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, 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.

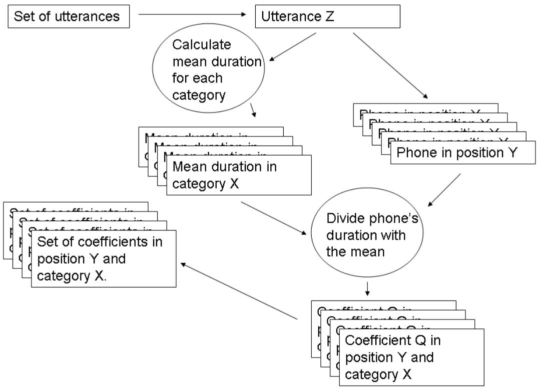
Set of utterances
Utterance Z
Calculate mean duration for each category
Phone in position Y
Mean duration in category X
Divide phone's duration with the mean
Set of coefficients in position Y and category X.
Coefficient Q in position Y and category X

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.

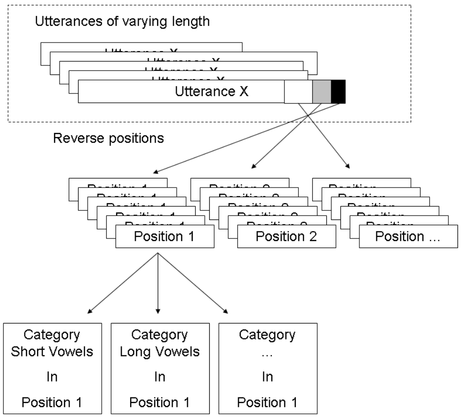
Utterances of varying length
Utterance X
Reverse positions
Position 1
Position 2
Position ...
Category Short Vowels In Position 1
Category Long Vowels In Position 1
Category ... In Position 1

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.

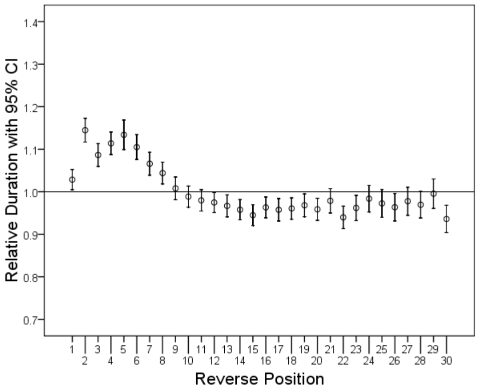

Figure 3: *Relative duration of short vowels*

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.

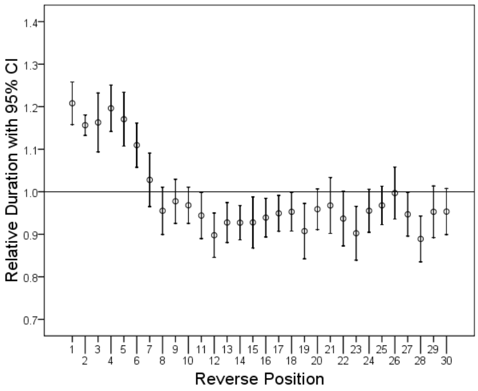

Figure 4: *Relative duration of long vowels*

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

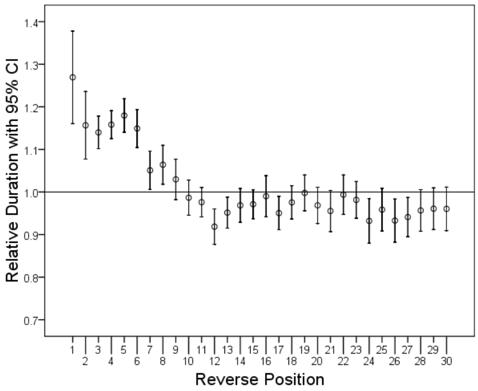

Figure 5: *Relative duration of diphthongs*

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.

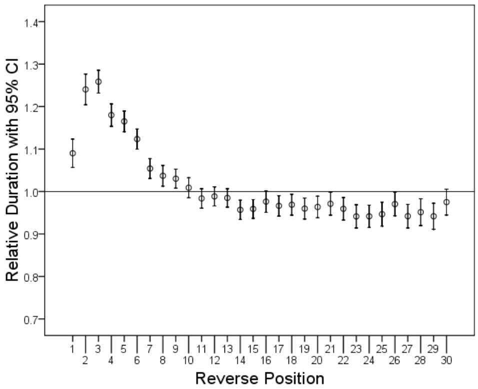

Figure 6: *Relative duration of short non-plosive consonants*

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.

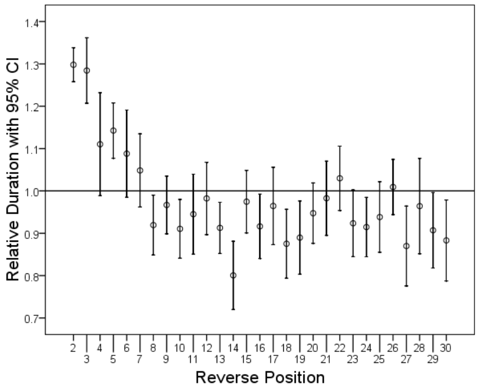

Figure 7: *Relative duration of long non-plosive consonants*

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.

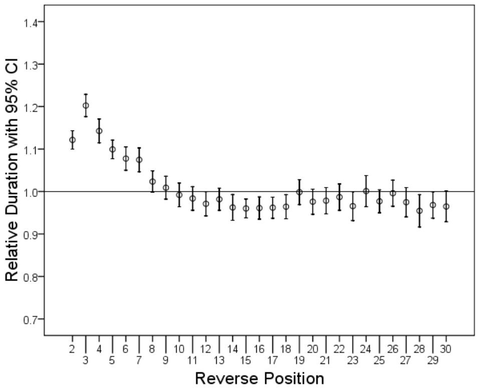

Figure 8: *Relative duration of short voiceless plosives*

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.

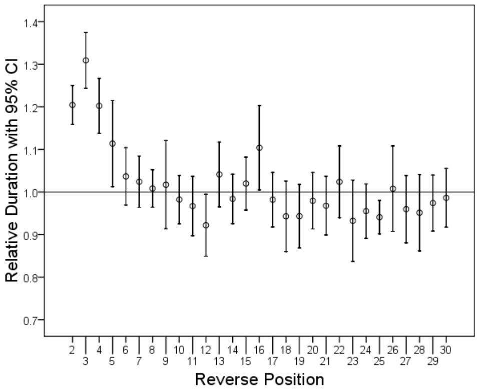

Figure 9: *Relative duration of long voiceless plosives*

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.