Predicting speech fluency in children using automatic acoustic features

Lionel Fontan, Shinyoung Kim, Verdiana De Fino, Sylvain Detey

Predicting speech fluency in children using automatic acoustic features

Lionel Fontan[*], Shinyoung Kim[†], Verdiana De Fino[*‡] and Sylvain Detey[†]
[*]Archean LABS, Montauban, France
E-mail: lfontan@archean.tech
[†]GSICCS, Waseda University, Tokyo, Japan
E-mail: joykim0324@akane.waseda.jp, detey@waseda.jp
[‡]IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
E-mail: verdiana.defino@irit.fr

***Abstract*— The present study aims at predicting the speech fluency of children using automatic acoustic measures derived from forward-backward divergence segmentation (FBDS). Thirteen Korean children were recorded while reading out loud a set of sentences. Three native-Korean speakers evaluated the fluency of each sentence on a five-point scale. A FBDS algorithm was used to segment speech recordings into sub-phonemic units and silent segments. In addition to the low-level acoustic features directly derived from FBDS segments, higher-level acoustic features were computed by clustering FBDS segments into pseudo-syllables and silent breaks. Both low- and higher-level features were used to predict average ratings of speech fluency, using a leave-one-speaker-out cross-validation scheme and three regression models: a multiple linear regression, a support vector regression, and a random-forest regressor. Highly accurate predictions were achieved, with average root-mean-square errors (RMSEs) as low as 0.3. Prediction accuracy did not significantly change as a function of regression model. Using higher-level features yielded lower RMSEs than using raw FBDS features. The results of a multiple linear regression using higher-level features (R^2 = 0.94) suggest that speech/silence ratio and pseudo-syllable rate are the two most important predictors of speech fluency.**

I. Introduction

Measures of speech fluency are useful tools for monitoring speech production skills during first (L1) and second language (L2) acquisition, as they indicate the extent to which speech "flows easily without pauses and other disfluency markers" [1]. However, subjective evaluations of speech fluency are time-consuming since raters (e.g., teachers) have no choice but to assess speech productions individually, that is, one speaker at a time. To cope with this issue, early studies proposed using automatic speech recognition (ASR) to compute rapid and objective estimates of speech fluency. For example, in the context of L2 acquisition, ASR systems were used to compute speech rate and speech/silence ratio estimates that were strongly correlated with subjective ratings of speech fluency [2,3].

However, an important drawback of ASR systems is that they can be challenged by noncanonical speech. ASR systems are usually trained using speech samples produced by healthy, native adult speakers, and their recognition accuracy decrease when processing nonnative (L2) speech [4], pathological speech [5], or speech produced by children [6]. Using ASR-based features to predict speech fluency in such populations can thus be hazardous, as word-recognition errors may bias the measurement of speech-fluency predictors (e.g., speech rate can be under- or overestimated depending on the length of the words recognized by the system). To overcome this shortcoming, some authors used low-level, temporal acoustic analyses to directly estimate predictors of speech fluency such as speech/silence ratio, speech rate, rate of silent breaks and presence of hesitations [7-9]. In particular, forward-backward divergence segmentation (FBDS [10]) was successfully used for predicting speech fluency in adult L2 speakers [8-9] as well as in adults with speech disorders [11]. The FBDS algorithm detects significant changes in the trajectory of the signal energy over time (e.g., abrupt increases or decreases in signal energy) and, when applied to speech, results in a subphonemic segmentation of the input signal. To measure speech fluency, authors either used low-level predictors directly based on FBDS segments (e.g., number of FBDS speech segments per second as a measure of speech rate [8,9]) or predictors that required a clustering of FBDS segments into higher-level units such as silent breaks and pseudo-syllables [12] (e.g., pseudo-syllable rate [13-15]).

