Speech rate estimation: how long should the utterance be?

Pablo Arantes

Universidade Federal de São Carlos, Brazil

pabloarantes@gmail.com

Abstract

The aim of the present work is to investigate how long should a speech sample be in order for the speaking rate derived from it can be considered representative of the whole utterance from which the sample has been taken. Eight Brazilian Portuguese speakers read a 144-word text in three rate levels, slow, normal and fast. Speech rate was measured cumulatively as the number of phonetic syllables (segments between consecutive vowel onsets) per second from the first to the last syllable. Change point analysis was used to determine the influence of rate level on the amount of time necessary for the cumulative estimate of speech and articulation rates to stabilize around the rate yielded by the whole utterance. Mean stabilization latencies are 8.9 seconds. Stabilization intervals take up a median number of 41 syllables. No effect of rate level was found on both stabilization time and number of syllables in the stabilization interval. Mean deviation between the global rate and the rate value at stabilization point is 7.8%.

Index Terms: prosody, speaking rate, speech rate, articulation

rate, forensic phonetics

1. Introduction

Speech rate is a variable that reflects how fast or slow speech is rendered in a given utterance. It is measured as the rate of linguistic units uttered per time unit [1]. Different linguistic units can be chosen as reference, such as the word, syllable or phone, resulting in a more coarse- or fine-grained measure. Common choices of time unit are the second or minute. When speech rate is estimated for a whole utterance, it is usually referred to as global speech rate. Pfitzinger [4] defines global speech rate as the measure obtained by "dividing the number of segments by the sum of their durations for a complete utterance".

Even though speech rate has been extensively studied and has proven a useful parameter for a lot of different purposes, ranging from linguistic analysis to speech technology, no discussion can be found on the literature about how long a speech sample has to be in order for the resulting global speech rate estimate can be seen as representative of the long-term behavior of a speaker. Assessments of minimum sample length can be useful in the planning of large databases of speaking rate [2, 3] and in the context of forensic casework, where it can be one of the parameters used in speaker comparison [1]. The aim of this paper is to suggest some directions on how to pursue an answer to the question of what is the minimum sample size necessary to estimate speech rate.

2. Materials and methods

2.1. Materials and measurements

Eight Brazilian Portuguese native speakers (5 males, aged between 18 and early 30s) read the 144-word long Lobato passage ("A Menina do Narizinho Arrebitado"), a phonetically rich text containing all BP phonemes. The duration of the samples varied from 23.69 to 53.98 seconds. The sound files were segmented and labelled by an expert phonetician. Consecutive vowel onsets were identified and defined vowel-to-vowel (VV) units. Vowel onset locations were stored in accompanying metadata files (TextGrid objects) for further processing. A custom Praat script was used to extract VV interval durations.

For the purposes of this experiment, speech rate was defined as the number of vowel-to-vowel (VV) units per second. VV intervals are syllable-sized units defined as all the segments uttered between two consecutive vowel onsets. See [6] for the rationale on the usefulness of VV grouping to reveal prosodic structure.

To determine how speech rate changed throughout a given speech sample, it was calculated cumulatively from the first to the last VV unit present in each sample. The cumulative speech rate up to the i^{th} VV unit, cSR_i, can be obtained by dividing the index of the VV unit, i, by the sum of the durations of the VV units from the first up to the i^{th}, as expressed in the formula below.

$$cSR_i = \frac{i}{\sum_{j=1}^{i} \text{dur}_j}$$

2.2. Statistical analysis

The time series defined by the consecutive values of cSR were analyzed using a statistical technique called changepoint analysis [8], implemented as a package for the R statistical computing environment. For the purposes of our study, we used a function that finds the point in time that separates the time series in two parts having significantly different underlying variance values. A parameter was passed to the function instructing it not to assume that the values follow a normal distribution, since a visual inspection of a number of histograms showed that the most of the samples are highly skewed. In the samples analyzed here, the variance always decreases over time, with a median variance reduction factor of 32 (minimum of 6.8 and maximum of 115.7). We call the point identified by the analysis the stabilization point because after it the speech rate estimate tends to stabilize around a much narrower range of values, approaching what could be called its long-term value. The analysis was able to identify a stabilization point in all the time series in the present sample.

