# REVISING PARAMETERS FOR PREDICTING L2 SPEECH FLUENCY AND PROFICIENCY

Heini Kallio & Mikko Kuronen

University of Jyväskylä
heini.h.kallio@jyu.fi, mikko.j.kuronen@jyu.fi

## ABSTRACT

The aim of the study was to investigate whether integrating parameters based on pause location improve the prediction of fluency and proficiency in L2 Finnish monologic speech. To answer this question, multiple linear regression models were fitted using two data sets containing L2 Finnish speech and expert assessments of fluency and oral proficiency. Separate models were derived for fluency and proficiency using combined data as well as the two separate data sets. The comparison of the models indicate that pause-by-location parameters can improve the prediction of L2 fluency and proficiency, but the relevant parameters and their significance in the regression models depend on the speech data. Parameters with low incidence work only in longer speech samples, while parameters with frequent occurrence can be used even in shorter samples. The results have implications for improving automatic assessment of L2 speech especially in low-resource languages.

**Keywords:** L2 speech fluency, L2 proficiency, pause locations, automatic assessment

## 1. INTRODUCTION

Speech fluency parameters have been investigated as predictors of fluency and proficiency for several second or foreign languages (L2). However, research on L2 fluency in low-resource languages is limited. The current study extends the research to a scarcely studied L2, Finnish, and explores new fluency parameters for automatic assessment of fluency and proficiency in L2 Finnish.

Speech fluency is one of the strongest indicators of L2 proficiency and often used as a diagnostic measure in language assessment. Previous studies that have attempted to quantify L2 speech fluency have measured features related to speed, pausing, and repair [1]. Especially the ones of speed and pausing have proven to be good predictors of L2 fluency. Recent findings of [2] support the importance of speed measures as fluency indicators in spontaneous L2 Finnish speech, but they also found the durations of so-called composite breaks and filled pauses as well as creaky voice to affect human ratings of L2 fluency and proficiency. However, there are conflicting results on the significance of some fluency parameters, such as pause rate and pause duration [3, 4, 5]. Reasons for the differing results can be looked for in variation in the number and type of parameters used, but also in the analyzed data.

The occurrence of disfluencies in speech can depend on various factors. For example, [4] found that the predictive power of different acoustic measures varied not only between beginner and intermediate level speakers, but also between speech styles: filled pauses seem to be more typical for spontaneous speech than read speech. The use of silent and filled pauses can also be language specific [6, 7]. Although studies on speech fluency in both L1 and L2 Finnish are limited, the findings of [8, 9] indicate that Finnish speakers may use silent pauses considerably longer in duration than speakers of other more frequently studied languages (English, French, Italian, German, and Spanish as studied in [7]). It is therefore important to study speech fluency in different languages in order to find out, which fluency features are global and which ones language-dependent.

The role of pause location to L2 fluency have been investigated only in the recent years. These studies suggest that pause location is an important aspect in perceived L2 fluency [10, 11], and pause locations have also been integrated into automatic assessment of L2 fluency [12]. However, only silent pauses have been taken into account with respect to their location. Moreover, pause locations have been generally defined in terms of grammatical clauses, which may limit the usefulness of the measure: in our recent study we found that silent and filled pauses within grammatical *phrases* were better indicators of (dis)fluency than pauses within clauses in L2 Finnish [13]. Pauses after incomplete words also significantly reduced the perceived fluency of L2 Finnish speech [13]. Based on these results we now investigate whether pause-

by-location parameters can improve the prediction of perceived fluency and proficiency in L2 Finnish (RQ1). We also study the possible differences in predictive fluency parameters between two similar data sets (RQ2). This study is among the first ones to integrate pause-by-location parameters into predicting L2 proficiency, and the first one to acknowledge the location of filled pauses in predicting L2 fluency and proficiency.

