Interspeech 2018
2-6 September 2018, Hyderabad

Automatically measuring L2 speech fluency without the need of ASR: a proof-of-concept study with Japanese learners of French

Lionel Fontan[1], *Maxime Le Coz*[1], *Sylvain Detey*[2]

[1]Archean LABS, Archean Technologies, Montauban, France
[2] Waseda University – SILS, Tokyo, Japan

{lfontan,mlecoz}@archean.tech, detey@waseda.jp

Abstract

This research work investigates the possibility of using automatic acoustic measures to assess speech fluency in the context of second language (L2) acquisition.

To this end, three experts rated speech recordings of Japanese learners of French who were instructed to read aloud a 21-sentence-long text. A Forward-Backward Divergence Segmentation (FBDS) algorithm was used to segment speech recordings (sentences) into acoustically homogeneous units at a subphonemic scale. The FBDS processing results were used — along with more classic measures such as raw percentage of speech and length/standard deviation of silent pauses — to estimate speech rate and regularity of speech rate, while a formant tracking algorithm was used to estimate speech fluidity (i.e., quality of coarticulation). A step-by-step multiple linear regression was finally computed to predict the experts' mean fluency ratings.

Results show that FBDS-derived measures, raw percentage of speech, and standard deviation of the first formant curve derivative can be combined together to calculate accurate estimates of speakers' fluency scores ($R = .92; P < .001$). As only low-level signal features were used in the study, the method could also be relevant for the assessment of speakers of other target languages, as well as for the assessment of disordered speech.

Index Terms: Speech fluency, Second language acquisition, Automatic measures, Japanese, French, Learner corpus

1. Introduction

The concept of speech fluency is a broad one, since it can relate not only to pronunciation, but also to lexical access, syntactic complexity, discourse planning or even overall linguistic proficiency. However, in the field of second language (L2) speech processing, fluency is usually defined in relation to accent, intelligibility and comprehensibility [1], with a specific focus on temporal features such as pauses and speech rate [2]. From a pedagogical perspective, it can be defined as "the degree to which speech flows easily without pauses and other disfluency markers" [3, p. 5]. Therefore, for L2 learners, the acquisition of a native-like fluency is an important goal, and its assessment usually a requirement. Yet, as is the case for other dimensions of pronunciation assessment [4, 5], fluency can be evaluated in different ways, either as perceived fluency, e.g. by human listeners on a given scale, or through phonetic-phonological analyses, with measures of speech rate, phonation time ratio, pruned syllables, articulation rate, mean length of runs, silent pause ratio or filled pause ratio [1].

Because perceptual evaluations are subjective and phonetic-phonological analyses are time-consuming, several attempts have been made to measure L2 speakers' fluency – as captured by experts' ratings – by means of automatic techniques. For example, automatic speech recognition (ASR) was used to calculate temporal variables such as rate of speech, phonation ratio, and mean length of pauses that proved to be strongly correlated with fluency ratings [6]. After being applied with success to read speech, the application of these objective techniques was extended to the more challenging evaluation of spontaneous speech fluency [7].

However, using ASR to assess the fluency of L2 speakers presents several limitations. The first and most obvious limit is that ASR-derived methods are dependent on the target language the ASR systems were set up for. Second, the performances of ASR systems – in terms of word error rate and phonetic alignment – depend on the canonicity of speech they are fed with; ASR results tend therefore to lack reliability if their acoustic models were not sufficiently trained on speech produced by L2 speakers [8]. Finally, ASR can be thought of as a rather heavy method: the sole creation of acoustic models generally relies on the annotation of hundreds of hours of speech.

For another purpose – namely the assessment of pathological speech – some researchers [9] recently set up a different approach that goes beyond the above-mentioned limits. The authors used only low-level signal (acoustic) measures – both temporal (e.g. total length of silent pauses) and spectral (number of abrupt spectral changes) – to predict fluency ratings for stutterers. Their results showed that human ratings could be predicted by the automatic acoustic measurements.

The present study investigates the use of similar low-level signal analyses to predict speech fluency ratings for learners of a second language with Japanese learners of French as a pilot population. The long-term objective of this work is to integrate an objective and rapid tool for the assessment of speech fluency into a computer-assisted pronunciation training (CAPT) software [10, 11].

