Loquens 9(1-2)
Diciembre 2022, e090
eISSN 2386-2637
https://doi.org/10.3989/loquens.2022.090

Comparison of intensity-based methods for automatic speech rate computation

Wendy Elvira-García[1], Mireia Farrús[1], Juan María Garrido Almiñana[2]

[1]Universitat de Barcelona
[2]Universidad Nacional de Educación a Distancia (UNED)
wendyelvira@ub.edu ORCID: https://orcid.org/0000-0001-7002-9851
mfarrus@ub.edu ORCID: https://orcid.org/0000-0002-7160-9513
jmgarrido@flog.uned.es ORCID: https://orcid.org/ 0000-0002-3310-8582

Enviado: 01/05 /2022; Aceptado: 05/12/ 2022; Publicado en línea: 09/06/2023

Citation / Cómo citar este artículo: Wendy Elvira-García, Mireia Farrús, Juan María Garrido Almiñana (2022). Comparison of intensity-based methods for automatic speech rate computation. *Loquens 9*(1-2), e090, https://doi.org/10.3989/loquens.2022.e090.

ABSTRACT: Automatic computation of speech rate is a necessary task in a wide range of applications that require this prosodic feature, in which a manual transcription and time alignments are not available. Several tools have been developed to this end, but not enough research has been conducted yet to see to what extent they are scalable to other languages.
In the present work, we take two off-the- shelf tools designed for automatic speech rate computation and already tested for Dutch and English (v1, which relies on intensity peaks preceded by an intensity dip to find syllable nuclei and v3, which relies on intensity peaks surrounded by dips) and we apply them to readand spontaneous Spanish speech. Then, we test which of them offers the best performance. The results obtained withprecision and normalized mean squared error metrics showed that v3 performs better than v1. However, recall measurement shows a better performance of v1, which suggests that a more fine-grained analysis on sensitivity and specificity is needed to select the best option depending on the application we are dealing with.

Keywords: Prosody, speech rate, syllablecount, automatic assessment.

RESUMEN: *Comparación de dos métodos basados en la intensidad para el cálculo automático de la velocidad de habla.*– El cálculo automático de la velocidad de habla es una tarea fonética útil y que además se hace indispensable cuando no hay disponible unatranscripción manual a partir de la cual determinar una tasa de habla manual. Se han desarrollado varias herramientas para este fin, pero todavía no se ha llevado a cabo suficiente investigación para ver hasta qué punto las herramientas son aplicables a lenguas distintas para las que fueron diseñadas. En este artículo probamos dos herramientas para el cálculo automático de la velocidad de habla ya evaluadas para el neerlandés y el inglés (v1,que se basa en la determinación de picos de intensidad precedidos de un valle para encontrar núcleos de sílaba, y v3, que se basa en picos de intensidad rodeados de valles) ylas aplicamos a un corpus de habla leída yespontánea del español para analizar cuál ofrece mejores resultados en español.
Los resultados de precisión y del error cuadrático mediano normalizado obtenidosmuestran que v3 funciona mejor que v1. No obstante, el *recall* muestra mejor rendimiento para la v1, lo que nos indica que se necesita unanálisis detallado de la sensibilidad y laespecificidad para seleccionar la mejor opción en función de los objetivos del análisis posterior que se quiera hacer.

Palabras clave: Prosodia, velocidad de habla, evaluación automática.

Copyright: © 2022 CSIC. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) License

1. INTRODUCTION

Automatic computation of speech rate has several applications in speech technologies, such as automatic evaluation of prosody. Several studies have explored, for example, its use for automatic evaluation of speech fluency (Cucchiarini, Strik, & Boves, 1998, 2000a, 2000b, 2002; Neumeyer, Franco, Digalakis, & Weintraub, 2000; Zechner, Higgins, Xia, & Williamson, 2009; Honig, Batliner, Weilhammer, & Nöth, 2010, among others). Usual approaches to speech rate computation use a phonetic aligner to obtain the necessary phonetic segmentation. This is so because the performance of speech recognition systems, if available, is not good enough to guarantee that the obtained phonetic segmentation is reliable.