## 2. DATA AND METHODS

### 2.1. Speech data and human assessments

In this study we used two speech and assessment data sets provided by the DigiTala project [14]: one consisting of semi-spontaneous narratives produced by 147 adult (aged 18–61) and another from 53 younger learners of Finnish (high school students aged 15–21), both with related expert assessments of language skills. The data sets are described in more detail in [2] and [15]. Both data sets consist of responses to narrative tasks, but the instructive response times and speakers' proficiency and fluency distributions differed between the data sets. The adult L2 Finnish speakers were instructed to speak for 1.5 minutes, and the mean duration of their speech samples was 83.6 seconds. The young L2 speakers were instructed to speak for one minute on the given topic, and the mean duration of the samples was 42.6 seconds. In the adult data, ratings are skewed towards lower proficiency and fluency levels, while in the young data the ratings are skewed towards higher proficiency and fluency levels. The two data sets were analyzed separately but also combined to get a more balanced distribution in terms of speaker proficiency and fluency.

Expert ratings were collected for the adult and young speech data using the same experimental settings, assessment criteria, and rating scales [2, 15]. For the current study, assessments of overall proficiency level (holistic scale 0–6) and fluency (analytic scale 1–4) were used. In [2] and [15], the inter-rater reliability was tested and found sufficient for research purposes.

### 2.2. Fluency parameters

The speech samples were annotated to word and utterance levels using previously prepared transcriptions and the online alignment tool WebMAUS [16]. Annotations were revised and disfluencies marked manually in Praat [17] by a phonetic expert and checked by another to avoid mistakes and differing interpretations of disfluencies.

We computed 44 fluency parameters from extracted annotation intervals using R [18]. 21 general fluency parameters included the following: speech and articulation rate; rate, ratio, and mean duration of silent pauses (SP) 50–250 ms and > 250 ms; rate, ratio, and mean duration of filled pauses (FP) as well as corrections or repetitions; mean length of run in words; rate and mean duration of utterances; rate and mean duration of utterance breaks; pause ratio; disfluency ratio. An utterance was defined as a continuous speech run, which is separated from the next by a break of 250 ms or longer. The utterance break, in turn, could contain silent or filled pauses, hesitations, corrections, or repetitions. Disfluency ratio refers to the relative proportion of all pauses and disfluencies in the response.

23 parameters were based on pause and disfluency location and included SPs and FPs further defined by their syntactic location: between clauses (BC), within clauses (WC), between phrases (BP), and within phrases (WP). The duration threshold of pauses-by-location was set to 250 ms, following [19, 5, 3, 20, 2]. We defined clause in Finnish as a constituent that links a predicate to a subject or object [21]. Phrase, in turn, was defined as a word or group of words that act together as a grammatical unit but do not necessarily include a predicate [21]. In addition, pauses within words, or between an incomplete and a corrected word (WW), were measured. All fluency parameters were further z-scored to 0 for statistical analysis.

### 2.3. Statistical analysis

The effects of the fluency parameters to human ratings were studied using multiple linear regression models with average ratings of fluency or proficiency as the dependent variable and fluency parameters as predictors. The simplest models were derived with a stepwise feature selection method using the stepAIC algorithm (implemented in the R package MASS [22]). To study the contribution of the pause-by-location parameters in predicting L2 fluency and proficiency ratings we fitted separate model with only general fluency parameters (referred to as Model 1) and with both general and pause-by-location fluency parameters (Model 2). The two models were compared with respect to their predictive power as well as with likelihood ratio tests (RQ1). To study the role of data in predicting L2 fluency and proficiency, the models were also fitted separately for the adult and young data sets (RQ2). For brevity, we do not report the full models in this paper but focus on the most

relevant results in terms of our RQs.

## 3. RESULTS

Using the combined data, pause-by-location parameters improved the predictive power of the fluency model from 59% (Model 1) to 60% (Model 2). The likelihood ratio tests showed that the fits of Model 1 and Model 2 did not differ significantly (Chi-Squared test-statistic = 9.6, p-value > 0.05). For proficiency, integrating pause-by-location parameters to the model improved the predictive power of the model from 59% (Model 1) to 62% (Model 2). The fits of Model 1 and Model 2 differed significantly (Chi-Squared test-statistic = 24.2, p-value < 0.01).

