Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

Creaky voice and utterance fluency measures in predicting perceived fluency and oral proficiency of spontaneous L2 Finnish

Heini Kallio[1], *Rosa Suviranta*[2], *Mikko Kuronen*[1], *Anna von Zansen*[2]

[1]University of Jyväskylä, Finland
[2]University of Helsinki, Finland
heini.h.kallio@jyu.fi

Abstract

While utterance fluency measures are often studied in relation to perceived L2 fluency and proficiency, the effect of creaky voice remains ignored. However, creaky voice is frequent in a number of languages, including Finnish, where it serves as a cue for phrase-boundaries and turn-taking. In this study we investigate the roles of creaky voice and utterance fluency measures in predicting fluency and proficiency ratings of spontaneous L2 Finnish (F2) speech. In so doing, 16 expert raters participated in assessing narrative spontaneous speech samples from 160 learners of Finnish. The effect of creaky voice and utterance fluency measures on proficiency and fluency ratings was studied using linear regression models. The results indicate that creaky voice can contribute to both oral proficiency and fluency alongside utterance fluency measures. Furthermore, average duration of composite breaks – a measure combining breakdown and repair phenomena – proved to be the most significant predictor of fluency. Based on these findings we recommend further investigation of the effect of creaky voice to the assessment of L2 speech as well as reconsideration of the utterance fluency measures used in predicting L2 fluency or proficiency.

Index Terms: creaky voice, utterance fluency, L2 speech, language assessment

1. Introduction

Fluency is a frequently used term in language pedagogy and testing, but it has several definitions: in its broad sense, L2 fluency is often synonymous with general L2 proficiency, while the narrow definition, also examined as *utterance fluency* [1], refers to the fluidity or temporal features of speech [2]. Researchers have found temporal fluency measures to be strong predictors of human assessments of fluency [2, 3, 4, 5, 6, 7] as well as oral proficiency [8, 9, 10, 11], and similar features have also been incorporated in automatic assessment systems [12].

While utterance fluency is well studied, the effect of voice quality in terms of *creakiness* (also referred to as vocal fry, laryngealization, glottalization, or pulse phonation) to the perception of speaker's language proficiency has gained little attention in the research literature. The present study investigates the relations of creaky voice and temporal fluency measures in spontaneous L2 Finnish (F2) speech to expert assessments of fluency and general oral proficiency.

We use the term creaky voice for voice quality that involves low subglottal pressure and high level of adductive laryngeal tension [13], which leads to low f0 and the perception of croak-like quality where the listener can hear the individual glottal pulses. In some languages, creaky voice is connected to f0 declination, phrase-boundaries as well as turn-taking (see [14] for English, [15] for Swedish, and [16, 17, 18] for Finnish). In Estonian, creaky voice is also associated with stress- and timing related phenomena [19]. Language learners may thus be perceived as more native-like, if they manage to use creaky voice appropriately, while the absence of creaky voice can make the L2 speaker sound less fluent. However, studies on L2 and the use of creaky voice indicate that it is easier for a language learner to reduce creakiness than include it in speech [20, 21], perhaps because creaky voice, in relation to modal voice, can be considered as an individual characteristic. Pillot-Loiseau et al. [20] found that English learners of French – a language where creaky voice is considered unusual – reduced the use of creaky voice when speaking French compared to their L1 English, although they still used more creaky voice than native speakers of French. French learners of English, in turn, showed no difference in their use of creaky voice between French and English. Skarnitzl et al. [21] studied the use of glottalization in word linking in Czech learners of Spanish, Italian, and Portuguese, where creaky voice is less frequent than in Czech. They found that more experienced language learners used less creaky voice and glottal stops in linking words than less experienced language learners.

