain speech sounds while lengthening the others, the conflicting results from the other languages may be based on biased material and call for a re-evaluation of the methods used. Or, we should accept the asymmetry and recognize that the initial processes are not so uniformly represented in natural languages as final lengthening. Our materials and method showed a combination of the lengthening and shortening, necessitating further studies before we can establish on what principles utterance-initial duration operates in Finnish.

Acknowledgements

This study was funded by the Finnish Funding Agency for Technology and Innovation (TEKES). Our studies in segmental duration are aimed at developing duration modeling for speech synthesis. We would like to thank Professor Martti Vainio (University of Helsinki) for providing us with the single-speaker corpus.

References

1. Campbell, N.: Segmental Elasticity and Timing in Japanese Speech. In Speech Perception, Production, and Linguistic Structure. Y. Tohkura, E.Vatikiotis-Bateson, andY. Sagisaka, Eds. IOS Press (1992) 403-418
2. Chung, H., Gim, G., Huckvale, M.: Consonantal and Prosodic Influences on Korean Vowel Duration. Proceedings of Eurospeech, Vol.2. Budapest, Hungary (1999) 707-710
3. Duez, D.: Acoustic Correlates of Subjective Pauses. Journal of Psycholinguistic Research, Vol. 22(1). (1993) 21-39
4. Greenberg, J.: Language Universals: with Special Reference to Feature Hierarchies. Mouton (1966)
5. Hakokari, J., Saarni, T., Salakoski, T., Isoaho, J., Aaltonen, O.: Determining Prepausal Lengthening for Finnish Rule-Based Speech Synthesis. Proceedings of Speech Analysis, Synthesis and Recognition: Applications of Phonetics (SASR 2005). Krakow, Poland (2005)
6. Hansson, P.: Prosodic Phrasing in Spontaneous Swedish. Academic Dissertation. Travaux de l'institut de linguistique de Lund 43. Lund: Lund University (2003)
7. Hockey, B.A., Fagyal, Zs.: Phonemic Length and Pre-Boundary Lengthening: an Experimental Investigation on the Use of Durational Cues in Hungarian. Proceedings of the XIVth International Congress of Phonetics Sciences, San Francisco (1999) 313-316.
8. Kaiki, N., Takeda, K., Sakisaga, Y.: Statistical Analysis for Segmental Duration Rules in Japanese Speech Synthesis. In proceedings of the 1990 International Conference on Spoken Language Processing. Kobe, Japan (1990) 17-20
9. Krull, D.: Prepausal Lengthening in Estonian: Evidence from Conversational Speech. In Lehiste, I., Ross, J. (eds.), Estonian Prosody: Papers from a Symposium, Proceedings of the International Symposium on Estonian Prosody, Tallinn, Estonia. Tallinn: Institute of Estonian Language (1997) 136-148
10. Nagano-Madsen, Y.: Temporal Characteristics in Eskimo and Yoruba: a Typological Consideration. In Papers from the Sixth Swedish Phonetics Conference. Göteborg (1992)
11. Vainio, M.: Artificial Neural Network Based Prosody Models for Finnish Text-to-Speech Synthesis. Academic dissertation, University of Helsinki. (2001)
12. White, L.S.: English Speech Timing: a Domain and Locus Approach. University of Edinburgh PhD dissertation (2002)
13. Zu, Y., Chen, X.: Segmental Durations of a Labelled Speech Database and its Relation to Prosodic Boundaries. In Proceedings of the 1st International Symposium on Chinese Spoken Language Processing (ISCSLP 1998) (1998)

Paper IV

Utterance-initial Duration of Finnish Non-plosive Consonants

Tuomo Saarni, Jussi Hakokari, Olli Aaltonen, Jouni Isoaho, Tapio Salakoski. NODALIDA 2007, the 16th Nordic Conference of Computational Linguistics, 25-26 May 2007 Tartu, Estonia.

Utterance-initial duration of Finnish non-plosive consonants

Tuomo Saarni
Department of Information Technology

University of Turku
FI-20014 TURKU
tuomo.saarni@utu.fi

Jussi Hakokari
Department of Information Technology/
Phonetics Laboratory
University of Turku
FI-20014 TURKU
jussi.hakokari@utu.fi

Olli Aaltonen
Phonetics Laboratory
University of Turku

Jouni Isoaho
Dept. of Information Technology
University of Turku

Tapio Salakoski
Dept. of Information Technology
University of Turku

Abstract

We have investigated utterance-initial duration of non-plosive consonants in two qualitatively different Finnish speech corpora. The goal has been to identify any possible lengthening or shortening effects the domain edge (here, the beginning of an utterance) might have on segmental duration. Duration was observed at phone level. The results indicate that cases of lengthening, shortening, and absence of any effect all occur. Those are determined by the speech sounds phonemic identity, and the results were similar in both corpora. For instance /s/ and /r/ are lengthened while /j/ and /m/ are shortened. Contrasted with previous research on various languages, the phonetic universality associated with final lengthening does not apply for initial duration processes.

1 Introduction

Several domains (levels of phrase or utterance) have been credited to show initial domain-edge processes in various languages. These processes have mainly been referred to as either initial lengthening or shortening. Lengthening refers to cases in which speaking rate is briefly decelerated right as the speaker commences articulation. Shortening is the opposite; the speaker accelerates (producing relatively short segments) before resuming normal pace. Final lengthening, a domain-edge process involving considerable slowing down at the ends of utterances, has been found in practically all languages investigated. Yet initial effects have produced contrasting results depending on the language in question and the methodology used.

Initial lengthening has been reported in Chinese (Zu & Chen 1998, Cao 2004; syllable duration). Languages with reported shortening include Swedish (Hansson 2003; syllable duration at word level), Japanese (Kaiki et al. 1990), and Eskimo (Nagano-Madsen 1992). Venditti & van Santen (1998; phone duration) report initial lengthening of consonants and shortening of vowels in Japanese. White (2002) has found differences between various English consonant sounds.

The phenomenon is likely to be related to initial strengthening, a stronger contact between associated articulators such as the tongue and palate. These two effects, however, have been connected in no uniform fashion. For instance, Fougeron & Keating (1997) found that while for American English /n/ there is spatially greater linguo-palatal contact initially than medially, the acoustic duration is in fact shorter. The opposite was found in Korean by Cho & Keating (2001); in Korean initial strengthening and lengthening appear to correlate. Fougeron (2001) has suggested strengthening may be language-specific. Fougeron also (2001) claims articulatory variations in initial position are not conditioned by pauses but occur also internally at boundaries. In another tradition of terminology, boundary-adjacency is the common name for finality and initiality.

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 160–166

Previous investigation (Saarni et al. 2006) into the matter has revealed what could be considered initial domain-edge effect on segmental duration. Lengthening was found in all utterance-initial vowels (diphthongs included) in syllables such as V or VC. The lengthening did not extend to the entire first syllable (such as CV or CVC); cf. Byrd (2000) for similar observation in English. There was also shortening of phonologically long plosives, but the general category of non-plosive consonants was hardly affected. The category contains many articulatorily diverse sounds, however. Since edge effects in them has been documented in other languages (cf. White 2002), we decided to examine them phoneme by phoneme to find out if there are contrasting qualities that were neutralized in the former categorical examination.

The potential of corpus studies and phone-level approach have been mostly overlooked in previous research. Syllable-level studies, mainly on traditional elicited laboratory speech, have dominated duration research. Our previous results have led us to believe a syllable-level examination will miss some of the finer details of domain-edge processes. Not only can the observed phenomenon operate on a finer time scale than the syllable, but phonemically specific behaviors do not show if syllables consisting of different sounds are treated equally.

The study at hand covers the native Finnish consonants with the exception of plosives and any phonemically long consonants. Plosives are usually impossible to measure in initial position since the sound signal carries no trace of the initiation of the implosive phase. Long consonants, on the other hand, are geminates and may not occur in initial position.

This study is limited to the paradigm of corpus-based speech acoustics, and cannot as such address the question of initial strengthening. However, acoustic duration of speech segments will be carefully examined, allowing us to contribute to the controversy around the seemingly language-dependent domain-initial edge effects.

At this point, when little conclusive has been presented on the subject, we need to recognize the possibility of two different kinds of initial edge effects. First, the first position is articulatorily peculiar in that there is no excitation sound until the contact between articulators is already made. For instance, plosives usually do not have audible implosion phases. Fricatives and approximants may also start out "half-way", at the point when a constriction of the vocal tract is already reached. Second, a longer lasting compression or expansion may take place (cf. Hansson 2003) independently of the above-mentioned effect, just like final lengthening usually increases segmental duration over a number of phones. White (2002) also points out that utterance-initial syllable onsets are shorter than word-initial onsets utterance-medially.

First, the speech corpora used in the study are briefly described. Second we explain the way in which the statistical analysis was run on the corpora. Both the numerical results and some description of the figures follow third.

2 Speech Material

Two kinds of Standard Finnish speech corpora were studied. The first one ('single-speaker', or SS) consisted of sentences picked from a periodical and read aloud by an adult male speaker. The reading was done with intent to prepare a corpus for research use. SS is comprised of 967 utterances with 41 306 phones. Of these 14 170 are non-plosive consonants and thus investigated here.

The second ('multi-speaker', or MS) consisted of television news reading, field and weather reports, and oral presentations by 9 men and 6 women, all of whom were professional speakers. Unlike the individual in SS, these speakers were not aware their speech would be used for research purposes. There were a total of 1 148 utterances and 31 414 phones including 10 584 short consonants.

All in all, there were about one and a half hours of continuous speech with any and all pauses eliminated. The corpora were annotated by hand and improved and rechecked several times both by a trained human annotator and by computer scripts designed to detect suspicious annotation. Scripts were designed for preparing the corpus information for phone-level statistical analysis, as ~73 000 phones cannot be entered manually.

3 Methods

To examine how duration in utterance-initial environment develops as closely as possible, we chose a phone-level approach. It is our conviction that researchers should not restrict themselves to syllable and word-level measuring exclusively, as

has been the trend. Our previous research has shown that not all phenomena of segmental duration operate on syllable level. Conversely, some information may actually be overlooked unless phone-by-phone calculations are run on the test material. We organized all the phones into separate data sets by their phonemic identity and their distance from the beginning of the utterance. For instance, 22 utterances in SS and 34 utterances in MS began with the phoneme /r/. All these were then put into their respective slot "position 1" (see graphs 1-8 in the results section). In 30 and 39 of the utterances the second phone was /r/, and all these were assigned into "position 2". This was done to the first 15 positions and all the phonemically short non-plosive consonants, /s, r, m, j, n, h, ʊ, ŋ, l, ɾ/. The few and far between non-native sounds (such as /ʃ/ and /f/) were not studied, and neither at this point the phonologically long variants (/sː, nː, .../) of native consonants. As geminates, the latter may occur at the earliest between the first and the second syllable (i.e. position 2). Finally, the mean duration and 95 % confidence interval were calculated for all positions. The confidence intervals are shown as error bar graphics; if two error bars do not overlap, their difference is statistically significant at $p<0.05$ level. For comparison, there is a horizontal line indicating the mean duration of the phoneme in question. It is the mean calculated from the entire corpus, not just the first fifteen positions in the figures.

A caveat on terminology is in order. We prefer to use the word utterance in the purely phonetic sense of a single, continuous flow of speech internally uninterrupted by pauses. There is no reference to a syntactic unit, such as sentence, made here. Terminal and non-terminal intonation units are treated equally, which is not necessarily the most informative alternative.

4 Results

The results show there is both significant lengthening and shortening in utterance-initial consonants, depending on what phoneme is examined. In the figures, the vertical axis shows the mean duration of applicable segments in milliseconds and their 95 % confidence intervals; the horizontal axis describes the position from the beginning of the utterance. The horizontal line is the mean duration of all the phonemes in question that can be found in the corpus, even those that are beyond the 15 phone scope of the graph. Please bear in mind that the overall mean is somewhat high due to segments that have been significantly affected by final lengthening (Hakokari et al. 2005).

The phonemes can be divided roughly into four groups. First, the sounds /s/ and /r/ are significantly lengthened in both corpora. Second, the sounds /m/ and /j/ are significantly shortened in both corpora. Third, the sounds /n/, /h/, and /ʊ/ are shortened to some degree in MS but not in SS. Fourth, the sounds /ŋ/ and /l/ are not affected in either one. The latter are not shown in the figures below.

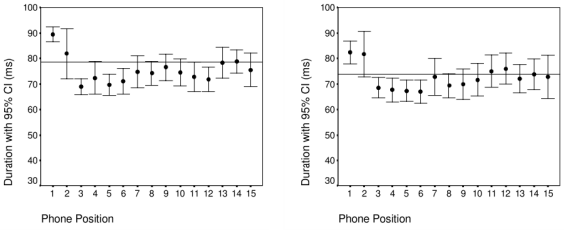


Figure 1. /s/ in SS (left) and MS (right) corpora.

The alveolar fricative /s/, of which there are distinct rounded and unrounded allophones, is lengthened initially in the first position and to some degree in the second. In both corpora there is a gentle shortening (of dubious significance, though) after the lengthening before the mean line is reached. There were a total of 177 utterance-initial items in both corpora combined, but only 38 in the second position.

