interspeech 2008 incorporating SST 08

Correlation of Utterance Length and Segmental Duration in Finnish Is Questionable

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

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

10.21437/Interspeech.2008-281

Abstract

This paper examines the way acoustic segmental duration correlates with utterance length in Finnish. It is commonly assumed that shorter utterances are characterized by greater segmental duration. However, that view has recently attracted some criticism. We conducted an explorative study on two linguistically uncontrolled Finnish-language speech corpora by examining segmental duration as a function of utterance length. We tested a hypothesis that the perceived differences in duration are in fact caused by short utterances containing a greater proportion of domain-edge effects, such as final lengthening. Pearson correlations were calculated for the timing information in different conditions. The results show that the weak correlation holds no more if domain edges are excluded from the material, suggesting there is no domain-span process at work.

Index Terms: segmental duration, Finnish, domain-span, domain-edge, final lengthening

1. Introduction

Common experience tells us that words articulated in isolation have considerably greater segmental duration than those in connected speech. The same may apply to very short utterances of two or so words. The tendency has provoked an idea that, by some mechanism, segmental duration gets shorter as a function of utterance length. In other words, a domain's (such as utterance) size has an inverse relation to its constituents' size (such as individual speech sounds). The same has been proposed for the smaller domain, word. A theory of polysyllabic shortening [1] claims that as a word grows in length (e.g. lone, lonely, loneliness), its constituent syllables become shorter.

Early on, Klatt [2] pointed out that polysyllabic shortening may be confounded by final lengthening. The issue is still debated and recent research has produced conflicting results. Turk [3] found evidence for polysyllabic shortening, while White's recent studies of domain-edge and domain-span processes [4] found little evidence for compression on word level or utterance-level. He suggests that perceived domain-span (something that affect the entire domain, such as utterance) effects may in fact be caused by domain-edge processes (something that affects the beginning or the end of the domain), such as final lengthening and utterance-initial shortening and lengthening. Suomi [5] has studied Finnish, again finding no compelling evidence for a word-level process. He concludes that "The results show that polysyllabic shortening does not operate in Finnish. Apart from the very shortest words, segments and syllables had the same duration irrespective of how many syllables followed these units in the same word, in both unaccented and accented versions. Lengthening of vowels in the very shortest words in turn may have independent motivation: these vowels were presumably lengthened relative to those in longer words to make room for the tonal rise–fall tune. Indeed, the lack of polysyllabic shortening may not require any explanation at all, as the burden of proof seems to rest on those who insist on it." Regardless of criticism towards polysyllabic shortening, assumptions about domain-span compression on the utterance level continue to surface in discussions about segmental duration.

We will employ a different kind of material and methodology to test the hypothesis that there is no independent domain-span effect that would account for the seemingly shorter segmental durations in longer utterances. The hypothesis is that short utterances and isolated words are mostly or entirely affected by lengthening processes such as final lengthening and prominent accent, producing longer segmental duration without a specific domain-span mechanism.

The study makes use of two small, uncontrolled Finnish speech corpora as opposed to traditional lab speech. Both corpora are known to exhibit considerable utterance-final deceleration of articulation rate [6] (amounting to final lengthening for present practical purposes), as well as various kinds of utterance-initial lengthening and shortening [7] [8]. Instead of measuring and comparing absolute duration at specific locations in utterances of varying length (as is customary), we will approach the problem in terms of correlation. If an utterance-level domain-span effect is expected, we should also find a significant negative correlation between segmental duration and utterance length.

2. Speech material

Two different small Finnish speech corpora were used in the study. One is elicited and the other can be considered not elicited, but not spontaneous either. They are spoken by more or less professional speakers in a formal, literary style. The corpora have the advantage of well structured, longer utterances without extensive hesitating, pausing, or incomplete sentences. On the other hand, the speech material is likely to have many qualities that are alien to spontaneous and colloquial Finnish, and the results should not be interpreted to represent those. The corpora were manually annotated at phone level by trained phoneticians.

