Minimum Sample Length for the Estimation of Long-term Speaking Rate

Pablo Arantes1, Anders Eriksson2, Verônica G. Lima1

1Languages and Linguistics Department, São Carlos Federal University, Brazil
2Department of Linguistics, Stockholm University, Sweden

pabloarantes@gmail.com, anders.eriksson@ling.su.se, vegomeslima@gmail.com

Abstract

In this study, we expand on previous experiments designed with the aim of determining the minimum length that an audio sample should have in order for the speaking rate derived from it to be representative of the sample as a whole. We compare two different approaches to establishing that the time series of the cumulative speaking rate calculated over the audio sample has reached stability. We also compare the effect on stabilization time of four other factors that may affect the way speaking rate is calculated. The results show that all factors tested have significant effects, although of limited practical concern. Overall, average stability time is 12.1 seconds, with the bulk of the distribution lying between 7.9 and 16.2 s.

Index Terms: speaking rate, speech rate, articulation rate, forensic phonetics

1. Introduction

In this paper we expand on a previous work [1] that tried to suggest some directions on how to give a principled answer to the question of determining what the minimum audio sample length for the estimation of long-term speaking rate should be so that the estimate derived from a part of the sample can be considered representative of the whole sample and maybe of other samples by the same speakers.

This question is relevant in a more theoretical sense because it broadens our understanding of basic phonetic properties of speaking rate. It also matters in a more practical sense in fields such as forensic phonetics, where forensic experts may want to use temporal information such as speaking rate as a parameter in speaker comparison tasks (see [2] for the discriminant power of articulation rate), but may be faced with speech samples that are limited in length. It is not uncommon that recordings in forensic casework are as short as 20 seconds (see [3]). Kendall [4] also points out that in research areas such as sociophonetics it is worth knowing if there is a real advantage in investing resources to collect large-scale corpora in terms of the patterns observed being more stable and representative of the speakers surveyed.

Discussion on the literature about the issue of minimum sample length seems to be limited to [4] and our previous study. Kendall cites one brief work [5] that estimates minimum sample length in about 5 minutes, although there is limited information on how the authors arrived at this estimate. The work developed by Kendall himself is presented here show that applying Kendall’s stabilization criterion with some modifications inspired by our own approach the estimates that we arrive at are similar to the ones we reported before and are much shorter, in the order of less than 30 seconds.

2. Material and methods

2.1. Speakers and speech material

The speech material analyzed here is the same as in [1]: a 110-word long text called “A Menina do Narizinho Arrebitado” by Brazilian writer Monteiro Lobato, read by 8 Brazilian Portuguese (BP) native speakers (3 female, 5 male) with ages ranging from 18 to about 30. This particular passage was chosen because it has the advantage of being relatively short and at the same time it contains at least one occurrence of all BP phonemes. The text was read first at a self-chosen normal level and then slow and fast relative to speaker’s normal with no help of external visual or auditory cues. Mean overall reading time is 33.5 s (min. 21 s, max. 54 s).

2.2. Linguistic units

Onset and offset location for each phone in the audio recordings were identified independently by the first and third authors and corrected when there was disagreement, following the criteria suggested in [6]. Phone boundaries were used as reference to further segment the recordings in syllables, VV units and words. Boundary locations for each unit type were stored in accompanying metadata files (TextGrid objects).

For the syllable segmentation, the authors referred to standard BP dictionaries whenever there was doubt about the syllable count or division for a particular word. VV units (or phonetic syllables) are syllable-sized units defined as all the segments uttered between two consecutive vowel onsets (see [7] for more details). For the word segmentation, the utterances were segmented in phonological rather than lexical words, following the rules proposed by [8] with modifications suggested by [9]. Fifty-seven phonological words were identified in the text passage analyzed.

Custom Praat scripts were used to extract duration information stored in the metadata files and do further processing.

2.3. Linguistic units counting

When analyzing speaking rate, one methodological decision that has to be made has to do with the criterion chosen to count the number of linguistic units in a given utterance. In most cases the decision comes down to choosing between a phonological or a phonetic criterion. In the first case one should count what is variously called grammatical, canonical or lexical units (be they phones, syllables etc) and in the second case one should count the number of units actually spoken as can be seen in the...
recorded waveform.