A typical cumulative speech rate time series is shown in Figure 1.

Rate level was the independent variable controlled in the experiment. The passage was read in three rate levels by all speakers: self-selected normal/habitual, slow and fast. Figure 2 shows mean VV duration per speaker as a function of rate level. One-way ANOVA tests conducted separately for each speaker were used to compare mean VV duration among the three rate levels. When there was a significant main effect, paired *t*-tests with Holm-corrected *p*-values were performed to test for differences among levels. Results show that there is no difference in mean VV duration among the three rate levels for one male speaker. Excluding this speaker, the fast-slow comparison always yields a significant difference, the normal-fast comparison is significant for one speaker and the normal-slow comparison is significant for two speakers.

2.3. Error measure

In order to estimate how well the rate value at the stabilization point (r_{st}) reflects the global rate (r_g), i.e. the value obtained considering all the VV units in the sample, an error measure was defined as follows:

$$\frac{r_{st} - r_g}{r_g} \cdot 100$$

In the example shown in Figure 1, the error is around 15% (r_g is 7.03 VV/s and r_{st} is 6.08 VV/s).

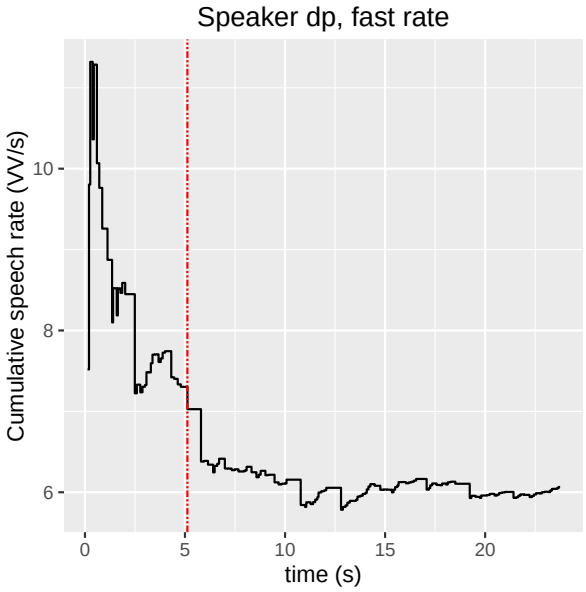

Figure 1: *Cumulative speech rate along a complete sample. Dashed vertical line indicates stabilization point location (5.12 s). Speech rate variance after the stabilization point is almost 70 times smaller than before it.*

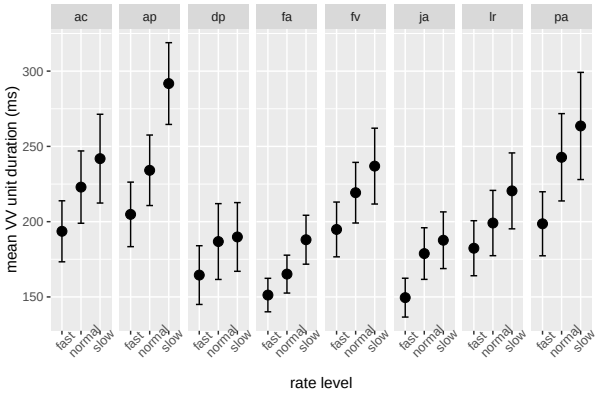

Figure 2: *Mean VV duration as a function of rate level. Vertical panels indicate different speakers. Vertical bars indicate 95% confidence intervals around the mean.*

3. Results

3.1. Stabilization time

Figure 3 presents the breakdown of stabilization times by rate level (slow, normal and fast). Mean stabilization time and standard deviation (shown in parentheses) is 8.9 s (3.19) for all levels collapsed – fast 7.9 (3.1), normal 8 (2.4) and slow 10.8 (3.4). A one-way ANOVA analysis carried out to test for differences in mean stabilization time among the three rate levels failed to yield a significant result [$F(2, 21) = 2.3$, *ns*].

