Speech Communication 148 (2023) 66–80

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

# Prosody and fluency of Finland Swedish as a second language: Investigating global parameters for automated speaking assessment

Heini Kallio [a,*], Maria Kautonen [b], Mikko Kuronen [a]

[a] Department of Language and Communication Studies, University of Jyväskylä, Jyväskylä, Finland
[b] School of Applied Educational Science and Teacher Education, University of Eastern Finland

ARTICLE INFO

*Keywords:*
Spoken L2 assessment
Finland Swedish
Fluency
Prosody

ABSTRACT

This study investigates prosody and fluency of Finland Swedish as a second language (L2). The main objective is to investigate global measures of prosody and fluency as predictors of overall oral proficiency, fluency, and pronunciation ratings.

We analyzed parameters related to temporal fluency, timing (based on syllable durations), and f0 change from spontaneous speech produced by 30 native and 235 non-native speakers of Finland Swedish representing proficiency levels from beginner to intermediate. We used pairwise comparisons to investigate the differences between native speech (L1) and L2 samples from different proficiency levels. To study the predictability of ratings with acoustic parameters, we fitted a multiple linear regression model for each assessed dimension of L2 skills.

The comparison of L1 and L2 samples as well as L2 samples with different proficiency and fluency levels showed clear differences in f0 change and fluency parameters. Standard deviation of syllable durations also showed differences with respect to L2 learners' fluency level. The results for multiple linear regression models, however, indicate contribution of rate-normalized standard deviation of syllable duration to fluency ratings, alongside traditionally used fluency parameters. As for proficiency ratings, f0 slope complemented fluency parameters in the prediction model. The predictive power of the parameters varied depending on the assessed dimension of L2 skills.

This study provides new information on the prosodic features of Finland Swedish as a second language and suggests new research on the assessment of non-dominant varieties of pluricentric languages. The results support previous findings on the importance of speed and pausing measures in predicting oral L2 skills. However, further investigation of language-specific f0 and timing parameters as part of automated or computer-assisted speaking assessment is called for.

## 1. Introduction

Prosodic systems of languages differ significantly from one another (Hirst and Di Cristo, 1998; Bruce, 2012), which can cause challenges to many language learners (Trofimovich and Baker, 2006; Mennen, 2007). Prosody has received increased attention among L2 researchers during the last couple of decades, since correct production of prosodic features have been proved to promote comprehensibility and intelligibility of an L2 speaker (Munro and Derwing, 1995; Derwing and Munro, 1997; Isaacs, 2018a). Most studies, however, have focused on the production of majority language as an L2 – including English (e.g., Derwing et al., 2004; Trofimovich and Baker, 2006; Kang et al., 2010), Dutch (e.g., Cucchiarini et al., 2002; 2010; Bosker et al., 2013), French (e.g., Préfontaine et al., 2016), and Hungarian (e.g., Kormos and Dénes, 2004), among others. Regarding Swedish as an L2, research has also focused on the dominant variety of this pluricentric language, namely the Central Standard Sweden Swedish (hereafter referred to as CS; see, e.g., McAllister et al., 1999; Bannert, 2004; Kuronen et al., 2016). Research on the non-dominant varieties of Swedish, such as Finland Swedish, is scarce.

The purpose of this study is to investigate the prosodic properties of Finland Swedish (hereafter referred to as FS) from spoken L2 assessment perspective. The overall goal is to investigate global acoustic parameters of f0 change, timing, and fluency that could be integrated into an automatic assessment algorithm. Automated/computer-assisted assessment methods have been recognized to have a great potential to meet

* Corresponding author.
*E-mail address:* heini.h.kallio@jyu.fi (H. Kallio).

https://doi.org/10.1016/j.specom.2023.02.003
Received 27 February 2021; Received in revised form 14 December 2022; Accepted 17 February 2023
Available online 25 February 2023

some obstacles in oral skills assessment by, e.g., decreasing the individual rater's work and by increasing the reliability of assessment (Cheng, 2011; Luo et al., 2016) as well as having a potential to enhance foreign language learning (Golonka et al., 2014; Wik, 2011). The development of automatic assessment system, however, requires knowledge about phonetic and prosodic features underlying the assessments.

What makes FS an interesting object for L2 speech research is its status as an official language in Finland as well as its own characteristics that differ from the varieties spoken in Sweden. As a non-dominant variety of Swedish, FS has its own characteristics in lexicon, phonology, morphology, syntax as well as in interaction and pragmatics (Norrby et al., 2011; Lindström et al., 2017). Perhaps the most evitable differences between FS and CS concern their phonetic and prosodic properties, which are partly due to the fact that FS has been affected by the majority language, Finnish (e.g., Ringen and Suomi, 2009, 2012; Helgason et al., 2013). Many of the students studying Swedish in Finland also learn FS, which makes research on spoken L2 Finland Swedish relevant for improving the assessment of spoken language skills in the Finnish education system.

Due to the official status of FS, it is a compulsory subject in basic education in Finland, and the national matriculation examination test of L2 Swedish is taken by over 15,000 upper secondary school students yearly (The Finnish Matriculation Examination Board, 2021), which makes FS the second most tested L2 in Finnish schools. The Ministry of Education and Culture in Finland has set a goal to include foreign and second language speaking exams into the national matriculation examination in the near future (Ministry of Education and Culture in Finland, 2017). This decision excites the need for research on spoken L2 skills in languages spoken in Finland as well as their reliable and effective assessment methods. The current study is part of the DigiTala research project that aims to develop automatic tools for assessing spoken language skills in large-scale, high-stakes contexts (Kautonen and von Zansen, 2020). The main goal of the DigiTala project is to develop automatic assessment systems for FS and Finnish as second languages. The systems will consist of an automatic speech recognition engine trained for L2 speech specifically, and several machine learning models that evaluate grammatical accuracy, vocabulary range, pronunciation, fluency, task accomplishment, and overall oral proficiency. This study contributes to the development of a feature-based assessment system for L2 Finland Swedish by investigating acoustic parameters potentially useful for the evaluation models of fluency, pronunciation, and overall oral proficiency.

### 1.1. Prosodic features of Finland Swedish

Compared to CS, the prosody of FS is studied to a lesser extent. FS, especially the Standard variety spoken in Southern Finland[1] (Ivars, 2015) is, in some phonetic respects, said to be more similar to Finland's majority language Finnish than CS, as it differs from CS regarding some phonemic characteristics (Leinonen, 2004), vowel quality (Kuronen, 2000), aspects of consonant quality and quantity (Leinonen, 2004; Ringen and Suomi, 2012; Helgason et al., 2013), and prosodic features such as realization of sentence and word stress (Tevajärvi, 1982; Vihanta et al., 1990; Kautonen and Kuronen, 2021). For example, the lexical pitch accents *acute* and *grave* that are characteristic for CS (Riad, 2014), are absent in FS, so that words such as ⟨ anden ⟩ (definite form of ⟨ and ⟩ 'duck') and ⟨ anden ⟩ (definite form of ⟨ ande ⟩ 'spirit') are pronounced similarly in FS with no difference in the number nor placement of f0 peaks. This lack of word accent opposition also affects FS sentence intonation, which is found to be similar in Finnish and FS in declarative sentences (Aho, 2010). In FS, f0 tends to be falling after the stressed syllable in a word, and f0 is often declining also in statements and questions (Vihanta et al., 1990; Aho, 2010). Melodically, FS is associated with those varieties of Sweden Swedish (spoken in Scania, Gotland, and Dalarna) that are characterized by a relatively simple melody with recurring f0 peaks and a level intonation between stressed syllables (Bruce, 2010).

Although Finnish has affected the pronunciation and prosody of FS in some respects, similarities between FS and CS still originates from the common linguistic properties. While Finnish, for example, has fixed word stress in the initial syllable (e.g., ⟨ kala ⟩ 'fish', ⟨ kalastus ⟩ 'fishing'), the placement of word stress varies in Swedish (e.g., ⟨ banan ⟩ 'the lane/path', ⟨ banan⟩ 'a banana'). The relation of phonological quantity and the use of duration as a stress marker in FS, however, has not been studied thoroughly. On one hand, duration contrast related to linguistic quantity is more consistent in Finnish than in Swedish, where it is strongly related to stress (Engstrand and Krull, 1994; Strangert, 1985). On the other hand, quantity realization of FS has been found to be affected by Finnish, causing FS speakers to exaggerate the Swedish quantity contrast (Helgason et al., 2013). Vihanta et al. (1990) noted that the marking of stress and focus is weaker in FS than in CS with respect to the use of f0, intensity, and duration. Heinonen (2019), in turn, found the duration ratios of stressed vs. unstressed syllables and words (proportional to utterance length) of FS speakers similar to the ones of CS speakers.

### 1.2. Previous studies on f0, rhythm, and fluency of L2 speech

Various studies on learning and assessing f0 in L2 speech have focused on specific tonal features, namely intonation and pitch accents (Broselow, 1988; Grosser, 1993; Grosser at al., 1997; Rasier and Hiligsmann, 2007; Wennerström, 2000), while fewer studies have investigated global f0 changes in L2 speech. Perhaps the focus on locally bound f0 phenomena has restrained researchers from quantifying intonation features as parameters applicable to an automatic assessment system, although some promising systems have been proposed (Arias et al., 2010; Cheng, 2011; Escudero-Mancedo et al., 2017; Li et al., 2017). The method of Arias et al. (2010) relied to reference intonation patterns that the language learner tried to imitate. The L2 productions were then compared to the reference L1 intonation patterns. Cheng (2011) received high correlations between automatic and human ratings by using canonical contour models at word-level f0 and energy and combining them with phone-level duration information. Cheng concluded, however, that duration information had the strongest predictive power for prosody evaluation. Escudero-Mancebo et al. (2017) successfully distinguished native and non-native speakers of Spanish with groups of automatically derived ToBI labels. Li et al. (2017) automatically classified L2 English intonation as either rising or falling based on deep neural networks.