Utterance fluency is often divided into three components: 1) speed fluency, referring to, for example, speech and articulation rate, 2) breakdown fluency, referring to the frequency or duration of silent and filled pauses, and 3) repair fluency, referring to the occurrence of self-corrections and repetitions [22]. Most studies have investigated these components as separate measures, and composite measures have generally combined speed and breakdown fluency (investigated as speech rate or mean length of run, see, e.g., [5, 22, 23]). A combination of breakdown and repair fluency, however, has not been studied before, as far as we know of. In the current study, instead of studying repair fluency separately, we use a composite measure that combines breakdown and repair fluency. We base this decision to the fact that no clear relationship has been found between repair fluency measures and perceived fluency or oral proficiency [3, 5, 8]. However, since small correlations to fluency ratings have been reported [6, 24], we want to take repairs into account in the new composite measure.

Research on the fluency features in F2 speech is limited to studies by Toivola et al. [25, 26]. In their analysis of read speech [25] they found that native speakers of Finnish tend to make longer pauses than F2 speakers, but F2 speakers pause more often than natives. In a longitudinal study on F2 fluency, in turn, they found that during a year-long stay in Finland, F2 speakers' articulation rate and pause duration increased, while the number of pauses decreased. In the current study, we analyze both the duration and rate of pauses in relation to the assessed proficiency and fluency of F2 speakers.

The objective of this study is to investigate the effect of creaky voice and several utterance fluency measures on fluency

10.21437/SpeechProsody.2022-158

and proficiency ratings of spontaneous F2 speech. This study is part of the DigiTala project that investigates and develops automatic tools for spoken L2 assessment [27]. The project aims to develop automated assessment for high-stakes language tests (see [28]) and for self-regulated learning purposes in the Finnish language contexts [27].

2. Material and Methods

2.1. Speech data and human assessments

The speech data was provided by the National Certificates of Language Proficiency tests for Finnish as a second language [29] for a wider assessment context [27]. Thus, the data of this study consists of one speech sample per speaker (N=160). The speech samples were responses to narrative tasks [30] from intermediate and advanced level tests [29]. In both tasks, the speakers had one minute time to prepare and 1.5 minutes to speak. The absolute duration of the responses varied from 33.5 to 90 seconds per speaker.

Sixteen experienced [29] raters each assessed 35 to 37 samples from the current speech data. Each speech sample was assigned for two raters in a way that enabled systematic overlap between the ratings. In addition, a control set of 9 samples was rated by all 16 raters in order to test inter-rater reliability.

The assessments were done online using Moodle 3.8.3. Before the assessments, the raters also participated in an online training session where we introduced the Moodle environment and the assessment criteria. The criteria consist of a 7-point holistic and five 3-4-point analytic rating scales of which only two were used in this study. The criteria were developed for purposes of the Digitala project, which aims at assessing the overall proficiency level, task completion, fluency, pronunciation, range and accuracy of F2 speech [27]. For the current study, assessments of only overall proficiency level (holistic scale) and fluency (analytic scale) were used.

2.2. Extraction and computation of acoustic parameters

The speech samples were prepared for analysis and acoustic parameters were extracted using Praat [31]. Speech samples that had poor signal quality due to microphone issues or disruptive background noise were discarded. The final data set thus contains 147 speech samples.

The speech samples were annotated using three interval tiers. The first tier divided the temporal changes into "utterances" and "composite breaks". Here we define utterance as a continuous speech run, which is separated from the next by a composite break of 250 ms or longer. The composite break is a measure that can contain silent or filled pauses, hesitations, corrections and repetitions. The break threshold (250 ms) has been commonly used in previous research on speech fluency to define pauses and separate speech runs [32, 5, 6, 33]. Here, each utterance can also include silent pauses (SP) and filled pauses (FP) shorter than 250 ms. Both types of pauses were annotated in the second tier, also when occurring within composite breaks. Moreover, we divided silent pauses longer than 250 ms into two duration categories: long silent pause (LSP, 0.25 – 5 s) and very long silent pause (ELSP, $>$ 5 s).