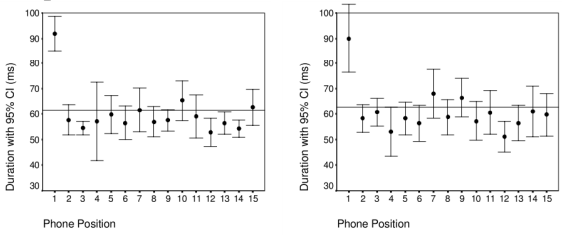


Figure 2. /r/ in SS (left) and MS (right) corpora.

The medioalveolar trill /r/ shows similar behavior in both corpora. It is considerably lengthened in the initial position, after which there is no effect. There were a total of 56 utterance-initial items in both corpora combined, and 64 in the second position.

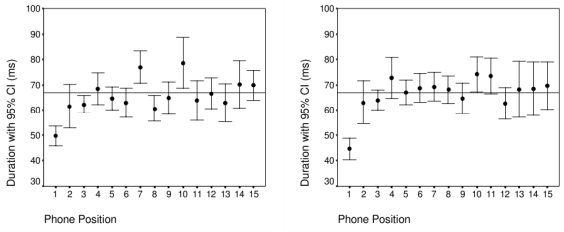


Figure 3. /m/ in SS (left) and MS (right) corpora.

The results for the bilabial nasal /m/ are near-identical for the three first positions in both corpora. Unlike with /s/ and /r/, the first position is significantly shorter than the following ones. There were a total of 199 utterance-initial items in both corpora combined, but only 16 in the second position.

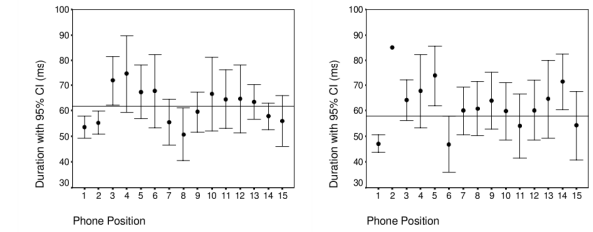


Figure 4. /j/ in SS (left) and MS (right) corpora.

The palatal approximant /j/ is shortened to some degree in SS for both initial and second position, but only for the initial MS. However, the second position has only one item in MS and three in SS, making it impossible to hypothesize anything. All in all, there is much variation in the sound's duration beyond the initial position, and the sound is very hard to segment objectively. Furthermore, /j/ may only occur syllable-initially and its sample size is relatively low. The first position has 257 items in both corpora combined; we can only conclude those are significantly shorter than the mean.

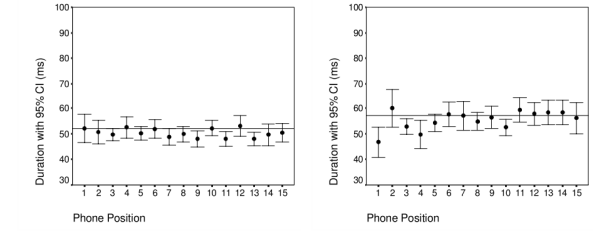


Figure 5. /n/ in SS (left) and MS (right) corpora.

The alveolar nasal /n/ is slightly shorter initially in MS than the rest, but its statistical significance is not clear. In SS corpus there is no shortening whatsoever. There were a total of 112 utterance-initial items in both corpora combined, and 62 in the second position.

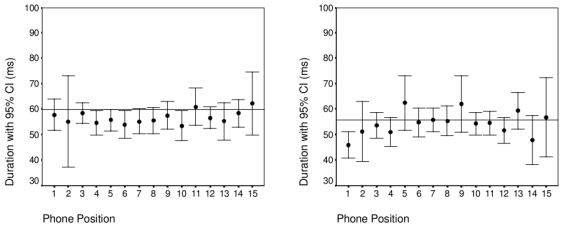


Figure 6. /ʋ/ in SS (left) and MS (right) corpora.

The initial labiodental approximant /ʋ/ is again slightly shorter than the rest in MS (significance unclear), but not in SS. There were a total of 144 utterance-initial items in both corpora combined, but only 12 in the second position. Also /ʋ/ can only occur in syllable-initial position.

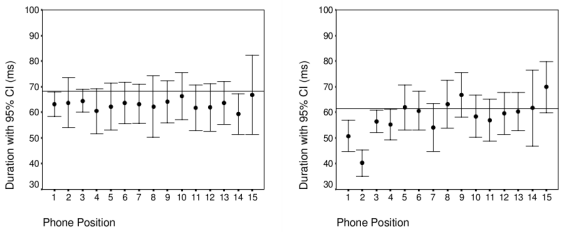


Figure 7. /h/ in SS (left) and MS (right) corpora.

Also /h/ shows no shortening in SS but some in MS. The second position is particularly pronounced. The sound is frustratingly difficult to segment accurately due to the variety of strategies that can be used to produce it, including a variety of non-modal phonations.

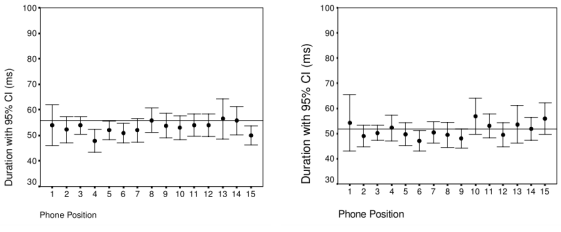


Figure 8. /l/ in SS (left) and MS (right) corpora.

The lateral approximant /l/ (fig. 8) showed no effect on duration in either of the corpora.

There was a slight downward trend for /ŋ/ in MS, though, which might marginally contribute to initial shortening in a word level examination. /ŋ/

does not occur in the initial position in Finnish due to phonotactic restrictions. In educated Standard Finnish /ɾ/ is mostly realized as a single trill (flap), but it cannot occur word-initially or finally. It is very often credited to be a voiced alveolar plosive /d/ mainly for orthographical reasons. It may be produced as a true voiced plosive by some speakers, especially in foreign names and recent loan words, but that is uncommon. In vernaculars it is either omitted or it has become assimilated with other sounds; consequently, a variety of strategies are used to produce the phoneme (see Suomi 1980 for a detailed account). In any case it is marginal in frequency (only ~2,9 % of all the non-plosive consonants in the material) and does not produce very reliable results. There were 6 counts of initial /ɾ/ in the material showing a high mean duration but little consistency.

5 Discussion

As described in the results section, many of the phonemes displayed deviant duration in the first (onset) or the first two positions within an utterance. The question of stress must be addressed first, as stressed syllables are generally expected to undergo lengthening. Unstressed utterance-initial speech sounds are next to impossible to study for reference, since Finnish has an invariable first syllable stress. Unlike in some other first syllable-stressing languages, such as the closely related Estonian, even foreign loan words are forced into the same stress pattern in Finnish. Furthermore, the lengthening as witnessed in /r/ applies to the first position only (syllable onset) while the second position represents overwhelmingly the syllable coda (syllable-initial consonant clusters are not native to Finnish). Thus, lengthening can be expected in words such as /ro.po/ (‘a coin, mite’) but not in words such as /or.po/ (‘an orphant’). Obviously, the fact that some phonemes are shortened and other lengthened is equally difficult to explain away in terms of stress or accent. On the other hand, a future study should be done on the corpora to determine whether any of these effects can be reproduced with a word-initial instead of an utterance-initial examination.

Another issue is segmentation. The corpora had different annotators, but the results are still mostly comparable. However, certain speech sounds are more difficult to objectively segment than others. The trill /r/ makes the following vowel r-colored making it often a subjective task to determine where the sound ends. The true approximants /ʋ/ and /j/, well described as glides, are notoriously difficult to segment, especially in a medial position. Neither has any fricative noise in Finnish. However, the shortening of /j/ and the lengthening of /r/ are so clear in both corpora it is fairly safe to say they are not solely products of segmentation strategies. /ʋ/ and /j/, being phonotactically restricted in Finnish, are unfortunately very rare in position two and not that common elsewhere either (in fact may only occur syllable-initially); hence the great variation in both corpora. The sample size for the first-position /j/ is considerably greater for the multi-speaker corpus; being of more informal a nature, it contains many utterances beginning with the word /jɑ/ (‘and’).

Nasal coarticulation may affect adjacent sounds as well, much depending on the speaker. True nasal articulation is still easy to tell apart from nasalized vowels because the oral closure and release may be pinpointed accurately in the speech signal. The shortening of /m/ is especially significant, since the two other nasal phonemes /ŋ/ and /n/ show little or no shortening. White (2002) has reported very similar results for /m/ in his English test material.

/h/ was slightly affected by shortening in multi-speaker but not to a slightest degree in the single-speaker corpus. Since segmenting the sound is an extremely subjective task, we hesitate to draw any conclusions on the subject. There is a variety of allophones and articulatory variation in how the sound is realized, ranging from breathy phonation to fricative noise.

The alveolar fricative /s/ was significantly lengthened initially. The sound also underwent an exceptional amount of final lengthening as both phonologically short and long in a past study by Hakokari et al. (2005). That suggests the sound is more liable to vary in duration according to its position in prosodic structure. Shadle & Scully (1995) have suggested the exact opposite; the fricative is presumably insensitive to vowel context. Fougeron (2001) has found the sound fairly insensitive to prosodic context as well, and characterized it as having “few degrees of articulatory

and acoustic freedom". The reason for such differences between languages cannot be answered at the moment. Differences in method alone do not feel adequate to count for the discrepancy. On the other hand, in English, there is no such allophonic variance (cf. "articulatory freedom") in /s/ as in Finnish. In Finnish the labialized allophone [sʷ] is acoustically very distinct, with energy at relatively low frequencies; it is easily interpreted as /ʃ/ by English speakers. In the absence of a voiced/voiceless distinction of consonants in the language it may be produced voiced.

Given that all vowels have been found to undergo lengthening in initial position (Saarni & al. 2006), it is worth noting that either lengthening or shortening are more common than no modification of duration at all.

6 Conclusion

This study has observed the duration of short consonants (plosives excluded) in two qualitatively different Standard Finnish speech corpora. The goal has been to identify any possible durational effects an utterance-initial position has on these speech sounds. Previous studies have indicated both language-specific and, within language, phoneme-specific initial manipulations of segmental duration.

The results suggest there is no reason to posit either a feature initial shortening or lengthening in Finnish, as both kinds of durational patterns occur. Lengthening or shortening seems governed by the phonemic identity of the segment occupying the initial position in an utterance. The voiceless sibilant and voiced trill were lengthened, while nasals and approximants showed various amounts of shortening. The lateral approximant was the only sound unaffected in both corpora, although there was considerable variation in its initial position that sample size alone does not explain.

The results are not all similar to those obtained in other languages, which supports the view that initial duration is not based on strictly universal premises. Some level of individual variation may be expected as well, since 2 of the consonants were, on average, shortened by the 15 speakers of the multi-speaker corpus, but not by the individual in the single-speaker corpus.

The instantly visible lengthening and shortening affect only the first or the first two phones of the utterance. Word-level examinations were not run on the corpora at this point, but none of the results rule out the possible shortening or lengthening of the entire first word or so. Statistically significant compression or expansion of the beginning of the utterance (as in narrow confidence intervals) may be established only with a great amount of data, since the intrinsic durations of speech sounds will induce variation.

Perhaps the greatest contribution of this study is pointing out that the most common approach used today, limiting oneself to the syllable level and making no distinction between different speech sounds (operating on "syllable duration"), is prone to miss even the most robust characteristics of duration near the edge of the domain. The results presented in this paper may be useful for instance in speech synthesis. On the other hand, the methods and analysis may be used by researchers in speech technology to produce viable speech scientific information (provided they have a corpus readily available), even when their primary concern is technological development.

References

Dani Byrd. 2000. Articulatory Vowel Lengthening and Coordination at Phrasal Junctures. Phonetica 57, pp. 3-16.

Jianfen Cao. 2004. Restudy of segmental lengthening in Mandarin Chinese. Proceedings of Speech Prosody 2004 (SP-2004), Nara, Japan, pp. 231-234.

Taehong Cho & Keating. 2001. Articulatory and acoustic studies on domain.initial strengthening in Korean. Journal of Phonetics 29 (2), pp. 155-190

Cécile Fougeron. 2001. Articulatory properties of initial segments in several prosodic constituents in French. Journal of Phonetics 29 (2), pp. 109-135.

Cécile Fougeron and Patricia A. Keating. 1997. Articulatory strengthening at edges of prosodic domains. Journal of the Acoustical Society of America 101 (6), pp. 3728-3740.

Jussi Hakokari, Tuomo Saarni, Tapio Salakoski, Jouni Isoaho, Olli Aaltonen. 2005. Determining prepausal lengthening for Finnish rule-based speech synthesis. Proceedings of Speech Analysis, Synthesis and Recognition, Applications of Phonetics (SASR 2005), Kraków, Poland.