Studies that have examined global f0 changes in L2 speech have found differences, for example, in f0 range and variation (Ullakonoja, 2007; Busà and Urbani, 2011; Kautonen, 2017; Kuronen and Tergujeff, 2018). Ullakonoja (2007) investigated the f0 of Finnish learners of Russian and found that L2 speakers had narrower f0 range than the native speakers. The findings of Busà and Urbani (2011) about Italian learners of English were similar, although the differences depended on sentence type. As for Finland Swedish, native speakers seem to vary their f0 less than Finnish-speaking L2 learners when speaking Swedish (Kautonen, 2017). For the L2 learners, the wider f0 range and higher values for mean f0 slopes seemed to be in line with producing too many stressed words and stressing the words more than native speakers of FS. Kuronen and Tergujeff (2018), in turn, studied both global and local f0 phenomena in Finnish learners of CS and found a connection between the acquisition of the two: speakers who acquired tone accent 2 achieved

[1] In the study, we focus on the characteristics of Standard Finland Swedish, which is said to resemble varieties spoken in Southern Finland, more specifically in Central Uusimaa region (Ivars 2015). There are, however, different dialects also within the Finland Swedish variety, which also differ in phonetic respects.

more native-like f0 contours as well as larger standard deviation and f0 range. Thus, both too narrow and too wide f0 range seem to be characteristic for L2 learners compared to L1 speakers.

Speech rhythm varies between languages and depends primarily on realizations of word and sentence stress (Ladefoged and Johnson, 2014). A traditional (but controversial, see, e.g., Eriksson, 1991) classification is to divide languages into stress-timed or syllable timed (e.g., Abercrombie, 1967). Many acoustic-phonetic measurements have been proposed for comparing rhythm in different languages, but most parameters depict the temporal organization of speech. We acknowledge that rhythm is a more complex phenomenon than what can be uncovered with measures related to timing and phonotactic patterns and will thus refer to such measures as timing parameters or measures of timing. Interval measures (Ramus et al., 1999) and so-called pairwise variability indices (Grabe and Low, 2002) have perhaps been the most widely used timing parameters. Both approaches have generally divided speech into vocalic and consonantal intervals, but also combinations of the two have been proposed (see, e.g., Liss et al., 2009). Ramus et al. (1999) computed proportions of vocalic proportions in a sentence (%V) and standard deviations of both vocalic intervals (ΔV) and consonantal intervals (ΔC). Dellwo (2006) introduced the rate-normalized parameters of ΔV and ΔC by dividing the standard deviations by the mean vocalic/consonantal durations. Low and Grabe (1995), in turn, introduced the pairwise variability index (PVI), which computes the sum of the durational differences between adjacent vocalic or consonantal intervals, taking the absolute value of each difference and dividing it by the duration of the pair. Later, Low et al. (2000) used a normalized pairwise variability index (nPVI), which divides the mean absolute difference between durations of neighbouring interval pairs by the mean duration of the interval pairs.

While the above-mentioned timing parameters have somewhat successfully been used for comparing temporal organization of native speech across languages, the few applications to L2 speech have provided differing results: White and Mattys (2007) found no significant differences between L1 and L2 Dutch (spoken by native English) nor between L1 and L2 English (spoken by native Dutch) in any measures of timing. Thomas and Carter (2006), in turn, found Spanish L2 speakers of English to produce vocalic NPI values that were significantly different from both L1 Spanish and L1 English. They saw this result as an interference from a syllable-timed language (Spanish) to a stress-timed language (English). Dutch, as far as the traditional rhythmic classification is concerned, falls into the same stress-timed group than English, which might explain the results of White and Mattys (2007). Other studies have, however, managed to distinguish between L1 and L2 in terms of timing measures: Jang (2008) found that Korean learners of English show a lower vocalic variability (rate-normalized ΔV) than native speakers. Moreover, Gut (2009) found significant differences between L1 and L2 German as well as L1 and L2 English by measuring syllable duration ratios from learners with various native languages. Timing parameters have also been applied in attempts to automatize the assessment of oral language skills (e.g., Hönig et al., 2010).

Thus, some transfer concerning temporal organization of speech can arise from L1 to L2 when the languages are different from one other, but if a speaker is acquiring a language that shares timing features with their first language, it is less likely that their speech timing in L1 and L2 differ significantly. The case of Finnish speakers learning Finland Swedish, however, is not yet clear. On one hand, Swedish is considered a stress-timed language (Bruce, 2010), while Finnish has been claimed to share features attributed to syllable-timed (Karlsson, 1983, p. 176) but also to mora- and stress-timed languages (O'Dell et al., 2007). On the other hand, FS is affected by Finnish, which may result in more similar temporal organization of speech between the two languages than between FS and CS (Vihanta et al., 1990). Studies have, however, found significant differences between native FS and L2 speakers' stress production (Kautonen, 2017; Kallio et al., 2020) and syllable durations (Heinonen, 2019). Kautonen (2017) examined Finnish speakers' intonation in declarative utterances in Finland Swedish on CEFR levels B1–B2 and found that the L2 speakers varied their f0 more than L1 speakers. This was seen as a result of L2 speakers' excessive stress production. Heinonen (2019) found that the production of sentence stress, especially with regards to syllable duration, was one of the most challenging features for L2 learners of FS, regardless of their proficiency level. Kallio et al. (2020), in turn, investigated L2 Finland Swedish syllable prominence with a novel, continuous wavelet transform (CWT) based method, and found significant correlations between the resemblance of L1 and L2 prominence patterns and expert ratings of prosodic proficiency. Moreover, the most important prosodic signal in FS prominence realizations proved to be duration, which supports the observations of Heinonen (2019). While the study of Kautonen (2017) investigated spontaneous monologue speech, Heinonen (2019) and Kallio et al. (2020) focused on L2 stress production in read speech. Based on their results, however, we can assume some L1 effect on the timing parameters in L2 speakers of Finland Swedish.

Fluency is one of the most commonly used terms in language pedagogy and testing and it is included in most assessment criteria of oral L2 skills such as IELTS (British Council, 2019), Pearson, 2017, and the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001; Council of Europe, 2020). There are, however, several ways to approach fluency (Chambers, 1997; Huhta et al., 2019). We approach fluency from a *narrow perspective* that refers to spoken performance instead of general proficiency (Lennon, 2000). More specifically, we examine utterance fluency, from which Tavakoli and Skehan (2005) have identified three components: (1) *speed fluency*, referring to the speed at which speech is delivered, (2) *breakdown fluency*, meaning the pausing phenomena in speech, and (3) *repair fluency*, referring to false starts, repetitions, and self-corrections the speaker has made during their utterance.

These temporal features have been studied widely in L2 speech, and most research have concluded that parameters related especially to speed and breakdown fluency correlate with assessments. For example, Cucchiarini et al. (2002) found *speech rate* as the best predictor of fluency scores, while Derwing et al. (2004) found significant correlations between fluency ratings and *pausing* as well as *articulation rate*. Kormos and Dénes (2004), in turn, found *speech rate, phonation-time ratio*, and *mean length of utterance* to be the best predictors of fluency ratings. Bosker et al. (2013) found that *filled* and *silent pauses* and the *mean length of pause* together with *mean length of syllable* are significant predictors of fluency ratings. Similar parameters have also proven to be strong predictors of prosodic competence (Hönig et al., 2010; Cheng, 2011; Kallio et al., 2017; Kang and Johnson, 2018, 2021).

The current research aims to contribute to two aspects of L2 speech assessment in particular: (1) Despite the recognized importance of prosodic features in L2 speech, other than fluency parameters are rare in automated assessment of oral language skills. In general, automatic systems are currently more accurate in assessing segmental features than prosodic features (Isaacs, 2018b). Therefore, more research is needed on the effect of prosodic features to assessments in order to meet the importance that prosody has for communicative skills. (2) Most studies have focused on the assessment of read-aloud speech, because automatic assessment tends to work better for more predictable speech samples, such as read-aloud speech, than for spontaneous speech (Qian et al., 2020: 66–67; Zechner et al., 2009). The ability to produce spontaneous speech is, however, essential in communication (as described in the CEFR; Council of Europe, 2001; Council of Europe, 2020) and therefore it is also important to be able to assess it with automated methods. This study aims at deepening our knowledge on the topic by exploring prosodic and fluency features in spontaneous speech.

### 1.3. Research objectives

The main objective of this study is to investigate automatically or semi-automatically measurable parameters of fluency, parameters of

timing, and f0 change as predictors of oral proficiency as well as analytic fluency and pronunciation ratings with Finland Swedish as an L2. We focus on global parameters that previous research has found important in assessing L2 speech and that could be integrated into an automatic assessment algorithm. Since Finland Swedish is only marginally studied with regards to these features, we first compare the selected parameters between native FS speech and L2 FS samples with different proficiency levels. The research questions that the current study aims to address are:

RQ1: Do L1 speakers differ from L2 speakers with regards to parameters of fluency, parameters of timing, and f0 change?
RQ2: Can the selected prosodic and fluency parameters distinguish L2 speakers with different proficiency levels?
RQ3: What is the relative contribution of the selected prosodic and fluency parameters to ratings of overall oral proficiency, fluency, and pronunciation?

To answer RQ1, we perform pairwise comparisons between speech samples produced by native speakers of Finland Swedish and samples produced by L2 speakers of FS with different proficiency levels. To answer RQ2, pairwise comparisons are also performed between L2 samples grouped by overall oral proficiency level. Based on previous findings by Ullakonoja (2007), Busà and Urbani (2011), Kautonen (2017), and Kuronen and Tergujeff (2018), we expect the f0 range and variation to differ at least between L1 and L2 speech, but also between L2 proficiency groups to some extent. As for the timing parameters, we believe that some transfer from L1 to L2 speech is possible, but previous knowledge of the measures of timing in Finland Swedish is too scarce for detailed predictions. However, based on the findings of Kautonen (2019), Heinonen (2019), and Kallio et al. (2020), we expect that the lower the proficiency of Finland Swedish, the larger the differences between the timing parameters of L1 and L2 groups. Also, based on several studies on L2 fluency (Cucchiarini et al. 2002; Derwing et al., 2004; Kormos and Dénes, 2004; Bosker et al., 2013; Kallio et al., 2017; Kang and Johnson, 2018, 2021), we expect the fluency parameters to successfully discriminate between different L2 proficiency level groups.