The single speaker corpus consists of excerpts from a periodical read aloud by a male speaker. The corpus is thus elicited, but there is no specific linguistic pattern; it contains utterances of various lengths. There are 967 utterances with 41 306 phones altogether. Median utterance length was 40; the distribution is shown in [Figure 1].

Accepted after peer review of full paper

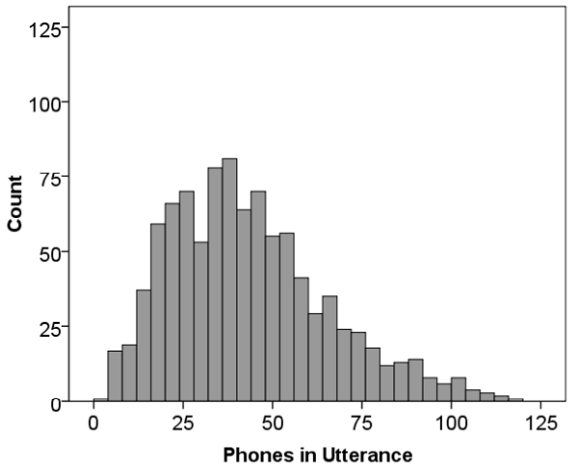

Figure 1: *Distribution of utterance length in the single-speaker corpus.*

The multi-speaker corpus consists of television news reading and field reports, a weather forecast and radio presentations. The corpus features 15 adult speakers of varying ages; 9 men and 6 women. There were 1156 utterances with 31 414 phones altogether. Median utterance length in the corpus was 24; the distribution is shown in [Figure 2].

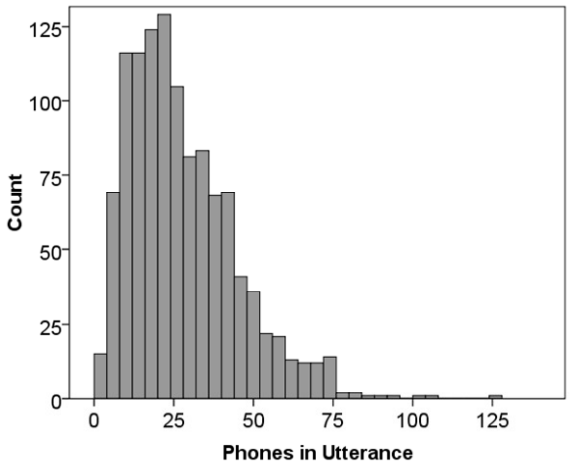

Figure 2: *Distribution of utterance length in the multi-speaker corpus.*

3. Methods

The idea was to determine whether there is a negative correlation between utterance length and the duration of individual segments, indicating a domain-span mechanism affecting articulation rate. The domain-edge vs. domain-span controversy will be approached by manipulating sampling; intact utterances' correlation will be compared to utterances' of which the edges have been eliminated. Finally, we construct an artificial setting by comparing very long (minimal influence of edge processes) and very short (maximal influence of edge processes) utterances to demonstrate how radically different the results can be if the influence of edge processes are overlooked entirely. In a quantity language such as Finnish there are considerable differences in inherent duration between phonologically short and long segments. To reduce resulting variation, the material was further divided into seven categories of phones. The seven categories were phonologically short vowels, non-plosive consonants, voiceless plosives, the phonologically long counterparts of the three, and diphthongs.

Utterance was defined as a continuous stretch of speech that is delimited by acoustic silence in left and right contexts. Utterance length was measured in how many phones it contains, not by the amount of words or syllables. No further linguistic or phonetic detail and structure was taken into consideration. Should the described domain-span process be at work, it will have to produce gradually shortening segments in the corpora as we increase utterance length, regardless of the content.

