10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Using the forward-backward divergence segmentation algorithm and a neural network to predict L2 speech fluency

Lionel Fontan[1], *Maxime Le Coz*[1], *Charlotte Alazard*[2]

[1]Archean LABS, Montauban, France
[2]Octogone-Lordat, EA 4156, Toulouse, France
lfontan@archean.tech, mlecoz@archean.tech,charlotte.alazard@univ-tlse2.fr

Abstract

This study aims at developing an automatic system for measuring speech fluency in a second language (L2). Eighteen learners of French, all of them native speakers of English, were recorded during read-aloud tasks in French. Six native-French speakers with a background in L2 acquisition and phonetics rated the recordings in terms of speech fluency.

Automatic measures of speech fluency were computed following four consecutive steps. First, (1) the forward-backward divergence segmentation (FBDS) algorithm was used to segment speech recordings into subphonemic units. Then, (2) the FBDS-derived segments were automatically clustered into higher-level units: pseudo syllables and silent breaks. (3) Four predictors of speech fluency were computed: pseudo-syllable rate, standard deviation of pseudo-syllable duration, rate of silent breaks, and percentage of speech. Finally, (4) the four predictors were combined together using either a multiple linear regression (MLR) or a neural network (NN) to predict human ratings of speech fluency.

A very strong correlation ($R = 0.89$) between the NN-based automatic scores and the average human ratings are achieved. The correlation coefficient achieved with the MLR is significantly lower ($R = 0.85$), but a ten-fold cross-validation indicates similar performances for the two models with regards to their behavior on unknown data.

Index Terms: speech fluency, L2, automatic assessment, forward-backward divergence segmentation, neural network

1. Introduction

Speech fluency can be defined as “the degree to which speech flows easily without pauses and other disfluency markers” [1]. In the context of second-language (L2) learning, measures of speech fluency can be used to monitor the learners’ production skills during both spontaneous oral interactions and read-aloud tasks; however, subjective assessments of speech fluency are time-consuming and prone to inter-rater variation [2].

Several attempts have been made to develop objective measures of speech fluency. In [3] and [4], automatic speech recognition (ASR) systems were used to calculate acoustic-phonetic predictors of perceived speech fluency such as speech rate, phonation/time ratio, and frequency of silent breaks. Strong correlations were observed between ASR-derived measures and native speakers’ ratings of speech fluency.

However, using ASR systems to measure L2 speech fluency present several disadvantages. First, the acoustic models used in ASR systems are usually trained (i.e., built) with native speech. As a consequence, ASR systems may be challenged by non-native speech, resulting in a higher rate of recognition errors [5] that might impact ASR-derived measures of speech fluency. The second and more obvious limit of ASR systems is that they are language-dependent. For example, the ASR systems used in [3] and [4] for the prediction of speech fluency in Dutch as a foreign language cannot be used for another target language.

Recently, several studies overcame these shortcomings by using ASR-free, low-level, purely acoustic measures to assess the speech fluency of speakers with motor-speech disorders [6] and of L2 learners [7,8]. In [7] and [8], the forward-backward divergence segmentation (FBDS) algorithm [9] was used to automatically segment L2 speech recordings. The FBDS results were used to compute several predictors of speech fluency, such as the rate of FBDS-derived segments or the standard deviation of the duration of FBDS-derived segments. Eventually, the authors used a multiple linear regression to combine the automatic measures and predict the speech-fluency performances of eight Japanese learners of French, as evaluated by three native-French speakers. The system could predict quite accurately human speech fluency ratings, with a mean R^2 in excess of 0.80 when considering more than four sentences per speaker.

The first goal of the present study is to extend this proof-of-concept to another population: native-English speakers of French as a foreign language. Also, one limit in [7] and [8] was that, as the FBDS algorithm results in a subphonemic segmentation of speech signals [9], the FBDS-derived predictors of speech fluency were not easily interpretable — in the sense that the segments they are based on differ from usual phonetic units such as phonemes or syllables. In the present study, we propose a slightly different approach consisting in (1) computing FBDS-derived segments and (2) clustering them into two phonetically relevant units: (pseudo) syllables and silent breaks. Several predictors are then derived from these units (e.g. rate of pseudo syllables, rate of silent breaks) and combined together to predict human ratings of speech fluency. Finally, a last goal of this study is to assess the accuracy of the speech-fluency prediction system when performing a linear combination of the automatic predictors (i.e., when using a multiple linear regression, as in [7] and [8]) and when using a neural network.