2. Speech materials

2.1. Speakers

Eight undergraduate Japanese university students (4 male, 4 female, age 18-22) participated in this study; at the time of the recordings, they were studying French in two universities in Tokyo (Japan). In order to get different levels of expected fluency, we selected 12 sets of data from the Corpus Longitudinal Interphonologique de Japonais Apprenants de Français (CLIJAF) [12], couched within the framework of the InterPhonologie du Français Contemporain recording protocol [13, 14]. The CLIJAF corpus is made up of two parts: a two-year long longitudinal study of beginner Japanese learners of French on the one hand, and a study of pre-advanced Japanese learners of

10.21437/Interspeech.2018-1336

French who have studied in a French-speaking environment at least once in their life on the other hand. Therefore, 4 students were recorded twice (respectively after 4 and 19 months of study), while the 4 other students were more advanced (they all had studied for at least 1 year in France, during their primary, secondary or higher education). In relation to the overall linguistic proficiency levels defined within the Common European Framework of Reference for Languages [15], our data included 4 speakers/levels of A1 level (beginners), 4 of B1 level (intermediate) and 4 of B2 level or above (advanced).

2.2. Recording task

Participants read out the text "Le Premier Ministre ira-t-il à Beaulieu" used in the recording protocol of the (I)PFC project ((Inter)Phonologie du Français Contemporain, [16]), an artificially constructed text similar to a short local newspaper article designed to cover all main aspects of French phonology, and used both with native and non-native speakers [17]. The text consisted of 21 sentences; to help the Japanese learners, it was divided into short paragraphs, each preceded by a brief summary in Japanese.

In total, the participants recorded 252 sentences, for an approximate length of 48 minutes. As the recordings took place in different locations (PC classrooms equipped with headphone and microphone sets, as well as soundproof recording studios), their quality is heterogeneous.

3. Human fluency ratings

3.1. Participants and procedure

Three native French speakers, with a solid background in second language acquisition and phonetics, took part in the rating tasks. They had a significant experience in perceptive assessment of non-native speech; they were instructed to judge the 252 sentences based on four different dimensions, each of which was scored on a 5-point scale:

- Global fluency: perceived ease of speech (from 0 – *non fluent* to 4 – *as fluent as a native speaker*);
- Speech rate: perceived speed of speech (from 0 – *very slow* to 4 – *as fast as a native speaker talking relatively fast*);
- Regularity of speech rate: perceived changes in the tempo (accelerations, decelerations and breaks; from 0 – *very irregular* to 4 – *totally regular*);
- Speech fluidity: perceived fluidity of coarticulation (smoothness of transitions between phones; from 0 – *not fluid at all* to 4 – *as fluid as speech produced by a native speaker*).

A web-based user interface, connected to a database, was created for the rating task. The sentence recordings were presented in a random order to the human raters, who could score each one into a dedicated JavaScript form. At any time during the rating procedure, the raters were free to replay any recording and could, if necessary, change the associated scores.

At the end of the rating procedure (which took approximately five hours for each rater) a total of 3,024 scores (252 sentences $\times$ 4 dimensions $\times$ 3 raters) were collected.

3.2. Inter-rater agreement

Table 1 shows inter-rater Spearman's rank correlation coefficients as a function of the four rating dimensions: global fluency, speech rate, regularity of speech rate, and speech fluidity. As can be observed, all correlations are highly significant ($P < .001$), and their strength ranges from moderate (for regularity of speech rate $.58 \geq \rho \leq .69$) to strong (for the three other dimensions $.72 \geq \rho \leq .84$).

Table 1: *Inter-rater agreement for speech fluency, speech rate, regularity of speech rate, and speech fluidity ratings, as measured by Spearman's rank correlation coefficients ($n = 252$)*

		Rater 2	Rater 3
Speech fluency	Rater 1	$\mathbf{.84^{***}}$	$\mathbf{.80^{***}}$
	Rater 2		$\mathbf{.80^{***}}$
Speech rate	Rater 1	$.77^{***}$	$.77^{***}$
	Rater 2		$.72^{***}$
Regularity of speech rate	Rater 1	$.58^{***}$	$.69^{***}$
	Rater 2		$.62^{***}$
Speech fluidity	Rater 1	$.72^{***}$	$.76^{***}$
	Rater 2		$.77^{***}$

$^{***}P < .001$

3.3. Computation of final scores

As we observed a good agreement between raters, the ratings were eventually averaged for each sentence. Some information on the distribution of the resulting scores is reported in Table 2. For each dimension the scores range from very low values (0 or .3) up to the maximum value of 4; mean scores are close to the center value (2) of the five-point scales used for the ratings. However, the Kolmogorov-Smirnov test shows that scores do not follow a normal distribution ($P < .001$ for all four dimensions); hence, for consecutive comparisons between human ratings and automatic scores only non-parametric tests are used.