Phonetic aligners appear then as an alternative to obtain a more accurate segmentation of the speech chain, but they need the orthographic transcription of the input discourse to be known. If the computation of the speech rate of unrestricted text —not previously known by the system— is attempted, there are some alternatives that do not require a full phonetic segmentation of the input speech to be available, such as the automatic detection of syllabic nuclei. With the aim of exploring this alternative, the current paper compares the performance of two different methods for the automatic computation of the number of syllabic nuclei using a similar technique based on intensity peak detection. The first one is a Praat script described in de Jong and Wempe (2009) and the second one is another Praat script developed by the same authors and other collaborators (de Jong, Pacilly, & Wempe, 2021), in which a different approach to detect intensity peaks is applied. The final goal is to determine which of them would perform better in a task of syllable detection oriented to speech rate calculation and to establish if any of these two methods is adequate to be used in an automatic prosody evaluation system for Spanish.

This paper is structured as follows: Section 2 briefly overviews the related work on this topic, Section 3 describes the experimental setup, Section 4 presents the assessment results, and finally, Sections 5 and 6 sketch the discussion and conclusions, respectively.

2. RELATED WORK

Most of the studies that deal with automatic computation of speech rate are based on the transcriptions obtained —either manually or automatically— from speech material. They mainly differ in the units used to compute speech rate. Most of them are based on counting the number of syllables within a specific segment of speech, providing the speech rate computation as the number of syllables per second, while some other works also provide other measures. Verhasselt and Martens (1996), for instance, defines speech rate as the number of phones per second and computes them over the sentences of the TIMIT corpus. Pfitzinger (1996) also used the number of phones per second as speech rate measure over a total of 240 sentences spoken by eight different speakers.

The literature on automatic speech rate computation tools without transcriptions, which is the goal of the current paper, is scarce. One of the most relevant works in this respect is Pfau and Ruske (1998), in which speech rate is computed by means of vowel detection, based on loudness in vowel regions, which tends to be higher than in consonant regions. Similarly, the method of Pellegrino, Farinas and Rouas (2004) is based on an unsupervised vowel detection algorithm scalable to any language. Validation was assessed on a spontaneous speech subset of the OGI Multilingual Telephone Speech Corpus. In Narayanan and Wang (2005) and Wang and Narayanan (2007), the authors present novel methods for speech rate estimation, measured as the number of syllables per second, analyzing the segments contained between pauses in the Switchboard database (Godfrey & Holliman, 1993). Both methods are based on an extension of signal correlation —essential for syllable detection— by including temporal correlation and prominent spectral sub-bands.

The work described in Dekens, Demol, Verhelst and Verhoeve (2007) is also based on the number of syllables per second, and the authors evaluate the performance of several speech estimators on a multilingual database covering Dutch, English, French, Romanian and Spanish, by using sub-band and time correlation to detect the number of vowels and diphthongs.

However, giving that speech rate can be computed using syllables or phones and total time of speech, any tool that identifies either syllable boundaries or vowels can be used for this task, for example, tools that syllabify conversational speech (Landsiedel et al., 2011; Mary et al., 2018) or tools that locate syllable nuclei (Sabu, Chaudhuri, Rao, & Patil, 2021). Using this last method, de Jong et al. (2007) and de Jong & Wempe (2009) compute speech rate over two corpora of spoken Dutch, by identifying peaks in intensity that are preceded by dips, which is then considered as a syllable nucleus. In Sabu et al. (2021), the authors use the TIMIT dataset (Garofolo et al., 1993) and a children's oral reading corpus created ad hoc, for which they identify vowel sonority by means of local peak picking on a frequency-weighted energy contour.