To answer RQ3, the predictive power of the distinguishing parameters to each assessed dimension (overall oral proficiency, fluency, and pronunciation) is further examined with stepwise multiple linear regression modeling. We expect several parameters to contribute to the prediction of overall oral proficiency, while the fluency ratings we expect to be predicted mainly with temporal fluency parameters. However, since the different dimensions of language skills are usually connected to the overall proficiency level of the speaker, we can expect the selected parameters to affect all ratings to some extent.

## 2. Material and methods

### 2.1. Speech data and assessments

The speech data for the present study is from the DigiTala corpus that was collected while piloting a computer-aided oral language test (Karhila et al., 2016). Groups of upper secondary school students (aged 16–17 years) took the pilot test in a classroom environment using headset microphones. The participants were native speakers of Finnish who had studied Swedish as a compulsory subject for 4–7 years. In order to obtain reference samples for analysis, the same pilot test was also taken by native FS speakers.

The test consisted of various tasks ranging from read-aloud sentences to picture narration and simulated dialog, and a random set of tasks was given to each examinee from a pool of trials. For the current study we selected answers from six tasks that elicited relatively spontaneous narrative speech using picture and/or written stimulus. In these tasks, the examinees were to describe the weather, order a meal, warn the police about an animal on the road, describe a person to a security guard, guide someone to the pharmacy using a map, and persuade a friend to come to a rock concert with them. In each of the tasks, the speakers had 30 s to react to the provided stimulus. In total, the data includes 235 samples from L2 speakers and 30 samples from native speakers of Finland Swedish.

Four expert raters with expertise in Swedish participated in assessing the selected samples using a holistic six-point assessment scale for the overall oral proficiency level (Appendix A) and four analytic three-point scales including criteria for fluency, pronunciation, grammar, and vocabulary (Appendix B). The assessment criteria for the overall proficiency level and for the four analytic dimensions were derived from the Finnish National Core Curriculum 2003 (Finnish National Agency for Education, 2003), which is based on the descriptions of language proficiency in the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001; Council of Europe, 2020). Since the current study focuses on acoustic measurements and disregards the linguistic content of the speech samples, we use the ratings for overall proficiency, fluency, and pronunciation.

The raters conducted the assessment individually using headphones in a quiet environment. Two raters assessed all the 235 samples included in this study. A random set of 36 samples was used as a control set and was assessed by all four raters. In addition, the two "control raters" assessed independent sets of 18 samples each. As a result, 36 samples are rated by 3 raters. Interrater reliability was examined with intraclass correlation coefficient (ICC) using the *irr* package in R (Gamer et al., 2012). Both mean ratings (rounding to one decimal place) and categories based on the ratings are used in further analysis. From the CEFR scale, levels below A1, A1, A2, B1, and B2 were applied to the selected speech samples and transformed into continuous numeric variable in order to compute the average ratings. In addition to the average ratings, the samples were grouped into four categories based on the ratings. The proficiency categories include natives ($N = 30$), B-level ($N = 42$), A-level ($N = 163$), and below A-level ($N = 30$). The fluency and pronunciation categories include natives and categories responding to grades 1–3. When the average rating took place in the middle of two levels, the sample was grouped into the higher level.

### 2.2. Extraction and computation of acoustic parameters

The speech samples were prepared for analysis and all acoustic measurements were performed using Praat (version 6.1.38, Boersma and Weenink, 2021). The three authors annotated the speech samples at syllable level using previously prepared transcriptions and the online alignment tool WebMAUS (Kisler, Reichel and Schiel, 2017), and annotations were revised and disfluencies marked manually. A control set of 15 L2 samples (three samples representing each proficiency level) was annotated by all three authors to ensure that markings were done systematically. For syllabification, we applied the maximum onset principle with the restrictions of Swedish phonotactics (Bruce, 2012: 30–32). As for disfluencies, silent pauses (SP), filled pauses (FP), and sections uttered with "wrong language" – some other language than Swedish – (WL) were marked in annotations.

As opposed to most studies on fluency, false starts, repetitions, and self-corrections were marked as normal syllables, when they were recognized as Swedish. We opted for this method for the following reasons: first, clear connections have not been found between repair phenomena and perceptions of fluency (see, e.g., Cucchiarini et al., 2002; Kormos and Dénes, 2004; Bosker et al., 2013, Kallio et al., 2017). Second, repair phenomena seem to occur infrequently, making it difficult to include in quantitative studies (Götz, 2013; Kahng, 2014; Peltonen and Lintunen, 2016). Our data consists of spontaneous speech, which has been found to have fewer false starts, repetitions, and corrections than read speech (Cucchiarini et al., 2002; Kallio et al., 2017). In addition, repeated words and syllables might be difficult for an automatic algorithm to recognize as disfluencies instead of parts of the intended utterance. Another decision that differs from previous research concerns pause thresholds: based on findings that many pauses shorter

than 250 ms (a commonly used threshold: see, e.g., Derwing et al., 2004; Kormos and Dénes, 2004; Préfontaine et al., 2016) cannot be attributed to articulation (Hieke et al., 1983; Campione and Véronis, 2002) and auditory observations of the current speech data, we opted not to use a threshold when marking pauses and disfluencies.

Acoustic measurements were performed with Praat scripts that derived durations of syllables and other labelled intervals (SP, FP, and WL) as well as various f0 measurements in both hertz and semitones. Since the speech data includes samples from both male and female students, we decided to use f0 parameters computed in semitones. Thus, three parameters of f0 change in semitones were used in this study: f0 range, standard deviation, and mean slope (a measure of the steepness of declination).

Two timing parameters were computed from syllable durations: rate-normalized standard deviation of syllable duration (nΔS) and normalized pairwise variability index (nPVI). The nΔS was computed as a function of mean standard deviation of syllable duration within a speech sample. Similarly, nPVI was computed as a function of mean durational difference between consecutive syllables within a speech sample.

Six parameters were computed from the disfluency intervals SP, FP, and WL: disfluency-time ratios (hereafter referred to as SP-ratio, FP-ratio, and WL-ratio, respectively) and frequency (hereafter referred to as SP-freq, FP-freq, and WL-freq, respectively). Frequency of a disfluency in a sample was computed as average number/second instead of more commonly used number/minute due to relatively short durations of the speech samples ($< 30$ s).

As for speed measures, both articulation rate and speech rate were computed. Articulation rate was computed by dividing the total number of syllables in a sample by the sample duration (in seconds), excluding silent and filled pauses as well as sections uttered in a wrong language from the sample duration. Speech rate was computed by dividing the total number of syllables in a sample with the total duration of the sample (including pauses and WL). All acoustic parameters and their operationalizations are presented in Table 1.

### 2.3. Statistical analyses

The differences in f0 change, timing parameters, and fluency parameters between the native speakers and proficiency levels below A, A, and B as well as fluency categories 1–3 were investigated with pairwise comparisons using Wilcoxon rank sum test (with appropriate Bonferroni corrections for multiple comparisons). The effects of f0 change, syllable timing, and fluency parameters to assessments were studied using multiple linear regression models (MLR) with average ratings of either proficiency, fluency, or pronunciation as a dependent variable and acoustic parameters as predictor variables. The simplest models were derived with a stepwise feature selection method using the stepAIC algorithm (implemented in the R package *MASS*, Ripley et al., 2013) that compares the Akaike Information Criterion (AIC) of all possible models and selects the one with the least information loss. StepAIC also avoids multicollinearity by removing highly correlating predictor variables. Inter-rater reliability was examined with intraclass correlation coefficient (ICC) using the *irr* package in R (Gamer et al., 2012).

**Table 1**
Acoustic parameters and their operationalizations.

| Parameter | Operationalization |
|---|---|
| f0 range | Pitch range in semitones |
| f0 std | Standard deviation of pitch in semitones |
| f0 slope | Mean pitch slope in semitones |
| nPVI | Normalized pairwise variability index: rate-normalized mean difference (ms) between consecutive syllables |
| nΔS | Rate-normalized mean standard deviation of syllable duration (ms) |
| SP-ratio | Silent pause ratio: total duration of silent pauses / total duration of response |
| SP-freq | Rate of silent pauses per second: number of silent pauses / total duration of response |
| FP-ratio | Filled pause ratio: total duration of filled pauses / total duration of response |
| FP-freq | Rate of filled pauses per second: number of filled pauses / total duration of response |
| WL-ratio | Wrong language ratio: total duration of wrong language intervals / total duration of response |
| WL-freq | Rate of wrong language intervals per second: number of wrong language intervals / total duration of response |
| ArtRate | Rate of syllables per second without pauses or other disfluencies |
| SpeechRate | Rate of syllables per second, pauses and disfluencies included |

## 3. Results

Since the test tasks elicited spontaneous speech, the speech samples varied notably with respect to their length (number of syllables produced per response varied between 1 and 92). Compared to features estimated from a longer response, short responses often have insufficient evidence to compute reliable parameters. Thus, in order to improve the reliability of our parameters, we discarded samples that remained below the first quartile with respect to number of syllables. As a result, samples with less than eight (Swedish) syllables were excluded from the statistical analyses. This reduced the number of samples rated below A level from 30 to six and samples rated as A level from 163 to 138. All the B level and native samples included 8 or more syllables and the size of these groups thus remained intact. Descriptive statistics about the data (the number of speech samples, syllables, and disfluencies per proficiency group) is shown in Table 2.

As for the multiple linear regression models, samples rated below A level of proficiency were excluded because the group size decreased significantly when short samples were pruned from the data. Moreover, the raters were instructed to assess fluency and pronunciation only for samples they perceived as A-level or higher. Thus, the contribution of the acoustic parameters on the average ratings for proficiency, fluency, and pronunciation was analyzed using only the samples rated as A level or higher (total amount of samples analyzed in MLR models = 180).