However, as none of the previous studies used both low-level and higher-level predictors, it is not clear if the clustering of FBDS segments into higher-level units provide any significant benefit for the prediction of speech fluency. Also, another limit of previous studies is that only linear combinations of features (e.g., multiple linear regressions) were used to predict speech fluency. Yet, it is possible that the importance of a given predictor depends on other predictors (e.g., one could assume that the presence of silent breaks has a lesser impact on speech fluency when speech rate is already very low) — in which case using nonlinear models could improve the prediction accuracy. Finally, so far, the FBDS algorithm was only used to predict speech fluency in *adult* speakers, despite the numerous clinical and educational applications that could be developed for children using such

techniques (e.g., for the automatic assessment of children literacy during read-aloud tasks).

Therefore, the main aim of the present study is to establish the proof-of-concept that the FBDS algorithm can be used to accurately predict speech fluency in children. A second objective of the study is to overcome the methodological limits of previous research works by 1) determining whether the clustering of FBDS segments into pseudo-syllables and silent breaks significantly improves the prediction of speech fluency by comparison with the use of low-level FBDS features alone, and 2) if nonlinear regression models yield more accurate predictions of speech fluency than a multiple linear regression. A final contribution of this work is to analyze regression results to determine which features contribute the most to the prediction of speech fluency.

II. Subjective Ratings of Speech Fluency

A. Speech material

In order to collect reference data bearing enough variability in terms of speech fluency, 13 Korean children (6 female) of different ages and different levels of language exposure were recruited. Their age ranged from 9 to 12 years (*mean*: 10 years and 11 months; *standard deviation*, *SD*: one year). All were born and raised in Tokyo (Japan) under Korean parents and were attending local Japanese schools at the time of their participation. They all used Korean and Japanese, with different amounts of each language used at home depending on each family.

During the recording task, the children were seated in a quiet room in their homes, at their desks. They were instructed to read out loud an excerpt from *The Giving Tree* [16], a children's book translated in Korean, which consisted of 9 sentences.

The children were recorded using an ECM-MS957 electret condenser microphone (SONY, Tokyo, Japan). The microphone was placed in front of each child, at a distance of approximately 20 cm. Prior to the study, which complied with the ethical guidelines of Waseda University (Tokyo, Japan), all children's parents provided their informed consent.

B. Rating procedure

The first five sentences pronounced by the children (*mean length of the sentences*: 7.6 words; *SD*: 3.6) were used for the fluency rating task, for a total of 65 sentences (5 sentences × 13 children). Three female raters, all of them native speakers of Korean, participated in the assessment. Their age was comprised between 25 and 28 years. None of them reported any history of hearing difficulties.

Speech recordings were presented to each rater in a random order, using the Prodigy software (ExplosionAI GmbH, Berlin, Germany) version 1.11.7. Each rater assessed twice the whole 65 speech recordings, using a five-point scale in which 1 and 5 corresponded to the lowest and highest degree of speech fluency, respectively.

Prior to the rating procedure, the raters were familiarized with the concept of speech fluency. In particular, they were made aware that their ratings of speech fluency should not be influenced by the ability of the speakers to pronounce Korean phonemes — a poor fluency being possibly associated with a correct segmental production, and vice-versa. The raters also listened to several examples of utterances illustrating the whole fluency range. These examples were not part of the material later used in the rating task.

III. Prediction of Speech Fluency

To predict ratings of speech fluency, the audio recordings were first segmented using FBDS. Then, FBDS segments were clustered into silent breaks and pseudo-syllables, and low-level and higher-level predictors were computed. Finally, three regression models were used to predict speech fluency ratings. The whole procedure is detailed in the following subsections.

A. Automatic segmentation of speech signals

Three steps were used to segment the speech signals into FBDS segments, pseudo-syllables and silent breaks (Fig. 1). During the first step, FBDS was applied. The principle of FBDS is as follows. The FBDS algorithm uses two windows: a short-term analysis window, and a longer-term window used as a buffer to store the current FBDS segment. The signals contained in the two windows are each modeled by an auto-regressive Gaussian model. As the short-term window is used to progressively scan the input signal, the distance between the two Gaussian models (measured by the Kullback-Leibler divergence measure [17]) is used to detect segment boundaries. Each time a boundary is detected, the long-term window buffer is flushed, and the analysis goes forward. This process is eventually carried out backwards, and the boundaries found during forward and backward analyses are merged together.