3. EXPERIMENTAL SETUP

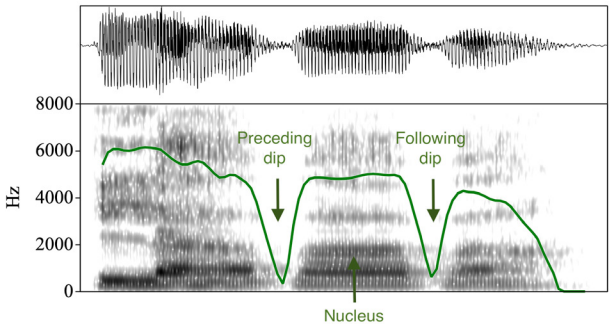
3.1. Evaluated tools for speech rate computation

The present paper analyses the performance of two tools distributed under a GNU General Public License (de Jong & Wempe, 2009; de Jong et al., 2021). Both of them are Praat-based scripts that use intensity in order to find syllable nuclei. More specifically, they extract an intensity object using the following parameters: 'minimum Pitch' set to 50Hz and the autocorrelation method. After this point, their behavior differs.

The first tool (v1), described in de Jong and Wempe (2009), applies a predefined threshold (2dB above the median intensity of the total sound file) to find peaks preceded by a dip in intensity (see Figure 1). Then, out of those peaks, it discards those that are unvoiced.

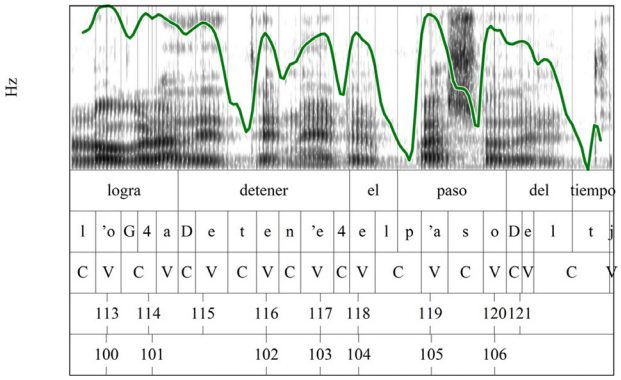
The second tool (v3) relies on a different method (de Jong et al., 2021). It detects every intensity peak above 25 dB and below 95 % of the highest peak (in order to disregard loud bursts in the signal). Then, it measures the intensity surrounding the peak and if it is a dip of at least 2dB at both sides the peak is labelled as syllable nucleus (de Jong et al., 2021).

Figure 1: Intensity curve of the Spanish phrase "La baba" (the slime) with a syllable nucleus and its preceding and following dips highlighted.



Therefore, the main difference between the tools is that one (v1) considers as syllable nuclei those intensity peaks preceded by a dip, and the other (v3) considers as syllable nuclei those intensity peaks that are surrounded by intensity dips. This difference results in the same judgments most of the time, however in some cases it does not. Discrepancies between v1 and v3 are usually related to approximants, whose dip is short enough to be considered a whole with the next one, and laterals (and nasals to a lesser degree) in coda position (Figure 2).

Figure 2: Waveform, spectrogram and intensity of the Spanish sentence "Logra detener el paso del tiempo" '(It manages to stop time)' depicting the vowel nuclei found by v1 (tier 4) and v3 (tier 5).



3.2. Materials

In order to test which method (preceding peak or surrounding peak) offers the best performance in Spanish, we used a subcorpus from the AHUMADA corpus (Ortega-Garcia, Gonzalez-Rodriguez, & Marrero-Aguiar, 2000) selected for the VILE project (Albalá et al., 2008; Battaner Moro et al., 2005), consisting of recordings of 30 male speakers, with a total of 3.5 hours of speech, recorded in three different sessions in different days (M1- M2- M3), and two different conditions: read speech (26984 vowels) and spontaneous speech (35366 vowels).

The read subcorpus consists of the reading of a phonologically and syllabically balanced text of approximately one minute read at a normal speech rate. All speakers read the same text in the three sessions.