Simple linear regression analysis was used to predict stabilization time based on mean duration of VV units. A significant regression equations was found [$F(1, 22) = 18.38$, $p < 0.001$], with an R^2 of 0.45. Stabilization time is equal to -3.6 + 0.06 · (mean VV duration), when VV duration in measured in milliseconds. Stabilization time increases 6.1 seconds for each 100 millisecond of mean VV unit duration. Figure 2 suggests that rate levels have an effect of VV duration and, as the regression analysis indicates, that the longer the mean VV duration, the longer the stabilization interval should be. The lack of effect in the ANOVA analysis may be due to the relatively small sample size and small effect size. Replications of the study with a bigger sample size may yield significant differences among levels.

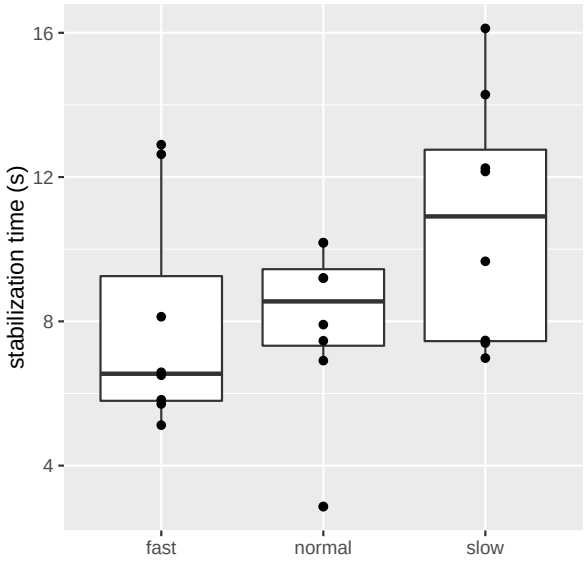

Figure 3: *Stabilization time as a function of speech rate level. There is no difference among levels.*

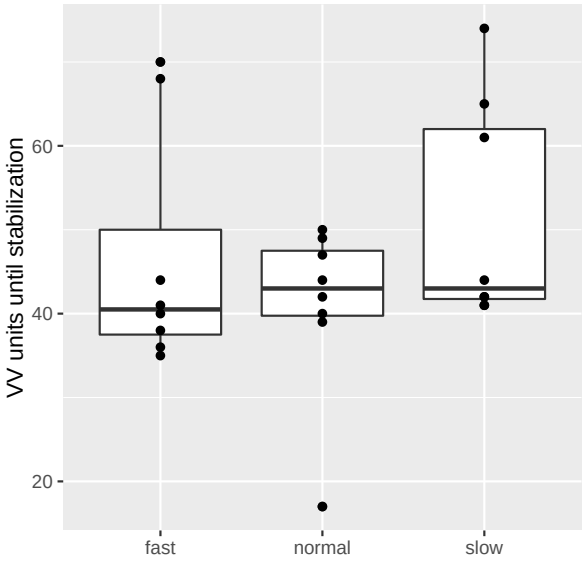

Figure 4: *Number of VV units in the stabilization interval as a function of rate level. There is no difference among levels.*

3.2. Number of VV units in the stabilization interval

Figure 3 presents the breakdown of the number of VV units encompassed by the stabilization interval by rate level. Mean and standard deviation (shown in parentheses) number of VV units is 46.2 (13) units - fast 46.5 (14.2), normal 41 (10.5), slow 51.2 (13.3). A one-way ANOVA analysis carried out to test for differences in mean number of VV units among the three rate levels failed to yield a significant result [$F(2, 21) = 1.3$, ns].

