TOWARDS A METHODOLOGY TO ESTIMATE MINIMUM SAMPLE LENGTH FOR SPEAKING RATE

Pablo ARANTES[1]
Verônica Gomes LIMA[2]

Abstract: The aim of the present work is to investigate how long a speech sample should be so that the speaking rate derived from it 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. Speaking rate was measured cumulatively as the number of phonetic syllables per second from the first to the last syllable in the sample. Two types of rates were measured, speech rate and articulation rate. Change point analysis was used to determine the influence of rate type and 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 9.2 seconds for speech rate and 8.7 s for articulation rate. The slow rate tends to stabilize later than fast and normal rates for both types of rate. Stabilization intervals take up a median number of 41 (speech rate) and 59 (articulation rate) syllables. Mean deviations between the global rate and the rate value at stabilization point are 7.8% (speech rate) and 4.2% (articulation rate). The stabilization times estimated here can be useful in sociophonetic research, forensic and clinical phonetics.

Keywords: Prosody. Speaking rate. Speech rate. Articulation rate. Forensic phonetics.

[1] UFSCar – Universidade Federal de São Carlos – Departamento de Letras. São Carlos – São Paulo – Brasil. 13565-905 – pabloarantes@gmail.com
[2] UFSCar – Universidade Federal de São Carlos – Departamento de Letras – Bacharelado em Linguística. São Carlos – São Paulo – Brasil. 13565-905 – vegomeslima@gmail.com

http://dx.doi.org/10.21165/gel.v14i2.1810

1. Introduction

Broadly considered, speaking rate is a temporal parameter that reflects how fast or slow speech is rendered in a given utterance. Technically, it can be defined as the rate of linguistic units uttered per time unit (KÜNZEL, 1997). Different linguistic units can be chosen as reference (e.g. word, syllable or phone), resulting in a more coarse- or fine-grained measure. Depending on the window of measurement, speaking rate will tend to reflect local changes that may function as a correlate of prominence (BARBOSA, 2007; PFITZINGER, 1998) or serve as a global estimate of how slow or fast whole utterances are spoken. Pfitzinger (1996) defines global rate as the measure obtained by "dividing the number of segments by the sum of their durations for a complete utterance". Also, the way pauses are handled defines two types of speaking rate. If pauses are included as part of the linguistic unit duration, the measure is known as speech rate. If pauses are not included, the resulting measure is called articulation rate. In this context, speaking rate is a superordinate term.

Studies reported in (KÜNZEL, 1997) suggest that speaking rate can be useful in speaker comparison tasks. There seems to be no discussion on the literature, though, on how long the speech sample should be so that the resulting global speaking rate can be said to be representative of the long-term behavior of a speaker. From the point of view of voice comparison methods, both Jessen (2008) and Gfroerer (2003) point out that limited quantity of speech material is one of the problems affecting forensic casework. The rationale is that short recordings may not contain enough data to be representative of the broad range of speech patterns of a speaker. Jessen (2008) also points out that there is no fixed lower boundary under which voice comparison is made impossible and states that "at least something like eight seconds of speech from the anonymous speaker and at least about double that time for the suspect is recommended" (p. 16), although the author provides no technical justification for the figures mentioned. Gfroerer (2003) mentions 20 seconds as being a typical duration encountered in forensic casework. Estimates of minimum sample length are also important in the planning of large multi-speaker databases of speech rate population values such as those described by (CAO; WANG, 2011; JESSEN, 2007; SILVA, 2016) that are of paramount importance in speaker comparison scenarios.

The aim of this paper is to suggest some directions on how to give a principled answer to the question of how to determine minimum sample size for speaking rate estimation.

2. Materials and methods

2.1 Speech materials and variables

Eight Brazilian Portuguese (BP) native speakers (3 female, 5 male, with ages ranging from 18 to 30, all college students) read the same text, the 144-word long Lobato passage, "A Menina do Narizinho Arrebitado", a phonetically rich text containing all BP phonemes. Mean reading time is 33.5 seconds (minimum of 21 s and maximum of 54 s).

Speaking rate was measured as the rate of vowel-to-vowel (VV) units per second. VV units are syllable-sized units defined as all the segments uttered between two consecutive vowel onsets (see BARBOSA, 2007) for the rationale on the usefulness of VV grouping to speech production and perception). Vowel onsets in the audio recordings were semi-automatically identified with the help of a Praat script and their positions were then checked by an expert phonetician and hand-corrected as needed. Vowel onset locations were stored in accompanying metadata files (*TextGrid* objects in Praat) for further processing. Custom Praat (BOERSMA, 2001) and R (TEAM, 2016) scripts were used to extract VV interval durations and do further processing needed.