There seems to be no definitive consensus in the literature about what choice is best. For a more complete account of the pros and cons of each criterion, see [2, 10] and references therein. Because of this lack of consensus we decided to apply both criteria and see if a significant effect on stabilization time arises.

2.4. Rate type

The relevant literature points to two different types or kinds of speaking rate that result from the decision of what to make of (filled or unfilled) pauses and other disfluencies. When pause and other disfluencies are included in the duration interval used to measure speaking rate we have speech rate and when only the duration of fluent speech is considered we have articulation rate. Following our previous work [1], we investigate both speech rate and articulation rate.

2.5. Stabilization criteria

In a previous work [1] we explored a methodology to determine the minimum amount of speech data necessary to yield a stable measure of speaking rate. It consisted of obtaining the time series defined by the cumulative speaking rate (speech or articulation) calculated over an entire audio sample and submitting it to a statistical technique called change point analysis, that detects the point in time where a significant change in the underlying variance of the time series takes place (see [11] for technical details). A similar procedure has been successfully applied by [12, 13] to find stabilization points in time series of cumulative measures of central tendency of fundamental frequency. A package for the R statistical computing environment [14] named changepoint [15] was used for this analysis.

Cumulative speaking rate ($cSR$) is calculated from the first to the last linguistic unit of choice (phone, syllable etc) in a speech sample following formula 1, where $cSR_i$ is the cumulative speaking rate up to the $i^{th}$ unit and $dur_i$ is the sum of the duration of all units ranging from the the first one up to the $i^{th}$.

$$cSR_i = \sum_{j=1}^{i} dur_j$$ (1)

Figure 1 shows the time series of cumulative speaking rate generated for one speech sample from the present experiment. The dashed vertical red line indicates the stabilization point identified by the change point analysis at 27.6 s. Time series variance in the stretch that goes from the start up to and including the stabilization point is 16 times bigger than the variance in the section that goes from there up to the last unit.

Kendall [4] tackles the same issue but with a different criterion to establish stability. To do that, the author also analyzes time series defined by measures of speaking rate (in the work cited he only looks into articulation rate) calculated in a cumulative way, although he calculates it not on a unit by unit basis as we do here, but accumulates the duration of all units in what he calls a “phonetic utterance”, or speech stretches between pauses. More important here is the fact that the criterion he uses to establish stabilization in the time series is a variation threshold of $\pm 5\%$ relative to the global rate, i.e., the rate value derived from all data from a given sample. This value corresponds to the just noticeable difference (JND) for speech tempo derived experimentally in [16] (henceforth, “perceptual threshold”).

In the example shown in figure 1, the global speech rate, i.e., the last value in the time series, is 8.28 phones/s. The green box horizontal lines indicate the lower and upper boundaries defined by the threshold values of $\pm 5\%$ relative to that value and the dashed vertical blue line at 5.4 s indicates the latest point in the time series after which all subsequent variation is within the limits of the green box.

![Figure 1: Cumulative speech rate (phones/sec) along a 54 s long utterance spoken in slow rate by a male speaker. The phonological criteria was used to count the number of phones. Dashed red line at 27.6 s indicates the stabilization time returned by the change point analysis. Dashed blue line at 5.4 s indicates the stabilization time suggested by the perceptual threshold criteria.](image)

2.6. Error measure

The error measure $e_{st}$, as defined in formula 2, was created in order to estimate how much the rate value at the stabilization point ($r_{st}$) approaches the global rate ($r_g$), i.e. the value obtained considering all the units in the sample. The value returned by the formula is a percentage and can be positive ($r_{st}$ overestimates $r_g$) or negative ($r_{st}$ underestimates $r_g$).

$$e_{st} = \frac{r_{st} - r_g}{r_g} \cdot 100$$ (2)

2.7. Statistical analysis

The experimental design consisted of five independent variables (IV) with differing number of levels:

- **Criterion** to define the stabilization point (2 levels): change point analysis or perceptual threshold.
- **Linguistic Unit** (4 levels): phonetic syllable, phonetic syllable and word.
- **Rate Type** (2 levels): speech rate and articulation rate.
- **Rate Level** (3 levels): slow, normal and fast.
- **Transcription type** (2 levels): phonological, phonetic.