10.21437/SpeechProsody.2020-189

2. Human ratings of speech fluency

2.1. Speech materials

Eighteen English-speaking learners of French (15 females; age range: 20-60 years; mean age: 32 years), who participated in a 8-week-long phonetic training course in Toulouse, France, were recruited for this study. Based on the common European framework of reference for languages [10], the teacher in charge of the training estimated that, at the start of the course, half of the learners had a beginner level in French (either A1 or A2), whereas the other half had an advanced level (B2).

All learners were recorded three times: at the beginning (T0), at the midterm (T0 + 4 weeks), and at the end (T0 + 8 weeks) of the phonetic training course. Each time, the learners were required to read aloud a text. Different texts were used for each recording session (i.e., for T0, T+4 and T+8); the tests were also selected from pedagogical materials in accordance with the learners' levels in French (i.e., different texts were used for beginner and advanced learners). The recordings took place in a double-walled sound-isolation booth (ambient noise level: 28 dB, A-weighted), using a computer, a TASCAM DM-3200 digital mixing console (TEAC Corporation, Tokyo, Japan), and an omnidirectional Sennheiser MD46 microphone (Sennheiser, Wedemark, Germany). For each learner and recording session, only one sentence (containing from 13 to 24 words; mean: 18 words) was used for the human ratings of speech fluency, for a total of 54 sentences (18 learners x 3 recording sessions).

2.2. Participants and rating procedure

Six native-French speakers (four females; age-range : 22-25 years) participated in the rating task. They all had a background in phonetics and second-language acquisition. Each rater completed the task individually in a quiet room. The software "Presentation" (Neurobehavioral Systems, Berkeley, USA), running on a PC, was used to present the speech recordings. The raters listened to the recordings through Sennheiser HD380 pro headphones (Sennheiser, Wedemark, Germany).

The rating task was split into three runs separated by 2mn breaks; during each run, all 54 sentences were presented to the raters in a random order. Each sentence was therefore presented three times to each rater, for a total of 162 presentations. After each presentation, the raters used a Cedrus RB-730 response pad (Cedrus, San Pedro, USA) to give a speech fluency (SF) rating on a 5-point scale ranging from 1—"Very poor speech fluency" to 5—"As fluent as speech produced by a native speaker".

Table 1: *Spearman correlation coefficients between each pair of ratings (R) given by each rater for the same speech recordings.*

	R#1, R#2	**R#2, R#3**	**R#1, R#3**
Rater #1	0.72**	0.60**	0.55**
Rater #2	0.67**	0.76**	0.66**
Rater #3	0.87**	0.83**	0.82**
Rater #4	0.71**	0.72**	0.72**
Rater #5	0.72**	0.71**	0.70**
Rater #6	0.62**	0.80**	0.74**

** $p < 0.01$, one-tailed

2.3. Intra-rater agreement and reliability

To check the intra-rater agreement and reliability, the three different SF ratings that were given by each rater for the same speech recordings were considered. Spearman correlations were computed for each couple of ratings (i.e., {rating #1, rating #2}, {rating #2, rating #3}, and {rating #1, rating #3}) given for the same recordings. As can be seen in Table 1, intra-rater correlation coefficients are highly significant and range from 0.55 to 0.87 (mean: 0.72).

The intra-rater reliability was then assessed by computing Cronbach's α coefficient for the different SF ratings (R#1, R#2, and R#3) given by each rater (Table 2).

Table 2: *Intra-rater reliability (Cronbach's α).*

Rater	Cronbach's α
Rater #1	0.83
Rater #2	0.88
Rater #3	0.94
Rater #4	0.89
Rater #5	0.91
Rater #6	0.89

Cronbach's α values are high to very high, ranging from 0.83 to 0.94 (mean: 0.89). Based on the strong-to-very-strong intra-rater correlations and on the high-to-very-high intra-rater reliability, we decided to average SF ratings for each rater and speech recording. These average ratings were then used to assess inter-rater agreement and reliability.