Table 2: *Distribution of mean scores among the 252 sentences for each annotated dimension (Std. dev.: Standard deviation; Reg.: Regularity)*

Variable	*Min*	*Max*	*Mean*	*Std. dev.*
Fluency	0	4	2.3	1.0
Speech rate	0.3	4	2.4	0.9
Reg. of speech rate	0.3	4	2.2	0.9
Fluidity	0	4	2.2	1.0

4. Automatic measures

4.1. Automatic segmentation of speech

The segmentation algorithm used in this study was proposed by [18]. In this approach the speech signal is conceived as a sequence of almost stationary segments, each one being modelled by an auto-regressive Gaussian model:

$$\begin{cases} y_n = \sum a_i y_{n-i} + e_n; \\ var(e_n) = \sigma_n \end{cases} \quad (1)$$

with y the speech signal and e a Gaussian white noise. The algorithm analyses the speech signal on both a long-term and a short-term scale (here a 10ms sliding window), and calculates the distance between the two resulting models, using the

Kullback-Leibler divergence measure [19]. When the divergence exceeds a given threshold, the lower limit of the short-term window is labeled as a segment boundary, and the analysis continues forward. As this analysis process is not symmetric, it is also performed backward and the results of the forward and backward segmentation processes are finally merged. This segmentation algorithm, referred to as Forward-Backward Divergence Segmentation (FBDS), results in a subphonemic segmentation [18].

4.2. Computation of individual predictors

4.2.1. Speech rate

For each sentence, speech rate was measured as the number of segments produced by the FBDS algorithm, divided by the sentence duration (in seconds). The underlying asumption is that the higher the speech rate, the larger the number of segments found by the FBDS algorithm.

4.2.2. Regularity of speech rate

The regularity of speech rate, for a given sentence, was computed as the standard deviation of the lengths of the segments found by the FBDS algorithm (in seconds). It was hypothesized that the presence of long pauses and hesitations in a sentence would increase the variability of speech segments duration, and therefore those of FBDS segments.

4.2.3. Speech fluidity

The only spectral analysis conducted for the automatic estimation of speech fluency was that of fluidity of coarticulation. Given that fluency is related to the anticipation of sounds to be produced [20], it was hypothesized that a lack of planning would result in abrupt transitions between phones, i.e., in a poor coarticulation.

In order to quantify this phenomenon, as the acoustic formants are linked to the movements of speech articulators, a tracking of speech formants was implemented. The analysis was however limited to that of the first formant. It was assumed that the less speech movements are anticipated, the most abrupt transitions would be found between speech sounds. As a consequence, the estimation of (lack of) speech fluidity was computed as the standard deviation of the first formant curve derivative.

4.2.4. Length and regularity of silent pauses, percentage of speech

Silent pauses were automatically detected as segments where amplitude was not exceeding 10% of the maximum amplitude found in the sentence, and lasting at least 250 ms. The regularity of silent pauses was computed for each sentence as the standard deviation of silent pauses duration.

Raw percentage of speech was simply calculated as 100 minus the silent pause percentage (i.e., 100 minus the total pause length divided by the total duration of recordings and multiplied by 100).

5. Results

5.1. Individual estimators of speech rate, regularity of speech rate, and fluidity

Table 3 presents the correlations between the automatic estimators of speech rate, regularity of speech rate, and fluidity, and the human annotations. As expected, the estimated speech rate (i.e., the rate of FBDS segments) is strongly and positively correlated with perceived speech rate, meaning that the higher the perceived speech rate, the larger the number of segments found by the FBDS algorithm.

Also, the estimator of the regularity of speech rate (i.e., the standard deviation of FBDS segment length) is negatively correlated with the human annotations. In other words, the more irregular the speech rate is perceived, the more variability is found in the length of FBDS segments – which is in line with our original assumption.