### 3.1. Inter-rater reliability

Inter-rater reliability was examined with intraclass correlation coefficient (ICC) using the *irr* package in R (Gamer et al., 2012). In our data, each speech sample was rated by the same two to four raters. Intraclass correlations were thus computed using a two-way mixed effect model with rater as a fixed effect (Koo and Li, 2016). Missing values were discarded and only samples with more than one rating were studied. Since we use the mean ratings as a dependent variable, we computed inter-rater consistency relative to the mean of all ratings/sample (Koo and Li, 2016). This ICC consistency value was 0.90 for proficiency (95% confidence interval 0.84–0.95), indicating excellent reliability, 0.62 for fluency (95% confidence interval 0.26–0.83), indicating moderate reliability, and 0.31 for pronunciation (95% confidence interval −0.34–0.69), indicating poor reliability. Based on these inter-rater consistency values, we compare speaker groups based on their proficiency and fluency levels, and pronunciation ratings are

**Table 2**
The number of pruned speech samples, total time, total number of syllables & total number of labelled disfluencies per proficiency group. The information concerning the original data (including samples with less than 8 syllables) is shown in brackets.

| Group | Samples | Total time (min) | Syllables | SPs | FPs | WLs |
|---|---|---|---|---|---|---|
| Below A | 6 (30) | 1.25 (2.86) | 59 (166) | 31 (62) | 5 (10) | 7 (12) |
| A | 138 (163) | 28.69 (31.03) | 2047 (2191) | 785 (863) | 114 (109) | 67 (71) |
| B | 42 | 12.79 | 1255 | 365 | 41 | 21 |
| native | 30 | 5.53 | 1176 | 141 | 25 | 4 |

studied only in the multiple linear regression models.

### 3.2. Comparison of L1 and L2 speech

First, we investigated the differences between the native speakers and proficiency levels below A, A, and B with pairwise comparisons using Wilcoxon rank sum test. A pruned dataset was used, including 186 non-native and 30 native speech samples. The mean parameter values per group with their 95% confidence intervals are presented in Table 3.

Fig. 1 shows the distributions of the f0 change parameters. The native samples' distributions of f0 range in semitones were significantly different from L2 samples below A and A-level ($p < 0.05$) as well as from B-level samples ($p < 0.001$). In addition, the difference was significant between A- and B-level samples ($p < 0.05$). The f0 ranges of the native samples were smaller than the ones of A- and B-level samples, while the B-level samples resulted in larger f0 ranges than the A-level samples. The standard deviation (in semitones), in turn, showed significant difference only between natives and B-level samples ($p < 0.01$). Again, the parameter values were higher for B-level than for other speech samples.

The distributions of mean f0 slope of the native samples differed significantly from all L2 samples ($p < 0.001$ for A- and B-level samples, and $p < 0.01$ for samples below A level). The parameter values were higher for natives than other groups, indicating that natives produced, on average, steeper slopes than L2 speakers.

For the timing parameters, nPVI showed no significant differences between the groups. In fact, the nPVI distributions of native and B-level samples were nearly equal, while the nPVI values within lower proficiency groups were very irregular. The nΔS differed significantly between native and A-level samples ($p < 0.01$) but showed no significant differences between other groups (between B-level and native, between B-level and below A-level, and between below A-level and native). Again, the parameter values within lower proficiency groups showed notable irregularity. Fig. 2 shows the distributions of the two timing parameters with respect to proficiency level.

Both articulation and speech rate differed significantly between native and L2 samples ($p < 0.001$). The native samples showed considerably higher values for articulation and speech rates, as shown in Fig. 3. The articulation rate of L2 samples, however, did not differ significantly between proficiency levels, and significant difference in speech rate was found only between B-level and below A-level samples ($p < 0.01$).

For the disfluency parameters, filled pauses (FP-ratio and FP-freq) showed no significant differences between the groups. Instead, the distributions of silent pause-time ratios (SP-ratio) of the native samples differed significantly from all L2 samples ($p < 0.001$). The distributions in frequencies of silent pauses (SP-freq), in turn, differed significantly only between A- and B-level samples (0.001). The relative amount of speech uttered in a wrong language, WL-ratio, showed significant differences between natives and samples below A-level ($p < 0.001$) as well as between B- and below A-level samples ($p < 0.05$). Similarly, the WL-freq distribution of below A-level samples differed significantly from natives ($p < 0.001$) as well as from A- and B-level samples ($p < 0.01$). The distributions of SP-ratio, SP-freq, WL-ratio, and WL-freq are presented in Fig. 4.

Pairwise comparisons were also performed for the data grouped by fluency categories. A pruned dataset without samples below A-level was used, including 180 L2 and 30 L1 samples. The mean parameter values per group and their 95% confidence intervals are shown in Table 4.

The native samples' distributions of f0 range in semitones were significantly different from L2 samples with fluency category 1 (FC1, $p < 0.001$) and fluency category 2 (FC2, $p < 0.05$). The parameter showed no significant differences between the non-native groups. The standard deviation of f0 differed significantly only between natives and FC1 ($p < 0.05$). The mean f0 slope, in turn, showed significant differences between natives and all non-native groups. In addition, fluency category 3 (FC3) differed significantly from FC1 ($p < 0.01$). Fig. 5 shows the distributions of the f0 change parameters with respect to the fluency category and in comparison to the native speakers.

For the timing parameters, the nPVI showed no significant differences between groups, while nΔS was significantly different between natives and FC1 ($p < 0.001$). Interestingly, a statistically significant difference was also found between FC1 and FC2 ($p < 0.05$), but not between FC3 and other non-native groups. Fig. 6 shows the distributions of the timing parameters by fluency groups.

As for speed measures, both articulation rate and speech rate differed significantly between all groups ($p < 0.001$ for all pairs, except for articulation rate between natives and FC3: $p < 0.01$). The mean values in FC3 are closer to the one of natives than in proficiency group B. For the disfluency parameters, SP-ratio showed significant differences between all groups ($p < 0.001$ for all pairs, except for natives and FC3: $p < 0.05$). SP-freq, in turn, showed no significant differences between groups. FP-ratio differed significantly between natives and fluency categories 2 ($p < 0.01$) and 3 ($p < 0.001$) but showed no significant differences between FC1 and other groups. WL-ratio showed a significant difference only between natives and FC3 ($p < 0.01$). FP-freq and WL-freq remained insignificant in differentiating the groups. Fig. 7 depicts the distributions of articulation rate, speech rate, SP-ratio, and FP-ratio with respect to fluency categories.

### 3.3. Multiple linear regression models

The contribution of f0 change, timing parameters, and fluency parameters to the average ratings of proficiency, fluency, and pronunciation was studied through the computation of a step-by-step multiple linear regression, using the stepAIC algorithm implemented in R package *MASS* (Ripley et al., 2013). The 180 average ratings of either proficiency, fluency, or pronunciation were used as the dependent variable, and the 13 acoustic parameters presented in Table 1 were used as

**Table 3**
Mean values and 95% confidence intervals for acoustic parameters by proficiency group.

| | Below A (N=6) | | A (N=138) | | B (N=42) | | Natives (N=30) | |
|---|---|---|---|---|---|---|---|---|
| Parameter | mean | 95% CI | mean | 95% CI | mean | 95% CI | mean | 95% CI |
| ArtRate | 2.25 | 1.49,3.02 | 2.87 | 2.74,3.00 | 3.14 | 2.91,3.38 | 4.79 | 4.46,5.11 |
| SpeechRate | 0.85 | 0.57,1.13 | 1.45 | 1.32,1.59 | 1.71 | 1.49,1.93 | 3.88 | 3.48,4.27 |
| f0 range | 95.7 | 91.6,99.9 | 89.8 | 88.0,91.5 | 93.3 | 90.6,96.0 | 85.6 | 82.2,89.0 |
| f0 std | 3.71 | 2.62,4.80 | 3.39 | 3.12,3.66 | 3.69 | 3.30,4.08 | 2.70 | 2.31,3.08 |
| f0 slope | 6.22 | 4.33,8.11 | 6.65 | 6.18,7.11 | 7.38 | 6.34,8.41 | 10.9 | 9.74,21.1 |
| nPVI | 49.5 | 34.6,64.4 | 50.9 | 48.4,53.4 | 53.7 | 50.6,56.9 | 53.0 | 50.3,55.6 |
| nΔS | 0.49 | 0.42,0.56 | 0.47 | 0.45,0.50 | 0.51 | 0.48,0.53 | 0.53 | 0.50,0.56 |
| SP-ratio | 0.48 | 0.44,0.53 | 0.45 | 0.43,0.48 | 0.43 | 0.39,0.46 | 0.20 | 0.17,0.24 |
| SP-freq | 0.09 | 0.06,0.12 | 0.11 | 0.10,0.13 | 0.06 | 0.05,0.07 | 0.08 | 0.06,0.10 |
| FP-ratio | 0.05 | −0.01,0.11 | 0.03 | 0.02,0.03 | 0.02 | 0.01,0.03 | 0.05 | 0.04,0.07 |
| FP-freq | 0.05 | 0.01,0.09 | 0.03 | 0.02,0.04 | 0.03 | 0.02,0.04 | 0.04 | 0.02,0.06 |
| WL-ratio | 0.09 | 0.02,0.16 | 0.02 | 0.02,0.03 | 0.02 | 0.01,0.03 | 0.06 | −0.10,0.23 |
| WL-freq | 0.08 | 0.03,0.12 | 0.02 | 0.02,0.03 | 0.02 | 0.01,0.03 | 0.01 | −0.002,0.02 |