The dependent variables (DV) analyzed are: stabilization point of the speaking rate time series, measured in seconds from the start of the sample; error rate, defined in formula (2); number of linguistic units present in the interval between the start of the sample and the stabilization point; number of pauses present in the interval between the start of the sample and the stabilization point. The first DV is ratio-scaled, the second, interval-scaled, and the last two are ordinal.
Significance tests were applied in order to establish the existence of statistical effects on DVs due IVs. Homogeneity of variance tests were used to test the null hypothesis that there are no differences in variance among levels within IV. For the first two DVs, parametric tests such as Analysis of Variance (ANOVA) or $t$-tests were used to check for significant effects on mean values if the sample is homoscedastic or their nonparametric counterparts, Kruskal–Wallis [17] or Mann–Whitney–Wilcoxon tests if the sample is heteroscedastic. In the case of the last two DVs, because they are counts, we always used the nonparametric tests. When the IV being tested had more than two levels, pairwise $t$-tests or Mann–Whitney–Wilcoxon tests with Holm-corrected $p$-values [18] were used to check for differences among levels. An $\alpha$ level of 5% was adopted for all tests. All analyses were run on R [14].

3. Results

3.1. Effect of stabilization criterion

Figure 2 shows boxplots of the distribution of the four DVs as a function of stabilization criterion. For this analysis, all the other IVs are collapsed. Since we are interested in the different results of the two criteria on the same audio samples, the paired version of the statistical tests are used. Statistical results show that stabilization criteria have a significant effect on all DVs. Mean values are given inside parentheses, the first for the change point criterion and the second for the threshold one. For the last two DVs, median values are also present, in the form mean/median: stabilization time (11.42 s, 13.59 s)

Perceptual threshold. Mean stabilization time: phone 10.47 (6.07), syllable 12.33 (5.44), VV 10.59 (4.11), word 20.97 (5.04). The sample is heteroscedastic $[\chi^2(3) = 156.8 \ p < 0.001]$. There is a significant effect of unit type on stabilization time $[\chi^2(3) = 26.1 \ p < 0.001]$. Pairwise comparisons show that VV is the linguistic unit that yields the shortest stabilization times and there is no difference among the other three types, although the observed differences are in the order of 3 s at most.

3.2. Effect of linguistic unit

Here we look into the effect of the four different linguistic units on mean stabilization time. Due to space restrictions we will not present results concerning the other DVs. The analysis is done in two parts, one deals with the data generated by the change point criterion and the other with the perceptual threshold criterion. Figure 3 shows the corresponding boxplots.

Change point. Mean stabilization time (in seconds, standard deviation in parentheses) for the four different linguistic units used to compute speaking rates are: phone 12.64 (6.79), syllable 11.15 (4.86), VV 9.44 (3.85), word 12.45 (3.23). The sample is heteroscedastic $[\chi^2(3) = 19.17 \ p < 0.001]$, indicating that the linguistic units used to group the segments do not have all the same variance. There is a significant effect of unit type on stabilization time $[\chi^2(3) = 26.1 \ p < 0.001]$. Pairwise comparisons show that VV is the linguistic unit that yields the shortest stabilization times and there is no difference among the other three types, although the observed differences are in the order of 3 s at most.

3.3. Effect of rate type

In this analysis we compare mean stabilization time as a function of rate type. The statistics tests are run separately for each stabilization criterion. All other IVs are collapsed. Figure 4 shows the corresponding boxplots.

Change point. Mean stabilization time (in seconds, standard deviation in parentheses) for the two types of rate: articulation 10.76 (5.19), speech 12.08 (4.76). The sample is homoscedastic $[\chi^2(1) = 0.3 \ ns]$ and the paired $t$-test shows a
significant effect \( t(191) = -3.6, p < 0.01 \).