Finally, measures of fluidity of coarticulation are negatively (though weakly) correlated with perceived fluidity, meaning that – as expected – the more fluid speech is perceived, the weaker the variations in first formant tend to be.

Table 3: *Bivariate correlations between automatic estimators and human annotations of speech rate and speech fluidity*

Variable	Spearman's ρ	*P*-value
Speech rate	.77	$< .001$
Regularity of speech rate	-.72	$< .001$
Fluidity	-.21	.001

5.2. Prediction of speakers' fluency ratings

To predict human fluency ratings, a step-by-step linear regression was computed for the 252 observations (i.e., the 252 sentences) with, as dependent variable, mean fluency score, and, as independent variables, all the automatic measures: speech rate, regularity of speech rate, speech fluidity, mean length of silent pauses, standard deviation of the length of silent pauses, and percentage of speech.

The best linear model ($R = .82$) rejected two out of the six independent variables: mean duration of silent pauses and standard deviation of silent pauses. This rejection might be due to the fact that both variables were significantly correlated with some other predictors ($\rho = -.66$ and $P < .001$ between percentage of speech and mean duration of pauses; $\rho = .44$ and $P < .001$ between standard deviation of pauses and regularity of speech rate), and thus did not bring enough predictive power to the model.

Table 4 presents the standardized coefficients of the four remaining predictors. Speech rate appears to be the main contributor for the model, followed by the regularity of speech rate, percentage of speech, and speech fluidity. The Kolmogorov-Smirnov test indicates that the distribution of the model residuals is not statistically different from a normal distribution ($P = .20$).

As the long-term objective of the present work is to create a CAPT tool for the automatic assessment of the fluency of individual L2 learners, prediction scores were aggregated for the 12 learners/levels involved in the study. Figure 1 illustrates the strong and positive relationship between the human fluency ratings and automatic scores ($R = .92; P < .001$).

However, in the context of L2 teaching, time constraints are often decisive, and the number of sentences used in this study for each learner/level ($n = 21$) might not be realistic in this regard. One question is thus to determine the amount of sentences needed to get reliable estimates of speech fluency.

In order to adress this point, correlations between automatic and human fluency ratings were computed for a varying number n of sentences considered for each learner/level (with

Table 4: *Standardized (beta) coefficients of predictors found by the multiple linear regression*

Predictor	β-coefficient	P-value
Estimator of speech rate	0.61	$< .001$
Est. of speech rate regularity	-0.19	.001
Percentage of speech	0.17	$< .001$
Est. of speech fluidity	-0.15	$< .001$

$1 \leq n \leq 21$). For each n number of sentences, the index of the first sentence considered varied from 1 to 21. As a consequence, for each number n of sentences, 21 correlations were calculated.

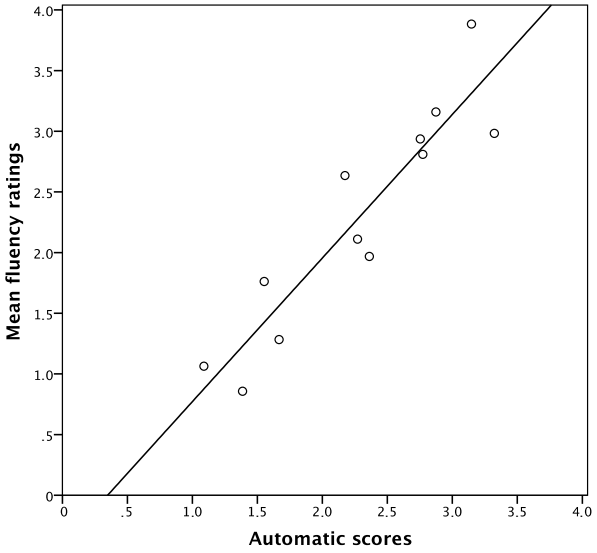

Figure 1: *Scatterplot relating, for the 12 learners/levels, mean human fluency ratings and automatic scores (equation for the regression line:* $y = 1.18x - 0.41$*;* $R = .92, P < .001$*)*

Figure 2 shows the mean coefficients of determination R^2 for each n number of sentences considered per learner/level, and the associated standard deviations. As can be seen, the mean coefficient of determination tends to follow a logarithmic growth as a function of the number of sentences considered. For $n > 4$, the mean R^2 value exceeds .80, which may be considered as a strong coefficient of determination.