Of the 44 fluency parameters, 12 were included in the final model predicting fluency ratings. Five of these were pause-by-location parameters: rate of filled pauses within phrases (WP-FP rate, $p < 0.05$), ratio of silent pauses between an incomplete and a corrected word (WW-SP ratio, $p < 0.05$), mean duration of silent pauses within phrases (mean WP-SP, $p < 0.1$), mean duration of silent pauses between phrases, (mean BP-SP, $p > 0.1$), and ratio of filled pauses within clauses (WC-FP ratio, $p > 0.1$). 15 parameters were included in the final model predicting proficiency ratings, and seven of these were pause-by-location parameters: WC-FP ratio ($p < 0.01$), WP-FP ratio ($p < 0.01$), WW-SP ratio ($p < 0.01$), mean duration of filled pauses within clauses (mean WC-FP, $p < 0.01$), mean WP-SP ($p < 0.01$), ratio of filled pauses between phrases (BP-FP ratio, $p < 0.1$) and rate of silent pauses within phrases (WP-SP rate, $p < 0.1$). Speech rate was expectedly the most significant predictor of both fluency and proficiency ratings ($p < 0.001$).

The pause-by-location parameters had differing improvements depending on the data: the predictive power of the fluency model improved from 68% to 81% for the young data set (Model 1 – Model 2 Chi-Squared test-statistic = 54.7, p-value < 0.001), while the predictive power of the fluency model for the adult data set improved from 58% to 60% (Model 1 – Model Chi-Squared test-statistic = 18.6, p-value < 0.05). For proficiency, the predictive power improved from 50% to 72% for the young data set (Model 1 – Model 2 Chi-Squared test-statistic = 49.7, p-value < 0.001) and from 61% to 66% for the adult data set (Model 1 – Model 2 Chi-Squared test-statistic = 31.2, p-value < 0.001).

The final fluency model for the adult data set included 22 fluency parameters, 10 of which were based on pause locations. The final fluency model for the young data set also included 22 parameters, 12 of which were pause-by-location parameters. The final proficiency model for the adult data set included 18 fluency parameters, 8 of which were based on pause locations. The final proficiency model for the young data set included 19 parameters, 8 of which were based on pause locations.

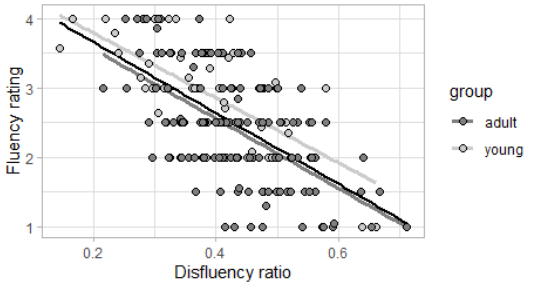


**Figure 1:** The relationship between disfluency ratio and fluency ratings for the two data sets "adult" and "young" as well as for the combined data (black regression line).

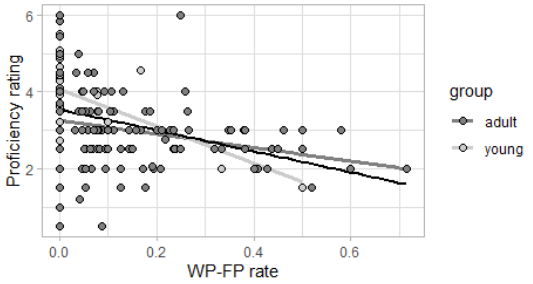


**Figure 2:** The relationship between WP-FP rate and proficiency ratings for the two data sets "adult" and "young" as well as for the combined data (black regression line).

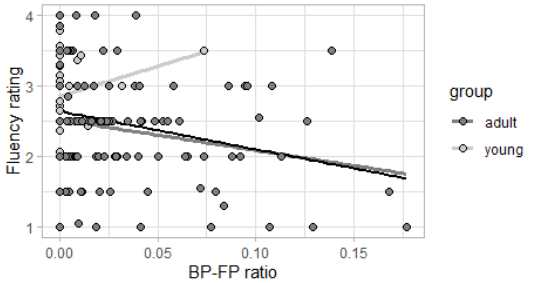


**Figure 3:** The relationship between BP-FP ratio and fluency ratings for the two data sets "adult" and "young" as well as for the combined data (black regression line).

Disfluency ratio proved to be a significant predictor of fluency in both data sets but was not included in the models predicting proficiency. Figure 1 illustrates the linear relationship between

disfluency ratio and fluency ratings of the two data sets. Of the pause-by-location parameters, WP-FP rate and WW-SP ratio had significant negative effects on the fluency as well as proficiency ratings. In the model for the combined data the effect was stronger for proficiency. Figure 2 shows the linear relationship between proficiency ratings and WP-FP rate for both data sets.