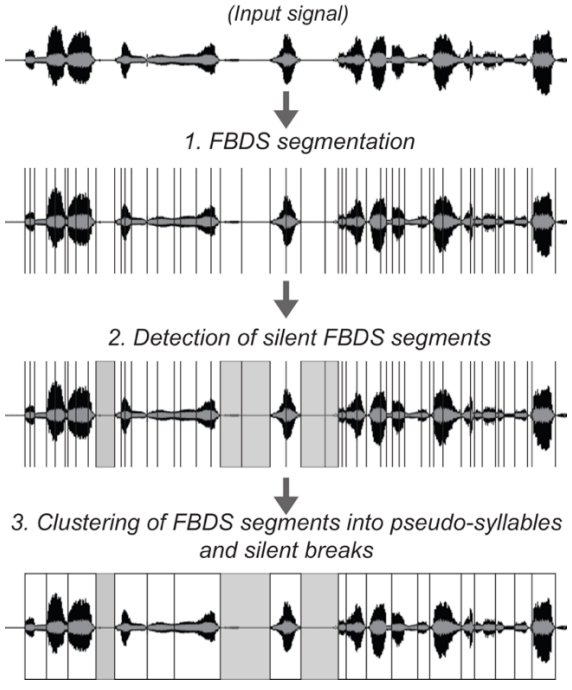

Fig. 1. Processing steps used to automatically segment speech signals into pseudo-syllables (transparent rectangles shown at step 3) and silent breaks (gray-shaded rectangles shown at step 3).

In a second step, the segments found by the FBDS algorithm were classified as either speech segments or silent segments, depending on their maximum energy exceeding 4% of the maximum energy found in the whole recording (in which case the segment was classified as speech), or not (in which case the segment was classified as silence).

In the third step, FBDS speech segments and FBDS silent segments were clustered into pseudo-syllables and silent breaks, respectively. Two consecutive FBDS speech segments were considered as part of the same pseudo-syllable if the average signal energy did not decrease more than a given ratio when switching from the first segment to the next. To define silent breaks, consecutive silent FBDS segments were merged together. All resulting silence clusters that lasted more than 250ms were considered as silent breaks, as this threshold was shown to be optimal for the measurement of speech fluency [20]. Fig. 2 shows, as an example, the results of the automatic segmentation of a speech signal corresponding to the three first words of the target text pronounced by one of the children.

B. Computation of predictors of speech fluency

Both low-level and higher-level predictors of speech fluency were computed. Low-level predictors were computed using FBDS speech segments and FBDS silent segments, that is, using the information directly available after the step 2 shown in Fig. 1. These low-level predictors included:

- Rate of FBDS speech segments: the number of FBDS speech segments divided by the duration of the recording. This feature provides an estimate of speech rate, and is thus expected to be positively correlated with speech fluency;
- Standard deviation of FBDS speech segments duration. As the standard deviation of speech segments increases with the presence of hesitations (filled pauses, i.e., sustained vowels), this feature is expected to be negatively correlated with speech fluency;
- Rate of FBDS silent segments: the number of FBDS silent segments divided by the duration of the recording. It is assumed that the higher the rate of FBDS silents, the lower the speech fluency;
- Speech ratio: the total duration of FBDS speech segments, divided by the duration of the recording. A high speech ratio is supposed to be indicative of a high speech fluency, and vice-versa.

Higher-level predictors were computed based on the information available after the clustering of FBDS segments into pseudo-syllables and silent breaks (step 3 in Fig. 1). Higher-level predictors included:

- Rate of pseudo-syllables: the number of pseudo-syllables divided by the duration of the recording;
- Standard deviation of pseudo-syllable duration;
- Rate of silent breaks: the number of silent breaks divided by the duration of the recording.