The third tier contained segments of creaky phonation that were marked only when occurring within words (that is, creakiness related to, e.g., hesitations were not marked). Possible silence in the beginning and at the end of the recordings were not included in the fluency measures. The annotations were performed by a trained annotator through auditory analysis and visual inspection of the spectrogram. The labelled intervals from the tiers were extracted using a Praat script.

Eleven utterance fluency measures were computed from extracted annotation intervals using R [34]: articulation rate (ART-rate), absolute number of utterances (U-freq), average duration of utterance (mean-U), average duration of composite break (C-break), relative proportion of all silent pauses (pause-ratio), rate of short silent pauses per minute (SP-rate), rate of long silent pauses per minute (LSP-rate), rate of very long silent pauses per minute (ELSP-rate), rate of filled pauses per minute (FP-rate), average duration of within-utterance silent pauses (mean-SP), and average duration of filled pauses (mean-FP). In addition, the relative proportion of creaky voice was computed (creak-ratio). Strongly correlating variables were avoided: since speech rate depends on the amount of pausing, we opted to use articulation rate as a speed measure. The final set of acoustic parameters and their operationalizations are presented in Table 1.

Table 1: *Acoustic parameters*

Measure	Operationalization
ART-rate	rate of phones produced per second without pauses or other disfluencies
U-freq	absolute number of utterances
mean-U	average duration of utterance
C-break	average duration of composite break
pause-ratio	relative proportion of all silent pauses (total duration of silent pauses / total duration of response)
SP-rate	rate of short silent pauses ($<$ 250 ms) per min
LSP-rate	rate of long silent pauses (0.25 – 5 s) per min
ELSP-rate	rate of very long silent pauses ($>$ 5 s) per min
FP-rate	rate of filled pauses per minute
mean-SP	average duration of short silent pause ($<$ 250 ms)
mean-FP	average duration of filled pause
creak-ratio	relative proportion of creaky voice (total duration of creaky segments / total duration of response)

2.3. Statistical analysis

Inter-rater reliability was tested with intraclass correlation coefficient (ICC) using the *irr* package in R [35]. The ICC was computed from a control set, where all 16 expert raters assessed the same 9 random speech samples. The ICC was computed with a one-way model with speakers as random effects and comparing 1) the raters' agreement to the mean rating of a speech sample and 2) the raters' consistency with respect to the individual ratings of other raters.

For fluency ratings, the inter-rater agreement ICC value was 0.99 and inter-rater consistency ICC value was 0.85, indicating an excellent inter-rater reliability. For proficiency ratings, the agreement ICC was 0.99 and the consistency ICC was 0.90. Since the consistency among raters was high for both fluency and proficiency assessments, we decided to use mean grades in studying the relation of fluency and proficiency to acoustic parameters.

The effect of acoustic parameters on fluency and proficiency ratings was studied using linear regression models (LM) with average ratings as dependent variables and acoustic parameters as predictors. The simplest models were derived using a feature selection method stepAIC (implemented in the R pack-

age MASS [36]) that selects the model with least information loss based on the Akaike Information Criterion (AIC).

3. Results

The contribution of acoustic parameters (explained in Table 1) on the ratings of proficiency and fluency was studied using a stepwise linear regression model with average ratings as a dependent variable and acoustic parameters as predictor variables. The models were fitted separately for fluency and proficiency ratings. Table 2 summarizes the results of the models with predictor t-values and respective significance levels based on p-values as well as the adjusted R^2 of the final models.