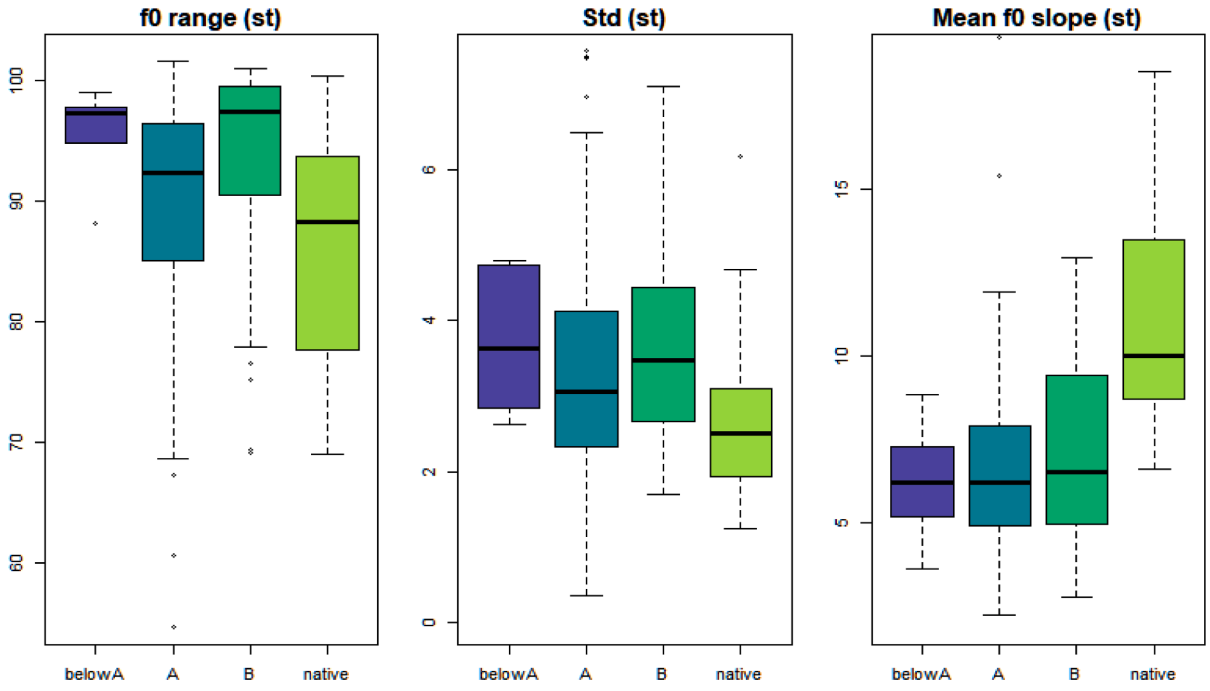


**Fig. 1.** F0 range, standard deviations, and mean slopes of analyzed speech samples in semitones by proficiency group (proficiency levels from left to right: below A, A, B, and native).

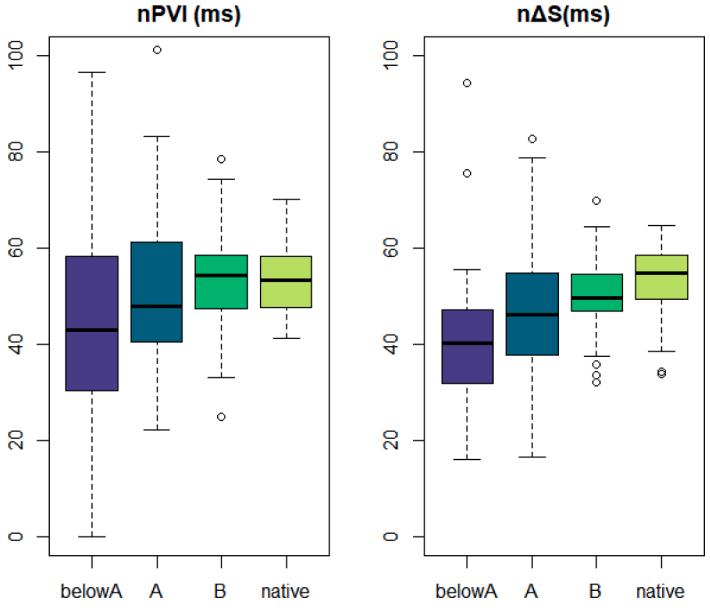


**Fig. 2.** The rate-normalized pairwise variability index for syllables (nPVI) and rate-normalized mean standard deviation of syllable duration (nΔS) of analyzed speech samples by proficiency group (proficiency levels from left to right: below A, A, B, and native).

predictor variables. The selection of predictors started with the full models, and the direction of the stepwise regression was set to default ("both"). The final models were compared with the full models with likelihood ratio tests that showed that the nested models fit the data as well as the full models ($p = 0.96$ for proficiency models, $p = 0.76$ for fluency models, and $p = 0.94$ for pronunciation models). Table 5 summarizes the results of the final MLR models with predictor t-values and respective significances based on p-values as well as the adjusted $R^2$ of the models.

The MLR model explained the greatest proportion of variation for the fluency ratings (multiple $R^2 = 0.46$ and adjusted $R^2 = 0.44$). The most significant predictor for fluency was speech rate with a positive t-value of 9.29, indicating that the higher the speech rate, the better the fluency rating. The disfluency parameters SP-freq, FP-ratio, and FP-freq also contributed to the fluency ratings, but the effect was significant only for SP-freq and FP-ratio. All disfluency effects were expected to be negative, indicating that the higher the disfluency parameter values, the lower the rating. This was the case for SP-freq and FPratio, but not for FP-freq (t-value = 1.52). The effect of FP-freq, however, remained statistically insignificant.

As for the timing parameters, normalized standard deviation of syllable durations (nΔS) showed a significant positive effect for fluency ratings (t-value = 2.38), indicating that the higher the nΔS, the better the rating. All f0 change parameters remained insignificant in predicting

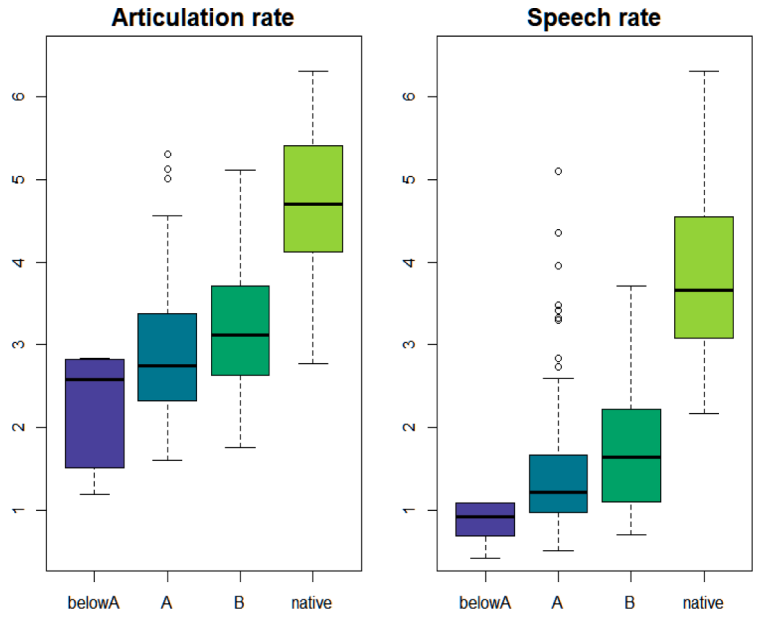


**Fig. 3.** Articulation and speech rate (syllables/second) of analyzed speech samples by proficiency group (proficiency levels from left to right: below A, A, B, and native).

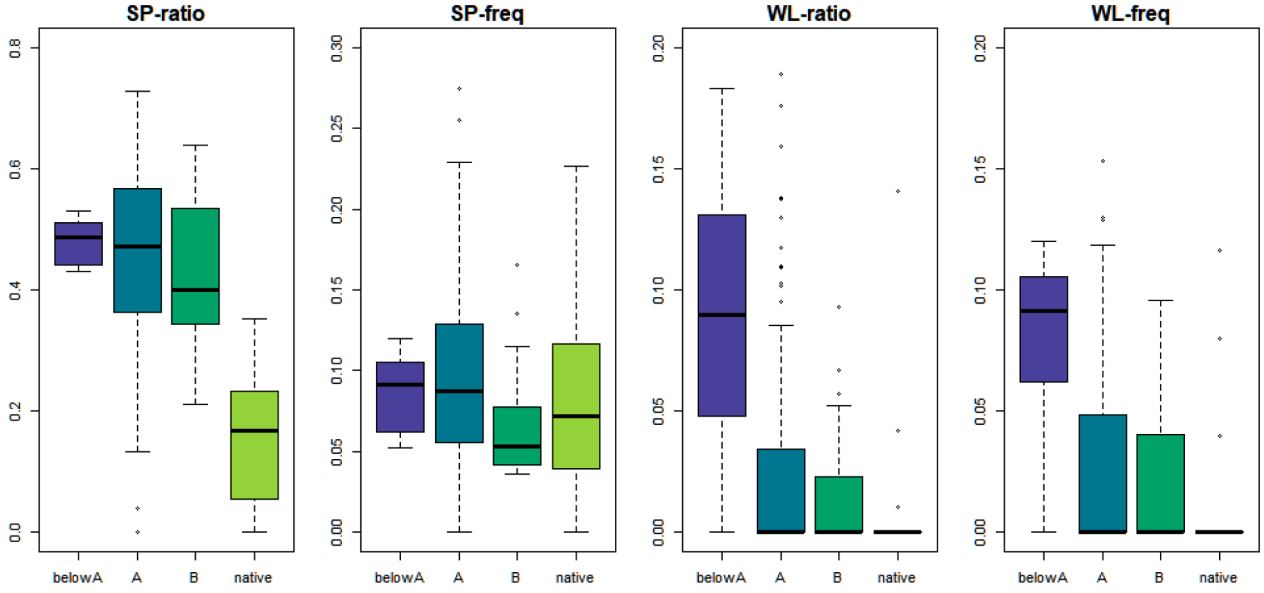


**Fig. 4.** SP-ratio, SP-freq, WL-ratio, and WL-freq of analyzed samples by proficiency group (proficiency levels from left to right: below A, A, B, and native).