6. Discussion

The main objective of this study was to use automatic and "low-level" acoustic measures in order to predict experts' fluency ratings for speech produced by Japanese learners of French. In this regard, the results are very encouraging, with strong correlations observed between automatic and human ratings. The two most contributive predictors are the automatic estimators of speech rate and of regularity of speech rate, both derived from the results of the Forward-Backward Divergence Segmentation (FBDS) algorithm [18].

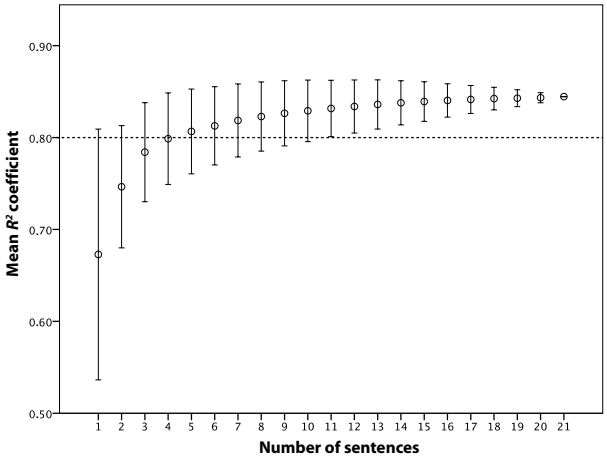

Figure 2: *Mean* R^2 *coefficient between predicted and observed fluency scores as a function of the number of sentences considered per learner/level. Error bars represent* $\pm$ *one standard deviation.*

Even if the correlations achieved are rather high, the contribution of other acoustic measures could be investigated. For example, the presence of filled pauses due to reading hesitations were not directly taken into account in this study. The presence of filled pauses have certainly influenced the measures of rate of speech and of regularity of speech rate – because filled pauses result in longer FBDS segments – but maybe the prediction of perceived fluency could be enhanced by direct measures of mean length and/or number of filled pauses.

Also, because the tracking of several formants at the same time can be tricky, in this study the speech fluidity measure was limited to the analysis of the first formant curve. A next step could be to integrate the tracking of other formants – which relate to other points of articulation such as tongue and lips – and to measure their benefit for the prediction of speech fluency.

From a more applied point of view, the results show that the measures can be combined to create a predictive system for gaining rapid and objective estimates of speech fluency, achieving a high reliability when considering more than four sentences per speaker. Future work will be devoted to the integration of these measures into a computer-assisted pronunciation training tool designed for Japanese learners of French [10, 11, 21]. However, as only low-level acoustic measures are used, the prediction system might prove useful for the assessment of speakers of other target/source languages, or even for the objective assessment of pathological speech; they could be used as a complement to automatic measures of speech comprehensibility [22] or segmental production [23] for pathologies causing speech fluency disorders such as Parkinson's disease [24].

7. Acknowledgements

This research has been supported by the Japanese Society for the Promotion of Sciences through Grants-in-Aid for Scientific Research (B) No. 23320121 and No. 15H03228 to Sylvain Detey. We wish to thank Yuji Kawaguchi, Mariko Kondo, Corentin Barcat, Isabelle Racine, Jacques Durand, Mito Matsuzawa, Jean-Luc Nespoulous, Tsuyoshi Umeno, Kaori Ohmura, Xavier Aumont, as well as all the students who participated in the study.

8. References

[1] R. I. Thomson, "Fluency," in *The Handbook of English Pronunciation*, M. Reed and J. Levis, Eds. Hoboken, New Jersey: Wiley, 2015, pp. 209–226.

[2] R. Ghanem and O. Kang, "Pronunciation features in rating criteria," in *Assessment in Second Language Pronunciation*, O. Kim and A. Ginther, Eds. London, U.K.: Routledge, 2018, pp. 115–136.

[3] T. M. Derwing and M. J. Munro, *Pronunciation Fundamentals. Evidence-based Perspective for L2 Teaching and Research*. Amsterdam, Netherlands: John Benjamins, 2015.

[4] T. Isaacs and P. Trofimovitch, *Second Language Pronunciation Assessment. Interdisciplinary Perspectives*. Bristol, U.K.: Multilingual Matters, 2017.