Table 2: *Summary of the linear regression models with predictor t-values and adjusted R^2s. p-values: 0.1–0.05', 0.05–0.01*, 0.01–0.001**, < 0.001***.*

Predictor	Proficiency	Fluency
ART-rate	5.59***	3.81***
U-freq	2.30*	-
mean-U	-	-
C-break	-2.02*	-6.28***
pause-ratio	-	-
SP-rate	-	-
LSP-rate	-3.44***	-2.19*
ELSP-rate	-2.56*	-
FP-rate	-	-
mean-SP	-	-
mean-FP	-1.97'	-2.28*
creak-ratio	2.99**	2.61*
Model R^2 (adj.)	0.51	0.46

The most significant predictors for proficiency ratings were ART-rate, LSP-rate, and creak-ratio. Articulation rate showed a significant positive effect for proficiency ratings (t-value = 5.59, $p < 0.001$), indicating that the faster the articulation, the higher the proficiency. The frequency of silent pauses > 250 ms, in turn, showed a significant negative effect (t-value = -3.44, $p < 0.001$), indicating that the more such pauses, the lower the proficiency. Interestingly, creak-ratio provided a significant positive effect for proficiency (t-value = 2.99, $p < 0.01$), indicating that the more the speaker used creaky voice, the higher the proficiency. Number of utterances/response, mean duration of composite break, rate of very long silent pauses (>5 s), and mean duration of filled pauses also contributed to the prediction of proficiency. The model predicting overall oral proficiency accounted for 54 per cent of the variance in the ratings (multiple R^2 = 0.54 and adjusted R^2 = 0.51). The adjusted R^2s for individual predictors are 0.31 for ART-rate, 0.27 for C-break, 0.11 for creak-ratio, 0.09 for mean-FP, 0.06 for ELSP-rate, 0.04 for U-freq, and 0.008 for LSP-rate.

Since the coefficient of determination is notably higher for C-break than other measures with similar significance levels, we illustrate the linear relationship between C-break and proficiency ratings in Figure 1. Figure 2, in turn, shows the relationship between creakiness and proficiency.

As for fluency ratings, the single most significant predictor was the mean duration of composite break (C-break) with significant negative effect on the ratings (t-value = -6.28, $p < 0.001$). This indicates, that the longer the break between continuous speech runs, the lower the fluency rating. Another significant predictor was articulation rate with a positive t-value of 3.81 and $p < 0.001$. Similarly to the prediction of proficiency, LSP-rate, mean-FP, and creak-ratio also contributed to the prediction of fluency. Both pause measures showed a significant negative effect for proficiency ratings (t-value = -2.19 and $p < 0.05$ for LSP-rate, t-value = -2.28 and $p < 0.05$ for mean-FP), indicating that the more silent pauses >250 ms and the longer the filled pauses, the lower the fluency rating. Creak-ratio showed a significant positive effect for fluency (t-value

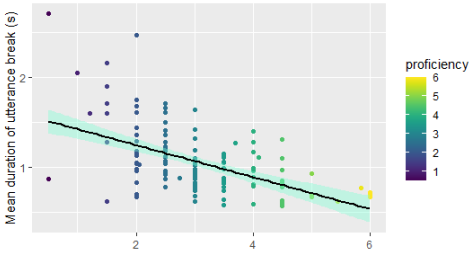

Figure 1: *The linear trend between mean duration of composite break (y-axis) and proficiency ratings (x-axis).*

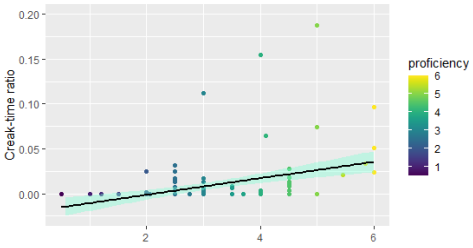

Figure 2: *The linear trend between relative amount of creak in response (y-axis) and proficiency ratings (x-axis).*

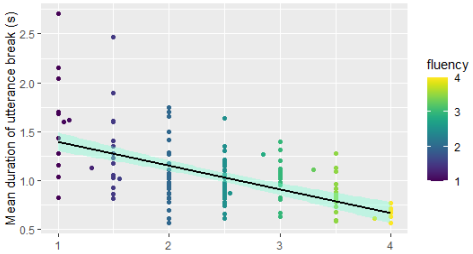

Figure 3: *The linear trend between mean duration of composite break (y-axis) and fluency ratings (x-axis).*