Consistent with the assumptions made for low-level features, the rate of pseudo-syllables is expected to be positively correlated with speech fluency, whereas standard deviation of pseudo-syllables and rate of silent breaks are assumed to be negatively correlated with speech fluency.

C. Application of regression models

Three regression models were finally used to predict speech fluency ratings: a multiple linear regression (MLR), and two nonlinear models — a support vector regression (SVR) using a Radial Basis Function, and a random forest regressor (RFR). All models were implemented and evaluated using the *Scikit Learn* Python library [19], version 0.24.2.

Two sets of predictors were used with each model (Table 1). The *low-level* predictor set contained the three predictors that were calculated based on FBDS segments. The *higher-level* predictor set contained the three predictors that were calculated based on pseudo-syllables and silent breaks. Both sets contained speech ratio, as this measure does not change when calculated using low-level or higher-level units (i.e., using FBDS speech segments or pseudo-syllables).

As the number of speech samples did not allow for the creation of a separate validation set, hyperparameters of the SVR (regularization parameter, gamma, and epsilon) and RFR

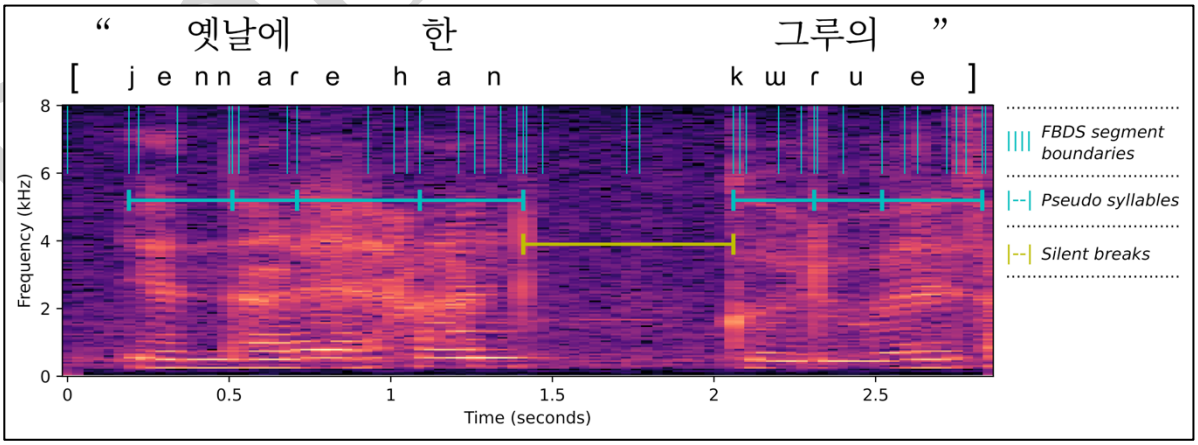

Fig. 2. Spectrogram corresponding to the first three words of the target text (“옛날에 한 그루의”– “Once upon a time there was a (tree)”) pronounced by one of the children, along with the corresponding phonetic transcription based on the International Phonetic Alphabet, and the FBDS segment boundaries, pseudo-syllables and silent breaks that were automatically identified.

TABLE I
PREDICTORS USED IN THE LOW-LEVEL AND HIGHER-LEVEL SETS

Low-level set	Higher-level set
Speech ratio	
Rate of FBDS speech segments	Rate of pseudo-syllables
Standard deviation of FBDS speech segment duration	Standard deviation of pseudo-syllable duration
Rate of FBDS silent segments	Rate of silent breaks

(number of trees and maximum depth of the trees) models were tuned using a nested cross-validation procedure.