[5] O. Kim and A. Ginther, *Assessment in Second Language Pronunciation*. London, U.K.: Routeledge, 2018.

[6] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–999, 2000.

[7] ——, "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech," *The Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2862–2873, 2002.

[8] M. Benzeghiba, R. D. Mori, and O. Deroo, "Automatic speech recognition and speech variability: a review," *Speech Communication*, vol. 49, p. 763–786, 2007.

[9] T. Lustyk, P. Bergl, and R. Cmejla, "Evaluation of disfluent speech by means of automatic acoustic measurements," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1457–1468, 2014.

[10] L. Fontan and M. Le Coz, "Correction automatique d'erreurs de prononciation en L2 : démonstration d'outils logiciels pour les apprenants japonais de FLE," in *Congrès CAP 2017 – Écologie du français et diversité des langues*, Kyoto, Japan, 2017.

[11] L. Fontan, M. Le Coz, S. Detey, C. Domin, and F. Hapel, "Présentation du logiciel d'entrainement à la prononciation CAPT-L2 : aspects phonétiques et lexicaux," in *Traitement automatique de la parole et ressources pour la didactique de l'oral en L2 : variation, corpus, techniques*, Toulouse, France, 2017.

[12] S. Detey, "CLIJAF: corpus longitudinal interphonologique de Japonais apprenants de français. projets Kakenhi (B) n°23320121 & n°15H03227," Japanese Society for the Promotion of Science, Tech. Rep., 2011-2019.

[13] S. Detey, I. Racine, Y. Kawaguchi, and F. Zay, "Variation among non-native speakers: the InterPhonology of Contemporary French," in *Varieties of Spoken French*, S. Detey, J. Durand, B. Laks, and C. Lyche, Eds. Oxford, U.K.: Oxford University Press, 2016, pp. 491–502.

[14] S. Detey and I. Racine, "Towards a perceptually-assessed corpus of non-native French: the InterPhonology of Contemporary French (IPFC) project illustrated with a longitudinal study of Japanese learners' /b-v/ production," *International Journal of Learner Corpus Research*, vol. 3, no. 2, pp. 223–249, 2017.

[15] Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, U.K.: Cambridge University Press, 2001.

[16] J. Durand, B. Laks, and C. Lyche, "La phonologie du français contemporain: usages, variétés et structure," in *Romanistische Korpuslinguistik – Korpora und Gesprochene Sprache/Romance Corpus Linguistics – Corpora and Spoken Language*, C. Pusch and W. Raible, Eds. Tübingen, Germany: Gunter Narr Verlag, 2002, pp. 93–106.

[17] S. Detey, J. Durand, B. Laks, and C. Lyche, Eds., *Varieties of Spoken French*. Oxford, U.K.: Oxford University Press, 2016.

[18] R. André-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 36, no. 1, pp. 29–40, 1988.

[19] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[20] H. M. Sussman, C. T. Byrd, and B. Guitar, "The integrity of anticipatory coarticulation in fluent and non-fluent tokens of adults who stutter," *Clinical Linguistics & Phonetics*, vol. 25, no. 3, pp. 169–186, 2010.

[21] S. Detey, L. Fontan, and T. Pellegrini, "Traitement de la prononciation en langue étrangère: approches didactiques, méthodes automatiques et enjeux pour l'apprentissage," *Traitement Automatique des Langues*, vol. 57, no. 3, pp. 15–39, 2016.

[22] L. Fontan, T. Pellegrini, J. Olcoz, and A. Abad, "Predicting disordered speech comprehensibility from Goodness of Pronunciation scores," in *SLPAT – Satellite worshop of Interspeech '15*, 2015. [Online]. Available: http://www.slpat.org/slpat2015/papers/fontan-pellegrini- olcoz-abad.pdf

[23] T. Pellegrini, L. Fontan, J. Mauclair, J. Farinas, and M. Robert, "The Goodness of Pronunciation algorithm applied to disordered speech," in *Proceedings of Interspeech '14*, 2014, pp. 1463–1467.

[24] A. M. Goberman and M. Blomgren, "Parkinsonian speech disfluencies: effects of l-dopa-related fluctuations," *Journal of Fluency Disorders*, vol. 28, no. 1, pp. 55–70, 2003.