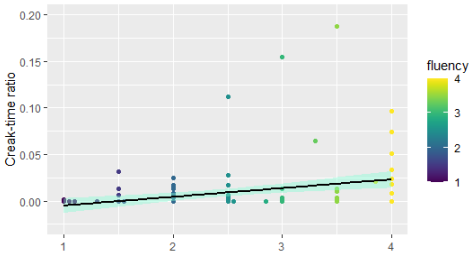

Figure 4: *The linear trend between relative amount of creak in response (y-axis) and fluency ratings (x-axis).*

= 2.61, $p < 0.05$), indicating that the more creaky voice was used, the better the fluency rating. The model accounted for 48 per cent of the variance in the fluency ratings (multiple R^2 = 0.48 and adjusted R^2 = 0.46). The adjusted R^2s for individual predictors were 0.29 for C-break, 0.25 for ART-rate, 0.11 for mean-FP, 0.09 for creak-ratio, and 0.007 for LSP-rate. The coefficient of determination R^2 was lower for models predicting fluency ratings than the ones predicting proficiency, possibly due to the narrower scale for fluency (1–4) compared to proficiency (1–6). Figure 3 illustrates the linear relationship between mean-U and fluency ratings, and Figure 4 shows the relationship between creakiness and fluency.

4. Discussion

This study investigated the role of creaky voice and utterance fluency measures in predicting perceived fluency and proficiency of L2 speakers of Finnish. Our results suggest that creaky voice contributes to the perception of fluency and proficiency in L2 Finnish, and that a composite measure of breakdown and repair fluency is more significant in predicting fluency ratings than previously used individual measures related to pausing or disfluencies.

Articulation rate was a significant measure in predicting both proficiency and fluency ratings, supporting previous results of speed fluency in other language contexts [37, 5, 6, 7]. However, the average duration of composite break (C-break) was even more significant predictor of fluency than articulation rate and explained most of the variance in fluency ratings. C-break is a composite measure combining silent pauses, hesitations, and repetitions or corrections as one interruptive break between continuous speech runs. In previous studies, these disfluencies have commonly been measured separately. Our result indicate that – at least in spontaneous L2 speech – such division between different disfluencies may not be necessary when estimating speakers' temporal fluency.

The frequency of long silent pauses (0.25 – 5 seconds) (LSP-rate) also proved significant for both proficiency and fluency ratings. The results for LSP-rate are partly in line with the findings of [25], where non-native speakers of Finnish were found to pause more often than native speakers. The results of [25] concern read speech, while our study indicates that the frequency of pauses $>$ 250 ms are significant also in spontaneous F2 speech. Both C-break and LSP-rate are measures related to prosodic phrasing, which is important to the comprehension of speech [38]. It should be noted, however, that in the current study the location of the pauses was not acknowledged. Yet, for the purposes of automatic assessment, such general measures have proved to be of use [23, 3, 10, 11].

Although we avoided strongly correlating variables, the LSP-rate is to some extent related to utterance frequency, since we opted to use the 250 ms break threshold in separating utterances. Moreover, since the response time was limited in the speaking tasks used in this study, the number of utterances might depend partly on the duration of utterances. These dependencies between the variables may have affected the results for LSP-rate, U-freq, and mean-U: U-freq was significant only for proficiency, and mean-U remained non-significant for both assessed dimensions. It should also be noted that the average duration of utterances used in the current study does not take into account possible differences in articulation rate, which can reduce the significance of this measure: a faster speaker is often considered more fluent than a slow speaker, but faster articulation can result in shorter utterance duration. A better measure would be the rate of utterances per minute or mean length of run, calculated as an average number of syllables produced in utterances [32, 5, 6].