Both the outer and inner cross-validation loops followed a leave-one-speaker-out (LOSO) scheme. The results of the inner loop were used to select optimal parameters. These parameters were fixed during the final LOSO cross-validation that was carried out to compare the performance of the three models.

IV. RESULTS

A. Reliability of subjective ratings of speech fluency

To assess intra-rater reliability, Spearman's rank correlation coefficients (*rhos*) and Cronbach's alphas (α) were computed on the two series of ratings provided by each rater for the 65 sentences (Table I). For the three raters, very strong correlations (all *rhos* ≥ 0.89), and very high alphas (all ≥ 0.95) are observed, indicating an excellent reliability.

To assess inter-rater reliability and agreement, the two series of ratings provided by each rater were first averaged. The results indicate a very high reliability, with a Cronbach's α equal to 0.97, and Spearman correlation coefficients ≥ 0.86 (Table II).

Inter-rater agreement was finally assessed by analyzing the distribution of fluency ratings (Table III). The aim of this analysis was to determine if all three raters used "the same yardstick" [2] when evaluating speech fluency — case in which a straightforward combination of ratings (i.e., the computation of mean fluency ratings across raters) could be carried out to produce the final reference ratings.

TABLE I
INTRA-RATER RELIABILITY (SPEARMAN'S *RHO* AND CRONBACH'S α)

Rater	*rho*	α
1	0.95 ($p < 0.001$)	0.98
2	0.89 ($p < 0.001$)	0.96
3	0.89 ($p < 0.001$)	0.95

TABLE II
INTER-RATER RELIABILITY: SPEARMAN CORRELATION COEFFICIENTS (*RHOS*) BETWEEN SPEECH FLUENCY RATINGS GIVEN BY EACH PAIR OF RATERS, AND ASSOCIATED ONE-TAILED P-VALUES

	Rater 2	Rater 3
Rater 1	*rho* = 0.88 ($p < 0.001$)	*rho* = 0.88 ($p < 0.001$)
Rater 2		*rho* = 0.86 ($p < 0.001$)

TABLE III
AVERAGE RATINGS ($\bar{x}$) AND ASSOCIATED STANDARD DEVIATION (*SD*) FOR EACH RATER

Rater	$\bar{x}$	*SD*
1	3.84	1.42
2	3.66	1.36
3	3.82	1.35

The descriptive statistics provided in Table III show that the distribution of fluency ratings is very similar across raters, with a maximum difference between mean ratings equal to 0.18 (for raters 1 and 2). As a Kruskal-Wallis test confirmed that there was no significant difference between the distributions of the ratings provided by the three raters ($H(2) = 1.04$; $p = 0.6$), the ratings were eventually averaged across raters. The final 65 speech-fluency ratings ranged from 1 to 5, with a mean value of 3.8 ($SD = 1.4$).

B. Bivariate correlations between individual predictors and speech fluency

The relationship between each predictor and average fluency ratings was assessed through nonparametric (Spearman) correlations (Table IV), as a Kolmogorov-Smirnov test indicated that the ratings were not normally distributed ($p < 0.01$). As was assumed, measures of speech rate and speech ratio are positively correlated with ratings of speech fluency: the higher the fluency, the faster the speech rate and the higher the speech ratio. Standard deviations of speech segments (FBDS speech segments and pseudo-syllables) are, on the contrary, negatively correlated with ratings of speech fluency. The same is true for the two measures of the rate of silences (FBDS silent segments and silent breaks): the higher the rate of silent segments, the lower the speech fluency. Altogether, the strongest correlations are observed for speech ratio, rate of silent breaks and rate of pseudo-syllables (all absolute *rhos* ≥ 0.88).

C. Prediction of speech fluency

Table V shows the average root-mean-square errors (RMSEs) found for each predictor set and regression model.