The spontaneous subcorpus consists of at least one minute of speech describing a picture, explaining speakers' last holidays, a well- known board game or simply something familiar to them.

This material was manually annotated at the phoneme, syllable and word levels for the VILE project (Albalá et al., 2008; Battaner Moro et al., 2005). The annotation procedure involved three steps: in the first one, a team of phoneticians orthographically transcribed intonational groups following the guidelines described in Llisterri, Machuca and Ríos, (2017); in the second one, EasyAlign (Goldman, 2011) was used to automatically align the annotation; finally, a human annotator revised the automatic segmentation.

3.3. Evaluation metrics

One of the main challenges when assessing systems dealing with the automatic computation of speech rate is the diversity and sparseness of evaluation metrics. The metrics used in the literature to evaluate the speech rate estimators vary among the different works and include a wide range of metrics such as the relative prediction error, the correlation coefficient between the estimated and actual syllables, the syllable error rate, the vowel error rate, the linear regression coefficient, the mean error, the standard deviation error, and F-score, among others. Moreover, these metrics are computed either over the number of syllables (or phones) as units of measurement, or directly over the speech rate measurement.

In the current paper, we present two different evaluations to compare the two tools addressed. Firstly, we show a performance analysis based on common metrics used for classification problems: accuracy, precision, recall, and F-score. For this assessment, we have considered the tier where vowel (syllable nuclei) and consonant intervals (non syllable nuclei) are labelled.

Additionally, we provide the root mean square error (RMSE) and normalized root mean square error (NRMSE) for the assessment analysis, based on the syllable annotation tier and, more specifically, the number of syllables of each file in the VILE corpus.

3.3.1. Performance metrics

For the first evaluation, we compare both tools using the standard performance metrics in classification problems:

- Accuracy: defined as the number of cases of the correctly predicted class, that is:

$$(1) \quad accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where *TP* = True Positives (detected syllable nuclei), *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.

- Precision: defined as the number of correctly detected syllable nuclei over the actual cases, that is:

$$(2) \quad precision = \frac{TP}{TP + FP}$$

- Recall: defined as the number of correctly detected syllable nuclei over the estimated cases, that is:

$$(3) \quad recall = \frac{TP}{TP + FN}$$

- F-score: defined as the combination of both precision and recall in the following form:

$$(4) \quad F\text{-}score = 2 \ \frac{precision \cdot recall}{precision + recall}$$

3.3.2. Assessment metrics

Farrús et al. (2021) explored the adequacy of several metrics commonly used in the literature, such as: (a) correlation coefficient between actual number of syllables —or speech rate— and estimated number of syllables —or speech rate measurement—, (b) mean error defined as the mean of the error in absolute values, (c) standard deviation error defined as the standard deviation of the previous mean, (d) coefficient of variation defined as (standard deviation error)/ (mean error), (e) mean square error (MSE), (f) root mean square error (RMSE), and (g) normalized root mean square error (NRMSE), by mean defined as:

$$(5) \quad RMSE = \sqrt{\frac{\Sigma_{i=1}^{N} (\hat{y}_i - \hat{y}_i)^2}{N}}$$

$$(6) \quad NRMSE = \frac{RMSE}{\bar{y}}$$

where N is the number of observations, y_i isthe ith reference (actual) value, $\hat{y}_i$ is its corresponding estimated value, and $\bar{y}$ is themean of the measured data.

Farrús et al. (2021) concluded that correlation coefficients were not adequate for this kind of assessment, and that, instead, the use of the relative error as a unit for the different metrics should be encouraged, since it homogenizes the assessment based on the number of syllables and speech rate, apart from exhibiting consistent and coherent results. In the current paper, we evaluate the performance of both tools by computing the number of syllables, the speech rate, and the relative error. Moreover, as suggested in our previous study, we compare both tools by means of RMSE as an assessment metric, together with its normalized value (NRMSE) for a better comparison between models computed over different scales.