**Perceptual threshold.** Mean stabilization time (in seconds, standard deviation in parentheses) for the the two types of rate: articulation 11.83 (6.72), speech 15.35 (6.34). The sample is homoscedastic \( \chi^2(1) = 2.1, n.s. \) and there is a significant effect of rate type \( t(191) = -11, p < 0.001 \).

There is a significant effect of rate type, with articulation rate having the fastest stabilization times. It is possible to hypothesize that the slightly longer stabilization times observed in the speech rate condition is due the inclusion of pause duration, that adds to the variability of the segments duration.

![Boxplots of stabilization time as a function of the four linguistic units, the two stabilization criteria and the two rate types.](image)

3.4. Effect of rate level

In this analysis we compare mean stabilization time as a function of rate level. The statistics tests are run separately for each stabilization criterion. All other IVs are collapsed. Figure 5 shows the corresponding boxplots.

**Change point.** Mean stabilization time (in seconds, standard deviation in parentheses) for the three rate levels: fast 9.92 (3.99), normal 11.56 (5.22), slow 12.77 (5.35). The sample is homoscedastic \( \chi^2(2) = 3.8, n.s. \) and the ANOVA shows a significant effect \( F(2, 381) = 11, p < 0.001 \). Pairwise comparisons indicate that all levels have significantly different means.

**Perceptual threshold.** Mean stabilization time (in seconds, standard deviation in parentheses) for the three rate levels: fast 11.79 (5.52), normal 13.82 (7.12), slow 15.17 (7.13). The sample is heteroscedastic \( \chi^2(1) = 9.4, p < 0.01 \) and there is a significant effect of rate level \( \chi^2(2) = 13.8, p < 0.01 \). Pairwise comparisons indicate that the fast level has a significantly lower mean stabilization time compared with the other two levels.

Overall, it seems that faster rates tend to stabilize faster regardless of the stabilization criterion considered. The slowing down of rate tends to induce a slight increase in mean stabilization time, accompanied also by an increase in its variability.

3.5. Effect of transcription type

In this analysis we compare mean stabilization time as a function of transcription type. The statistics tests are run separately for each stabilization criterion. All other IVs are collapsed. The paired version of the statistical tests are used.

![Boxplots of stabilization time as a function of the four linguistic units, the two stabilization criteria and the three rate levels.](image)

**Change point.** Mean stabilization time (in seconds, standard deviation in parentheses) for the two types of transcription are: phonetic 11.07 (5.68), phonological 11.08 (5.23). The sample is homoscedastic \( \chi^2(1) = 0.9, n.s. \) and the paired t-test does not yield a significant effect \( t(143) = -0.02, n.s. \).

**Perceptual threshold.** Mean stabilization time (in seconds, standard deviation in parentheses) for the two types of transcription are: phonetic 11.13 (5.49), phonological 11.13 (5.16). The sample is homoscedastic \( \chi^2(1) = 0.9, n.s. \) and the paired t-test does not yield a significant effect \( t(143) = -0.03, n.s. \).

4. Discussions and conclusions

The most notable finding brought by the present experiment is that minimum sample length (or what we are calling stabilization time) for both speech and articulation rates are much shorter than the few previous works suggest. While previous estimates range from 5 minutes [5] to as long as 9 minutes [4], we were able to show that both types of speaking rate can reach stabilization much sooner, consistently in less than half a minute, even using the same stabilization criteria employed by [4].

The results show that all the factors that we chose to control in the experiment have an effect on stabilization time, with the exception of the criterion used to count the number of linguistic units (transcription type, as we called it). Even though the effects on stabilization time are statistically significant, they are small enough not to be of practical concern. In the worst case, the effect is in the order of less than a dozen seconds (see section 3.2).

Of the elicited effects, the one that is more concerning is the one that shows that the error rates yielded by the use of the change point criterion to establish stabilization, around 6.8%, is slightly above the just noticeable difference for speaking tempo, 5% according to [16]. That may be of concern, because it could imply that the speaking rate at the stabilization point would be perceived as different from the global rate for the same utterance.

5. Acknowledgements

This work has been supported by STINT (Sweden) grant IB2015-6488. The third author was supported by FAPESP (Brazil) grant 2016/12646-0 between November 2016 and December 2017.
6. References