Two independent variables were controlled in the experiment: rate type and rate level. The rate type variable was measured in two ways (CRYSTAL; HOUSE, 1990): speech rate (silent pauses within VV units are computed as part of the unit's duration) and articulation rate (silent pauses are not included as part of VV duration).

F-tests were used to compare the variances of the two rate types separately for each speaker and yielded significant results in every case. Welch two sample *t*-tests were then used to compare the means of the two rate types and turned significant results for each speaker (an alpha level of 5% was adopted for all tests). The results are in line with what the literature suggests: mean speech rate values are slightly higher and variable than those for articulation rate.

Table 1: Articulation rate. Statistical comparisons among rate levels. Main effect and paired comparisons. Six speakers show significant differences between at least two levels

Speaker	Number of VV units		ANOVA	Multiple comparisons
AC	Fast	180	F(2, 541) = 2.7, $p < 0.1$	Normal-Fast: ns
	Normal	182		Normal-Slow: ns
	Slow	182		Fast-Slow: $p < 0.1$
AP	Fast	182	F(2, 569) = 34.8, $p < 0.001$	Normal-Fast: $p < 0.001$
	Normal	199		Normal-Slow: $p < 0.001$
	Slow	191		Fast-Slow: $p < 0.001$
DP	Fast	144	F(2, 482) = 2.7, $p < 0.1$	Normal-Fast: ns
	Normal	166		Normal-Slow: ns
	Slow	175		Fast-Slow: $p < 0.1$
FA	Fast	170	F(2, 532) = 9.3, $p < 0.0001$	Normal-Fast: ns
	Normal	186		Normal-Slow: $p < 0.01$
	Slow	179		Fast-Slow: $p < 0.001$
FV	Fast	178	F(2, 544) = 6.9, $p < 0.001$	Normal-Fast: $p < 0.05$
	Normal	185		Normal-Slow: ns
	Slow	184		Fast-Slow: $p < 0.001$
JA	Fast	164	F(2, 513) = 9.3, $p < 0.0001$	Normal-Fast: $p < 0.01$
	Normal	179		Normal-Slow: ns
	Slow	173		Fast-Slow: $p < 0.001$
LR	Fast	166	F(2, 507) = 3.5, $p < 0.05$	Normal-Fast: ns
	Normal	170		Normal-Slow: ns
	Slow	174		Fast-Slow: $p < 0.05$
PA	Fast	154	F(2, 485) = 5, $p < 0.01$	Normal-Fast: $p < 0.01$
	Normal	166		Normal-Slow: ns
	Slow	168		Fast-Slow: $p < 0.01$

Both rate types were measured cumulatively from the first to the last VV unit in each speech sample following formula 1, where cSR_i is the cumulative speaking rate up to the i^{th} VV unit in each speech sample. The consecutive values produced by the formula

constitute a time-series that starts in the first VV unit and at each step the next VV unit is added up to the last one in the sample.

$$cSR_i = \frac{i}{\sum_{j=1}^{i} \mathrm{dur}_j} \quad (1)$$

The passage was read at three rate levels by all speakers: self-selected normal/habitual, slow and fast. To elicit the slow and fast rates, speakers were verbally instructed to speak slower and faster than their self-chosen normal rate. Figure 1 shows mean VV duration for the articulation rate condition. The reciprocal of the mean VV value corresponds to the mean articulation rate. Figure 2 shows the same information for speech rate. One-way ANOVA tests conducted for each speaker separately 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 check for differences among rate levels (an alpha level of 5% was adopted for all tests). For articulation rate (see table 1 for detailed results), there is no difference in mean VV duration among rate levels for two speakers (AC and DP). Excluding these two, the fast-slow pair is always significant, the normal-fast pair is significant for four speakers and the normal-slow pair is significant for two speakers. For speech rate (see table 2 for detailed results), there is no difference in mean VV duration among rate levels for one speaker (DP). Excluding this speaker, the fast-slow pair is always significant, the normal-fast pair is significant for one speaker and the normal-slow pair is significant for two speakers. This analysis shows that the procedure successfully induced rate level change for both types of rate. The two-way normal-slow contrast is the most reliable difference for most speakers.