4. SYSTEM COMPARISON

4.1. Performance analysis

The two tools analyzed provide the number of syllables detected via a TextGrid with a point tier (in which the syllable nuclei are indicated as points in time). The Spanish databases are labeled sound-by-sound using interval tiers. In order to make the results comparable, we have combined the automatic point tier with the manual interval tier. We considered that the system succeeded when either there is a point in the time range of a manual interval labelled as vowel (true positive, TP) or there is no point within the time range of a manual interval labelled as consonant (true negative, TN). The system fails when we have a point within a consonant time range (false positive, FP), we have more than a point within a vowel time range (as many false positives as surplus points) or there is no point within a vowel time range (false negative, FN).

This comparison method is accurate for our purpose (computing the number of syllables detected by the script). However, it would not be accurate for tasks where the interest was the actual center of syllable nuclei, since the method counts as correct any point that falls within the vowel range without taking into account whether the script has placed the point in the vowel mid-point.

Table 1 shows the comparison of the results obtained by both tools in terms of performance analysis. It details the number of syllable nuclei (vowels) correctly detected (True Positives), wrongly detected (False Positives), missed (False Negatives) and correctly dismissed (True Negatives) by the two tools (v1 and v3) in the two analyzed conditions (read data and spontaneous data). The same results expressed in percentage are illustrated in Figure 3.

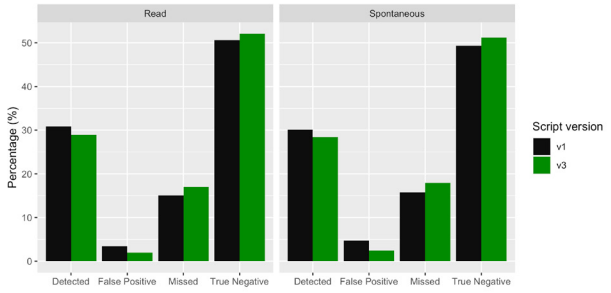
Figure 3 shows that v1 results are better for detected syllables (higher value) and missed syllables (lower value), whereas v3 performs better for true negatives (a higher value) and false positives (a lower value), taking as a reference the number of manually annotated vowels (nucleus) and consonants (non-nucleus) in both subcorpora (26984 nuclei for read, 35366 for spontaneous speech).

Table 1: Number of syllables detected (TP), false positives (FP) and missed (FN) and true negatives by the two scripts (v1 and v3 in the two situations (read corpus and spontaneous corpus)).

	read		spontaneous	
	V1	V3	V1	V3
TP	18137	16989	23208	21671
FP	2034	1147	3655	1851
FN	8847	9995	12158	13695
TN	29740	30601	37990	39049

Table 2 shows the main performance metrics for tool v1 and v3 for read and spontaneous speech with the best result highlighted in bold. Results show that, in general, v3

Figure 3: Percentage of detected (TP), False Positive, Missed (FN), and True Negative syllables in v1 and v3.



is the best tool when we consider precision and v1 shows a better performance in recall and F-score. Accuracy reveals contradictory results, having best results for tool v1 in read speech and the best result for tool v3 in spontaneous speech. However, accuracy is a discouraged metric in cases of heavily imbalanced cases (big difference between the number of false positives and false negatives or missed cases) (e.g. Mortaz, 2020) and in those cases performance analysis should rely on F-score metrics.

Table 2: Performance metrics for tool v1 and v3 for read and spontaneous speech.

	read		spontaneous	
	v1	v3	v1	v3
accuracy	**0.815**	0.810	0.795	**0.796**
precision	0.899	**0.994**	0.864	**0.921**
recall	**0.672**	0.629	**0.656**	0.613
F-score	**0.769**	0.753	**0.746**	0.733

4.2. Assessment metrics

In this section, we present the assessment metrics obtained for the following units of analysis: number of syllables, speech rate, and relative error. Speech rate is defined as number of syllables per second, and the relative error is defined as:

$$(7) \quad \varepsilon_r = \frac{|\#syll_a - \#syll_m|}{\#syll_m}$$

where $\#syll_a$ is the estimated (automatic) count of syllables, and $\#syll_m$ in the actual (manual) count. Since speech rate is obtained using the number of syllables along the entire speech duration, and the length of the spurt analyzed is the same in both evaluations (automatic and manual), the relative error applied to the number of syllables and to speech rate coincides, making it a homogenized measurement.