The significant parameters (excluding speed related parameters) and the predictive power of the models differed notably between the two data sets. For example, repair-related parameters (rate, ratio, and mean duration of corrections and repetitions) were significant only for the fluency model of the young data set, while mean length of run contributed only in the prediction models of the adult data set. Some parameters common for the models of adult and young data sets provided opposite effects: for example, ratio of silent pauses >250 ms had significant positive effects for the fluency and proficiency in the adult data set but significant negative effects in the young data set. BP-FP ratio, in turn, had a significant negative effect on the fluency ratings in the adult data set but significant positive effect in the young data set. The reason for the conflicting results might be due to rare occurrence of the parameterized features in the data. Figure 3 illustrates the relationship between BP-FP ratio and proficiency ratings of the two data sets.

## 4. DISCUSSION

The main aim of this study was to investigate whether integrating pause-by-location parameters improve models predicting fluency and proficiency of L2 Finnish monologic speech (RQ1). We also investigated the differences in prediction models between two similar data sets (RQ2).

Integrating pause-by-location parameters improved the predictive power of the fluency model only one percentage point, and the revised Model 2 (including pause-by-location parameters) did not differ significantly from Model 1 (including only general fluency parameters). The predictive power of the proficiency model, in turn, improved three percentage points and the revised Model 2 differed significantly from Model 1. Our results indicate that pause-by-location parameters are more strongly related to the perceived proficiency than fluency in L2 Finnish. This is an important finding, because pausing patterns have previously been studied in relation to perceived fluency rather than proficiency. It is also noteworthy that filled pauses within phrases contributed significantly to the prediction of both fluency and proficiency ratings. In previous studies, only silent pause locations have been under investigation. Our results encourage to study the role of filled pauses in L2 speech in more detail.

Of the general fluency measures, speech rate was expectedly the most significant predictor of both fluency and proficiency, regardless of the data. The less common parameter, disfluency ratio, proved significant in fluency prediction, which suggests that fluency models could be simplified by combining silent and filled pauses as well as corrections and repetitions into one parameter. However, the role of pauses and disfluencies seems to depend on how they occur in the data at hand. Although we studied two seemingly similar data sets (spontaneous L2 Finnish narratives), the final models of these data sets differ in the predictive parameters. This indicates that there are differences in the occurrence of disfluencies between the two data sets. These differences may also affect the results of the models for the combined data, as some parameters gained opposing effects in the two data sets. One reason for the differences can be looked for in the duration of the responses: the mean duration of the young L2 Finnish learners' responses was 42.6 seconds, while the mean duration of the adult L2 Finnish learners' duration was 83.6 seconds. The probability of disfluencies, especially filled pauses and repairs, increases with the duration of the response. The differences between the data sets manifest clearly in the occurrence of these disfluencies defined by location: for example, mean frequency of filled pauses within phrases was 0.13 for the young L2 learners and 2.4 for the adult L2 learners. Since the young data set is smaller (N = 53) and the speech samples are shorter than in the adult data set, some disfluencies may occur with only few speakers, which can distort the effect of these disfluencies, as shown in Figure 3. Therefore the statistical results for parameters with sparse occurrence must be interpreted with caution. With regards to pause locations, research on the use of silent and filled pauses in native Finnish speech is needed.

The current study contributed to the research of fluency parameters for automatic assessment of L2 fluency and proficiency in a low-resourced language, Finnish. The results support further investigation of pause-by-location parameters in predicting oral proficiency in L2 Finnish. Data-specific features may be an issue for training automatic assessment systems: using the same parameters regardless of the speech type and language can leave the prediction models weak or even biased.

## 5. ACKNOWLEDGEMENTS

The authors would like to thank the following people: Rosa Suviranta and Liisa Koivusalo for their help in annotating and analyzing the speech data, all the expert assessors, the Finnish National Certificates of Language Proficiency for their help in enabling the use of part of the speech data and recruiting expert assessors, and Aalto University's DigiTala team for their help in collecting human ratings. The DigiTala project is funded by the Academy of Finland and the research consortium includes University of Helsinki (grant number 322619), Aalto University (grant number 322625), and University of Jyväskylä (grant number 322965).

