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.

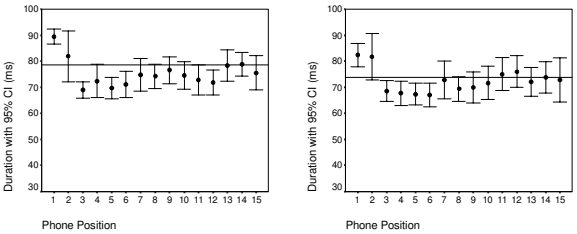

Figure 1. /s/ in SS (left) and MS (right) corpora.

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.

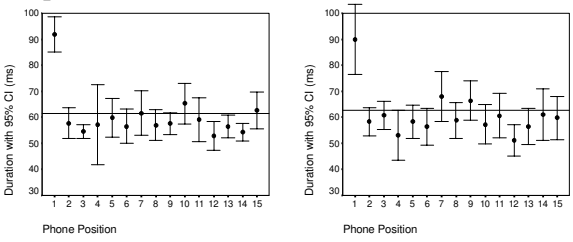

Figure 2. /r/ in SS (left) and MS (right) corpora.

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.

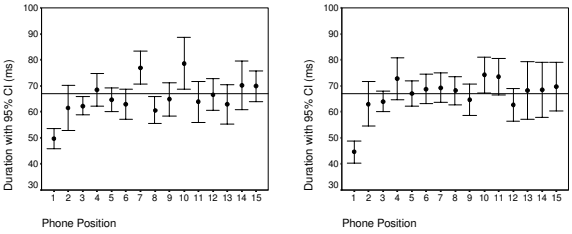

Figure 3. /m/ in SS (left) and MS (right) corpora.

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.

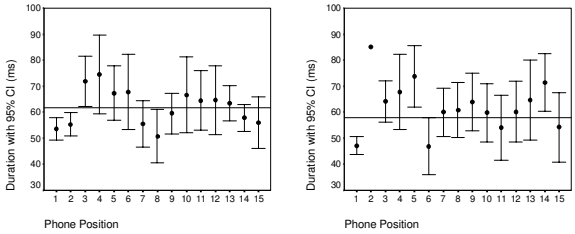

Figure 4. /j/ in SS (left) and MS (right) corpora.

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.

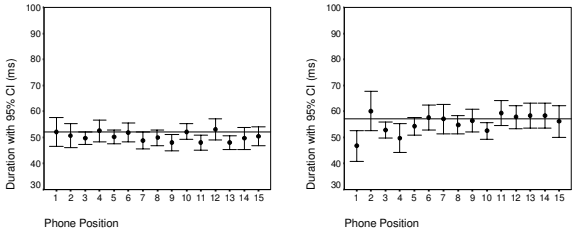

Figure 5. /n/ in SS (left) and MS (right) corpora.

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.

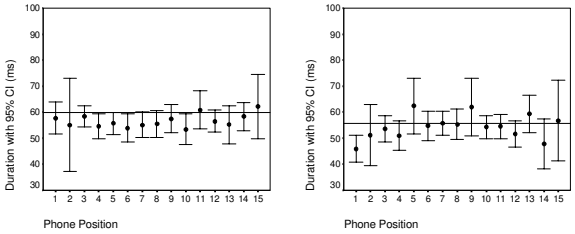

Figure 6. /ʋ/ in SS (left) and MS (right) corpora.

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.

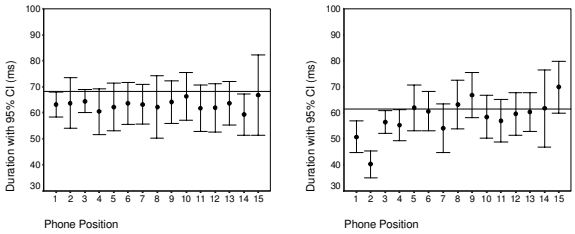

Figure 7. /h/ in SS (left) and MS (right) corpora.

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.

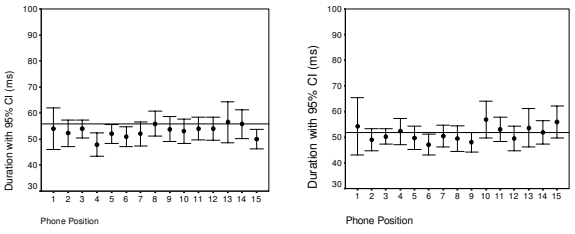

Figure 8. /l/ in SS (left) and MS (right) corpora.

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 [s^w] 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).