All the following procedures involve a Pearson correlation test for correlation and determination (adjusted R^2). The durations of individual phones (dependent variable) in the given categories were compared against utterance length (independent variable) distinguishing the single-speaker and the multi-speaker corpus.

3.1. Procedure A

The first test was run on the entire, unmodified speech material at hand. The purpose was to investigate if a correlation exists, and if so, to establish a point of comparison for procedures B and C.

3.2. Procedure B

Procedure B was identical to A in all respects, but the data itself was manipulated to exclude the effect of domain-edge processes. The data was cropped so that the first 3 phones and last 10 phones were eliminated in all utterances. We have observed some amount of final lengthening taking place as early as 10 phones from the end of the utterance, whereas there are no initial effects after the 3rd phone. The manipulation thus eliminated all the utterances with 13 or less phones entirely and left the rest truncated. The only remaining phone of a 14-phone utterance would still be considered a segment of a 14-phone utterance in the correlation test, not 1. The multi-speaker corpus had a greater proportion of very short utterances due to its more spontaneous nature, and was consequently trimmed more in the process.

3.3. Procedure C

Finally we studied only the longest and shortest of the utterances. Out of the original, unmanipulated data only utterances with 70 or more phones and 10 or less phones were included in the test. Recognizing that such is by no means a justifiable research method itself, we wanted to make a "garbage in, garbage out" demonstration of what kind of results are possible with biased speech material and by ignoring the influence of edge effects. The data is no longer valid for Pearson correlation as it is not normally distributed.

A practical misapplication would be to compare mean durations in isolated words and short expressions directly against those in longer utterances in connected speech. The results would presumably show a great disparity in segmental durations, but still not reflect a domain-span process.

4. Results

The results are presented procedure by procedure. The tables below show Pearson correlation and the adjusted R^2 scores by the two corpora and the seven sound categories.

4.1. Results A

The results of procedure A show that there is a very weak yet statistically significant (apart from long non-plosive consonants in the multi-speaker corpus) negative correlation between segmental duration and utterance length. Interpreting scores that low calls for caution, but we may assume that somehow utterance length influences acoustic duration marginally (cf. the coefficient of determination), and perhaps not in a very linear fashion. If the lengthening takes place only at domain-edges, the medial durations will not be affected. If that is correct, we will be able to neutralize all of the correlation or to boost it slightly by applying manipulations as described in Methods.

Table 1: *Results of unmodified corpora.*

	Multi-speaker Corpus		Single-speaker Corpus	
	Pearson Correlation	Adjusted R Square (%)	Pearson Correlation	Adjusted R Square (%)
Diphthongs	-0.114**	1.2	-0.138**	1.9
Long Vowels	-0.159**	2.5	-0.124**	1.5
Short Vowels	-0.131**	1.7	-0.080**	0.6
Long Voiceless Plosives	-0.225**	4.9	-0.148**	2.1
Short Voiceless Plosives	-0.146**	2.1	-0.070**	0.5
Long Non-plosive Consonants	-0.057	0.2	-0.153**	2.3
Short Non-plosive Consonants	-0.086**	0.7	-0.063**	0.4

* Correlation is significant at 0.05 level
** Correlation is significant at 0.01 level

4.2. Results B

With the utterance-edge process removed, the correlation is now even weaker and no longer uniformly negative. It is safe to say that there is no inverse relationship between utterance length and segmental duration in the manipulated data.

Table 2: *Results of manipulated corpora.*

	Multi-speaker Corpus		Single-speaker Corpus	
	Pearson Correlation	Adjusted R Square (%)	Pearson Correlation	Adjusted R Square (%)
Diphthongs	0.055	0.2	-0.072**	0.5
Long Vowels	0.085*	0.6	-0.019	<0.1
Short Vowels	-0.004*	<0.1	-0.021**	<0.1
Long Voiceless Plosives	-0.065	0.2	-0.063	0.2
Short Voiceless Plosives	-0.064**	0.4	0.009	<0.1
Long Non-plosive Consonants	0.070	0.3	-0.051	0.1
Short Non-plosive Consonants	-0.008	<0.1	-0.012	<0.1

* Correlation is significant at 0.05 level
** Correlation is significant at 0.01 level

4.3. Results C

The results for procedure C confirm that the correlation will become stronger with selecting only the longest and the shortest utterances.