Table 2: Speech rate: Main effect and paired comparisons among rate levels. Seven speakers show significant differences between at least two levels

Speaker	Number of VV units		ANOVA	Multiple comparisons
AC	Fast	180	F(2, 541) = 3.6, $p < 0.05$	Normal-Fast: ns
	Normal	182		Normal-Slow: ns
	Slow	182		Fast-Slow: $p < 0.05$
AP	Fast	178	F(2, 559) = 12.5, $p < 0.001$	Normal-Fast: $p < 0.1$
	Normal	199		Normal-Slow: $p < 0.01$
	Slow	185		Fast-Slow: $p < 0.001$
DP	Fast	144	F(2, 481) = 1.3, ns	Normal-Fast: ns
	Normal	166		Normal-Slow: ns
	Slow	175		Fast-Slow: ns
FA	Fast	171	F(2, 532) = 6.9, $p < 0.001$	Normal-Fast: ns
	Normal	186		Normal-Slow: $p < 0.05$
	Slow	178		Fast-Slow: $p < 0.001$
FV	Fast	178	F(2, 544) = 3.7, $p < 0.05$	Normal-Fast: ns
	Normal	185		Normal-Slow: ns
	Slow	184		Fast-Slow: $p < 0.05$
JA	Fast	164	F(2, 513) = 5.4, $p < 0.01$	Normal-Fast: $p < 0.05$
	Normal	179		Normal-Slow: ns
	Slow	173		Fast-Slow: $p < 0.01$
LR	Fast	166	F(2, 507) = 2.9, $p < 0.1$	Normal-Fast: ns
	Normal	170		Normal-Slow: ns
	Slow	174		Fast-Slow: $p < 0.05$
PA	Fast	154	F(2, 485) = 4.7, $p < 0.01$	Normal-Fast: $p < 0.1$
	Normal	166		Normal-Slow: ns
	Slow	168		Fast-Slow: $p < 0.01$

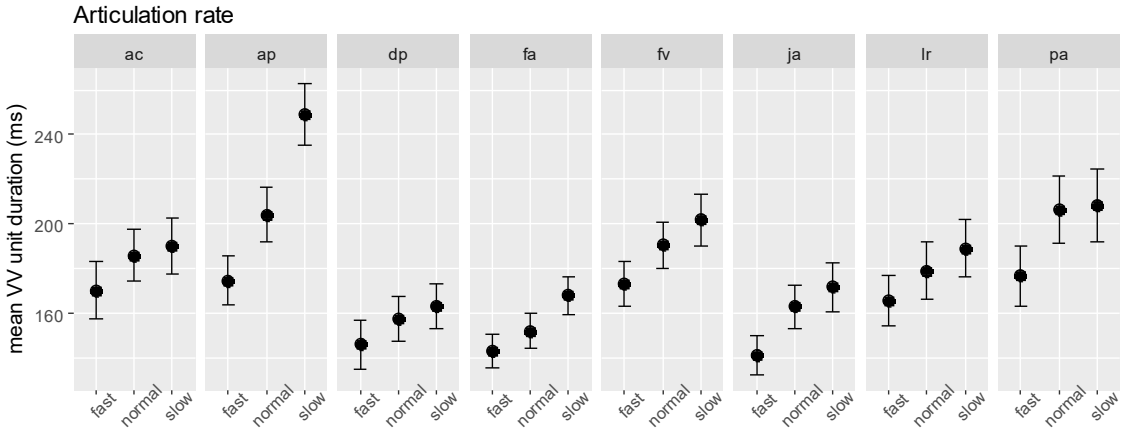

Figure 1: Mean VV duration for each speaker, broken by rate level. Whiskers indicate 95% confidence intervals around the mean