TABLE IV
SPEARMAN'S CORRELATION COEFFICIENT (*RHO*) BETWEEN EACH PREDICTOR AND RATINGS OF SPEECH FLUENCY, AND ASSOCIATED ONE-TAILED *P*-VALUE

Predictor	*rho*	*p*-value
Rate of FBDS speech segments	0.85	< 0.001
Rate of pseudo-syllables	0.88	< 0.001
Standard deviation of FBDS speech segment duration	–0.60	< 0.001
Standard deviation of pseudo-syllable duration	–0.37	0.001
Rate of FBDS silent segments	–0.66	< 0.001
Rate of silent breaks	–0.89	< 0.001
Speech ratio	0.89	< 0.001

TABLE V
AVERAGE ROOT-MEAN-SQUARE ERROR (RMSE) AND ASSOCIATED STANDARD DEVIATION AS A FUNCTION OF PREDICTOR SET AND REGRESSION MODEL

Predictor set	Regression Model	Average RMSE	Standard deviation
Low-level set	MLR	0.46	0.11
	SVR	0.41	0.17
	RFR	0.36	0.12
Higher-level set	MLR	0.35	0.11
	SVR	0.30	0.13
	RFR	0.37	0.12

In general, predictions of speech fluency are accurate, with average RMSEs close to, or even inferior to one tenth of the speech-fluency scale (i.e., to 0.4). This accuracy is higher than in previous studies that used comparable methods to predict the speech fluency of adult L2 learners, in which the lowest RMSE was 0.51 [13-15]. In the present study, the smallest RMSEs are found for the SVR model and the higher-level predictor set.

D. Effects of predictor set and regression type on prediction accuracy

As can be observed from Table V, for MLR and SVR models, the average RMSE tends to be lower when using higher-level predictors. The RFR model follows a different trend with a slight increase of the average RMSE when switching from low-level to higher-level predictors. Depending on the predictor set, the ranking of the regression models in terms of accuracy changes (RFR > SVR > MLR and SVR > MLR > RFR for the low-level and higher-level predictor sets, respectively).

To assess the statistical significance of the RMSE differences observed as a function of the predictor set and regression model, a linear mixed model was computed using SPSS Statistics version 23.0 (IBM, Armonk, NY). *Predictor set* and *Regression model* factors were treated as fixed effects, while *Speaker* factor (i.e., the speaker for whom the prediction was made in the LOSO setup) was considered as a random effect. The results indicate that *Predictor set* has a significant effect on the RMSE ($F(1, 72) = 5.3$; $p = 0.02$), contrary to *Regression model* ($F(2, 72) = 1.2$; $p = 0.3$). The interaction between *Predictor set* and *Regression model* just falls below the 5% level of significance ($F(2, 72) = 3.1$; $p = 0.049$).

E. Individual contribution of higher-level predictors

To evaluate the role of each higher-level feature for the prediction of speech fluency, a MLR was computed on the whole dataset (Fig. 2). The very high coefficient of determination indicates that the model explained 94% of the variability in fluency ratings. The standardized coefficients (*std. coef.*) of the regression equation show that *speech ratio* and *rate of pseudo-syllable* are the two most important predictors of speech fluency (*std. coefs*. of 0.52 and 0.49, respectively). *Rate of silent breaks* had a smaller predictive power (*std. coef.* = –0.31), and *standard deviation of pseudo-syllable duration* only played a marginal role for the prediction of speech fluency (*std. coef.* = –0.08).

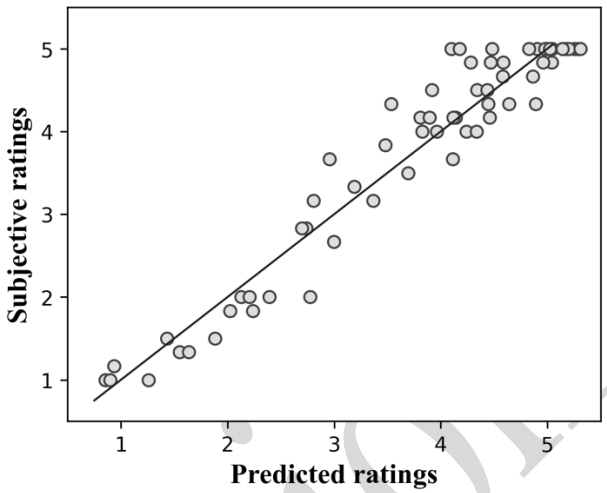

Fig. 2. Scatterplot relating average subjective ratings to automatic measures of speech fluency obtained with the MLR model, and associated regression line ($R^2 = 0.94$; *RMSE* = 0.34).