Table 3: *Results of the short and long utterances only.*

	Multi-speaker Corpus		Single-speaker Corpus	
	Pearson Correlation	Adjusted R Square (%)	Pearson Correlation	Adjusted R Square (%)
Diphthongs	-0.348**	11.6	-0.105*	0.9
Long Vowels	-0.344**	10.6	0.177**	3.0
Short Vowels	-0.361**	12.9	-0.091**	0.8
Long Voiceless Plosives	-0.628**	38.6	-0.109	0.7
Short Voiceless Plosives	-0.333**	10.9	-0.085**	0.7
Long Non-plosive Consonants	0.132	1.0	-0.228**	4.9
Short Non-plosive Consonants	-0.203**	4.1	-0.065**	0.4

* Correlation is significant at 0.05 level
** Correlation is significant at 0.01 level

There is a striking difference between the two corpora. The multi-speaker corpus correlations have grown considerably stronger, reaching particularly strong correlation and determination scores for the phonologically long voiceless plosive consonants. The single-speaker results, on the other hand, show hardly any quantitative change from the unrestricted material in procedure A. The correlation for long vowels is now positive yet still weak. The disparity is easily explained with the different utterance length distribution in the materials. The single-speaker data now consists mainly of very long utterances [Figure 3] that may or may not have a small fraction of them affected by final lengthening. However, there is hardly a linear relationship between the position of a segment and its duration.

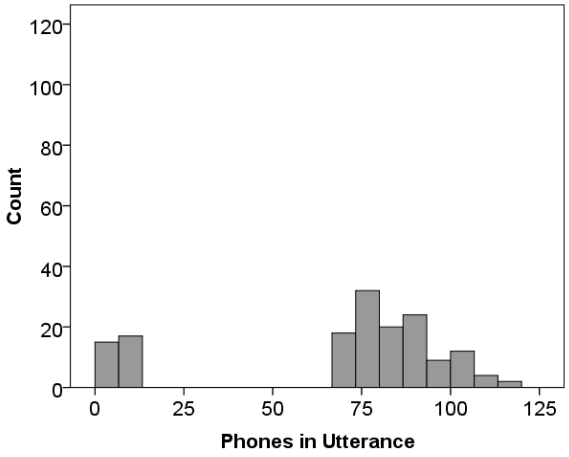

Figure 3: *Distribution of utterance length in the single-speaker corpus cropped for procedure C.*

Conversely, the multi-speaker has a great number of very short utterances that can not be considered connected speech. The number of long utterances is very small by comparison [Figure 4]. It is not surprising that in this material the correlation is stronger, as the calculation will encounter so many long segments in the short utterances, while the short segments are mainly in the small number of long utterances.

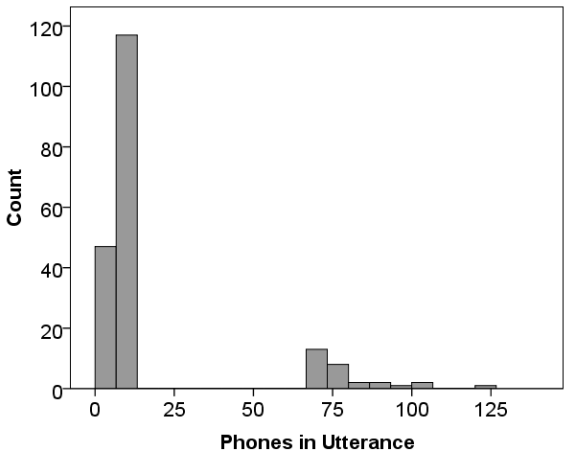

Figure 4: *Distribution of utterance length in the multi-speaker corpus cropped for procedure C.*

To summarize, the unmodified corpus produced a very weak but consistent negative correlation between utterance length and segmental duration. The material from which the interfering effects of domain edge processes had been eliminated produced no ($r<|0.09|$) correlation. The demonstration, in which only the longest and the shortest utterances were included, produced stronger negative

correlation (r≈-0.63 at most) depending on the utterance length distribution of the two corpora.