In Table 3, we show the total number of syllables obtained with the manual transcriptions in the entire corpus, as well as the number of syllables obtained in both automatic tools (v1 and v3) for read and spontaneous modalities. The results clearly show that v3 fails more than v1 when detecting syllable nuclei, although both tools underestimate the actual number of syllables.

Table 3: Total number of syllables obtained with the manual transcriptions and the automatic tools for both read and spontaneous speech.

read			spontaneous		
manual	v1	v3	manual	v1	v3
27005	20165	18155	35408	27491	24003

Tables 4 and 5 show the root mean square error (RMSE) and normalized root mean square error (NRMSE) respectively, obtained for both tools, the different units of analysis (number of syllables, speech rate, and error rate), and both read and spontaneous modalities.

Table 4: RMSE obtained for the different unitsof analysis.

	read		spontaneous	
	v1	v3	v1	v3
#syllables	**78.3**	99.8	**99.2**	135.3
speechrate	**1.420**	1.789	**3.604**	4.032
error rate	**0.261**	0.333	**0.233**	0.326

Table 5: NRMSE obtained for the different units of analysis.

	read		spontaneous	
	v1	v3	v1	v3
#syllables	1.030	**1.015**	1.128	**1.068**
speechrate	1.050	**1.033**	1.136	**1.105**
error rate	1.031	**1.015**	1.063	**1.030**

The best result within both tools, with each assessment metric and for both read and spontaneous speech is highlighted in bold. The results mainly show that, while tool v1 performs better when it is evaluated by means of RMSE, tool v3 performs better if we consider NRMSE.

5. DISCUSSION

The performance analysis (see 4.1) shows that both tools are reliable finding syllable nuclei (precision> 0.8

and recall > 0.5, in all cases). Also, both tools perform better with read speech than with spontaneous speech. However, they share a common problem in classification tasks: an imbalanced classification, with more false negatives than false positives, which complicates the assessment. For our data, this is a foreseeable result given that finding a syllable nucleus that is not preceded or followed by an intensity dip is more usual in speech than it is for a voiced intensity peak to be a consonant. This means that, if we want the tool to correctly disregard peaks that are not vowels, we need to use a more restrictive system —which v3 does by requiring a preceding and following intensity dip in order to consider an interval as a syllable nucleus— and that will give us a better result in accuracy and precision, which is exactly what is shown in Table 2, given that precision is as a measure of quality, meaning that the vowels that are marked as vowels with v3 are more likely to be real vowels. However, if we consider the global result of correctly identified and disregarded syllable nuclei (the quantity) a less restrictive rule (i.e., v1, which only considers previous intensity dips) has a better performance as illustrated by Table 2 recall and F-score.

For the aim of this paper, which is the automatic computation of speech rate, quantity measures can prove more relevant than quality measures given that, when computing speech rate, we are not interested in knowing whether the segment is a syllable nucleus but rather in getting a number of syllable nuclei as close as possible to the actual one. That is, if a false positive is later compensated by a missed nucleus the system is still accurate. This is the exact scenario when in a real syllable the automatic tool places the syllable nucleus within the onset instead of in the actual nucleus, but then does not label the vowel as nucleus.

In Table 3, we can clearly see that the number of syllables counted by v1 is closer to the actual number of syllables counted by v3. In other words, v3 is missing a larger number of syllables, which results in larger values of RMSE for v3, both in the read and the spontaneous modalities (Table 4). These results are consistent with those shown in Table 1, also illustrated in Figure 2: the number of detected (true positives) and false positive syllables is greater in v1. The number of missed syllables (false negatives) also contributes to enlarge the underestimation in the syllable counting.