Petra Hansson. 2003. Prosodic phrasing in spontaneous Swedish. An academic dissertation. Travaux de l'institut de linguistique de Lund 43. Lund University.

Nobuyoshi Kaiki, Kazuya Takeda, Yoshinori Sagisaka. Statistical analysis for segmental duration rules in Japanese speech synthesis. Proceedings of the 1990 International conference on Spoken Language Processing, Kobe, Japan, pp. 17-20.

Yasuko Nagano-Madsen. 1992. Temporal characterristics in Eskimo and Yoruba: a typological consideration. In Huber (ed.): Papers from the Sixth Swedish Phonetics Conference held in Gothenburg. Technical Report No. 10, Department of Information Theory, School of Electrical and Computer Engineering, Chalmers University of Technology, Gothenburg. pp. 47-50.

Tuomo Saarni, Jussi Hakokari, Jouni Isoaho, Olli Aaltonen, Tapio Salakoski. 2006. Segmental duration in utterance-initial environment: evidence from Finnish speech corpora. Advances in Natural Language Processing: 5th International Conference on NLP, FinTAL 2006, Turku, Finland. Published as a volume in Springer series "Lecture notes in Artificial Intelligence". pp. 576-584.

Christine H. Shadle and Celia Scully. 1995. An articulatory-acoustic-aerodynamic analysis of [s] in VCV sequences. Journal of Phonetics 23, pp. 53-66.

Kari Suomi. 1980. Voicing in English and Finnish stops. A typological comparison with an interlanguage study of the two languages in contact. Publications of the Department of Finnish and General Linguistics of the University of Turku 10.

Jennifer J. Venditti and Jan P.H. van Santen. 1998. Modeling segmental durations for Japanese text-to-speech synthesis. Proceedings of the Third ESCA Workshop on Speech Synthesis 1998.

Laurence S. White. 2002. English Speech Timing: a Domain and Locus Approach. University of Edinburgh PhD dissertation.

Yiqing Zu and Xiaoxia Chen. 1998. Segmental durations of a labeled speech database and its relation to prosodic boundaries. Proceedings of the 1st International Symposium on Chinese Spoken Language Processing (ISCSLP 1998).

Paper V

Measuring Relative Articulation Rate in Finnish Utterances

Jussi Hakokari, Tuomo Saarni, Tapio Salakoski, Jouni Isoaho, Olli Aaltonen. ICPhS The International Congress of Phonetic Sciences 2007, 6-10 August 2007 Saarbrucken, Germany.

MEASURING RELATIVE ARTICULATION RATE IN FINNISH UTTERANCES

Jussi Hakokari[1,3], Tuomo Saarni[2], Tapio Salakoski[2], Jouni Isoaho[3], and Olli Aaltonen[1]

[1]Department of Phonetics, University of Turku
[2]Turku Centre for Computer Science
[3]Department of Information Technology, University of Turku
jussi.hakokari@utu.fi, tuomo.saarni@utu.fi

ABSTRACT

This paper presents two investigations into articulation rate, or the distribution of segmental duration, in a Finnish language speech corpus. The first study, rank ordering of short utterances according to their component words' articulation rate, reveals that 75 % or more of Finnish utterances can be expected show some level of final lengthening. Also initial shortening, or accelerated speaking rate in the beginning of utterances, is present in amounts clearly above chance level. The second study, an investigation into how relative duration progresses in utterances, confirms the observations mentioned before. Furthermore, the second study shows the initial and final effects are statistically significant. Importantly, the results are near-identical to those obtained independently from Southern Swedish, even though the languages and corpora in question are entirely different.

Keywords: final lengthening, initial shortening, speaking rate, corpus, segmental duration

1. INTRODUCTION

Final lengthening (FL), the tendency to slow down articulation at the ends of utterances, has been observed in almost any language in which the matter has been adequately investigated. The only exceptions mentioned in the literature are quantity languages such as Finnish, Hungarian, Estonian, and Japanese (cf. [8]). Even in these the tendency has been confirmed later on [1][4][5][7]. Our own recent studies into Finnish FL [9] reveal how much lengthening different phases of finality induce at phone level, but do not address the question of pervasiveness. Subjective examination of the speech signal clearly gives the impression that FL, although common, does not affect every utterance. This study will answer the question how often we can expect FL in formal Standard Finnish. Furthermore, the study at hand will follow the progress of articulation rate from the beginning to the end in utterances of varying length.

We have adopted the methodology of Hansson [2][3] with a few adjustments. As one of many experiments in her dissertation, she studied utterance-level effects on speaking rate in Southern Swedish (Scanian). Unlike most studies, that have relied on laboratory speech and even nonsense words, her materials consisted of dialect recordings made in natural settings (10 speakers, 518 phrases) in addition to elicited speech. Her methodology was to measure syllable duration within the domain of word. The average syllable duration in final and initial words was thus contrasted by that in medial words. Since the number of syllables in any word was not of concern, the approach could be called a "syllable duration on word level" study. The approach presented in this study could be equally labeled a "phone duration on word level", as the level of observation is word but the level of measurement is the phone. Additionally, the method could be described as semi-phonemic, as phones are measured according to their membership in broad phonemic categories as explained in methods and materials.

2. METHODS AND MATERIAL

2.1. Speech material

The speech material consists of two speech corpora previously studied separately. Since they both have been deemed very similar in aspects of segmental duration (FL included) [9][10], they were considered fit to be combined and studied as a single corpus in the study at hand. The first corpus consists of individual sentences of various lengths, picked out from a Finnish-language periodical Suomen Kuvalehti and read aloud by an adult male speaker. The second consists of television news, weather broadcasts and radio presentations. Being

of roughly the same size, together the corpora add up to 2115 utterances, 72720 phones, or ~93 min of continuous speech flow with any acoustic pauses other than voiceless plosives removed.

The corpora were manually annotated at phone and word levels. The annotation was done with duration studies in mind; great emphasis was put on temporal accuracy (e.g. segment boundaries were determined at the level of a single waveform).

2.2. Methods

The first experiment was designed to give an estimate on how often in fact FL occurs. While our previous study [9] confirmed FL exists in Finnish, subjective experience tells us it is not unavoidable and it does not occur in every utterance. To rank order words of a single utterance according to their speaking rate is a methodological dilemma. We first divided all the phones in the corpus into the following categories: non-plosive consonants, plosives, and vowels. These were further divided into phonologically short and long consonants, as Finnish is a quantity language with short/long distinction in consonants and vowels alike. Monophthong long vowels were also separated from diphthongs. The term 'utterance' refers to any continuous speech flow with no internal pauses. An utterance may thus be either a terminal or non-terminal intonation unit, but the criterion for the end of an 'utterance' was acoustic pause. Internal syntactic boundaries alone did not delimit utterances in this study.

Second, each phone's duration was compared against the mean duration of its respective category. For instance, a 80 ms short vowel was compared against mean duration of all the short vowels in the material (64.9 ms), and given a factor of 80 ms/64.9 ms ≈ 1.23. Third, the entire word was given a mean factor based on all of its component phones. Finally, each sentence was reordered (rank ordered) according to the factors of their component words without losing the information of the original word order. The operation was performed on all 2, 3, 4, and 5-word utterances found in the speech material (n=1020).

The second experiment was designed to show how word-level speaking rate progresses in utterances. The experiment used the same information as the first one, but this time the factors assigned to words were brought together. For instance, all 5-word utterances were merged so that the factor for its component initial word was a mean of every first word in every 5-word utterance in the material. Utterances from 2 to 9 words in length were investigated for the second experiment (n=1708).

Durations were calculated and processed for statistical analysis using specially designed scripts which extract information from Praat TextGrid files. TextGrid files contain the word and phone level annotation.

3. RESULTS AND DISCUSSION

3.1. Rank Ordering

The results from the rank ordering are summarized in table 1. The columns represent utterances of various lengths. The ordinals on the left hand side indicate the words with the slowest articulation rate (greater segmental duration). For instance, of the 240 two-word utterances only 36 were ones in which the 1st word is in fact longer in segmental duration than the second. FL % stands for the percentage of utterances in which the final word is the longest, IS % (e.g. initial shortening) for the proportion of utterances in which the initial word is the shortest. The final row ("à Hansson") shows the frequency of FL as with Hansson's [3] criteria; any utterance in which the final word is longer than the penultimate is considered to exhibit FL.

Table 1: Summary of rank ordering results.

	2-word (n=240)	3-word (n=255)	4-word (n=281)	5-word (n=244)
1st word	36	17	0	2
2nd word	204	26	23	5
3rd word		212	42	18
4th word			216	36
5th word				183
FL %	85.00	83.14	76.87	75.00
IS %	85.00	63.13	55.87	48.77
FL % à Hansson	85.00	89.01	86.97	87.29

The results show that in as much as 75-85 % of the utterances there is evidence supporting FL and in ~49-63 % evidence supporting initial shortening. It must be noted that at chance level the percentages would be in the order of 50%, 33 %, 25 %, and 20 % for the utterance lengths examined. Hansson's methodology put the Swedish figures at 71-82 %; in our material FL was somewhat more pervasive (85-89 %). For reference, Mihkla [6] reported Estonian (a language closely related to Finnish) newsreaders

producing FL 60 % of the time depending on the context, but the details of their methodology remain obscure.

3.2. Relative Duration

The figures below (1-4) show the mean relative durations and 95 % confidence intervals ($p<0.05$) of 2 to 9-word utterances. Average articulation rate is represented by 1.0 on the left hand side. Above 1.0 their segmental durations exceed that average, consequently slowing down articulation rate. Below 1.0 speaking rate is accelerated and shorter phones are produced. In the legend below, 1st always refers to the first (i.e. initial) word in the utterances; the last (i.e. final) word is always the rightmost one.

Figure 1: Relative articulation rates in 2 (n=240) and 3-word utterances (n=255)

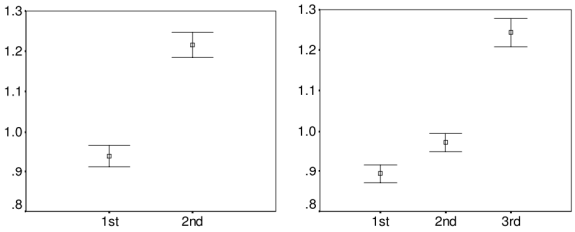


Figure 2: Relative articulation rates in 4 (n=284) and 5-word (n=244) utterances.

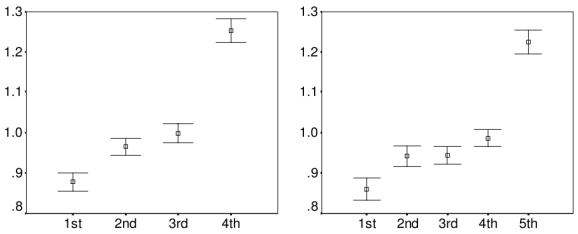


Figure 3: Relative articulation rates in 6 (n=239) and 7-word (n=205) utterances.

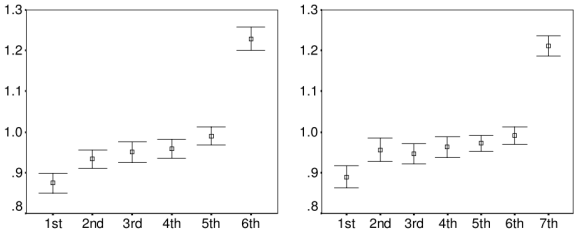
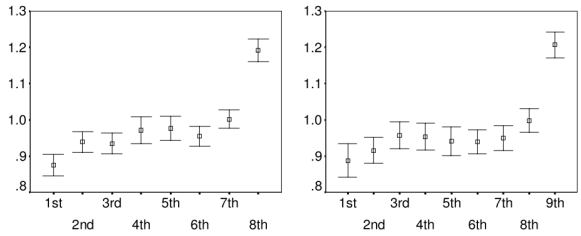


Figure 4: Relative articulation rates in 8 (n=142) and 9-word (n=99) utterances.



The articulation rate, as calculated in relative durations of their constituent phones, show results uniform across the various utterance lengths and are near-identical to Hansson's results. The results follow a simple pattern. The initial word is characterized by modest yet statistically significant shortening. The medial words are of roughly the same length, yet there is a non-significant tendency for phones to grow in duration towards the end of the utterance. Penultimate words expectedly stand out as longer than the rest of the medial words (cf. [1]), but final words are significantly lengthened. Our unpublished results further suggest that FL is not limited to the last one or two syllables, but in fact may commence earlier and gradually grow towards the boundary.

Utterance-initial shortening or compression, as it may be called, is present in each sentence type (2 to 9). In 9-word sentences it is no longer significant, though. There is a similar tendency in all phrase lengths in Hansson's [2][3] material, although the effect is not significant at 0.05 level. Nevertheless, since initial shortening is present in all Swedish utterance lengths (2 to 5 words) as well as in Finnish ones (2 to 9 words), it is fairly safe to assume the difference between the first and the second word is not a product of chance but some real articulatory or linguistic phenomenon. The factor is remarkably uniform with means between 0.85 – 0.90. For Hansson [3], the factor can be calculated to vary from 0.82 to 0.86, the 5-word phrases with a figure as low as 0.70 excluded.