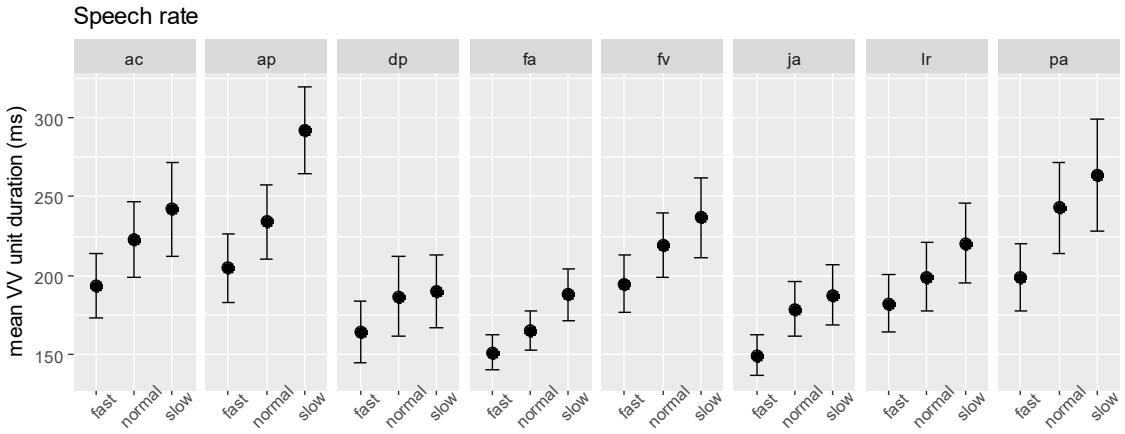

Figure 2: Mean VV duration for each speaker, broken by rate level. Whiskers indicate 95% confidence intervals around the mean

2.2 Statistical analysis

The time series defined by the cumulative rate estimates were submitted to a statistical technique called changepoint analysis (KILLICK; ECKLEY, 2014), which was implemented as a package for the R statistical computing environment named *changepoint* (KILLICK; HAYNES; ECKLEY, 2016). It detects the time point where a significant change in the underlying variance of the time series takes place. A parameter was passed to the relevant function instructing it not to assume that the values in the time

series follow a normal distribution, since a visual inspection of several histograms of cumulative values of speaking rate showed that the most of the samples are highly skewed. In the case of the data analyzed here, the variance always decreases over time, i.e., when more VV units are added to the speaking rate estimate. We call the time point identified by the analysis as the stabilization point, because after it the variability of the rate estimate is not greatly affected by the addition of more phonetic material and converges towards the global estimate. A similar procedure was successfully applied by (ARANTES; ERIKSSON, 2014), to find stabilization points in time series of cumulative measures of central tendency of fundamental frequency.

As an example of the procedure in action, figure 3 shows a time series of cumulative speech rate produced by one speaker in fast rate and the stabilization point location.

In the corpus analyzed here, variance always decreases after the stabilization point. Mean reduction factors (minimum and maximum values in parentheses) are 46 times (15, 182) for articulation rate and 40 times (7, 116) for speech rate.

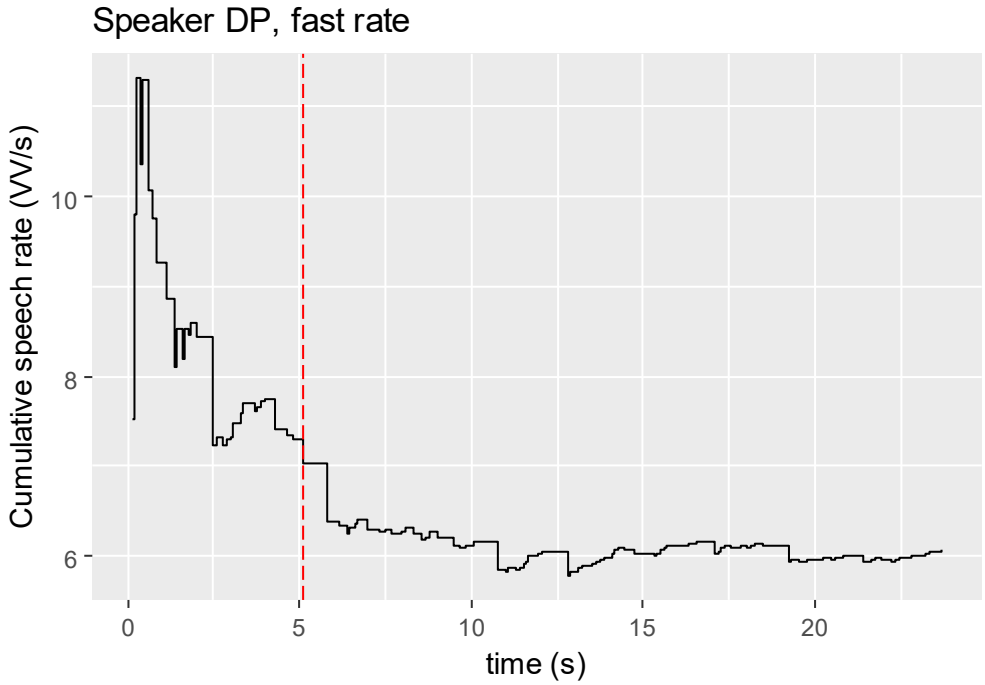

Figure 3: Cumulative speech rate along a complete reading. Dashed vertical line indicates the stabilization point location (5.12 s). Speech rate variance after the stabilization point is almost 70 times smaller than before it