## 6. REFERENCES

[1] P. Tavakoli and P. Skehan, "Strategic planning, task structure and performance testing," in *Planning and Task Performance in a Second Language*, R. Ellis, Ed. John Benjamins, 2005, pp. 239–273.

[2] H. Kallio, R. Suviranta, M. Kuronen, and A. von Zansen, "Creaky voice and utterance fluency measures in predicting perceived fluency and oral proficiency of spontaneous L2 Finnish," *Speech Prosody: Proceedings of the 11th International Conference on Speech Prosody*, pp. 777–781, 2022.

[3] H. R. Bosker, A.-F. Pinget, H. Quené, T. Sanders, and N. H. De Jong, "What makes speech sound fluent? The contributions of pauses, speed and repairs," *Language Testing*, vol. 30, no. 2, pp. 159–175, 2013.

[4] C. Cucchiarini, H. Strik, and L. Boves, "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] J. Kormos and M. Dénes, "Exploring measures and perceptions of fluency in the speech of second language learners," *System*, vol. 32, no. 2, pp. 145–164, 2004.

[6] D. Duez, "Silent and non-silent pauses in three speech styles," *Language and Speech*, vol. 25, no. 1, pp. 11–28, 1982.

[7] E. Campione and J. Véronis, "A large-scale multilingual study of silent pause duration," in *Speech prosody 2002, international conference*, 2002.

[8] M. Toivola, M. Lennes, and E. Aho, "Speech rate and pauses in non-native Finnish," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[9] N. Penttilä, A.-M. Korpijaakko-Huuhka, and R. D. Kent, "Tavallista sujumattomuutta: Aikuisten puheen sujuvuuden kvantitatiivinen analyysi. [Normal disfluency: Quantitative analysis of adults' speech fluency.]," *Puhe ja kieli*, no. 3, pp. 153–173, 2018.

[10] J. Kahng, "The effect of pause location on perceived fluency," *Applied Psycholinguistics*, vol. 39, no. 3, pp. 569–591, 2018.

[11] K. Saito, M. Ilkan, V. Magne, M. N. Tran, and S. Suzuki, "Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency," *Applied psycholinguistics*, vol. 39, no. 3, pp. 593–617, 2018.

[12] C.-N. Hsieh, K. Zechner, and X. Xi, "Features measuring fluency and pronunciation," in *Automated Speaking Assessment*, K. Zechner and K. Evanini, Eds. Routledge, 2019, pp. 101–122.

[13] H. Kallio, M. Kuronen, and L. Koivusalo, "The role of pause location in perceived fluency and proficiency in L2 Finnish," *Proceedings of ISAPh 2022, 4th International Symposium on Applied Phonetics*, pp. 22–27, 2022.

[14] M. Kautonen and A. von Zansen, "DigiTala research project: Automatic speech recognition in assessing L2 speaking," *Kieli, koulutus ja yhteiskunta*, vol. 11, no. 4, 2020.

[15] L. Koivusalo, "Phonetic Fluency in Finnish as a Second Language: Acoustic Analysis of High School Students' Spontaneous Speech," *Master's Thesis, University of Helsinki*, 2022.

[16] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[17] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer. [Computer program]. Version 6.0. 19," *Online: http://www. praat. org*, 2016.

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org/

[19] R. Towell, R. Hawkins, and N. Bazergui, "The development of fluency in advanced learners of French," *Applied Linguistics*, vol. 17, no. 1, pp. 84–119, 1996.

[20] M. Kautonen and M. Kuronen, "Kvantitativa perspektiv på L2-tal på olika färdighetsnivåer [Quantitative perspectives on L2 speech on different skill levels]," *Folkmålsstudier*, vol. 59, pp. 11–40, 2021.

[21] A. Hakulinen, M. Vilkuna, R. Korhonen, V. Koivisto, T. R. Heinonen, and I. Alho, *VISK: Iso suomen kielioppi. Helsinki: Verkkoversio [The Great Grammar of Finnish. Online version]*. Suomalaisen Kirjallisuuden Seura, 2004, http://scripta.kotus.fi/visk.

[22] B. Ripley, B. Venables, D. M. Bates, K. Hornik, A. Gebhardt, D. Firth, and M. B. Ripley, "Package 'mass'," *Cran r*, vol. 538, pp. 113–120, 2013.