However, the NRMSE metric (Table 5) shows otherwise: the RMSE normalized values by the mean of the measured data in v3 outperform those obtained in v1. The fact is that, although v3 fails more in detecting syllables than v1, such failure is more stable. This is strengthened by the measurement of other metrics such as the standard error (standard deviation of the mean error) and the coefficient of variation —or relative standard deviation— defined as (standard deviation)/mean). For both measurements, v3 shows a better performance than v1 for both read and spontaneous modalities.

On the one hand, this shows that, although v3 fails largely in missing syllables, such failure could be better compensated by a correction factor. On the other hand, and since v3 appears to be more restrictive in the detection conditions of syllables — we need an intensity dip in both side of the vowel and not only one as in v1—, but we can also ensure that the detected syllables come more often from actual syllable nuclei in v3 than in v1, in which the detected syllable could come more often from false nuclei. This is also strengthened by the larger number of true negatives in v3 encountered in Table 1 for both modalities.

6. CONCLUSIONS

The results presented and discussed in the previous sections indicate, on the one hand, that both methods of syllable detection are not fully reliable yet to face a speech rate analysis task: both detect a number of syllables which is remarkably lower than the number of syllables obtained from a manual annotation. However, v1 seems to offer a better performance for this task than v3, as the number of detected syllables is closer to the manually obtained value, which compensates the fact that it is less precise in the detection of actual syllables, a fact that is secondary in a speech rate calculation task if the number of detected syllables is close enough to the number of manually annotated ones.

On the other hand, the results also show that, although v3 detects in general less true syllables than v1, it seems more adequate for tasks in which it is important that detected syllables correspond to actual syllables, such as automatic acoustic measurements of corpora involving the detection of syllabic nuclei.

ACKNOWLEDGEMENTS

This work has been partially funded by the project "Métodos, modelos, métricas y herramientas para la evaluación de la prosodia (ProA)", reference number PGC2018-094233-B-C21. The first author is a "Serra Húnter Fellow". The authors would like to thank Dr. María Machuca for providing the VILE corpus used in these experiments.

7. REFERENCES

Albalá, M. J., Battaner, E., Carranza, M., Mota Gorriz, C. d. l., Gil, J., Llisterri, J., ... others (2008). VILE: Análisis estadístico de los parámetros relacionados con el grupo de entonación. *Language Design: Journal of Theoretical and Experimental Linguistics (Special Issue)*, 15–21.

Battaner Moro, E., Gil Fernández, J., Marrero Aguiar, V., Carbo Marro, C., Llisterri Boix, J., Machuca Ayuso, M. J., ... Ríos Mestre, A. (2005). VILE: estudio acústico de la variación inter- e intralocutor en español. *Procesamiento del Lenguaje Natural*, *35*, pp. 435-436.

Cucchiarini, C., Strik, H., & Boves, L. (1998). Quantitative assessment of second language learners' fluency: An automatic approach. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98)*, pp. 2619–2622. Sydney, Australia. http://dx.doi.org/10.21437/ICSLP.1998-754

Cucchiarini, C., Strik, H., & Boves, L. (2000a). Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication*, *30*(2-3), 109–119. http://dx.doi.org/10.1016/S0167-6393(99)00040-0

Cucchiarini, C., Strik, H., & Boves, L. (2000b). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America, 107*(2), 989–999. http://dx.doi.org/10.1121/1.428279

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America, 111*(6), 2862–2873. http://dx.doi.org/10.1121/1.1471894

de Jong, N. H., Pacilly, J., & Wempe, T. (2021). Praat scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy and Practice*, *28*(4), 456-476. http://dx.doi.org/10.1080/0969594X.2021.1951162

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. http://dx.doi.org/10.3758/BRM.41.2.385

de Jong, N. H., Wempe, T., et al. (2007). Automatic measurement of speech rate in spoken Dutch. *ACLC Working Papers*, *2,* 51–60.