2.3 Error Measure

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 shown in formula 2:

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

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

3. Results and discussion

3.1 Stabilization time

Figure 4 presents the breakdown of stabilization times of rate type (articulation and speech) and level (slow, normal and fast). Mean stabilization times and standard deviation (shown in parentheses) are 9.35 s (3.21) for articulation rate and 8.9 s (3.19) for speech rate. A paired *t*-test was used to compare the two speaking rate types and no significant difference was found [$t(23) = 1$, ns]. Mean stabilization time for the three levels of rate level are 7.81 s (2.88) for the slow, 8.41 s (2.72) for the normal and 11.15 s (3) for the fast rate. A one-way ANOVA performed on the merged sample of both types of speaking rate to test for differences in mean stabilization time yielded a significant result [$F(2, 45) = 6.2$, $p < 0.01$]. Pairwise *t*-tests with Holm-corrected *p*-values indicate that mean stabilization time for the slow rate is longer than the fast ($p < 0.01$) and normal ($p < 0.05$) rates.

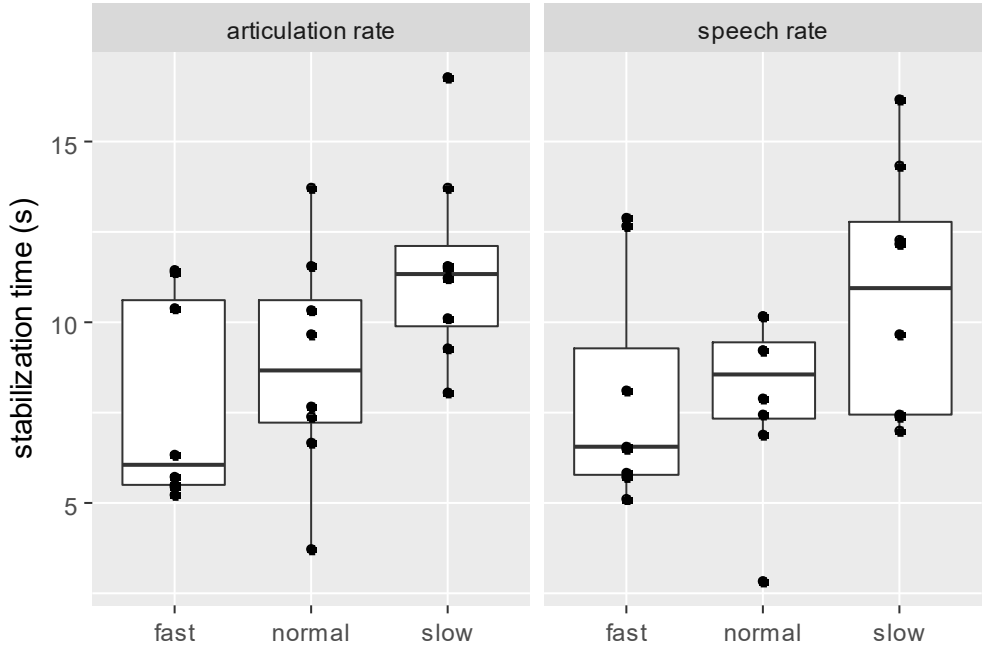

Figure 1: Stabilization times broken by rate type and rate level. Slow rates have longer stabilization latencies

Simple linear regression analysis was used to predict stabilization time based on mean duration of VV units per speech sample. Significant regression equations were found for articulation rate, speech rate and for the two samples merged. Relevant parameters of the three regression models are listed in table 3. None of the intercepts differ significantly from zero. Stabilization time increases 8.9 seconds (articulation rate), 6.1 s (speech rate) or 5 s (both types) for each 100 milliseconds of mean VV unit duration.

Table 3: Linear model parameters of stabilization time (seconds) as a function of mean VV unit duration (milliseconds). * $p < 0.001$, • $p < 0.1$.