2.4. Inter-rater agreement and reliability

Inter-rater agreement was assessed by computing Spearman correlations between the SF ratings given by each couple of raters for the 54 sentences. The results (Table 3) show highly significant correlations with coefficients ranging from 0.78 to 0.92 (mean: 0.84).

Table 3: *Spearman correlation coefficients between speech fluency ratings given by each pair of raters.*

	Rater #2	**Rater #3**	**Rater #4**	**Rater #5**	**Rater #6**
Rater #1	0.79**	0.86**	0.82**	0.78**	0.79**
Rater #2		0.84**	0.87**	0.78**	0.82**
Rater #3			0.89**	0.86**	0.92**
Rater #4				0.85**	0.88**
Rater #5					0.90**

** $p < 0.01$, one-tailed

Finally, the inter-rater reliability was assessed by computing the Cronbach's α coefficient with the six raters' SF scores. The resulting α is very high, with a value of 0.97. As a consequence, we decided, for every speech recording, to average the six raters' SF scores. According to a Kolmogorov-Smirnov test, the resulting average SF ratings are normally distributed (p = 0.2); they range from 1.00 up to 4.94, with a mean value of 2.89 and a standard deviation of 0.97.

3. Computation of predictors of speech fluency

3.1. Overall procedure

Several acoustic predictors of SF ratings were automatically computed by following three consecutive steps:

1. The segmentation of speech signals, using the FBDS algorithm;

2. The clustering of FBDS-derived segments into pseudo syllables and silent breaks;

3. The computation of SF predictors, based on the rate and duration of pseudo syllables and silent breaks.

3.2. Automatic segmentation of speech signals using the FBDS algorithm

The FBDS algorithm is designed to split an acoustic signal into segments bounded by articulatory changes [9]. To this end, the algorithm uses two analysis windows: (1) a short-term sliding window (in the present study, a 10-ms-wide window) for processing current signal information, and (2) a long-term "growing" window that models all the signal that was analyzed since a segment boundary was last identified.

The signals contained in both windows are modelled using two autoregressive Gaussian models of the form:

$$\begin{cases} y_n = \sum_{i=1}^{p} a_i y_{n-1} + e_n \\ var(e_n) = \sigma^2 \end{cases} \quad (1)$$

where y represents the acoustic signal, and e a Gaussian white noise.

Each time the short-term sliding window moves a step forward, the divergence between the two autoregressive models is measured using the Kullback-Leibler distance [11]. If the divergence exceeds a given threshold, then a segment boundary is set, the long-term window is flushed and the analysis continues forward.

As this analysis is not symmetric, it is also conducted backwards. The final segmentation of the signal is determined by the union of the boundaries found during the forward and backward analyzes. When applied to speech signals, the FBDS algorithm results in a subphonemic segmentation [9] (e.g., typically the attack, sustain and release parts of an isolated vowel would be split into three segments).

The FBDS algorithm was applied on all 54 speech files and the segmentation results were then used to detect pseudo syllables and silent breaks.

3.3. Automatic detection of pseudo syllables and silent breaks

To cluster the segments identified by the FBDS algorithm into pseudo syllables and silent breaks, for each recording the FBDS-derived segments were first classified as silent segments (i.e., segments only containing silence) if their average energy was lower than 4% of the maximum energy of the whole recording. In a second step, adjacent non-silent segments were grouped together into pseudo syllables if, when passing from one to another, the mean energy was not decreasing over a given threshold.