V. CONCLUSIONS

The present study demonstrated that FBDS-based automatic acoustic measures can be used to achieve highly accurate predictions of read-speech fluency in children. The study also demonstrated that clustering FBDS segments into higher-level units — namely, pseudo-syllables and silent breaks — significantly improves the prediction of speech fluency by comparison with the use of FBDS segments alone. This is all the more interesting since the method used for clustering FBDS segments into pseudo-syllables is theoretically relevant for all the languages in which syllables consist of vocalic nuclei, that is, for the vast majority of existing languages [20].

Another assumption was that using a nonlinear regression model might yield more accurate predictions than an MLR. The results failed to validate this hypothesis. This is not very surprising when considering that the MLR model could already account for 94% of the variance in speech-fluency ratings and achieved very low RMSEs (0.46 and 0.30 using low- and higher-level predictors, respectively), which left very little room for improvement. The analysis of the MLR coefficients showed that the two most important predictors of speech fluency were speech ratio and speech rate.

Altogether, the results of this study open the way to the use of FBDS-derived measures of speech fluency for numerous applications dedicated to children, whether educational (e.g., to assess reading skills at school [21]), clinical (e.g., to diagnose and monitor fluency disorders in children [22]), or scientific (e.g., to investigate the factors influencing the development of speech fluency during language acquisition [23]).

VI. ACKNOWLEDGMENTS

The authors are indebted to Prof. Mariko Kondo who helped with data collection, as well as to Dr. Julien Eychenne for his assistance with rater recruitment and his useful insights on Korean phonology. The authors also thank Dr. Saïd Jmel for his helpful statistical advices.

References

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

[2] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *J. Acoust. Soc. Am.*, vol. 107, no. 2, pp. 989–999, Jan. 2000, doi: 10.1121/1.428279.

[3] ——, "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech," *J. Acoust. Soc. Am.*, vol. 111, no. 6, pp. 2862–2873, Jun. 2002, doi: 10.1121/1.1471894.

[4] N.T. Vu, Y. Wang, M. Klose, Z. Mihaylova, and T. Schultz (2014), "Improving ASR performance on non-native speech using multilingual and crosslingual information," in *Proc. INTERSPEECH '14: 15th Proc. Annu. Conf. Int. Speech Commun. Assoc.,* Singapore, Sep. 2014, pp. 11–15, doi: 10.21437/Interspeech.2014-3.

[5] L. De Russis and F. Corno, "On the impact of dysarthric speech on contemporary ASR cloud platforms," *J. Reliab. Intell. Environ.*, vol. 5, no. 3, pp. 163–172, 2019, doi: 10.1007/s40860-019-00085-y.

[6] L. Gelin, M. Daniel, J. Pinquier, and T. Pellegrini, "End-to-end acoustic modeling for phone recognition of young readers," *Speech Commun*., vol. 134, pp. 71–84, Nov. 2021, doi: 10.1016/j.specom.2021.08.003.

[7] T. Lustyk, P. Bergl, and R. Cmejla, "Evaluation of disfluent speech by means of automatic acoustic measurements," *J. Acoust. Soc. Am.*, vol. 135, no. 3, pp. 1457–1468, 2014, doi: 10.1121/1.4863646.