Rate type	**Intercept**	**Coefficient**	R^2
Articulation	-6.4 •	0.089 *	0.47 *
Speech	-3.7 *ns*	0.061 *	0.45 *
Both	-1.5 *ns*	0.05 *	0.34 *

3.2 Number of VV units in the stabilization interval

Figure 5 presents the breakdown of the number of VV units encompassed by the stabilization interval by rate type (articulation and speech) and rate level (slow, normal and fast). Mean and standard deviation (shown in parentheses) number of VV units are 54 (14.4) units for articulation rate and 46.2 (13) for speech rate. A paired *t*-test was used to compare the two speaking rate types and a significant difference was found [$t(23) = 3.86$, $p < 0.001$]. One-way ANOVA tests were then applied separately to the samples of the two rate types to test for differences among rate levels and revealed no significant differences: articulation rate [$F(2, 21) = 2.06$, ns]; speech rate [$F(2, 21) = 1.3$, ns].

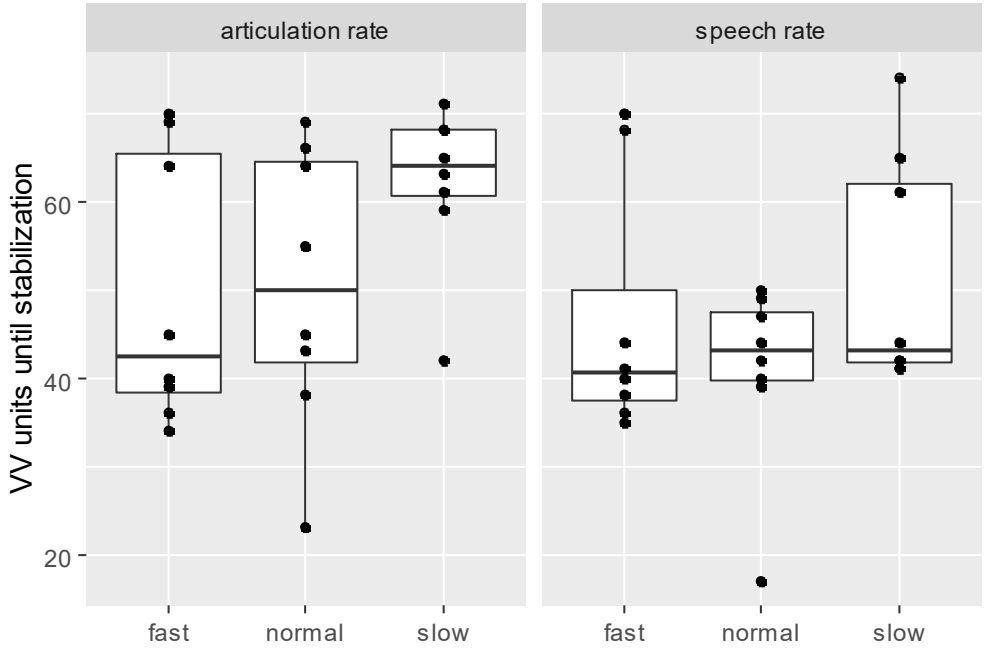

Figure 2: Number of VV units in the stabilization interval broken by rate type (articulation or speech) and rate level (fast, normal, slow). Articulation rate intervals comprise more VV units

3.3 Error measure

Figure 6 presents the breakdown of the error measure by rate type (articulation and speech) and rate level (slow, normal and fast). Mean estimation error and standard deviation (shown in parentheses) are 4.2% (2.7) for articulation rate and 7.6% (3.1) for

speech rate. A paired *t*-test revealed a significant difference between the means [$t(23) = -6.46$, $p < 0.001$]. One-way ANOVAs were then applied separately to the samples of the two rate types to test for differences among rate levels and revealed no significant differences: articulation rate [$F(2, 21) = 1.72$, ns]; speech rate [$F(2, 21) = 0.854$, ns].