Medial words are clearly discernible from the initial and the final. They are mostly level and between the initial and final words, but closer to initial than final. Hansson's Swedish data compares well. Medials in penultimate environment, having longer durations than the other medials, show a similar but weaker tendency in the Swedish data. Since Hansson has only studied phrases with 5 or fewer words, penultimate words can be compared against other medials only

in two phrase types, the 4 and 5-word ones. It is necessary to examine longer Swedish utterances before any reliable comparison can be made.

FL makes the final word considerably longer than the others ($p<0.001$). Across utterance types, the duration factor is in the excess of 1.2. From Hansson's [3] results we can deduce the corresponding factor vary from ~1.24 to ~1.38.

4. CONCLUSIVE REMARKS

We have been able to identify two phenomena related to articulation rate in our corpora. There is a considerable amount of utterance-final lengthening in the domain of utterance-final word. Second, there is a minor yet statistically significant and systematic shortening of the utterance-initial word. No conclusive effects were observed in medial words; articulation rate remains fairly stable mid-utterance regardless of the utterance's length. In other words, individuals studied here sharply began speaking with an accelerated speaking rate, and then slowed down to 'normal' until they finally slowed down considerably before pausing.

While the corpus was not annotated for stress or accent, there is no reason to attribute the findings to stress or prominence. The formal Finnish news speech is typically not accentuated to a great extent. The entire speech material used in the study is practically devoid of strong contrastive accent. The first two authors examined 21.4 % of the speech corpus and found noticeable prominence in only 13.9 % (first author) or 11.9 % (second author) of final words. That amount cannot contribute much to FL.

Perhaps the most important observation is that the results are strikingly similar between this study and the reference study on Southern Swedish dialects. The differences between the two are limited to certain details. In our material, both FL and initial shortening are more frequent but lesser in magnitude. That may have to do with individual variation or speaking style; ours is strictly formal literary style while the Swedish dialect recordings are a combination of natural and elicited speech. At this point there is no reason to suggest a language-specific background for the phenomenon. It ought to be noted that, in despite of geographical vicinity, there is no particular direct contact between Finnish and Southern Swedish (Scanian) speakers to speak of. The two groups are rarely exposed to each others speech, the languages are not genetically related, and they are very different both in terms of segmental and suprasegmental phonology. Final lengthening, as found in practically every language investigated so far, is likely to be a product of muscular and pulmonary mechanics involved in articulation, not a learned linguistic feature with a self-purposeful communicative function.

5. REFERENCES

[1] Hakokari, J., Saarni, T., Salakoski, T., Isoaho, J., Aaltonen, O. 2005. Determining prepausal lengthening for Finnish rule-based speech synthesis. *Proceedings of Speech Analysis, Synthesis and Recognition: Applications of Phonetics (SASR 2005)*, Kraków.

[2] Hansson, P. 2002. Articulation rate variation in South Swedish phrases. In: Bel, b. Marlien, I. (eds), *Proceedings of Speech Prosody 2002,* Aix-en-Provence, 371-374.

[3] Hansson, P. 2003. Prosodic phrasing in spontaneous Swedish. An academic dissertation. *Travaux de l'institut de linguistique de Lund 43*. Lund University.

[4] Hockey, B.A., Fagyal, Zs. 1999. Phonemic length and pre-boundary lengthening: an experimental investigation on the use of durational cues in Hungarian. *Proceedings of the XIVth International Congress of Phonetics Sciences*, San Francisco, 313-316.

[5] Krull, D. 1997. Prepausal lengthening in Estonian: evidence from conversational speech. In: Lehiste, I. & Ross, J. (eds.), *Estonian Prosody: Papers from a Symposium, Proceedings of the International Symposium on Estonian Prosody,* Tallinn, 136-148.

[6] Mihkla, M. 2005. Modelling pauses and boundary lengthenings in synthetic speech. *Proceedings of the Second Baltic Conference on Human Language Technologies (HLT'2005)*, Tallinn, 305-310.

[7] Vaissière, J. 1983. Language independent prosodic features. In: Cutler, A. & Ladd, R. (eds.) *Prosody: Models and Measurements*, 53-65.

[8] Venditti, J., van Santen, J. 1998. Modeling segmental durations for Japanese text-to-speech synthesis. Third ESCA Workshop on Speech Synthesis (SSW3-1998), 31-36.

[9] Hakokari, J., Saarni, T., Isoaho, J., Salakoski, T., Aaltonen, O. Prepausal lengthening in Finnish: further evidence for a phonetic universal. Submitted.

[10] Saarni, T., Hakokari, J., Aaltonen, O., Isoaho, J., Salakoski, T. 2007. Utterance-initial duration Finnish non-plosive consonants. *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, Tartu, 160-166.

Paper VI

Utterance-level Normalization for Relative Articulation Rate Analysis

Tuomo Saarni, Jussi Hakokari, Jouni Isoaho, Tapio Salakoski. Interspeech 2008 incorporating SST'08, 22-26 September 2008, Brisbane, Australia.

Utterance-level Normalization for Relative Articulation Rate Analysis

Tuomo Saarni[1,2], *Jussi Hakokari*[2], *Jouni Isoaho*[2], *Tapio Salakoski*[1,2]

[1]Turku Centre for Computer Science, Turku, Finland
[2]Department of Information Technology, University of Turku, Finland
tuomo.saarni@utu.fi

Abstract

This study describes a computational method for studying variation in articulation rate in a qualitatively mixed speech corpus. The method works within the scope of individual utterances, replacing each single speech sound's time information with a coefficient based on its duration relative to its environment. It can be used to generalize and determine points of acceleration and deceleration in articulation at the phone level, even when the general speaking rate varies greatly due to speaker, style, and utterance length related effects. To demonstrate the usability of the proposed method, we track observed deceleration of articulation rate (a form of final lengthening) towards the ends of utterances in a linguistically uncontrolled Finnish-language speech corpus with several speakers and styles.

Index Terms: final lengthening, normalization, segmental duration, Finnish, articulation rate, speaking rate

1. Introduction

Elicited laboratory speech has long dominated the field of prosody research. A controlled experimental setting with no unwanted linguistic variation works in the researcher's advantage, so that one can concentrate on phonetic detail. The researcher can safely ignore much of the consideration of what has and has not affected the phonetic signal, provided all the participants have read aloud the very same, carefully planned, utterances.

In despite of its well-established advantages, elicited speech does warrant criticism, however. Speech comes in many levels of formality, ranging from entirely unplanned, spontaneous conversations to planned speeches and monologues, or reading aloud written text. Elicited sentences with clearly marked, contrastive stresses ("I said the HOUSE burned down, I did not say the MOUSE burned down") represent an extreme end of the formality scale. One could argue that such a manner of speaking is marginal in everyday human interaction. Certain aspects of prosody, such as F0 contours, are challenging to study without a controlled paradigm. Some other aspects, such as segmental duration in a readily annotated data, are more discrete and more easily studied by automatic means.

Manners of speaking that approach the spontaneous can be studied only if we are willing to compromise controllability. Corpora with varying speaking styles are available, and automatic processing allows us to easily convert large quantities annotated speech data into statistical data. The errors induced by lack of controllability have to be addressed, however. Certain measures have to be taken to reduce inaccuracy by contextual differences; in duration research, a central issue is normalization.

This paper examines speech timing from the point of view of varying articulation rate. While people speak, they tend not to articulate at the same speed but constantly change their articulation rate. That tendency is usually treated analytically as specific lengthening (such as final lengthening; for discussion in quantity languages genetically related to Finnish, see [1], [2]) and shortening (for general discussion, see [3]; for occurrence in the corpus at hand, see [4] [5]) processes. Our previous study [6] examined the relative duration of word-level units in Finnish speech corpora, using intra-corpus normalization. The results suggest that, on a very general level, speakers tend to start out articulating an utterance somewhat faster, then gradually slow down a little, and finally slow down considerably in the end.

In this paper we demonstrate our method and continue to examine the final lengthening or deceleration of articulation rate. However, we strive for a greater level of detail, namely phone-level units, in a mixed corpus featuring a number of speakers, distinctive styles, and utterance lengths. Consequently, we must apply a different normalization routine to exclude the effect of the considerable variation in speaking rate while retaining domain-edge processes such as final lengthening. We use a normalization technique that allows us to study the development of speaking rate within an utterance, and produce a generalized, phone-level tracking that does not rely on absolute segmental duration.

What is usually understood as normalization involves applying manipulations directly to the data to make comparison of quantitatively different data sets meaningful. For instance, recordings of different speakers are manipulated so that articulation rate becomes more or less equal. Inter-speaker normalization, however, cannot necessarily account for unintended variation the speaker might produce, such as the possible influence of utterance length. Our method models speaking rate in one utterance at a time, and compares its constituent segments to the model. No inter-utterance comparison is needed; each segment gets a coefficient that represents its relative duration within the very same utterance.

2. Normalization

The normalization process aims [Fig. 1] to eliminate the influence of varying speaking rates between individual phonetic utterances (continuous speech sample delimited by silence left and right) regardless of who has produced them. Whereas inter-speaker normalization is useful in providing comparable results for both fast and slow speakers, the method at hand eliminates absolute durations altogether and transforms the segmental timing information in a corpus from milliseconds to relative speed coefficients. Value 1.0 represents average articulation rate in one utterance, and anything else is either relatively faster or slower (i.e. shorter or longer) than that.

The first step is to establish comparison. We will want to give each phone in the utterance its own coefficient,

Accepted after peer review of full paper

depending on whether it is longer or shorter than what would be considered average in its context. The simplest solution would be to calculate the mean duration of all the segments and compare each phone to the mean. That, however, would render inherently or phonologically long segments considered slowly articulated and short ones articulated fast. The results would be contaminated by the phonemic content of the utterance and would not reflect articulation rate very accurately. Such is especially pronounced in a quantity language (e.g. Finnish) with distinctive phonological length. The other extreme, comparing each phone only against representatives of the same phoneme is obviously out of the question, as many phonemes are expected to occur in an utterance only once.

The solution here was to establish broad categories of phones that share similar properties. The seven categories were phonologically short vowels, non-plosive consonants, voiceless plosives, their long counterparts, and diphthongs. Plosives were separated as they tend to be longer in duration than other consonants and generally immeasurable from the acoustic signal alone in utterance-initial and final positions. However, the category of non-plosive consonants still remains varied, containing nasals, fricatives, etc. Another approach would be to categorize by similar inherent (mean) durations exclusively and ignore the sounds' nature altogether.

To illustrate, the script will calculate the mean duration of all the short vowels in a given utterance and then divide each individual short vowels' (in that utterance) duration with the mean to get a coefficient (eg. a 40 ms phone divided by 63 ms mean = ~0.63). Once all the sounds in the utterance are treated this way, the script moves to the next utterance in the corpus. Finally, the original time intervals have been replaced by coefficients in the entire corpus. Should the corpus contain any number of very slowly or fast articulated utterances, the method would prevent them from carrying any extra weight in the data.

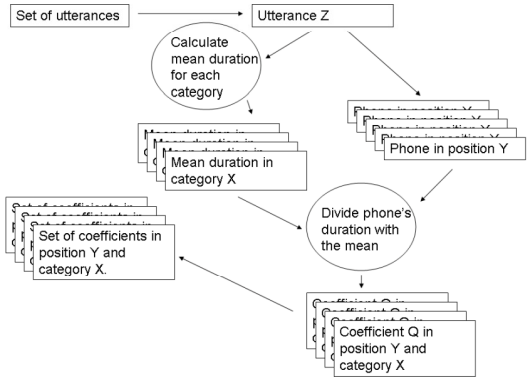


Figure 1: *Steps in calculating normalized duration coefficients for each sound category and position.*

3. Speech Material and Procedure

The Finnish speech corpus used contained reading aloud sentences (~60 %), television news and field reports, a weather broadcast, and oral presentations on the radio. None of the material could be considered spontaneous, but none of it was elicited test sentences in the traditional speech science sense. There were 16 more or less professional adult speakers (10 male, 6 female) of Standard Finnish, who produced very few dysfluencies or hesitations. The corpus was manually annotated at phone level by trained phoneticians.

Since the normalization is essentially done within individual utterances, as if there was nothing else in the corpus, it is unavoidable that the method produces increasingly balanced results with longer utterances. Hence, all the utterances consisting of less than 10 phones (mainly words in isolation) were excluded, leaving a total of 1960 utterances remaining. The rare cases in which a phone was the single representative of its category were ignored to avoid giving those an automatic 1.0 coefficient due to comparing a value against itself.

The durational change towards the ends of utterances observed in the corpora was finally examined. The traditional method of comparing durations of (syllable or word-size) units in various environments does not work well here with an uncontrolled corpus and phone-level detail. It is necessary count back phone by phone from the end of the utterance in order to place utterances of different length on the same line. Hence, all the utterances with their coefficients were placed in a reverse order. The final phone of each utterance was considered as position 1, the penultimate one position 2, the third last position 3, and so on. The operation results in data with coefficients stacked together position by position [Fig. 2]; it can now be statistically analyzed for instance by the given seven sound categories.