3.3. Estimation error

Figure 5 presents the breakdown of the error measure by rate level. Mean estimation error and standard deviation (shown in parentheses) are 7.6% (3.1). A one-way ANOVA analysis carried out to test for differences in estimation error among the three rate levels failed to yield a significant result [$F(2, 21) = 0.8$, *ns*]. All rate values obtained at the stabilization points overestimate the global rate. Rate values at the stabilization point respect the same level ordering defined by the global rate values for 7 out of 8 in the sample.

If mean rate values are used to order the speakers from slowest to fastest (rate levels collapsed), comparing the ordering obtained when using global rate values and the values estimated at stabilization points shows that only a pair of adjacent speaker swap places.

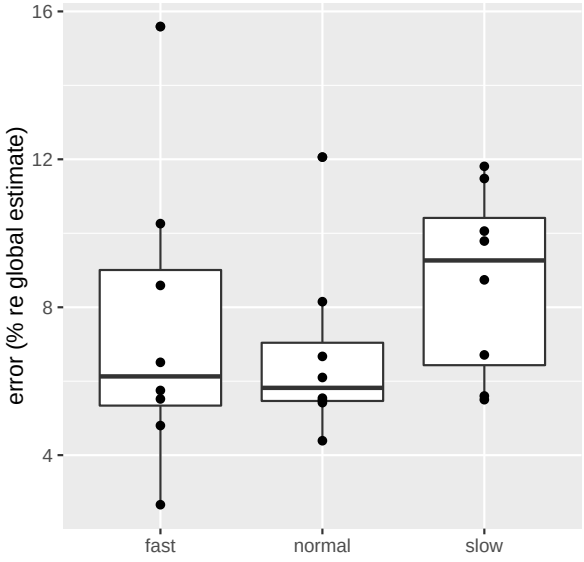

Figure 5: *Estimation error as a function of rate level. There is no difference among levels.*

4. Conclusions

To the extent of our knowledge, there has been no systematic investigation on how to determine the minimum speech sample length necessary to derive a reliable estimate of global rate. One of the contributions of the present paper is to outline an objective method to approach this subject. Although there were no main effects related to the independent variable controlled, that may be due to the relatively small sample.

Nevertheless, the results obtained in this pilot study serve as a guideline for further research.

Overall, the results are encouraging. The statistical technique employed provides an objective way of estimating minimum sample length for determining speech rate. The results obtained may be used as reference values for future work and by forensic experts in their casework. The methodology developed here yields reasonably low error rates and the speech rate values obtained at stabilization points roughly preserve the same speaker ranking obtained when using the values estimated by the whole samples.

In follow-up studies, stabilization times for word and phone rate could be investigated, as well as independent variables other than rate level, such rate type (articulation rate vs. speech rate) and speaking style (spontaneous vs. read speech). It also seems interesting to investigate within-speaker and between-language variability of speaking rate stabilization points.

5. References

[1] Künzel, H. "Some general phonetic and forensic aspects of speaking tempo", Forensic Linguistics, 4(1), 48–83, 1997.

[2] Jessen, M. "Forensic reference data on articulation rate in German", Science and Justice, 47, 50–67, 2007.

[3] Cao, H. and Wang, Y. "A forensic aspect of articulation rate variation in chinese", Proceedings of the XVIIth ICPhS, 396–399, Hong Kong, 2011.

[4] Pfitzinger, H. R. "Two approaches to speech rate estimation", Proceedings of the 6th Australian Int. Conf. on Speech Science and Technology (SST 96), 421-426, Adelaide, 1996.

[5] Pfitzinger, H. R. "Local speech rate as a combination of syllable and phone rate", Proceedings of the 5th ICSLP, vol. 3, 1087–1090, Sydney, 1998.

[6] Barbosa, P. A. From syntax to acoustic duration: a dynamical model of speech rhythm production, Speech Communication, 49, 725–742, 2007.

[7] Arantes, P. and Eriksson, A. "Temporal stability of long-term measures of fundamental frequency", Proceedings of Speech Prosody 7, 1149–1152, 2014.

[8] Rebecca Killick and Idris Eckley. "changepoint: An R package for changepoint alysis". R package version 1.1.5. http://CRAN.Rproject.org/package=changepoint, 2013.