**Table 4**
Mean values and 95% confidence intervals for acoustic parameters by fluency group.

| | FC1 (N=43) | | FC2 (N=118) | | FC3 (N=19) | | Natives (N=30) | |
|---|---|---|---|---|---|---|---|---|
| Parameter | mean | 95% CI | mean | 95% CI | mean | 95% CI | mean | 95% CI |
| ArtRate | 2.51 | 2.33,2.70 | 2.93 | 2.81,3.05 | 3.90 | 3.48,4.31 | 4.79 | 4.46,5.11 |
| SpeechRate | 0.98 | 0.88,1.08 | 1.51 | 1.40,1.63 | 2.72 | 2.28,3.16 | 3.88 | 3.48,4.27 |
| f0 range | 93.8 | 91.9,95.7 | 90.0 | 88.1,91.9 | 86.8 | 80.9,92.7 | 85.6 | 82.2,89.0 |
| f0 std | 3.59 | 3.15,4.04 | 3.47 | 3.18,3.76 | 3.11 | 2.52,3.70 | 2.70 | 2.31,3.08 |
| f0 slope | 5.71 | 5.05,6.38 | 7.02 | 6.46,7.59 | 8.05 | 6.90,9.19 | 10.90 | 9.74,12.10 |
| nPVI | 48.7 | 45.0,52.4 | 51.9 | 49.3,54.6 | 55.7 | 49.8,61.7 | 53.0 | 50.3,55.6 |
| nΔS | 0.45 | 0.39,0.50 | 0.49 | 0.47,0.51 | 0.50 | 0.44,0.56 | 0.53 | 0.50,0.56 |
| SP-ratio | 0.54 | 0.50,0.57 | 0.44 | 0.41,0.46 | 0.30 | 0.24,0.35 | 0.20 | 0.17,0.24 |
| SP-freq | 0.07 | 0.06,0.08 | 0.10 | 0.90,0.11 | 0.16 | 0.09,0.24 | 0.08 | 0.06,0.10 |
| FP-ratio | 0.04 | 0.02,0.05 | 0.02 | 0.02,0.03 | 0.01 | 0.003,0.2 | 0.05 | 0.04,0.07 |
| FP-freq | 0.04 | 0.03,0.05 | 0.03 | 0.02,0.04 | 0.02 | 0.01,0.04 | 0.04 | 0.02,0.06 |
| WL-ratio | 0.03 | 0.01,0.04 | 0.02 | 0.02,0.03 | 0.005 | −0.002,0.01 | 0.06 | −0.10,0.23 |
| WL-freq | 0.02 | 0.01,0.04 | 0.02 | 0.02,0.03 | 0.004 | −0.002,0.01 | 0.008 | −0.002,0.02 |

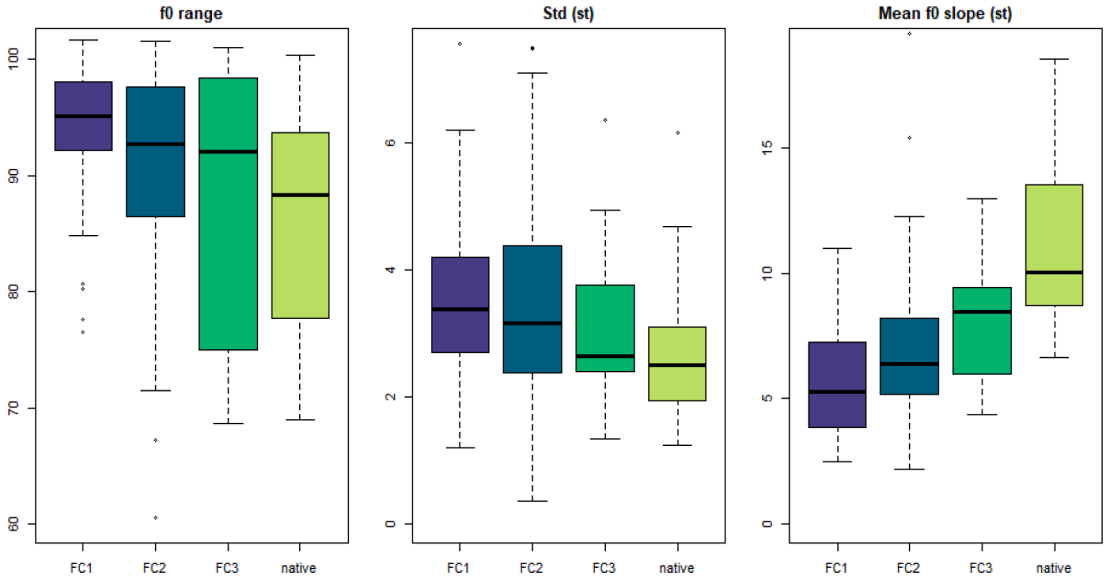


**Fig. 5.** F0 range, standard deviations, and mean slopes with respect to fluency categories.

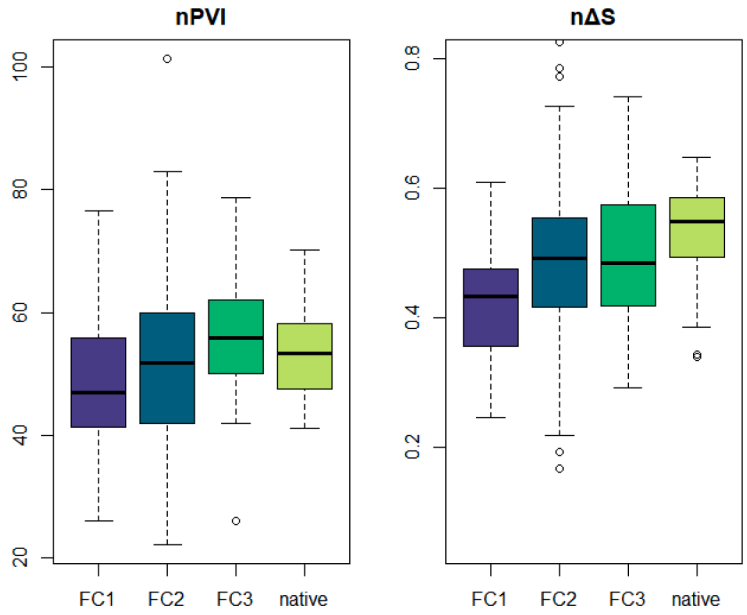


**Fig. 6.** The rate-normalized pairwise variability index for syllables (nPVI) and rate-normalized mean standard deviation of syllable duration (nΔS) of analyzed speech samples by fluency category.

fluency ratings. The standardized coefficients of the final model were 0.69 for speech rate (t-value = 13.70), −0.12 for SP-freq (t-value = −0.25), 0.12 for FP-freq (t-value = 0.11), −0.19 for FP-ratio (t-value = 0.16), and 0.13 for nΔS (t-value = 0.62).

The MLR model predicting overall oral proficiency accounted for 34 percent of the variance in the ratings (multiple $R^2 = 0.34$ and adjusted $R^2 = 0.33$). The most significant predictors of oral proficiency were speech rate (t-value = 7.99) and SP-freq (t-value = −8.06). Speech rate showed a positive effect, indicating that the faster the speech and the less breaks, the higher the proficiency. Conversely, as expected, the frequency of silent pauses showed a negative effect for proficiency ratings. From the disfluency parameters, also WL-ratio (t-value = −1.51) contributed to the prediction model but its effect remained statistically insignificant.

From the f0 change parameters, only mean f0 slope in semitones was included in the prediction model, although it did not provide a significant effect. The effect of f0 slope, however, was negative (t-value = −1.50), indicating that the smaller the mean slope, the better the proficiency rating. This result is in contrast with the comparisons between native and L2 groups in the previous section, where the mean f0 slope was significantly larger for native speech samples than for L1 speech samples. Both timing parameters related to syllable duration were excluded from the proficiency model. The standardized coefficients of the final model were 0.76 for speech rate (t-value = 8.01), −0.11 for f0 slope (t-value = −5.72), −6.43 for SP-freq (t-value = −0.84), and −0.09 for WL-ratio (t-value = −0.09).

Pronunciation ratings were the hardest to predict with the acoustic variables parameterized in this study. Only two predictors remained in the final MLR model: speech rate (t-value = 3.04) and WL-freq (t-value = −3.06). These predictors accounted only for eleven percent of the variance in the pronunciation ratings (multiple $R^2 = 0.11$ and adjusted $R^2 = 0.10$). The t-values for the standardized coefficients of the final

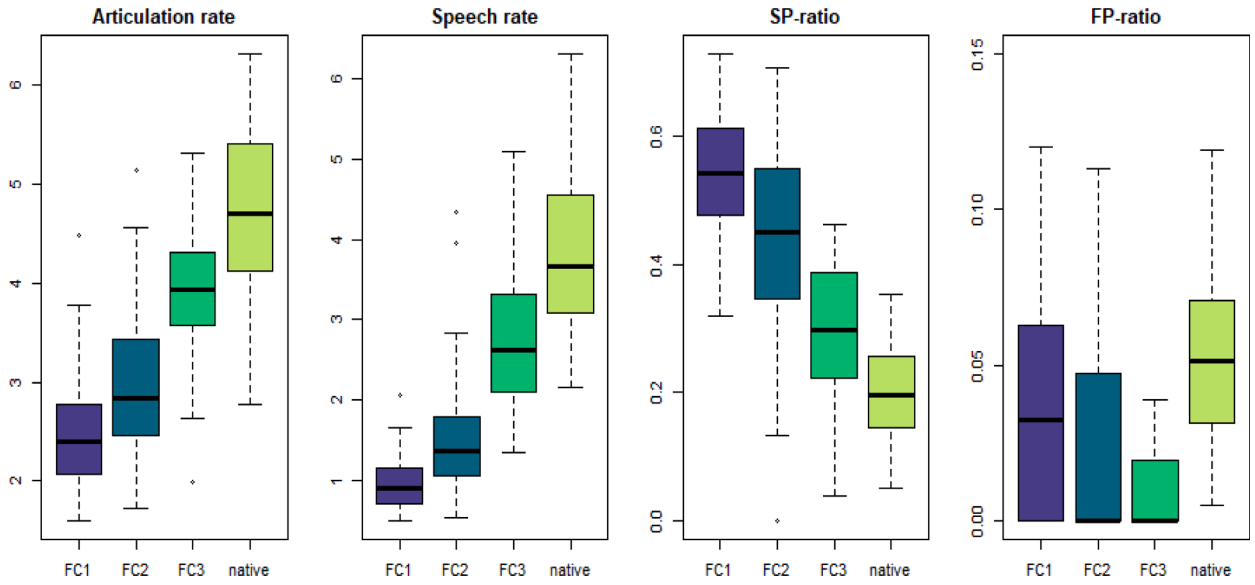


**Fig. 7.** Articulation rate, speech rate, SP-ratio, and FP-ratio of analyzed speech samples by fluency category.

**Table 5**
Summary of the MLR models with predictor t-values and model-specific adjusted $R^2$s. p-values: 0.1–0.05 ‘, 0.05–0.01*, 0.01–0.001**, < 0.001***.

| Predictor | Proficiency | Fluency | Pronunciation |
|---|---|---|---|
| f0 range | - | - | - |
| f0 std | - | - | - |
| f0 slope | −1.50 | - | - |
| nPVI | - | - | |
| nΔS | - | 2.38* | |
| SP-ratio | - | - | - |
| SP-freq | −8.06 *** | −1.65 | - |
| FP-ratio | - | −2.32 * | - |
| FP-freq | - | 1.52 | - |
| WL-ratio | −1.51 | - | - |
| WL-freq | - | - | −3.06 ** |
| ArtRate | - | - | - |
| SpeechRate | 7.99*** | 9.29 *** | 3.04** |
| Model $R^2$ (adj.) | 0.33 | 0.44 | 0.10 |

model were (t-value = 5.18) for speech rate and (t-value = −0.26) for WL-freq.