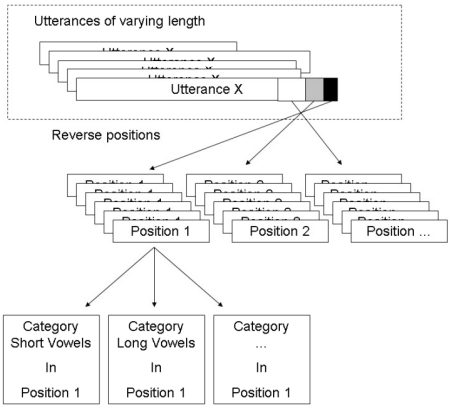


Figure 2: *An illustration of how the utterances are reversed and stacked for analysis.*

The described kind of examination allows us to track the development of relative duration phone by phone, from the final towards the medial positions in utterances of varying length, while avoiding much of the external influence on speaking rate by speakers, utterance length, and content.

4. Results and Discussion

The results show relative duration or articulation rate in utterance-final environment. They are arranged in separate figures for each phoneme category. The horizontal axis represents the distance from the end of utterance (position 1 equals final), measured in phones. The vertical axis represents the coefficient that measures articulation rate. For instance, vertical value of 1.2 may be interpreted as articulation rate 20 % lower than the mean. The mean coefficients are accompanied by 95 % confidence intervals for statistical reference. While the longest utterances in the corpus are in the excess of 100 phones (and thus 100 positions), the following results will feature only positions 1-30, which is enough for

showing both the baseline duration and the final deceleration. There are upward of 47 000 phones in positions 1-30 altogether.

Wider confidence intervals are here primarily the result of small sample size rather than great variation. Phonologically short phonemes are roughly tenfold more frequent than long ones in Finnish. For example, an utterance-final (position 1) diphthong (N=12), showing a particularly wide interval, is very uncommon in Finnish. Conversely, position 1 short vowels (N=1141; there were 1141 utterances ending in a short vowel in the material) and short non-plosive consonants (N=575) are commonplace. Some of the variation in the results below and the seemingly dynamic nature of articulation rate change make inferring phonetic significance somewhat difficult, necessitating some vagueness in interpretation.

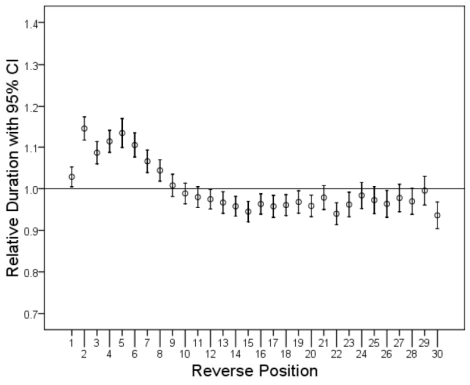


Figure 3: *Relative duration of short vowels*

Short vowels display a trend of lengthening fairly early on. Already the 8th position can be considered significantly longer. Final position, while still longer than baseline, is significantly shorter than the second position. Many Finnish speakers tend to reduce the final short phones of unstressed syllables.

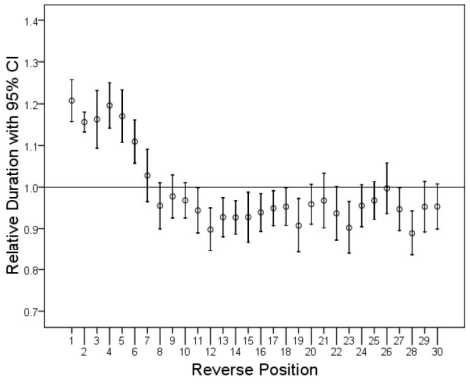


Figure 4: *Relative duration of long vowels*

Long vowels become significantly longer by the 6th last position, and very little changes after the 5th.

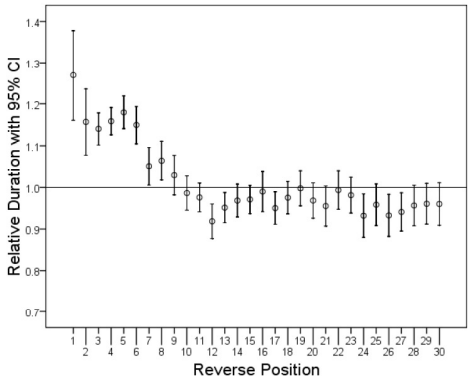


Figure 5: *Relative duration of diphthongs*

Diphthongs become significantly longer by the 6th last position, although a trend of lengthening can be observed from the 9th position onwards. The shape of the lengthening curve is remarkably similar to that of long vowels, possibly reflecting similar manner of articulation.

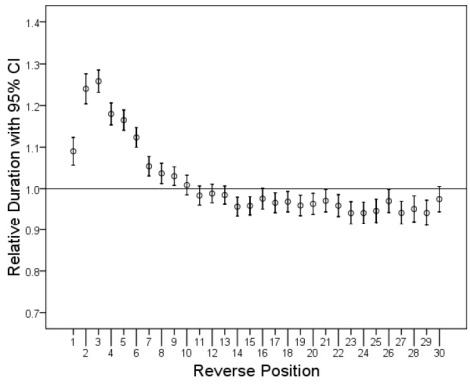


Figure 6: *Relative duration of short non-plosive consonants*

Short non-plosive consonants begin a steady climb around the 10th last phone, becoming significantly longer by the 8th or 9th last. The reduction of short final phones applies to consonants as well.

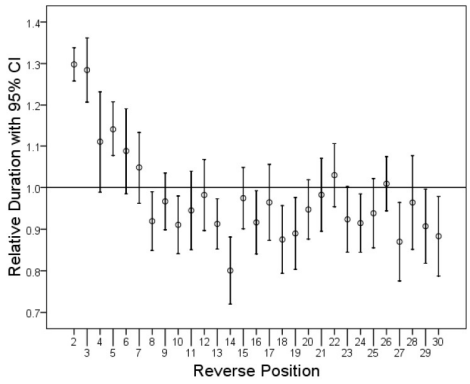


Figure 7: *Relative duration of long non-plosive consonants*

Long non-plosive consonants tend to grow longer from the 7th or 6th position onwards, with the 3rd and 2nd clearly above the typical. Position 1 is missing, as phonotactic restrictions preclude long consonants in all final environments. The amount of long non-plosives is the smallest of all the categories. Furthermore, the category is the most diverse in

terms of inherent duration; hence the great confidence intervals.

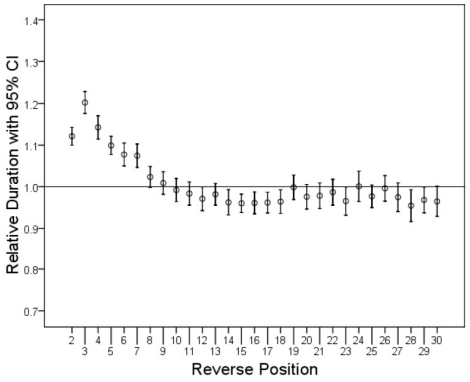


Figure 8: *Relative duration of short voiceless plosives*

Short voiceless plosives begin a steady climb around the 11th last phone, becoming significantly longer by the 7th last. While final voiceless plosives occur and are frequent, they are difficult to measure reliably and have been excluded from the data.

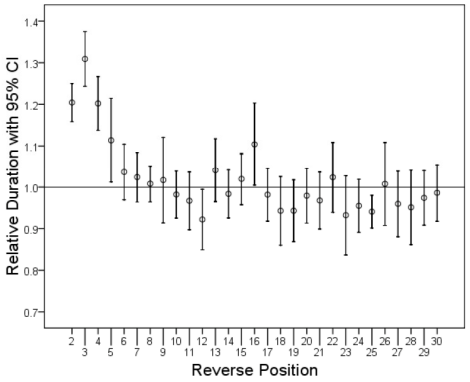


Figure 9: *Relative duration of long voiceless plosives*

Long voiceless plosives become clearly longer by the 4th last position. Long plosives may not occur in a final position.

To conclude, the slowing down of articulation rate witnessed at word level [6] can be observed in the current phone-level examination in all the established broad phoneme categories. There appears to be minor differences in between them both in when an actual segmental lengthening takes place and in how great the relative change is at most. It is unclear whether these differences would narrow down if sample size was increased. Most of the baseline duration is below 1.0, indicating relatively faster articulation; that is a consequence of the amount of lengthening present the end of utterances.

In the linguistic domain, the onset of lengthening roughly coincides with one word form. Finnish is a highly inflecting and compounding language with a small phoneme inventory and some relatively long lexical words; the mean size of a word form in the corpus is ~7.8 phones. However, as it operates on phone level exclusively, the present approach cannot predict whether lengthening is tied to lexical units. The previous study [6] suggested the penultimate word has longer segmental duration than the antepenultimate, but the greatest deceleration takes place during the final word.

Whether the observed phenomenon is the product of final lengthening alone, cannot be deduced without further experiments with proper distinction of prominent and non-prominent items in the corpus. As the speech material contains diverse clause and information structures, it is most likely that both a general physiological motor tendency (final lengthening) and the syntactic and semantic structure (accent, prominence) are contributing factors. However, the methodology used should rule out utterance length and any associated effect on segmental duration.

5. Conclusion

We have presented a method for normalizing acoustic duration of individual speech sounds within the immediate utterance context they were produced in. The method converts acoustic timing information in a speech corpus into coefficients that allow studying relative changes in articulation rate across different speakers, across varying speaking rates produced by a single speaker, and across utterances of varying length and content. The demonstration of the method describes in phone-level detail the deceleration of articulation rate towards the end of utterance, a prevalent feature in the Finnish-language speech corpus at hand. The deceleration is an effect of associated phenomena that can be collectively called utterance-final deceleration. The contribution of individual factors, such as prominence and the independent notion of final lengthening, will have to be investigated in further detail.

6. References

[1] Hockey, B.A. and Fagyal, Zs., "Phonemic length and pre-boundary lengthening: an experimental investigation on the use of durational cues in Hungarian", in Proceedings of the XIVth International Congress of Phonetics Sciences (ICPhS XIV), 313-316, 1999.

[2] Krull, D., "Prepausal lengthening in Estonian: evidence from conversational speech", in Lehiste, I. and Ross, J. [Eds], Estonian Prosody: Papers from a Symposium, Proceedings of the International Symposium on Estonian Prosody, 136-148, 1997.

[3] White, L.S., "English speech timing: a domain and locus approach", University of Edinburgh PhD dissertation, 2002.

[4] Saarni, T., Hakokari, J., Aaltonen, O., Isoaho, J., Salakoski, T., "Utterance-initial duration of Finnish non-plosive consonants", in the Proceedings of the the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007), 160-166, 2007.

[5] Saarni, T., Hakokari, J., Isoaho, J., Aaltonen, O., Salakoski, T., "Segmental duration in utterance-initial environment: evidence from Finnish speech corpora", in Proceedings of the 5th International Conference on Natural Language Processing (FinTAL), 576-584, 2006.

[6] Hakokari, J., Saarni, T., Salakoski, T., Isoaho, J., Aaltonen, O., "Measuring Relative Articulation Rate in Finnish Utterances", in the Proceedings of The 16th International Congress of Phonetic Sciences (ICPhS XVI), 1105-1108, 2007.

Paper VII

Right to ones voice?

Kai K. Kimppa & Tuomo Saarni. Ethicomp 2008, The Tenth ETHICOMP International Conference on the Social and Ethical Impacts of Information and Communication Technology, 24-26 September 2008, Mantua, Italy.

RIGHT TO ONES VOICE?

Kimppa, Kai K.
Information Systems, Department of Information Technology,
University of Turku
Joukahaisenkatu 3-5
FIN-20520 Turku
Finland
Telephone +358 2 333 8665
Fax +358 2 333 8600
Mobile +358 44 060 1321
e-mail: kai.kimppa@utu.fi

Saarni, Tuomo I.
Bioinformatics, Department of Information Technology, University of Turku
Joukahaisenkatu 3-5
FIN-20520 Turku
Finland
Telephone +358 333 6943
Fax +358 2 333 8600
Mobile +358 40 528 7167
e-mail: tuiisa@.utu.fi

Abstract

Speech synthesis is getting to a level where synthesized voice cannot be distinguished from the voice of the person it originally belongs to. This raises several ethical questions; both real ethical questions on what is right and what is wrong and questions of clear misuse for the designers and legal professionals to solve. In this paper we will first introduce the technology and current applications, then look at the ethical questions raised and finally suggest both technical and political/legal remedies to some of these problems.

Terms used

Donor = the person to whom the voice belongs
Producer = the producer of the speech synthesis software
User = the party using the software for synthesizing speech (of the donor).

1 Introduction

A research group at the University of Turku is developing a speech synthesis software which can synthesize the voice of anyone based on a set of rules used. The project has raised certain ethically relevant questions, such as who owns a voice of a person, if anyone? What kinds of rights, if any to a voice can anyone have? The development of the software and other similar softwares is active and these questions are bound to rise in the near future; if not now.