5. Discussion

Our findings are not in conflict with the common perception that very short utterances, effectively words in isolation or short expressions, exhibit greater segmental duration than what we would find in connected speech. However, there is nothing in our findings that would indicate a mechanism that affects duration over the entire domain of utterance. Edges, i.e. the beginnings and ends, are known to influence duration; eliminating them leaves no medial environment shortening of duration that could be explained in terms of utterance length. When we compare mean segmental duration (as opposed to calculating correlation) in utterances of varying lengths in our corpus, we will find the segments getting gradually shorter as the utterance length grows. That looks deceptively as if a domain-span process was at work. However, it is only the result of a greater proportion of the short utterance being affected by domain-edge processes. If the edges are removed, as in procedure B, there is no longer any systematic change.

From a practical point of view, there is nothing extraordinary here. Very short utterances can be considered to be under the influence of edge effects such as final and initial lengthening. Furthermore, they may be considered accented as well, as they carry all or most of the prominence; there is little else to share it with. Philosophically speaking, if the exclusive role of domain-edge processes holds, it may turn out to be a matter of taste whether what goes on should be called final lengthening, non-final shortening, or connected speech shortening. Nevertheless, our Finnish-language material has provided no support for a domain-span process that would expand segments in a short utterance and compress them in a long one. All the changes in articulation rate here appear to be influenced by the edges of the domain, and markedly the end of the domain.

The kind of results reported for English [4] and now for Finnish have a couple of important implications. First, the existing evidence ought to be considered before using an independent, inverse relationship between the length of a constituent and the duration of its subconstituent as an argument, especially regarding those languages. Second, the influence of domain edges should be taken into account when constructing experimental designs.

6. Conclusion

We have studied correlation between utterance length and the duration of its segments. The study was an explorative one and made use of two different, uncontrolled Finnish speech corpora. The purpose was to investigate whether an assumed domain-span process induces compression in segmental duration as utterances grow in length. We calculated Pearson correlation and adjusted R square scores for three sets of data. The results show that there is a weak correlation between utterance length and its segmental duration in an unaltered, raw timing data extracted from the corpora. Once final and initial portions of utterances (domain edges) that are known to undergo lengthening and shortening processes were cropped out of the data, there was no longer any correlation. The third manipulation left only the shortest and the longest of the utterances remaining and, expectedly, produced a stronger correlation. Our interpretation is that final lengthening, accent, and related phenomena may give the misleading impression of an inverse relationship between a constituent length (utterance) and its subconstituent duration (individual speech sounds). However, nothing in our results points toward an independent domain-span process influencing segmental duration.

7. References

[1] Lehiste, I., “The timing of utterances and linguistic boundaries”, Journal of the Acoustical Society of America, Volume 51, 2018-2024, 1972.

[2] Klatt, D.H., “Linguistic uses of segmental duration in English: Acoustic and perceptual evidence”, The Journal of the Acoustical Society of America, Volume 59, Issue 5, 1208-1221, 1976.

[3] Turk, A.E. and Shattuck-Hufnagel, S., “Word-boundary-related duration patterns in English”, Journal of Phonetics, Volume 28, Issue 4, 397-440, 2000.

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

[5] Suomi, K., ”On the tonal and temporal domains of accent in Finnish”, Journal of Phonetics, Volume 35, Issue 1, 40-55, 2007.

[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.

[7] 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.

[8] 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.