## 4. Discussion

This study examined the prosodic and fluency features of Finland Swedish as L2. The main findings of the study can be summarized in the following points, answering the research questions proposed in Section 1.3:

RQ1: L1 speakers of Finland Swedish differ from L2 speakers (with proficiency levels ranging from below A1 to B2) with respect to f0 change and fluency features, but timing parameters showed no clear division between native and L2 speakers. However, nΔS showed significant differences between speakers with respect to *fluency ratings*. The f0 parameters as well as fluency features also distinguished the speakers better with respect to fluency ratings than overall oral proficiency.

RQ2: F0 range, speech rate, and the frequency of silent pauses as well as “wrong language” sections in speech can distinguish A-level from B-level speakers.

RQ3: The contribution of f0 change, timing parameters, and fluency parameters in predicting average ratings varied depending on the assessed dimension. The predictive model for fluency ratings received the highest explanatory power with the following significant predictors: speech rate, rate-normalized standard deviation of syllable duration, and filled pause-time ratio. Overall oral proficiency ratings were slightly more difficult to predict than fluency ratings, while pronunciation ratings received a notably small coefficient of determination.

The reliability of the ratings varied with respect to the assessed dimension: proficiency ratings provided the highest ICC values that indicate excellent inter-rater consistency, and the ICC values for fluency ratings indicate moderate consistency. The ratings for pronunciation, however, provided a low ICC value indicating poor consistency, and therefore we opted not to use pronunciation ratings in grouping the L2 speakers for the pairwise comparisons. Moreover, the results of the MLR model predicting pronunciation ratings should be interpreted with caution.

Differences between L1 and L2 speakers occurred in all f0 change parameters, but there were some inconsistencies in the results. The f0 range of native samples was on average smaller than the ones of L2 speakers. Interestingly, however, the f0 range of B-level samples was larger than the one of A-level samples. Similarly, the standard deviation of f0 was larger for B-level samples than for all other groups (significant difference found only for native – B-level pair). However, the distributions of f0 change parameters were more linearly distinct when the data was grouped based on fluency ratings, as seen in figures in Section 3.2. As discussed in Section 1.2, both too narrow and too wide f0 range can be typical for L2 learners (see, e.g., Kautonen, 2017; Kuronen and Tergujeff, 2018). The phenomena can be related to difficulties found in L2 stress production: on one hand, too subtle stress contrasts can result in a monotonous intonation, and on the other hand, marking stress too strongly in speech can result in inappropriately wide f0 range.

Of the f0 parameters, the mean f0 slopes provided the clearest tendency: the native f0 slopes were significantly larger (or steeper) than the ones of all L2 groups. Although here was not enough variation within the f0 slopes of L2 samples in order to differentiate between proficiency levels, a significant difference was found between fluency categories 1 and 3. A clear tendency in mean f0 slopes can be seen in Fig. 5. Interestingly, however, f0 slope was not included in the MLR model for fluency, while it did improve the predictive power of the MLR model for overall oral proficiency. The role of f0 slope remained statistically non-significant in the unstandardized model for proficiency, but its beta coefficient gained a t-value of −5.78, indicating that the steeper the f0

slope of an L2 speaker, the lower the proficiency. This result is, however, inconsistent with the proficiency group comparisons, as the mean values and confidence intervals of B-level speakers were higher than the ones of A-level speakers. We will thus draw no conclusions on the role of f0 slope in L2 proficiency of Finland Swedish. Nevertheless, the clear difference between natives and non-natives indicate that the L2 speakers might fail to produce certain language-dependent intonational features of FS, encouraging f0 usage in L2 FS to be studied in more detail. Similar implications of f0 usage difficulties in L2 FS have been provided earlier by Kautonen (2017).

For the timing parameters, we expected at least some differences to be found between L1 and L2 samples despite the differing results found previously on FS rhythm (Vihanta et al., 1990; Heinonen, 2019). In our data, however, nPVI failed to distinguish speakers with respect to the assessed dimensions, but nΔS provided significant differences between natives and A-level samples as well as between fluency categories FC1 and FC2. Consequently, nΔS proved to be a significant predictor of fluency ratings with a positive t-value, supporting the findings of Heinonen (2019) and Kallio et al. (2020) on the syllable durations of FS learners. However, the t-value for the beta coefficient of nΔS was 0.62, indicating non-significant relationship with fluency ratings. Moreover, the effect size of nΔS remained very small.

The challenges with the timing parameters can stem from several aspects. For example, the use of duration as the dominant feature in PVI measures has received criticism, since duration cannot be assumed to be either the exclusive correlate of speech rhythm (see, e.g., White and Malisz, 2020) or to act independently from other cues in perception (Nolan and Asu, 2009). Rhythm is strongly related to stress, which in Swedish is found to be a combination of duration, f0, and intensity (Vihanta et al., 1990). In the study by Kallio et al. (2020), FS stress realizations were analyzed with a continuous wavelet transform (CWT) method that allows simultaneous use of duration, f0, and intensity (for more details, see Suni et al., 2017). With this method, Kallio et al. (2020) successfully predicted prosodic proficiency ratings of FS learners using correlations of L2 to L1 prominence estimates computed from sentences read by both native and non-native speakers of FS. In their study, duration proved to be the most significant single feature, while f0 had a detrimental effect on the prediction model. This result supports the previously acknowledged importance of duration as a stress signal in FS (Heinonen, 2019), which can also explain the contribution of nΔS in our results. However, it seems that in order to gain a comprehensive picture of L2 stress features one should consider not only temporal, but also the dynamic and tonal properties of stress production. A compositional analysis method, such as the CWT method in Kallio et al. (2020), is strongly recommended. However, the CWT method has been applied only for read L2 speech, enabling an exact comparison of L1 and L2 syllable prominence realizations and reducing the effect of linguistic and semantic context to the stress production (Suni et al., 2019; Kallio et al., 2020; 2021). Moreover, the method seems to work better with controlled speech data especially designed for studying stress production, such as in Kallio et al. (2020) where the data consisted of read sentences with minimal pairs differing only in word stress. In a more recent study, Kallio et al. (2021) compared the predictive power of the CWT method and traditional fluency measures with cross-linguistic data and found the contribution of the local prominence-based measurements to be relatively small compared to the temporal fluency measures. Although Kallio et al. (2021) also investigated read speech, the original material was not designed for the purposes of studying stress realizations. Since the data in the present study consists of spontaneous monologue speech that differ both in length and linguistic content, we opted for the traditional analysis of durational features.

Another issue regarding the use of duration-lead rhythm parameters might stem from the speech material: Krull and Engstrand (2003) discovered that syllable durations in CS and Spanish were more similar in spontaneous than in read speech. The characteristics of spontaneous speech may have affected our results as well: we ignored possible pauses between consecutive syllables in the computation of both timing parameters, which may have caused irregularities in parameter values especially in lower proficiency groups, where pauses and other disfluencies are relatively common. Also, hesitations in the L2 speech in our material sometimes lead to lengthening of syllables or words, which may have affected the timing parameters of the speakers. Moreover, increasing articulation rate can cause native (or native-like) speakers to simplify (and thus shorten) more complex syllable structure (Barry, 2007), which can result in smaller PVI values in native speakers than L2 speakers with significantly slower articulation rate. Although our timing parameters were rate-normalized, the normalization methods did not eliminate the effect of articulation rate.

Another result that supports further scrutinization of stress features in FS is that, despite failing to distinguish proficiency groups in pairwise comparisons, nΔS proved to be one significant predictor of fluency in the MLR models and showed significant differences between FC1 and FC2. This indicates that rate-normalized standard deviation of syllable durations indeed plays a role in overall temporal organization of Finland Swedish, but our parametrization failed in capturing the complexity of the phenomenon.

The parameters of speed fluency differed expectedly between native and L2 samples: both articulation and speech rate were significantly higher with native speakers than L2 speakers, and both parameters successfully distinguished speaker groups with respect to fluency ratings. With respect to proficiency, however, significant differences between L2 groups were found only in speech rates for B- and below A-level samples, while articulation rate did not distinguish the proficiency groups. The results of the MLR models further supported this observation: speech rate was found an extremely significant parameter in the MLR models predicting proficiency and fluency, following the results by, e.g., Cucchiarini et al. (2002), Kormos and Dénes (2004), and Kang and Johnson (2018), while articulation rate was not included in the prediction models. Speech rate also played a significant role in predicting pronunciation ratings, although the predictive power of the model was relatively low and the results concerning pronunciation should thus be interpreted with caution. Also the t-values for beta coefficients indicate that speech rate can significantly predict the human ratings in our data (t-value = 13.70 for fluency, t-value = 8.15 for proficiency, and t-value = 5.18 for pronunciation). This result supports the importance of speech fluency as universal fluency indicator.

The results of the present study differ slightly from some previous studies that found articulation rate to be a significant feature in predicting L2 proficiency or fluency (e.g., Cucchiarini et al., 2002; Kallio et al., 2017; 2021). This variation in the results can be due to differences in the selection of methods and variables: Cucchiarini et al. (2002) used correlations to examine the relations between individual acoustic parameters and perceived fluency, but they did not consider correlations between the acoustic parameters. In the current study, the stepAIC method used for feature selection avoids highly correlating variables, which has likely resulted in keeping only one speed measure in the MLR models: the correlation between articulation rate and speech rate in the present data was 0.74, indicating strong relationship between the two parameters. In the studies of Kallio et al. (2017, 2021), in turn, articulation rate was the only speed measure used.