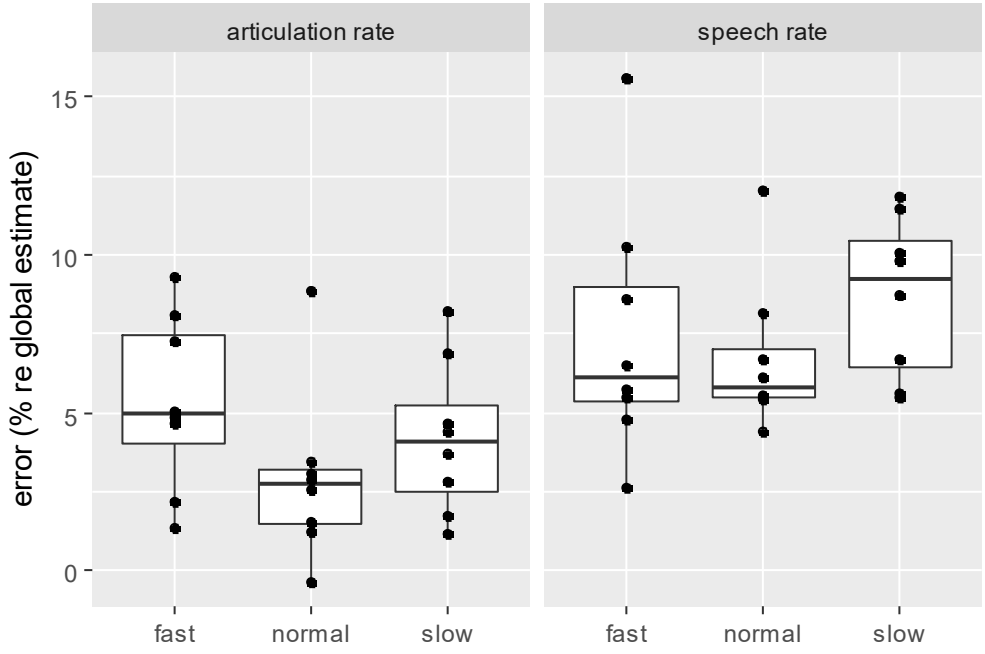

Figure 6: Estimation error broken by rate type (articulation or speech) and rate level (fast, normal, slow). Articulation rate shows smaller errors

All rate values obtained at the stabilization points overestimate the global rate, except for one data point. Rate values at the stabilization point respect the same level ordering defined by the global rate values for 6 out of 8 speakers in the articulation rate sample (speakers AC and PA are the exceptions) and 7 out of 8 speakers in the speech rate sample (speaker DP is the exception).

If mean rate values are used to order the speakers from slowest to fastest (rate levels collapsed), when comparing the ordering obtained when using global rate values and the values estimated at stabilization points, only a pair of adjacent speakers swaps places both in articulation rate (speakers DP and FA) and speech rate (speakers DP and JA).

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 speaking rate. One of the contributions of the present paper is to outline an objective method to approach this subject and the data presented here may serve as a guideline for further research.

The study uncovered the existence of two effects due to the independent variables manipulated: a rate level effect on stabilization time and a rate type effect on the number of VV units comprised by the stabilization interval.

The causal factor behind the rate level effect seems to stem from the fact that the number of VV units needed to achieve stabilization is relatively constant among the three rate levels, and so the fast and normal rates gather that number in less time than the slow rate. That is so because VV units in slow rate are on average longer than normal and fast rates (see figures 1 and 2).

The rate type effect on the number of VV units contained in the stabilization interval can be explained by the fact that VV units in articulation rate are shorter and less variable than those in speech rate, as mentioned in section 2.1. If the mean VV unit duration is multiplied by the mean number of units necessary to achieve stabilization, the total time obtained is roughly the same (9.6 s for articulation rate and 9.5 s for speech rate), a fact that is in line with the lack of rate type effect on stabilization time.

Overall, the results are encouraging. The statistical technique employed provides an objective way of estimating minimum sample length for determining speaking rate. The results obtained here indicate that speech and articulation rate estimates stabilize after 9.2 and 8.7 s, respectively (or after slightly less than 30% of the duration of the speech samples investigated). These values may be used as reference for future work and by forensic experts in their casework. Furthermore, the methodology developed here yields reasonably low error rates and the speaking 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 as speaking style

(spontaneous vs. read speech). It also seems interesting to investigate within-speaker and between-language variability of speaking rate stabilization points, as well as increasing the speaker sample size.

Acknowledgments: The authors would like to thank Plinio A. Barbosa for generously sharing the sound files analyzed here and for his comments on an early draft of the manuscript. The second author is supported by FAPESP grant 2016/12646-0.