3.4. Computation of outcome measures

Eventually, for each recording, the pseudo-syllable and silent-break boundaries were used to calculate four predictors of SF:

- The pseudo-syllable rate (PS_RATE), measured as the number of pseudo syllables divided by the total speech duration (i.e., the duration between the first and last PS, in seconds). This measure is assumed to be representative of the perceived rate of speech, and thus positively correlated with ratings of SF;
- The standard deviation of the duration of the pseudo syllables, in seconds (STD_PS_LENGTH). As this measure should increase with the presence of filled pauses, a negative correlation with ratings of SF is expected;
- The rate of silent breaks (SIL_RATE), measured as the number of silent breaks that exceeded 250ms (this threshold was defined according to the study described in [12]), divided by the total number of words in the recording. As STD_PS_LENGTH, this measure is expected to reflect speech disfluencies and should therefore be negatively correlated with ratings of SF;
- The percentage of speech (PCT_SPEECH), measured as the total duration of pseudo syllables divided by the total duration of the recording, in seconds. As this measure decreases with the presence of silence in the recordings, it is expected to be positively correlated with ratings of SF.

4. Bivariate correlations between automatic predictors and human ratings of speech fluency

To assess the strength and significance of the relationship between each predictor and average human ratings of SF, bivariate correlations were computed. Pearson's correlations were used for all predictors except for the rate of silent breaks, for which a Kolmogorov-Smirnov test indicated a non-normal distribution (p = 0.002); in this latter case, a non-parametric correlation was computed. The results are presented in Table 4.

Table 4: *Pearson's (r) or Spearman's (ρ) correlation coefficients between each predictor variable and average human ratings of speech fluency.*

Predictor	Correlation coefficient
PS_RATE	r = 0.76**
STD_PS_LENGTH	r = -0.42**
SIL_RATE	ρ = -0.80**
PCT_SPEECH	r = 0.66**

** $p < 0.01$, one-tailed

Highly significant, moderate-to-strong correlations are observed for all four predictors. PS_RATE is strongly and positively correlated with human ratings, indicating that, as expected, the higher the rate of pseudo syllables, the more fluent speech is perceived. STD_PS_LENGTH is moderately and negatively correlated with PS ratings, suggesting that the more variation in the duration of pseudo syllables, the lower the ratings of SF. A strong and negative correlation is observed between the rate of silent breaks (SIL_RATE) and ratings of SF, indicating that the presence of silent breaks is associated with lower scores of SF — which is in line with the positive and strong correlation found between PCT_SPEECH and ratings of SF.

5. Prediction of human ratings of speech fluency

In order to compute automatic ratings of speech fluency, two models were created and evaluated using the WEKA framework [13]. First, a multiple linear regression (MLR) was computed with the average human ratings of speech fluency as the dependent variable and the four automatic predictors (PS_RATE, STD_PS_LENGTH, SIL_RATE, and PCT_SPEECH) as independent variables. This model achieves a correlation of 0.85 with human ratings, with a root-mean-square error (RMSE) of 0.51.

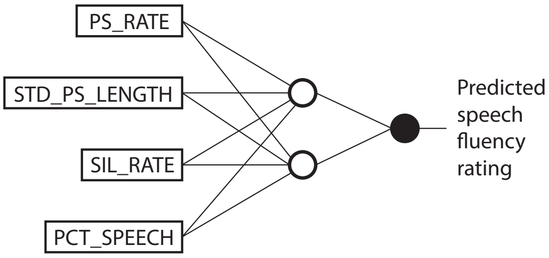

Figure 1: *Neural network used for the prediction of human ratings of speech fluency.*

A second model was created under the form of a neural network comprising a single hidden layer with two neurons, and taking as its input the four automatic predictors of speech fluency (Figure 1). As for the MLR, the neural network (NN) was trained on the full dataset. The NN model achieves a correlation of 0.89, with a RMSE of 0.54. The Figure 2 presents a scatterplot relating the actual SF ratings to the predicted ratings when using the NN. A Kolmogorov-Smirnov test indicates that the residuals of both the MLR and the NN models are normally distributed (both $p \geq 0.15$).