Interestingly, the relative amount of creak (creak-ratio) proved significant for both proficiency and fluency ratings. It should be noted though, that not all F2 speakers with high ratings used creaky voice. To our knowledge, this study is the first in which creakiness has been shown to contribute to the assessment of L2 proficiency and fluency. This is perhaps because the use of creaky voice is often seen as an individual characteristic related to, for example, social status rather than as a feature of a language [39, 40]. The few studies that have investigated creakiness in L2 focus on how well speakers can avoid creakiness in the target language [20, 21]. Our results, however, suggest that using creaky voice in L2 Finnish can affect the perception of fluency and proficiency positively. The reasons for this result should be studied further and can be looked for, for example, in the language-specific prosodic features: sentence intonation in Finnish is typically declining [18], often resulting in utterance-final creak in phrase-boundaries [16]. Based on these findings, we consider the possibility that creakiness contributes to prosodic chunking in Finnish, alongside intonation and pausing. Thus, the occurrence of creak might be perceived as part of native-like intonation by the raters. However, the use of creaky voice in L1 Finnish should be studied in more detail: existing literature focus either in conversational speech [17] or auditive observations illustrated with single cases [18]. It is possible that creaky voice serves many purposes in Finnish, such as in Estonian, where creak has been associated with secondary word stress and timing-related properties [19]. It should be taken into account that while the amount of creak varies between F2 speakers with different proficiency levels, the occurrence of creak with respect to utterance position might also differ. In the future, we plan to study the role of creaky voice in F2 proficiency with respect to its location in speech.

5. Conclusions

This study shows that creaky voice can contribute to the prediction of F2 fluency and proficiency alongside traditional and novel fluency measures. Our findings further suggest that a composite measure of breakdown and repair fluency, composite break, can be used in predicting fluency ratings instead of individual fluency measures used in many previous studies. We suggest that the use of creaky voice in L2 speech should be studied closer in order to understand why it seems to affect proficiency and fluency ratings. In future studies, the position of creak and its possible relation to f0 movements and thus to the language-specific intonational patterns should be scrutinized.

6. Acknowledgements

The authors would like to thank Sari Ohranen, Ari Huhta, Tuija Hirvelä, and Sari Ahola from the Finnish National Certificates of Language Proficiency for their help in enabling the use of the speech data and recruiting expert assessors. We also want to thank Yaroslav Getman, Ragheb Al-Ghezi and Ekaterina Voskoboinik from the Aalto University for their help in creating the Moodle environment for 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).

7. References

[1] N. Segalowitz, *Cognitive bases of second language fluency*. Routledge, 2010.

[2] P. Lennon, "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: Comparisons between read and spontaneous speech," *the Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2862–2873, 2002.

[4] T. M. Derwing, M. J. Rossiter, M. J. Munro, and R. I. Thomson, "Second language fluency: Judgments on different tasks," *Language Learning*, vol. 54, no. 4, pp. 655–679, 2004.

[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] 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.

[7] Y. Préfontaine, J. Kormos, and D. E. Johnson, "How do utterance measures predict raters' perceptions of fluency in French as a second language?" *Language Testing*, vol. 33, no. 1, pp. 53–73, 2016.

[8] N. Iwashita, A. Brown, T. McNamara, and S. O'Hagan, "Assessed levels of second language speaking proficiency: How distinct?" *Applied linguistics*, vol. 29, no. 1, pp. 24–49, 2008.

[9] H. Kallio, J. Šimko, A. Huhta, R. Karhila, M. Vainio, E. Lindroos, R. Hildén, and M. Kurimo, "Towards the phonetic basis of spoken second language assessment: temporal features as indicators of perceived proficiency level," *AFinLA-e: Soveltavan kielitieteen tutkimuksia*, no. 10, pp. 193–213, 2017.

[10] O. Kang and D. Johnson, "The roles of suprasegmental features in predicting English oral proficiency with an automated system," *Language Assessment Quarterly*, vol. 15, no. 2, pp. 150–168, 2018.