[8] L. Fontan, M. Le Coz, and S. Detey, "Automatically measuring L2 speech fluency without the need of ASR: A proof-of-concept study with Japanese learners of French," in *Proc. INTERSPEECH '18: 19th Proc. Annu. Conf. Int. Speech Commun. Assoc.,* Hyderabad, India, Sep. 2018, pp. 2544–2548, doi: 10.21437/Interspeech.2018-1336.

[9] S. Detey, L. Fontan, M. Le Coz, and S. Jmel, "Computer-assisted assessment of phonetic fluency in a second language: a longitudinal study of Japanese learners of French," *Speech Commun.*, vol. 125, pp. 69–79, Dec. 2020, doi: 10.1016/j.specom.2020.10.001.

[10] R. André-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. Signal Process.*, vol. 36, no. 1, pp. 29–40, Jan. 1988, doi: 10.1109/29.1486.

[11] D. Sztahó, D. and I. Valálik, I., "Speech Fluency Measurement of Patients with Parkinson's Disease by Forward-Backward Divergence Segmentation," in *Proc. CogInfoCom*, Naples, Italy, Oct. 2019, pp. 295–300, doi: 10.1109/CogInfoCom47531.2019.9090001.

[12] J. Farinas and F. Pellegrino, "Automatic rhythm modeling for language identification," in *Proc. Eurospeech*, Aalborg, Denmark, Sep. 2001, pp. 2539–2542.

[13] L. Fontan, M. Le Coz, and M. Kondo, "Building an ASR-free automatic tool for measuring the speech fluency of Japanese learners of English," presented at the 9th Int. Symp. Acquis. Sec. Lang. Speech – NEW SOUNDS '19*,* Tokyo, Japan, Aug. 30–Sep. 1, 2019.

[14] M. Kondo, L. Fontan, M. Le Coz, T. Konishi, and S. Detey, "Phonetic fluency of Japanese learners of English: automatic vs native and non-native assessment," in *Proc. SPEECH PROSODY '20: 10th Int. Conf. Speech Prosody*, Tokyo, Japan, May/Aug. 2020, pp. 784–788, doi: 10.21437/SpeechProsody.2020-160.

[15] L. Fontan, M. Le Coz, and C. Alazard, "Using the forward-backward divergence segmentation algorithm and a neural network to predict L2 speech fluency," in *Proc. SPEECH PROSODY '20: 10th Int. Conf. Speech Prosody*, Tokyo, Japan, May/Aug. 2020, pp. 925–929, doi: 10.21437/SpeechProsody.2020-189.

[16] S. Silverstein, *The giving tree*. New York, NY, USA: Harper, 1964.

[17] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat*, vol. 22, no. 1, pp. 79–86, Mar. 1951, doi: 10.1214/aoms/1177729694.

[18] N. H. De Jong and H. R. Bosker, "Choosing a threshold for silent pauses to measure second language fluency," in *Proc. DiSS '13: 6th Workshop on Disfluency in Spontaneous Speech*, Stockholm, Sweden, Aug. 2013, pp. 17–20.

[19] F. Pedregosa *et al*., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, 2825–2830, Oct. 2011.

[20] C Anderson, B. Bjorkman, D. Denis, J. Doner, M. Grant, N. Sanders, and A. Taniguchi. *Essentials of Linguistics*, 2nd ed. Toronto, Canada : eCampusOntario.

[21] S. H. Chae and M. Kim, "Semantic/phonological error characteristics and reading fluency for school-aged children with language learning disabilities in 3rd to 4th Grades," *Commun Sci Disord.*, vol. 24, no. 2, 387–401, Jun. 2019, doi: 10.12963/csd.19598.

[22] K. Eggers, K. and M. M. Leahy, *Clinical Cases in Dysfluency*. Abingdon-on-Thames, United Kingdom: Taylor & Francis, 2022.

[23] N. H. De Jong, "Fluency in second language testing: Insights from different disciplines," *Lang. Assess. Q*., vol. 15, no.3, pp. 237-254, Jun. 2018, doi: 10.1080/15434303.2018.1477780.