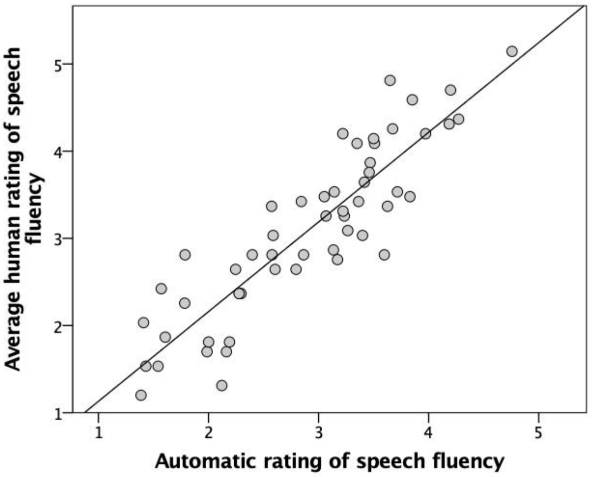

Figure 2: *Scatterplot relating average human ratings to the automatic measures of speech fluency obtained with the neural network, and associated regression line (*R^2 *= 0.79;* RMSE *= 0.54).*

In order to check the significance of the difference in the strength of the correlations achieved with the two models, a Fisher's *r*-to-*z* transformation (as implemented by [14]) was applied. As a directional hypothesis was being tested, a one-tailed test was used. The result indicates that the correlation obtained with the NN is significantly higher than the correlation obtained when using the MLR ($Z = -1.94$; $p = 0.026$). However, as the data sample used in this study is rather small, a ten-fold cross-validation was also conducted with both models to assess their behavior on new data. Results indicate similar performances for the two models, with, in both cases, an average correlation coefficient of 0.84 and an average RMSE of 0.53.

6. Discussion and conclusion

This study aimed at using the forward-backward divergence segmentation (FBDS) algorithm [9] to measure the speech fluency, in French as a foreign language, of 18 native-English speakers. The moderate-to-very-strong correlations observed between FBDS-derived measures and human ratings of speech fluency were all in line with our original assumptions. The pseudo-syllable rate was found to be strongly related to the perceived ease of speech production, a faster rate being associated with higher ratings of speech fluency. The other three FBDS-derived measures, all related to the presence of filled or silent pauses in the recordings, were also significantly associated with ratings of speech fluency: an increase in the presence of pauses was associated with lower ratings of speech fluency.

All FBDS-derived measures were finally combined together using (a) a multiple linear regression (MLR) and (b) a neural network (NN) in order to predict the human ratings of speech fluency. Very strong correlations were achieved with both models, with correlation coefficients of 0.85 and 0.89 for the MLR and NN models, respectively. This result is very promising, especially when considering that the predictions were made for speech recordings that contained only one single sentence. By comparison, in [7] and [8], the coefficient of the correlations between automatic and human ratings of speech fluency could only reach an average value of 0.89 if four sentences were considered (i.e., if the automatic and human ratings of speech fluency were averaged over four sentences; see figure 2 in [7]).

The NN model appears to achieve a significantly stronger correlation with human ratings of speech fluency than the MLR model. This could indicate that the perception of overall speech fluency is more complex than a simple linear combination of the perception of speech rate and of the perception of disfluencies (filled and silent pauses). In this case, a neural network might thus be better adapted to the prediction of perceived speech fluency than a MLR. However, as a cross-validation of both models indicated similar performances with regards to the prediction of new data, this result should be confirmed by a larger-scale study.

7. 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] Lennon, P. "Investigating fluency in EFL: a quantitative approach", *Language Learning*, vol. 40, no. 3, pp. 387–417, 1990.

[3] 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.

[4] ——, "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.

[5] M. Benzeghiba, R. D. Mori, and O. Deroo, "Automatic speech recognition and speech variability: a review," *Speech Communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[6] 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.

[7] 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 *INTERSPEECH 2018 — 19th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings,* 2018, pp. 2544-2548.

[8] 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," in revision.

[9] 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.

[10] Council of Europe. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume With New Descriptors*. 2018.

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

[12] De Jong, N. H., & Bosker, H. R. "Choosing a threshold for silent pauses to measure second language fluency," in R. Eklund (Ed.), *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech (DiSS)*, 2013, pp. 17–20.

[13] E. Frank, M. A. Hall, and I. H. Witten. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques," Burlington, USA: Morgan Kaufmann, Fourth Edition, 2016.

[14] I. A. Lee and K. J. Preacher. *Calculation for the test of the difference between two dependent correlations with one variable in common* [Computer software]. Available from http://quantpsy.org.