[11] H. Kallio, A. Suni, and J. Šimko, "Fluency-related temporal features and syllable prominence as prosodic proficiency predictors for learners of english with different language backgrounds," *Language and Speech*, 2021. [Online]. Available: https://doi.org/10.1177/00238309211040175

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

[13] J. Laver, *The phonetic description of voice quality*. Cambridge: Cambridge University Press, 1980.

[14] J. Laver and L. John, *Principles of phonetics*. Cambridge university press, 1994.

[15] R. Carlson, J. Hirschberg, and M. Swerts, "Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates," *Speech communication*, vol. 46, no. 3-4, pp. 326–333, 2005.

[16] P. Hirvonen, "Finnish and English communicative intonation." Ph.D. dissertation, University of Turku, 1970.

[17] R. Ogden, "Turn transition, creak and glottal stop in Finnish talk-in-interaction," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 139–152, 2001.

[18] A. Iivonen, "Finnish sentence accent and intonation," in *Phonetics of Russian and Finnish. General description of phonetic systems. Experimental studies on spontaneous and read-aloud speech.*, V. De Silva and R. Ullakonjoa, Eds. Frankfurt am Main: Peter Lang, 2009, pp. 67–73.

[19] K. Aare, P. Lippus, and J. Simko, "Creak as a feature of lexical stress in Estonian," *Proceedings of Interspeech 2017*, 2017.

[20] C. Pillot-Loiseau, C. Horgues, S. Scheuer, and T. Kamiyama, "The evolution of creaky voice use in read speech by native-French and native-English speakers in tandem: a pilot study," *Anglophonia. French Journal of English Linguistics*, no. 27, 2019.

[21] R. Skarnitzl, P. Čermák, P. Šturm, Z. Obstová, and J. Hricsina, "Glottalization and linking in the L2 speech of Czech learners of Spanish, Italian and Portuguese," *Second Language Research*, p. 02676583211015803, 2021.

[22] 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.

[23] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[24] A.-F. Pinget, H. R. Bosker, H. Quené, and N. H. De Jong, "Native speakers' perceptions of fluency and accent in L2 speech," *Language Testing*, vol. 31, no. 3, pp. 349–365, 2014.

[25] 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.

[26] M. Toivola, M. Lennes, J. Korvala, and E. Aho, "A longitudinal study of speech rate and pauses in non-native Finnish," in *Proceedings of the 6th international symposium on the acquisition of second language speech, new sounds*, 2010, pp. 499–504.

[27] 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.

[28] The Matriculation Examination Board. The Finnish Matriculation Examination. [Online]. Available: https://www.ylioppilastutkinto.fi/en/

[29] Finnish National Agency for Education. National Certificates of Language Proficiency. [Online]. Available: https://www.oph.fi/en/national-certificates-language-proficiency-yki

[30] S. Luoma, *Assessing speaking*. Ernst Klett Sprachen, 2004.

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

[32] 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.

[33] M. Kautonen and M. Kuronen, "Kvantitativa perspektiv på l2-tal på olika färdighetsnivåer," *Folkmålsstudier*, vol. 59, pp. 11–40, 2021.

[34] 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/

[35] M. Gamer, J. Lemon, M. M. Gamer, A. Robinson, and W. Kendall's, "Package 'irr'," *Various coefficients of interrater reliability and agreement*, vol. 22, 2012.

[36] 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.

[37] 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.

[38] A. A. Sanderman and R. Collier, "Prosodic phrasing and comprehension," *Language and Speech*, vol. 40, no. 4, pp. 391–409, 1997.

[39] I. P. Yuasa, "Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women?" *American Speech*, vol. 85, no. 3, pp. 315–337, 2010.

[40] N. B. Abdelli-Beruh, L. Wolk, and D. Slavin, "Prevalence of vocal fry in young adult male American English speakers," *Journal of Voice*, vol. 28, no. 2, pp. 185–190, 2014.