ARANTES, Pablo; LIMA, Verônica Gomes. Rumo a uma metodologia para estimar o comprimento mínimo de amostra para velocidade de fala. **Revista do Gel,** v. 14, n. 2, p. 183-197, 2017.

Resumo*: O objetivo deste trabalho é investigar qual seria a prolongação necessária de uma amostra de fala para que a velocidade de fala decorrente dela seja considerada representativa do enunciado total do qual a amostra foi retirada. Oito falantes de português-brasileiro leram um texto de 144 palavras em três níveis de velocidade: devagar, normal e rápido. A velocidade de fala foi medida cumulativamente pelo número de sílabas fonéticas por segundo, da primeira à última sílaba na amostra. Foram medidos dois tipos de taxas, a taxa de elocução de fala e a taxa de articulação. Foi usada a análise de rupturas para determinar a influência do tipo e do nível de taxa na quantidade necessária de tempo para estabilizar a estimativa cumulativa das taxas de elocução de fala e de articulação à taxa fornecida pelo enunciado total. As latências médias de estabilização são de 9,2 segundos por taxa de elocução de fala, e 8,7 s para a taxa de articulação. A velocidade "devagar" tende a estabilizar depois das velocidades rápida e normal, para os dois tipos de taxa. Os intervalos de estabilização levam um número médio de 41 (taxa de elocução de fala) e 59 sílabas (taxa de articulação). Os desvios médios entre a taxa global e o valor da taxa no ponto de estabilização são de 7,8% (taxa de elocução de fala) e 4,2% (taxa de articulação). Os períodos de estabilização estimados neste trabalho podem ser úteis em pesquisa sociofonética, em fonética forense e em fonética clínica.*

Palavras-chave*: Prosódia. Velocidade de fala. Taxa de elocução de fala. Taxa de articulação. Fonética forense.*

Submetido em: 12/03/2017.
Aceito em: 18/05/2017.

References

ARANTES, P.; ERIKSSON, A. **Temporal stability of long-term measures of fundamental frequency**. 2014, Dublin: ISCA, 2014. p. 1149-1152.

BARBOSA, P. A. From syntax to acoustic duration: a dynamical model of speech rhythm production. **Speech Communication**, v. 49, p. 725-742, 2007.

BOERSMA, P. Praat, a system for doing phonetics by computer. **Glot International**, v. 5, n. 9/10, p. 341-345, 2001.

CAO, H.; WANG, Y. **A forensic aspect of articulation rate variation in chinese**. [S.l: s.n.], 2011. p. 396-399.

CRYSTAL, T. H.; HOUSE, A. S. Articulation rate and the duration of syllables and stress groups in connected speech. **Journal of the Acoustical Society of America**, v. 88, n. 1, p. 101-112, 1990.

GFROERER, S. **Auditory-instrumental forensic speaker recognition**. [S.l: s.n.], p. 705-708, 2013.

JESSEN, M. Forensic reference data on articulation rate in German. **Science and Justice**, v. 47, p. 50-67, 2007.

_____. Forensic Phonetics. **Language and Linguistics Compass**, v. 2, n. 4, p. 671-711, 2008.

KILLICK, R.; ECKLEY, I. A. changepoint: An R Package for Changepoint Analysis. **Journal of Statistical Software**, v. 58, n. 3, p. 1-19, 2014. Disponível em: <http://www.jstatsoft.org/v58/i03/>.

KILLICK, R.; HAYNES, K.; ECKLEY, I. A. **changepoint: An R package for changepoint analysis.** [S.l: s.n.], 2016. Disponível em: <http://CRAN.R-project.org/package=changepoint>.

KÜNZEL, H. Some general phonetic and forensic aspects of speaking tempo. **Forensic Linguistics**, v. 4, n. 1, p. 48-83, 1997.

PFITZINGER, H. R. **Local speech rate as a combination of syllable and phone rate.** [S.l: s.n.], 1998. p. 1087-1090.

_____. **Two approaches to speech rate estimation**. [S.l: s.n.], 1996. p. 421-426.

SILVA, G. A. Proposta de construção de um banco de dados de amostras de fala para uso forense em um arcabouço bayesiano. **Revista Brasileira de Criminalística**, v. 5, n. 1, p. 35-45, 2016.

TEAM, R. Core. **R: A Language and Environment for Statistical Computing.** Vienna, Austria: R Foundation for Statistical Computing, 2016. Disponível em: <https://www.R-project.org/>.