In this paper we aim to look at the problem from an ethical perspective. The aim is to show that the rights of a person to their voice, rights of someone using such software and rights (typically copyrights and patents) of the producer of the software seem to create tensions for the use of such softwares – tensions, which need to be resolved. The supposed level of the software is that of synthesizing voice which is indistinguishable from the actual voice of a person, either by hearing the synthesizer producing sentences or, even by technical means.

We will consider the questions from the perspective of the rights of the person whose voice is being used, from the perspective of the rights (if any) of the producer and the user of the software. We will also look into the duties such software undoubtedly puts to the users of the software and the possible duties which fall to the producer. Finally, we will look into the consequences, both in a utilitarian and economic sense such software would and will introduce.

Many similar issues arise with the use of someone's voice as do with the use of a picture of someone, although some ethical issues are also specific and novel to voice synthesis. Questions such as are similar to those rising with pictures: Who owns the product of the voice synthesizer? What role does consent play in the use of someone's perfectly imitated voice? What moral rights must be considered? In this paper we will look at these issues amongst others and compare them to the ones presented for image usage (see e.g. [Weckert and Adeney 1994, or Evans and Mahoney, 2004]).

2 Technical background

Speech synthesis systems have been available for decades, and several ways to produce synthesized speech have emerged. Our research concerns an older method of creating artificial speech, a rule-based speech synthesis. A rule-based speech synthesis may be considered a truly synthetic way to produce speech since it does not make use of any samples of natural, human speech as most of the other synthesis methods do. On the other hand, the rule-based speech synthesis is commonly considered to be the most challenging way to produce high-quality synthetic speech.

Our on-going research has achieved promising results. Our latest studies have shown promising results to overcome the unnatural quality of rule-based artificial speech. At the moment, the synthesizer is able to produce imitations based on any speaker's example. There are still few acoustical errors in the sound quality, but the level of imitation can be considered to be very natural. However, if the imitated sentence is heard thru a telephone line it might be impossible to tell different the original speaker and the imitation. We hope to overcome these acoustical problems in the near future.

The imitating synthesizer is not currently a text-to-speech application, since it can only imitate. This means that the synthesizer can only repeat a sentence it has been given. However, this is a good result since the imitation is still done using the same methods as in the text-to-speech version, although the result is not done by using the same rules used in the

normal synthesis. We need to create new set of rules based on the imitated phrase. This means, that some amount of imitated data must be collected. We hope to be able to extract a set of rules by using algorithms used in machine learning.

The following diagram demonstrates the steps included in the speech synthesis. Transcription translates an orthographic text into phonetic alphabets. After the transcription the written text is in the form of pronunciation. The transcription with setting of durations and fundamental frequency is often referred as the syntagmatic part of creating speech synthesis. The intonation is generated in the syntagmatic steps. The following phoneme rules are acoustic models for each speech sound. The synthetic speech is produced and played back after the acoustic model of each sound is generated. The audio signal is generated using a Klatt-type signal generator [Klatt, 1980]. The system design is described in greater detail in [Saarni et al, 2006].

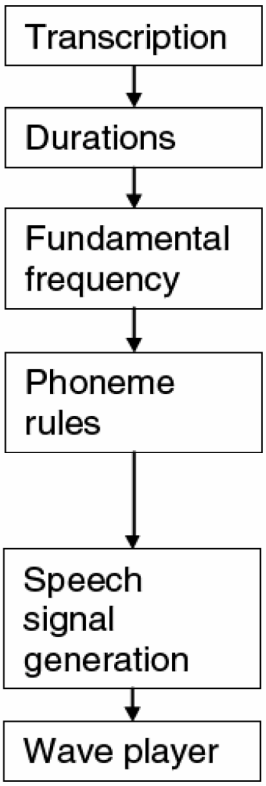


Diagram 1. Steps in generating a rule-based synthetic speech

There are some other speech synthesis techniques that already achieve a very high level of naturalness. These techniques can be divided harshly in two categories; concatenative and Hidden Markov Model based (HMM) synthesizers. Concatenative synthesizers use excerpts of real voice, which are then concatenated into continuous speech samples. One might consider concatenative synthesizer not to be actual speech synthesizers since they withhold real voice samples. The HMM synthesizers use neural networks, which are taught with a certain voice or set of voices. The method is described in more detail in [Tokuda et al., 2000]. The HMM synthesizers may include a high level of prosody modelling with several levels of detail resulting on high level of naturalness (see e.g. [Vainio et al., 1998]). Modern methods

of HMM synthesis include speaker adaptation techniques, which makes it easy to implement a new voice into the synthesizer (see e.g. [Qin et al, 2006]).

Since there are two good methods to produce artificial speech with high naturalness one might ask why we use rule-based methods. The rule-based speech synthesis is very light computationally and uses very little memory capacity compared to other speech synthesis methods. On the other hand, the rules in the synthesis are visible models of prosody and other acoustic features of speech. These models can be studied itself; while they would be lost in the neural networks of HMM synthesizers. And last but not least, we do find it fascinating to do it differently than the others.

3 Speech technology applications

Speech technology includes several applications other than speech synthesis. Most common are speech recognition, voice recognition (who's talking), speech synthesis and their compositions such as user interfaces and second language educational systems. Since the speech synthesis is mainly used in user interfaces we are going to concentrate more on this matter.

3.1 Artificial speech applications

The usage of voice in user interfaces is finally becoming more general. This due to the fact that the level of speech recognition is acceptable and the synthetic speech is intelligible and it sounds natural. People have used to use computer via mouse and eyes so it will take a while that we'll see voice user interface in every computer. However, voice based interfaces may become more common faster in such places where there was minimal communication needed previously between the user and machine, e.g. cars, household electronics.

Modern car navigators are already using concatenative speech synthesizers, which actually use only set of pre-recorded voice clips. In the near future, we might expect to see complete text-to-speech synthesizers integrated to the navigators. After which they can pronounce all the streets and cities in the map. Some car manufacturers include speech synthesizers to their top-end models to pass information to the driver. The speech synthesizer in a car is a feature only offered in the best models at the moment. We will see when this trend moves to cheaper models and brands.

Several other technologies may have advantage in markets after they include speech synthesizers. Implementation of a speech synthesizer might be the thing that separates the product from its rivals. The applications of speech synthesizers are limitless. Who would not want to have a coffee maker waking you up with scent of fresh coffee and Marilyn Monroe's or Cary Grant's voice whispering in your ear? The possibilities are here today. All we need is the products to surface.

3.2 Voice recognition

Another interesting field of study is the speech recognition and its applications, such as voice recognition. This means finding out who is talking. Latest news from Siemens reports new software that is used to recognize caller for faster service [Siemens AG, 2008]. In this case especially, the user is recognized for renewal of a lost password. It is not very difficult to imagine the problems raised with the idea of speaker recognition and speech synthesizer with

speaker adaptation capabilities. All it needs is a sample from the speaker and fast fingers to write requested words, and the system is bypassed.

4 Ethical questions raised

The ethical and social questions related to the use of a voice synthesis application can be divided into three categories:

1) Is there a right to ones voice, and if, what kind of right?
2) What are the borderline cases if any, and how to solve these?
3) What uses are clearly permissible and what uses are clearly not permissible?

The basic problems and some examples of these are considered. As the technology is novel, other problems not presented here are bound to appear. Thus the authors aim to show the kinds of areas in which ethical questions can be expected to arise.

4.1 Right to ones voice

The main question raised is whether a person has a natural right to their voice, and if so, what kind of a right? Whether we have rights to our voice through ownership or whether it is a natural human right or neither.

Ownership could be comparative to Intellectual Property Rights (IPRs). Three possibilities from the current IPR legislation could be candidates for ownership of ones voice: trademark, copyright and patent.

Even though Harley-Davidson abandoned their trademark application for the sound of V-Twin motorcycle engine [USPTO, 1994], as things such as UPS colour “brown” can be trademarked [USPTO, 2002], ownership to ones voice could be seen as a comparable. The main problem with trademarks is that the voice would then actually need to be trademarked. The process itself is not free of costs nor something that could be expected to be done by all potential people whose voice could be used for various purposes such as targets for imitation, answering machine messages, advertisement voices or for reading books etc.

Another option for protecting ones voice could be a copyright. The aspect of a copyright being a right to control reproductions would undoubtedly suit this situation. Artistic creations however would rather form rights to the users of the application, as the original voice would only be used as “inspiration”, i.e. the source for the actual product. Some kind of automatic licensing system could of course be considered, but that would not be according to the purpose of copyright; namely to control the use of the immaterial resource used.

Third option from an ownership angle would be a patent. If genes can be patented, as genes are found [Bioteknologia Info, 2005; Human Genome Program, 2006], not invented, and yet can be patented and as they are undoubtedly an innate feature of a person they are in, still fall under the patentable “inventions”, voice could be considered to be similar. Patent process is arguably even more cumbersome than a trademark process, however, and as ones “voice” is even less an invention than a gene, the use of patent would be troublesome at best.

As the three previous models have clear problems, a fourth option would of course be to invent a new form of IPRs. How practical such a new form then would be, is questionable,

especially if it only included ones voice. It could, however, be broadened to include *all* natural properties of human beings. This could then extend to cover all kinds situations, such as taking pictures of one (for commercial purposes) at public places etc. Certain exceptions could easily be made, such as taking pictures of political figures or their voices for satire purposes (similar to current human imitation uses).

The last option considered in this paper is that of an innate, unquestionable right to any uses of the voice of a person. As a comparison, one typically has no right to ones picture – pictures taken in public places do not need authorisation from the target of the picture to be published. Would a voice sample on a public place be similar? Intrinsic part of oneself? Why is a picture, i.e. "form", especially a face, not one then?

4.2 Misuse of ones voice

The second apparent question is whether the speech of a person can be misused. Should it be, that for example Osama Bin Laden actually is dead (as no video independently confirmed to be current of him has been seen for quite some time now) a synthesis software could easily enough be used to propagate the same ideas he has been known to support – or to make up new causes using his authority over certain groups. The authenticity of most recordings which are claimed to be made by him after late 2002 have been questioned [The Guardian, 2002; BBC, 2002, 2006]; for example BBC often uses terms such as "attributed to Osama Bin Laden" or "purporting to be from Bin Laden" – and even in cases where it is claimed that CIA analysts have verified the tape to be from Bin Laden, can they actually know that considering the level of even current speech synthesis software?

The previously reported case of voice recognition for password renewal is also a clear example of a misuse of the voice synthesis applications. Of course, password renewal would hardly be the only possible misuse for such applications. Simple examples of misuse of this nature rise easily to ones mind. Misuses such as calling a boss claiming to be someone else and having their voice to verify this and then misusing the situation either to get information from the person called or for example to slander co-workers or just have a go at the boss in someone else's name.

These cases, however, do not pose a "real ethical question" as they are undoubtedly misuses of the application. They are, of course relevant and ought be solved through legal measures (as of course slander and denigration already are), but also, if at all possible through technical measures. Some preliminary solutions are offered in a later chapter.

4.3 Borderline cases

Some exceptions to exclusions of the use of someone else's voice have apparently been made, e.g. imitator's use the voice of political figures often enough in their shows, often in a satirical context. However, the context in which it is used is also clearly identifiable as entertainment. Applications such as answering machine voices recorded by an imitator to imitate celebrity voices are closer to the problem presented. Do the persons whose voices are being used have a right to control the use of their voice in this kind of applications? What about situations in which a book is now read by the author; what if that is done 'with the author's voice' instead of 'by the author'? Is this misleading the customer of a voice book?

What kinds of contracts can and should be made with a person whose voice is being used? Should they be compensated for the use of their voice if the book is written by them or not? *Prima facie,* it would seem clear that they should be, but the copyright to the product read with their voice goes to the one using the application to create the read version of the book or article, not to the person whose voice is actually used.

Another more borderline case is "creating" advertisement or reader voices which strongly resemble that of known current voices of people who read children's programs or animated or digitally created movies, announce sports events or read nature programs.

5 Technical solutions

Having these questions in mind, does the producer have some ethical responsibilities, and if, what are these responsibilities? Should the product of the synthesizer be water marked in some way to distinguish it from the original speech of the person? Is the responsibility producer's or user's?

Although Howley et al. [2002] consider specifically privacy enhancing technologies, in a similar manner system design personnel need to be aware of the possible uses and consequences of these uses of speech simulation software (for Stakeholder theory, see e.g. [Bowern et al., 2004]). The software cannot be released before a study of the effects is done. In a competitive environment, the pressures to release early, without considering all the implications of the release are considerable [Powers, 2002]. However, the consequences of a release of an unfinished product that does not support abuse of the software can also be considerable.

A simple technical solution to the problem would be to insert a watermark to the artificially created sound. Inserting a wave length not heard by human ear, higher or lower than the human ear can decipher, which could be easily picked up by software could be an option. Another possible solution would be to insert sounds which do not muddle the original voice but are none the less present at same time as a more loud sound which the human ear would decipher primarily from the voice produced.

