On the Need of Standard Assessment Metrics for Automatic Speech Rate Computation Tools

Mireia Farrús[1], Wendy Elvira-García[2] and Juan María Garrido-Almiñana[2]
[1]*Universitat de Barcelona,* [2]*Universidad Nacional de Educación a Distancia, Spain*

Fluency is a relevant feature to assess speech, covering a wide range of linguistic abilities [1]. Among other elements, speech rate –measured as a specific length unit (usually syllables, phonemes or syllable nuclei) per unit of time–, has been shown to be one of the most prominent elements to measure fluency in speech pathologies, language acquisition, or bilingualism, among others [2-4]. However, a manual annotation of speech rate is costly and very time consuming, which poses a high difficulty in annotating large corpora for variational studies. To overcome it, several automatic tools for speech rate measurement have been proposed [5-8]. Nevertheless, these tools usually differ in the way how they are evaluated with respect to human annotations: whereas most of them rely on the correlation between human and automatic annotations [5,6,8], others use mean squared error, error rates, or insertion/deletion errors in the detection of specific units [7,8]. Moreover, these evaluations also differ on the elements being assessed: the evaluation of some tools is directly based on the speech rate measured in syllables, nuclei or phonemes per second [5,6], but others are based on the count of these units in a specific speech segment [7,8]; and whereas some of them compute the speech rate over the whole speech segments, others exclude intermediate pauses. Besides, as far as we know, only [7] distinguishes between read and spontaneous speech.

In the current work, we present a preliminary study on the assessment of a Praat-based tool that detects syllable nuclei and provides an automatic measure of speech rate [5], initially created and tested for Dutch. We used the off-the-shelf Praat script to evaluate its accuracy over a Spanish corpus from the VILE project [9,10], consisting of 30 male speakers, 3.5 hours of speech recorded in three different sessions, and two different conditions: read (22904 vowels) and spontaneous speech (32853 vowels). The corpus was manually annotated at the phoneme, syllable, and word levels. Then, we computed three different assessment metrics: (a) accuracy and recall of detected vowel nuclei, (b) interannotator agreement using Cohen's kappa (κ) coefficient (not weighted) for vowel nuclei manual/automatic detection, and (c) Pearson's correlation between manual and automatic measurements of number of syllables count, speech rate, and articulation rate (defined as speech rate excluding internal pauses).

Our results, shown in Table 1, show promising results in accuracy and recall (see also Figure 1) for both read spontaneous speech, and a rather good kappa coefficient. However, it largely fails in the detection of number of syllables in read speech, and such correlations in the number of syllables detected are not consistent with the correlation coefficients for speech and articulation rates. Such results clearly show that the assessment of the automatic tools depends on the evaluation metrics: while precision and recall metrics and kappa coefficient show a rather good accuracy, the correlation coefficient does not provide promising results. The number of speech samples, the length of the speech segments and the segment units used might be some of the factors for such discrepancies.

The tool submitted to evaluation is a relevant example on the use of phonetic knowledge to foster speech technologies and vice versa, which make possible variational studies using large corpora. However, the evaluation of this kind of tools must be properly assessed in a standard way to allow a fair comparison; therefore, robust standard assessment metrics are needed. With this work, we raise this necessity through the automatic calculation of speech rate and some of its related elements, although the problem can be extended to other phonetic and prosodic measurements for large corpora.

				correlation coefficient		
	precision	recall	κ	# syllables	speech rate	articulation rate
read	0.979	0.732	0.770	0.497	0.669	0.557
spontaneous	0.944	0.717	0.667	0.808	0.595	0.454

Table 1. *Evaluation metrics for both spontaneous and read corpus.*

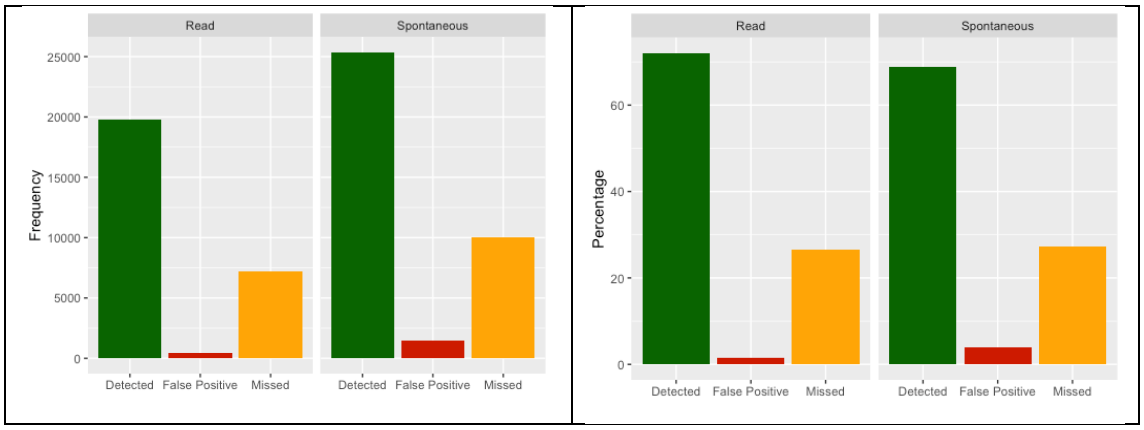

Figure 1. *Correctly and incorrectly detected vowel nuclei in both absolute (left) and relative measurements (right).*

[1] Fillmore, C. J. 1979. *On fluency. In Individual differences in language ability and language behavior*, 85-101. Academic Press.

[2] Trofimovich, P., & Baker, W. 2006. Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, 1-30.

[3] Kalinowski, J., Armson, J., Stuart, A., & Gracco, V. L. 1993. Effects of alterations in auditory feedback and speech rate on stuttering frequency. *Language and Speech*, 36(1), 1-16.

[4] Gordon, J. K., & Clough, S. 2020. How fluent? Part B. Underlying contributors to continuous measures of fluency in aphasia. *Aphasiology*, 34(5), 643-663.

[5] De Jong, N. H., & Wempe, T. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2), 385-390.

[6] Pellegrino, F., Farinas, J., & Rouas, J. L. 2004. Automatic estimation of speaking rate in multilingual spontaneous speech. In Proceedings of the *Speech Prosody Conference*.

[7] Pfitzinger, H. R., Burger, S., & Heid, S. 1996. Syllable detection in read and spontaneous speech. In Proceeding of *4th International Conference on Spoken Language Processing (ICSLP)*, vol. 2, 1261-1264.

[8] Wang, D., & Narayanan, S. S. 2007. Robust speech rate estimation for spontaneous speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8), 2190-2201.

[9] Battaner, E., Gil, J., Marrero, V., Carbó, C., Llisterri, J., Machuca, M. J., ... & Ríos, A. (2005). VILE: Estudio acústico de la variación inter e intralocutor en español. *Procesamiento del Lenguaje Natural*, (35), 435-436.

[10] Albalá, M. J., Battaner, E., Carranza, M., Mota Gorriz, C. D. L., Gil, J., Llisterri, J., ... & Ríos Mestre, A. 2008. VILE: Análisis estadístico de los parámetros relacionados con el grupo de entonación. *Language Design: Journal of Theoretical and Experimental Linguistics*, (special issue), 0015-21.