Dekens, T., Demol, M., Verhelst, W., & Verhoeve, P. (2007). A comparative study of speech rate estimation techniques. In *Proceedings of the Eighth Annual Conference of the International Speech Communication Association (INTERSPEECH 2007)*, 510–513. http://dx.doi.org/10.21437/Interspeech.2007-237

Farrús, M., Elvira-García, W., & Garrido- Almiñana, J. M. (2021). On the need of standard assessment metrics for automatic speech rate computation tools. In *4th Phonetics and Phonology in Europe 2021 Conference* (*PAPE 2021*).

Garofolo, J.-S., et al. (1993) *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Web Download. Philadelphia: Linguistic Data Consortium.

Godfrey, J.-J., Holliman, E. (1993). *Switchboard-1 Release 2 LDC97S62*. Web Download. Philadelphia: Linguistic Data Consortium.

Goldman, J.-P. (2011). Easyalign: an automatic phonetic alignment tool under Praat. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*. Florence, Italy. 28-21 August, 2011.

Honig, F., Batliner, A., Weilhammer, K., & Nöth, E. (2010). Automatic assessment of non-native prosody for English as L2. In *Speech Prosody 2010*, Chicago, IL, USA.

Llisterri, J., Machuca, M., & Ríos, A. (2017). VILE-P: un corpus para el estudio prosodico de la variación inter e intralocutor. Comunicación presentada en *SUBSIDIA: Herramientas y recursos para las ciencias del habla*, Málaga, Spain. June, 2017.

Mortaz, E. (2020). Imbalance accuracy metric for model selection in multi-class imbalance classification problems. *Knowledge-Based Systems, 210*, 106490. http://dx.doi.org/10.1016/j.knosys.2020.106490

Narayanan, S., & Wang, D. (2005). Speech rate estimation via temporal correlation and selected sub-band correlation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* Vol. 1, pp. 1–413. http://dx.doi.org/10.1109/ICASSP.2005.1415138

Neumeyer, L., Franco, H., Digalakis, V., & Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication, 30*(2–3), 88-93. http://dx.doi.org/10.1016/S0167-6393(99)00046-1

Ortega-García, J., González-Rodríguez, J., & Marrero-Aguiar, V. (2000). Ahumada: A large speech corpus in Spanish for speaker characterization and identification. *Speech Communication, 31*(2-3), 255–264. http://dx.doi.org/10.1016/S0167-6393(99)00081-3

Pellegrino, F., Farinas, J., & Rouas, J.-L. (2004). Automatic estimation of speaking rate in multilingual spontaneous speech. In *Speech Prosody 2004* (pp. 517–520).

Pfau, T., & Ruske, G. (1998). Estimating the speaking rate by vowel detection. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*. Vol. 2, pp. 945–948. http://dx.doi.org/10.1109/ICASSP.1998.675422

Pfitzinger, H. R. (1996). Two approaches to speech rate estimation. In *Proceedings of the 6th Australian International Conference on Speech Science and Technology (SST, 96).*Vol. 96, pp. 421–426.

Sabu, K., Chaudhuri, S., Rao, P., & Patil, M. (2021). An optimized signal-processing pipeline for syllable detection and speech rate estimation. In *National Conference on Communications (NCC, 2020)*. https://doi.org/10.48550/arXiv.2103.04346

Verhasselt, J. P., & Martens, J.-P. (1996). A fast and reliable rate of speech detector. In *Proceedings of Fourth International Conference on Spoken Language Processing (ICSLP'96)*. Vol. 4, pp. 2258–2261. http://dx.doi.org/10.21437/ICSLP.1996-577

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. http://dx.doi.org/10.1109/TASL.2007.905178

Zechner, K., Higgins, D., Xia, X., & Williamson, D. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*(10), 883-895. http://dx.doi.org/10.1016/j.specom.2009.04.009.