6 Legal remedies

If ownership of even such minor immaterial objects as items in a computer game (see e.g. [Reynolds, 2002, or Kimppa and Bissett, 2005]) can be considered to be a major issue, surely ownership of ones voice and its (perfect) imitation is a central question for ones self image.

Should the donor be automatically compensated for the use of their voice? If an automatic licensing system is taken into use, it would clearly be problematic for many reasons. Ones voice could be used e.g. in advertising products one would not commend. This would seem to indicate that the person should be able to control how and for what purposes their voice is used. A possibility to enforce the uses of ones voice could be a system which resemble the current IPR systems, in which the main argument is the right to exclude others from using the 'work', here the voice of the person in question. This, however, would run into the previously presented problems.

7 Discussion

Preliminarily, it would seem evident that misuses of such software are possible, if not even probable. Thus, the design of the software should already aim to minimize the possibility of such misuse, be it natural rights of the donor, duties toward the donor by the users and producers or just plain consequences of the use of such software.

What is more problematic is what kind of personal or innate right, if any, does the donor have to their voice. *Prima facie* it would seem plausible, that ones voice cannot be used without some form of authorization from the 'owner' (for a lack of a better term) of the voice. Whether this authorization can then be either an automatic license with a compensation for the use or whether the permission needs to be obtained through a contract is something which remains for the legal scholars and politicians to solve.

Acknowledgements

The authors would like to acknowledge the help of Tomi Orre in finding real world examples of the issues presented.

References

BBC, (2002), Bin Laden tape 'not genuine', Friday, 29 November, 2002, http://news.bbc.co.uk/1/hi/world/middle_east/2526309.stm, accessed 19.6.2008.

BBC, (2006), Timeline: The search for Bin Laden, http://news.bbc.co.uk/1/hi/world/south_asia/2827261.stm, accessed 19.6.2008.

Bioteknologia Info (2005), Bioteknologia tuo maapallolle uusia mahdollisuuksia, mutta myös tarpeita yhteisille pelisäännöille. Siksi se kiinnostaa koko kansainvälistä yhteisöä, Bioteknologia Info, 4/2005, available at: http://www.bioteknologia.info/etusivu/maailma/fi_FI/maailman_nakokulmasta/, accessed 27.6.2008.

Bowern, M., McDonald, C. and Weckert J. (2004), Stakeholder theory in practice: Building better software systems. Ethicomp 2004, University of the Aegean, Syros, Greece, 14 to16 April 2004, pp. 157—169.

Evans, Jill, and Mahoney, John (2004), Ethical and Legal Aspects of Using Digital Images of People: Impact on Learning and Teaching. Ethicomp 2004, University of the Aegean, Syros, Greece, 14 to16 April 2004, pp. 289—297.

The Guardian (2002), Swiss scientists 95% sure that Bin Laden recording was fake, Saturday November 30 2002, http://www.guardian.co.uk/world/2002/nov/30/alqaida.terrorism, accessed 19.6.2008.

Howley, Richard, Rogerson, Simon, Fairweather N. B. and Pratchett, Lawrence (2002), The Role of Information Systems Personnel in the Provision for Privacy and Data Protection in Organisations and Within Information Systems. Ethicomp 2002, Universidade Lusíada, Lisbon, Portugal, 13-15 November 2002, pp. 169—180.

Human Genome Program (2006), Genetics and Patenting, available at http://www.ornl.gov/sci/techresources/Human_Genome/elsi/patents.shtml, accessed 27.6.2008.

Kimppa, K. K. and Bissett, A. (2005), Is cheating in network computer games a question worth raising? CEPE 2005, July 17-19, Enschede, The Netherlands, pp. 259—267.

Klatt, D. H. (1980): Software for a Cascade/Parallel Formant Synthesizer. Journal of the Acoustical Society of America 67, 971—995.

Powers, Thomas M. (2002), Responsibility in Software Engineering: Uncovering an Ethical Model. Ethicomp 2002, Universidade Lusíada, Lisbon, Portugal, 13-15 November 2002, pp. 247—257.

Reynolds, Ren (2002), Intellectual Property Rights in Community Based Video Games. Ethicomp 2002, Universidade Lusíada, Lisbon, Portugal, 13-15 November 2002, pp. 455—470.

Qin, L., Wu, Y.-J., Ling, Z.-H., and Wang, R.-H. (2006), *Improving the performance of HMM-based voice conversion using context clustering decision tree and appropriate regression matrix,* Proc. of Interspeech, pp.2250—2253. For examples, see http://mail.ustc.edu.cn/~qinlong/demo.htm, accessed 5.26.2008.

Saarni, T., Paakkulainen, J., Mäkilä, T., Hakokari, J., Aaltonen, O., Isoaho, J. & Salakoski, T. (2006): Implementing a Rule-based Speech Synthesizer on a Mobile Platform. T. Salakoski et al. (Eds.): FinTAL 2006, LNAI 4139, pp. 349—355.

Siemens AG (2008), Listen who's talking: Siemens offers voice recognition solution, Press release, Heinz, A. http://www.siemens.com/index.jsp?sdc_sectionid=5&sdc_langid=1&sdc_unitid=2&sdc_contype=2&sdc_contentid=1478053, accessed 5.26.2008

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura T. (2000), Speech parameter generation algorithms for HMM-based speech synthesis, Proc. of ICASSP, pp.1315—1318.

USPTO (1994), US Serial No: 74485223, Original documents not available, withdrawn, http://tarr.uspto.gov/servlet/tarr?regser=serial&entry=74485223, accessed 27.6.2008.

USPTO (2002), Brown (single color used for the entire goods/services), US Serial No: 76408109, 13-May-2002, available at: http://tarr.uspto.gov/servlet/tarr?regser=serial&entry=76408109, accessed 27.6.2008.

Vainio, M., Altosaar, T., Karjalainen, M. and Aulanko R. (1998), Modeling Finnish Microprosody with Neural Networks, Linguistica Uralica, 34(3):199—204.

Weckert, John and Adeney, Douglas (1994), Ethics in Electronic Image Manipulation. Ethics in the computer age, Galtinburg, Tennessee, United States, pp. 113—114.

Paper VIII

Implementing a Rule-Based Speech Synthesizer on a Mobile Platform

Tuomo Saarni, Jyri Paakkulainen, Tuomas Mäkilä, Jussi Hakokari, Olli Aaltonen, Jouni Isoaho and Tapio Salakoski. FinTAL 2006, 5th International Conference on Natural Language Processing, 23-25 August 2006.

Implementing a Rule-Based Speech Synthesizer on a Mobile Platform

Tuomo Saarni[1], Jyri Paakkulainen[2], Tuomas Mäkilä[2], Jussi Hakokari[3], Olli Aaltonen[3], Jouni Isoaho[1], and Tapio Salakoski[1]

[1] Turku Centre for Computer Science, FI-20014, Finland
{Tuomo.Saarni, Jouni.Isoaho, Tapio.salakoski}@it.utu.fi
[2] Department of Information Technology, University of Turku, FI-20014, Finland
{Jyri.Paakkulainen, Tuomas.Makila}@it.utu.fi
[3] Phonetics Laboratory, University of Turku, FI-20014, Finland
{Jussi.Hakokari, Olli.Aaltonen}@utu.fi

Abstract. This paper describes the structure of a Finnish speech synthesis system developed at the University of Turku and evaluates the preliminary results of its implementation and performance on a platform with limited computing power. A rule-based approach was selected due to its high adaptability, low memory and computational capacity requirements. The speech synthesis system is written in Java™ MIDP 2.0 and CLDC 1.1. The synthesis is implemented on Nokia 6680 mobile device as a 65 kilobyte MIDlet. The system produces artificial speech at the sampling rate of 16 kHz. The results show that for a second of synthesized speech it takes 2.66 seconds for the system to produce it. Although the implementation was successful, improvements are needed to achieve a more acceptable level of time consumption.

1 Introduction

Speech synthesis systems have been available for decades, and several ways to produce synthesized speech have emerged. We have set out to study an older method of creating artificial speech, a rule-based speech synthesis. A rule-based speech synthesis may be considered a truly synthetic way to produce speech since it does not make use of any samples of natural, human speech as most of the other synthesis methods do. On the other hand, the rule-based speech synthesis is commonly considered to be the most challenging way to produce high-quality synthetic speech.

The development of embedded systems introduces whole new platforms with less memory capacity and computing power than in personal computers. Although technology evolves fast, we are encouraged to study possibilities in creating synthesized speech with less computational capacity than usually available. A rule-based synthesis system may be considered a small and computationally light system, and therefore suitable especially for platforms with limited capacity.

To examine the performance of the system, we have measured the time consumption of producing a second of synthesized speech signal (from hereon referred to as time cost ratio). I.e. the time consumed in creating the synthetic speech signal was

T. Salakoski et al. (Eds.): FinTAL 2006, LNAI 4139, pp. 349 – 355, 2006.

divided by the duration of the resulting signal. A real-time system would then produce more than a second of artificial speech in less than a second (the time cost ratio being less than 1). This would enable a real-time streaming of the synthesized speech simultaneously when created.

The study at hand investigates the possibility of implementing a rule-based speech synthesis on a mobile device supporting Java™ MIDP 2.0 [7] and CLDC 1.1 [1]. We are also interested of the system's time consumption on the chosen platform. The synthesis software was originally written in Java™ for personal computers; a MIDlet was the most obvious choice for implementation. The system was built at the University of Turku and was initially used to produce synthetic speech stimuli for behavioral experiments in speech sciences. Later on the system was re-evaluated by a joint project by the Phonetics Laboratory and the Department of Information Technology.

This paper first shows the overall structure of the speech synthesis program, followed by the testing procedure and the results. Finally, the results are discussed before the conclusive remarks.

2 Program Structure

This paper presents the system in two phases: the high-level synthesizer and the speech signal generator. The high-level synthesizer handles transcription, sets the segmental durations, models the fundamental frequency contour, and implements the phoneme-level rules. The signal generation phase creates the sound signal from an information matrix it receives from the high-level synthesizer. The structure is described in figure 1.

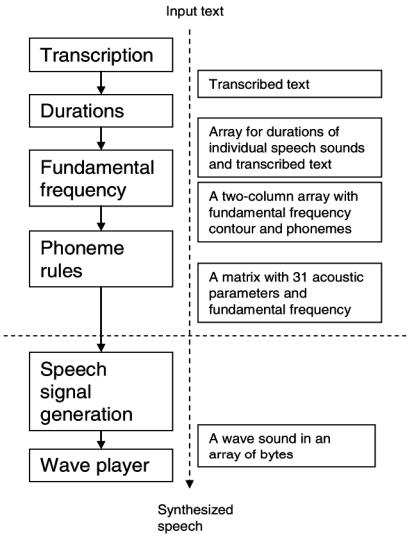


Fig. 1. The structure of the software is divided into two phases; the high-level synthesizer and the signal generation, which are separated by a horizontal dashed line. Each step in the synthesizing procedure is shown on the left and the changes in the information state is described on the right.

Implemented on a mobile device, the system does not differ much from the one on a PC. The software on the mobile device has fixed parameters that are adjustable on the PC version. Both systems operate on a time resolution of 10 milliseconds. Therefore, actual changes in the signal can only take place a hundred times per second. The synthesis software is currently used like text messaging with mobile devices. The input text is first typed and can then be synthesized.

Transcription refers to the conversion of the orthographic text to a phonetic representation. Finnish has a great advantage over many other languages from the point of view of speech synthesis. Finnish has a very high level of one-to-one correspondence between spelling and pronunciation. There are only few exceptions and most of them are possible to deal with simple rules; there is practically no need for an exception dictionary.

The durations are fixed to 70 milliseconds for short phonemes. Phonemically long phonemes are 140 ms long. A comma results in a 150 millisecond pause; a full stop introduces a 350 millisecond pause. The system has a prepausal lengthening implemented [2] [9]. Prepausal lengthening refers to the human tendency to slow down articulation right before a pause. The module now lengthens speech sounds' duration in the last word of each phrase, producing 90 ms and 180 ms long speech sounds instead of 70 ms and 140 ms. The durations are set in array which include the transcribed text and the duration of each character in transcription (including pauses).

The fundamental frequency is set to cascade model meaning that the frequency starts at 100 Hz and falls down to 70 Hz by the end of the sentence, crudely imitating a typical male speaker. Within each word of the sentence, the frequency raises 40 Hz during the first speech sound of each word. That represents the lexical stress always found on the first syllable in Finnish. If there's a comma within a sentence, this causes an additional rise of 20 Hz in the fundamental frequency.

The phoneme rules comprise of 34 acoustical parameters. Duration of each speech sound is needed to calculate its transition to the next. It has its own acoustical parameters which needs to be reached during a transitional phase between two speech sounds. The transition is currently done linearly and is usually very short, typically in the order of 30 milliseconds. The acoustical parameters are listed in Table 1. Most of the parameters are fixed in the current version so that they are shared by every speech sound.

The signal generator receives the matrix of parameters from the high-level synthesizer. The speech signal is generated by a formant synthesizer that resembles the ones described in [5] and [6]. The signal generator consists of a vocal tract model and sound sources for voicing, frication and aspiration. The vocal tract model consists of a cascade branch and a parallel branch. The cascade branch is used for generating

Table 1. Acoustical parameters used in signal signal generation with explanations

Parameter	Explanation	Parameter	Explanation
F0	Fundamental frequency	FNP	Frequency of nasal formant
AV	Amplitude of voicing	BNP	Bandwidth of nasal formant
TL	Voicing source low frequency emphasis	FNZ	Frequency of nasal antiformant
AF	Amplitude of frication	BNZ	Bandwidth of nasal antiformant
AH	Amplitude of aspiration	A1F…A6F	Amplitudes of parallel branch formants
F1…F6	Frequencies of first six formants	B1F…B6F	Bandwidths of parallel branch formants
B1…B6	Bandwidths of cascade branch formants	AB	Amplitude of bypass frication

sounds that consists of voicing and/or aspiration noise. The parallel tract is used mainly for fricative and plosive sounds. The signal generator is controlled by parameters shown in Table 1.

3 Implementation Platform

The implementation was conducted with Nokia 6680 mobile device using software version 4.04.07 (dated 22-08-05) and firmware version RM-36. The mobile device supports CLDC 1.1 (Connected Limited Device Configuration) [1] and Java™ MIDP 2.0 (Mobile Information Device Profile) [7]. The jar and heap size are only restricted by the available memory of the device [8].

The Nokia 6680 mobile device was selected as the platform due to its average MIDP 2.0 performance according to the result database of the JBenchmark J2ME benchmarking tool [4]. It should be noted that the benchmarking tool emphasizes graphical performance. Nevertheless, the results give guidelines on the overall performance of the device.

Compared to the previous versions both CLDC 1.1 and MIDP 2.0 offered several important features needed for the rule-based speech synthesis. Especially the floating-point support introduced in the CLDC 1.1 and the built-in Media API of MIDP 2.0 were valuable. However, the obvious problem in the MIDP 2.0 Media API was the lack of streaming of sounds. Speech is currently synthesized by first creating the signal in full and then playing it afterwards.

4 Testing Procedure and Time Consumption Measurements

The performance of the system was examined on the chosen platform. All non-relevant options and add-on devices were eliminated to minimize any interference. The testing was done in the following environment:

- No SIM card inserted
- No other programs installed
- No unnecessary memory usage (no calendar markings, phone numbers etc.)
- No memory card installed
- No optional devices connected
- No other programs running except the ones the device itself uses automatically when on
- The device fully charged and connected to the charger
- The device on before synthesizing a text and shut down after each synthesized passage
- Two-minute wait after the device was turned on to ensure it is fully functional
- Half-minute wait after the synthesis software was started to ensure the program has loaded
- After the text is typed, it is synthesized immediately
- The elapsed time is reported by the software itself and shown on the display

The synthesized sentences are random pickings from a Finnish periodical Suomen Kuvalehti. They represent Standard Finnish and are of varying length to provide information on the effects of varying input. Each input sentence was synthesized several times to achieve a more reliable average of the time cost ratios.

The average time consumption of the high-level synthesizer (phase 1) was 0.34 seconds per second of synthesized speech, ranging from 0.26 to 0.45. The high-level synthesis is therefore a real-time phase. The average time cost ratio of the signal generator (phase 2) was 2.32, ranging from 1.30 to 4.18. The total time cost ratio is 2.66 on the average. Consequently, the program is not a real-time system.

The results show that the time consumption increases in the signal generation phase as the input text grows longer. The same effect does not occur in the high-level synthesis. However, the time consumption of the entire synthesis consists mainly of the signal generation.

Table 2. The columns contains the input text, duration of the resulting synthetic sentence, standard deviations of time consumption of both phases (the high-level synthesis and the signal generation), the average time usage and the time cost ratios of both phases and the entire operation

Synthesized text	Dur. of the synth. speech (s)	St.dev.		Average time usage (s)			Time cost ratio		
		Phase 1	Phase 2	Phase 1	Phase 2	Total	Phase 1	Phase 2	Total
Lopulta kaipaatte tilaisuutta tunnustaa.	3.06	0.17	0.38	0.90	3.97	4.87	0.29	1.30	1.59
Muussa tapauksessa hän suunnittelee yliopistoon menemistä.	4.34	0.12	0.06	1.93	6.27	8.20	0.45	1.44	1.89
Nyt oli kiire, sillä kohta vihollinen ampuisi kaikilla pilleillä ja putkilla.	5.61	0.26	0.06	2.30	9.83	12.13	0.41	1.75	2.16
Nimityksiä perusteltiin aluksi sillä, ettei hallinnossa juuri ollut vasemmistolaisia virkamiehiä.	7.35	0.21	0.20	2.23	16.80	19.03	0.30	2.29	2.59
Sotien jälkeen suomalaisilla ei enää ole ollut mahdollisuutta käydä Valamossa, aniharvaa poikkeusta lukuunottamatta.	8.79	0.12	0.15	2.93	26.17	29.10	0.33	2.98	3.31
Nykymaailmassa ihmisviidakoita löytyy monista yhden sallitun puolueen tai sotilaiden hallitsemista yksinvaltaisista tai harvainvaltaisista maista.	10.68	0.06	0.17	2.77	44.60	47.37	0.26	4.18	4.44

The translations of the synthesized samples are as follows:

- Finally you will be longing for the chance to confess.
- He is planning on going to the university in any other case.
- We were in a hurry now, because soon the enemy would be firing with all their might.
- The nominations were first rationalized by the fact that there were really no leftist officials in the government.
- After the war the Finns had no more the opportunity to visit [the monastery island of] Valamo, with very few exceptions.
- Jungles of men are found in the modern world in countries of tyranny or oligarchy run by the military or a single allowed party.

5 Discussion and Conclusion

Our goal to implement a rule-based speech synthesis to a mobile platform was successful. The time cost ratio was close to a real-time system with the shortest input.

The real-time goal is not realized when the input grows longer. If the time cost ratio would be close to one the synthesized speech could start to play at the very moment the first waveforms are generated. Naturally, this would require streaming the audio signal and parallel synthesizing on the background. The MIDP 2.0 and CLDC 1.1 did not support streaming, which was the main reason the real-time goal was not achieved. The current version writes the sound into a buffer in 10 ms samples, which makes it easier to develop a streaming solution with the existing APIs. Unfortunately, the current buffering of samples slows down the system with higher usage of memory. The parallel synthesis on the background is also feasible with threading. On the other hand, if time consumption ratio exceeds one, it can be used to determine how many seconds must be produced before the signal can start to play (the rest of the phonemes being processed on the background). Of course, the time consumption of the wave player must be examined.

The high-level synthesizer is language-dependent, while the low-level system is not. The high-level system, on the other hand, is fully modular. Any single module that does not fit with a new language can be modified, replaced, or inactivated. Each new phoneme can be added easily, and the existing ones can be adjusted to fit the specific pronunciation of the target language.

The current version of the speech synthesis is based on a version made for non-mobile platforms. The original version was not optimized for low time consumption and therefore the solutions made affect to the mobile version. We expect a revision of the code to improve the time cost ratio significantly. Another considerable benefit would be the lowering of the sampling rate from 16 kHz to 8 kHz. This would halve the time cost ratio of the phase 2.

Java is not considered the best possible solution for real-time systems on embedded platforms [3]. We have considered the possibility to change the coding from Java™ to Symbian™, or we might use the most time critical parts in the Symbian™ code, which might solve the real-time problem completely. However, we are more interested to develop the current implementation and the solutions within.

This study did not include an examination of the memory consumption. The testing platform (Nokia 6680) has 8 MB of memory but it is expandable with a memory card. Our MIDlet takes only 65 kilobytes (the size of the jar-file) of memory and the memory usage can be considered small, though it has to be confirmed in a separate study.

We expect to achieve the real-time goal with rule-based synthesis in the near future. Another goal is to put the modular design to test by implementing a second language into the synthesizer.

Acknowledgements

This study was partially funded by the Finnish Funding Agency for Technology and Innovation (TEKES).

References

1. Connected Limited Device Configuration Specification – Version 1.1. Sun Microsystems. JSR-139. (2003)
2. Hakokari, J., Saarni, T., Salakoski, T., Isoaho, J., Aaltonen, O.: Determining Prepausal Lengthening for Finnish Rule-Based Speech Synthesis. Speech Analysis, Synthesis and Recognition, Applications of Phonetics. AGH University of Science and Technology, Kraków, Poland (2005)
3. Higuera-Toledano, M.T., Issarny, V., Banatre, M., Cabillic, G., Lesot, J.-P., Parain, F.: Java Embedded Real-Time Systems: an Overview of Existing Solutions. Object-Oriented Real-Time Distributed Computing, Third IEEE International Symposium on 15-17 March (2000) 392–399
4. JBenchmark Home Page. http://www.jbenchmark.com/, Kishonti Informatics LP. Accessed on March 31st (2006)
5. Klatt, D. H.: Software for a Cascade/Parallel Formant Synthesizer. Journal of the Acoustical Society of America 67 (1980) 971–995
6. Klatt, D. H., Klatt L. C.: Analysis, synthesis, and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America 87 (1990) 820–857
7. Mobile Information Device Profile for Java™ 2 Micro Edition – Version 2.0. Motorola and Sun Microsystems JSR-118 (2002)
8. Nokia 6680 Developer Home Page. http://www.forum.nokia.com/devices/6680, Nokia. Accessed on March 31st (2006)
9. Vainio, M.: Artificial Neural Network Based Prosody Models for Finnish Text-to-Speech Synthesis. Academic dissertation, University of Helsinki (2001)

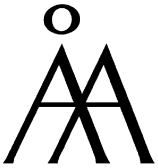
Turku Centre for Computer Science
TUCS Dissertations

94. Dubravka Ili, Formal Reasoning about Dependability in Model-Driven Development
95. Kim Solin, Abstract Algebra of Program Refinement
96. Tomi Westerlund, Time Aware Modelling and Analysis of Systems-on-Chip
97. Kalle Saari, On the Frequency and Periodicity of Infinite Words
98. Tomi Kärki, Similarity Relations on Words: Relational Codes and Periods
99. Markus M. Mäkelä, Essays on Software Product Development: A Strategic Management Viewpoint
100. Roope Vehkalahti, Class Field Theoretic Methods in the Design of Lattice Signal Constellations
101. Anne-Maria Ernvall-Hytönen, On Short Exponential Sums Involving Fourier Coefficients of Holomorphic Cusp Forms
102. Chang Li, Parallelism and Complexity in Gene Assembly
103. Tapio Pahikkala, New Kernel Functions and Learning Methods for Text and Data Mining
104. Denis Shestakov, Search Interfaces on the Web: Querying and Characterizing
105. Sampo Pyysalo, A Dependency Parsing Approach to Biomedical Text Mining
106. Anna Sell, Mobile Digital Calendars in Knowledge Work
107. Dorina Marghescu, Evaluating Multidimensional Visualization Techniques in Data Mining Tasks
108. Tero Säntti, A Co-Processor Approach for Efficient Java Execution in Embedded Systems
109. Kari Salonen, Setup Optimization in High-Mix Surface Mount PCB Assembly
110. Pontus Boström, Formal Design and Verification of Systems Using Domain-Specific Languages
111. Camilla J. Hollanti, Order-Theoretic Mehtods for Space-Time Coding: Symmetric and Asymmetric Designs
112. Heidi Himmanen, On Transmission System Design for Wireless Broadcasting
113. Sébastien Lafond, Simulation of Embedded Systems for Energy Consumption Estimation
114. Evgeni Tsivtsivadze, Learning Preferences with Kernel-Based Methods
115. Petri Salmela, On Commutation and Conjugacy of Rational Languages and the Fixed Point Method
116. Siamak Taati, Conservation Laws in Cellular Automata
117. Vladimir Rogojin, Gene Assembly in Stichotrichous Ciliates: Elementary Operations, Parallelism and Computation
118. Alexey Dudkov, Chip and Signature Interleaving in DS CDMA Systems
119. Janne Savela, Role of Selected Spectral Attributes in the Perception of Synthetic Vowels
120. Kristian Nybom, Low-Density Parity-Check Codes for Wireless Datacast Networks
121. Johanna Tuominen, Formal Power Analysis of Systems-on-Chip
122. Teijo Lehtonen, On Fault Tolerance Methods for Networks-on-Chip
123. Eeva Suvitie, On Inner Products Involving Holomorphic Cusp Forms and Maass Forms
124. Linda Mannila, Teaching Mathematics and Programming – New Approaches with Empirical Evaluation
125. Hanna Suominen, Machine Learning and Clinical Text: Supporting Health Information Flow
126. Tuomo Saarni, Segmental Durations of Speech

Turku
Centre *for*
Computer
Science

Joukahaisenkatu 3-5 B, 20520 Turku, Finland | www.tucs.fi

University of Turku
- Department of Information Technology
- Department of Mathematics



Åbo Akademi University
- Department of Information Technologies

Turku School of Economics
- Institute of Information Systems Sciences

ISBN 978-952-12-2391-4
ISSN 1239-1883