The disfluency parameters also provided varying results between pairwise comparisons and multiple linear regression models. The clearest differences between speaker groups were found in silent pause-time ratio (SP-ratio) with respect to fluency categories (Fig. 7). The results indicate the possibility of SP-ratio thresholds for distinguishing speakers with different fluency levels. However, SP-ratio was not included in the prediction models. This is very likely due to strong multicollinearity between speech rate and SP-ratio (correlation –0.74) and was therefore expected. Interestingly, however, the frequency of silent pauses (SP-freq) proved extremely significant in predicting proficiency ratings as well as improved the prediction model for fluency. The effect of SP-freq to the ratings was negative, indicating that the more

often silent pauses occur, the lower the ratings. In the proficiency model, the standardized effect size for SP-freq was relatively high (−0.67, indicating that the increase of one SP per second decreases the proficiency rating by 0.67 grade). Moreover, SP-freq did not differ significantly between native and L2 speech, but instead distinguished A- and B-level samples from each other. The reason for this might be our decision not to apply a minimum pause threshold in the annotations, which enabled us to consider shorter pauses than the traditionally used 250 ms in fluency research (Derwing et al., 2004; Kormos and Dénes, 2004; Préfontaine et al., 2016). Shorter pauses increase SP-ratio less than longer pauses, but when occurring frequently they might disturb the speech flow and thus contribute to the perception of proficiency in L2 speakers. Another interesting result concerns filled pauses and fluency ratings: while the effect of FP-ratio on the fluency ratings proved significantly negative, FP-freq showed a positive effect. This anomaly might be due to the very rare occurrence of this feature in the speech data. It should be noted, however, that filled pauses were on average more common in the native speech data than in L2 speech, as shown in Fig. 7.

The linear regression model for fluency ratings provided the highest explanatory power (multiple $R^2 = 0.46$ and adjusted $R^2 = 0.44$). This was expected, since many of the parameters that we used have previously been found important in L2 fluency (Cucchiarini et al., 2002; Derwing et al., 2004; Kormos and Dénes, 2004; Bosker et al., 2013). However, the relatively small size of data as well as narrow rating scale for fluency can reduce the statistical significance of the results.

Pronunciation ratings were the hardest to predict: the selected acoustic parameters explained only 11% of the rating variation. Low explanatory power was expected for the pronunciation model since our parameters measured utterance-level global phenomena instead of segmental variation. Moreover, the low inter-rater reliability in pronunciation ratings make the prediction of the ratings difficult.

Overall, the selected parameters were more powerful in predicting average ratings than categorizing speech samples into different proficiency levels. Although the assessments were proven to be sufficiently consistent, variation in the ratings could have affected the categorization into proficiency levels, and analysis done with average ratings might provide more reliable results. Moreover, different parameters contributed to predicting different dimensions of oral proficiency. The results promote the complexity of oral language skills and indicate that in automatic assessment of L2 speech, the interaction of rating criteria and acoustic predictors should be considered carefully.

## 5. Conclusions

This study investigated f0 change, timing parameters, and fluency parameters as predictors of oral language skills with Finland Swedish as L2. The main motivation of the study is the development of an automatic assessment system for L2 Finland Swedish. Our results also contribute to filling a clear gap in the phonetic research on non-dominant varieties of pluricentric languages.

The results support the language-independent importance of speed measures, particularly speech rate, in predicting oral proficiency and fluency in L2. Our results further indicate that there can be applicable thresholds for silence-speech time ratio in spontaneous L2 speech for distinguishing speakers by their fluency levels. More research with cross-lingual data is recommended to find out suitable thresholds as well as whether language-independent, global thresholds exist. The role of f0 and timing parameters, in turn, are very likely dependent on the language context, but can provide valuable information on L2 learners' oral skills. While the f0 parameters did not prove significant in predicting oral L2 skills in the current study, a clear tendency was found between fluency categories and mean f0 slopes. A further scrutinization of the role of f0 slopes in FS is thus recommended. Moreover, mean standard deviation of syllable duration showed potential in predicting fluency ratings, supporting the integration of this feature in the development of automatic assessment of L2 Finland Swedish speech. However, a more precise method that considers the effect of pauses to syllable durations should be used in measuring this timing parameter.

## CRediT authorship contribution statement

**Heini Kallio:** Conceptualization, Formal analysis, Investigation, Methodology, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Maria Kautonen:** Conceptualization, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Mikko Kuronen:** Conceptualization, Investigation, Methodology, Data curation, Project administration, Writing – original draft, Writing – review & editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The authors do not have permission to share data.

## Acknowledgments

This work was supported by the Academy of Finland [grant number 322965]. The rating scales used in assessments were compiled by the second author and Anna von Zansen in the DigiTala research project. The authors would also like to thank the transcribers of the speech data and the raters who rated the speech samples.

## Appendix A. Holistic assessment criteria

Assessment criteria for overall oral proficiency based on National Core Curriculum 2003 (Finnish National Agency for Education, 2003). (Translated from the original Finnish version used in the assessment)

| Proficiency level | Description of speaking proficiency (based on NCC 2003) |
|---|---|
| Below A1 | • Cannot produce even short answers in the target language (words / phrases) |
| A1 | • Can speak briefly about himself / herself and his / her living environment, copes with the simplest dialogues and service situations<br>• Breaks, repetitions and interruptions are common<br>• Pronunciation can produce intelligibility problems<br>• Memorized expressions<br>• Grammatical errors can occur a lot<br>***Can speak more about familiar topics, pronunciation generally understandable, broader basic vocabulary.*** |

(continued on next page)

(*continued*)

| | |
|---|---|
| ***DIFFERENCES A1 > A2*** | |
| A2 | • Copes with simple social encounters, is able to start and end a short dialog<br>• A lot of breaks and false starts<br>• Pronunciation is intelligible, occasional problems with intelligibility due to pronunciation and pronunciation errors<br>• Masters basic vocabulary, some idiomatic expressions, past tense forms and conjunctions<br>• Several mistakes in basic grammar |
| ***DIFFERENCES A2 > B1*** | ***Also communicates in more demanding situations mainly fluently, pronunciation does not cause intelligibility problems, wider vocabulary and structures, but grammatical errors occur.*** |
| B1 | • Description of specific topics, copes with the most common everyday situations, expression may not be very accurate<br>• Can maintain fluent speech<br>• Pronunciation is intelligible, but pronunciation errors, atypical intonation and stress occur<br>• Fairly extensive vocabulary and common idioms, different structures and sentences<br>• Grammar errors occur, but they do not prevent the message from being communicated |
| ***DIFFERENCES B1 > B2*** | ***Expression more accurate also spontaneous, including conceptual topics, pronunciation and intonation more typical, broader vocabulary and control over structures, discretion, occasional grammatical errors.*** |
| B2 | • Is able to express himself/herself confidently, clearly and politely in the way required by the situation, sometimes needs paraphrasing<br>• Communicates fluently and also spontaneously, rarely longer breaks or hesitation<br>• Pronunciation and intonation are clear and natural<br>• Concrete and conceptual, familiar and unfamiliar topics, extensive vocabulary. Versatile structures<br>• Errors do not affect intelligibility, corrects them sometimes himself/herself |
| ***DIFFERENCES B2 > C1*** | ***Also competent in complex conceptual and detailed situations, speech almost effortless, expresses nuances of meaning through pronunciation (intonation and stress), control over vocabulary and structures does not restrict expression, corrects errors himself/herself if necessary.*** |
| C1 | • Participates actively in complex conceptual and detailed situations, copes with a wide range of social interactions as required by the situation<br>• Communication is smooth, spontaneous, almost effortless<br>• Varies intonation and masters sentence stress<br>• Vocabulary and structures are extensive, do not restrict expression<br>• Errors do not affect intelligibility, can correct them himself/herself |

## Appendix B. Analytic assessment criteria

Assessment criteria for analytic dimensions based on the Finnish National Core Curriculum (Finnish National Agency for Education, 2003, Finnish National Agency for Education, 2019)

(Translated from the original Finnish version used in the assessment)

**Fluency (sentence prosody)**

0 = no performance or no target language
1 = really disfluent; several breaks, repetitions and interruptions, hesitation
2 = fairly fluent; just some shorter breaks, repetitions and interruptions, hesitation
3 = really fluent, effortless; no disturbing breaks, repetitions, interruptions, hesitation

**Pronunciation (segments, syllable / sound durations, word stress)**

0 = no performance or no target language
1 = weak, difficult to understand, a lot of pronunciation errors
2 = moderate, fairly easy to understand, but some problems and pronunciation errors
3 = good, fully intelligible, no major pronunciation errors

## References


Abercrombie, D., 1967. Elements of general phonetics. Edinburgh University Press, Edinburgh.

Aho, E., 2010. Spontaanin puheen prosodinen jaksottelu. Doctoral dissertation. University of Helsinki, Helsinki. http://urn.fi/URN:ISBN:978-952-10-6405-0.

Arias, J.P., Yoma, N.B., Vivanco, H., 2010. Automatic intonation assessment for computer aided language learning. Speech Communication 52 (3), 254–267.

Bannert, R., 2004. På väg mot svenskt uttal. Studentlitteratur, Lund.

Barry, W., 2007. Rhythm as an L2 problem: how prosodic is it? In: J. Trouvain & U. Gut (Eds.), Non-native prosody: Phonetic descriptions and teaching practice. Mouton de Gruyter, Berlin, pp. 97–120.

Boersma, P., Weenink, D., 2021. Praat: doing phonetics by computer. University of Amsterdam, Amsterdam. http://www.fon.hum.uva.nl/praat/.

Bosker, H.R., Pinget, A.F., Quené, H., Sanders, T., De Jong, N.H, 2013. What makes speech sound fluent? The contributions of pauses, speed and repairs. Language Testing 30 (2), 159–175.

British Council. IELTS Speaking: band descriptors. https://www.ieltsessentials.com/global/results/assessmentcriteria.

Broselow, E., 1988. Prosodic phonology and the acquisition of a second language. Linguistic theory in second language acquisition. In: S. Flynn, W. O'Neil (Eds.), Springer, Dordrecht, pp. 295–308.

Bruce, G., 2010. Vår fonetiska geografi. Om svenskans accenter, melodi och uttal. Studentlitteratur, Lund.

Bruce, G., 2012. Allmän och svensk prosodi. Studentlitteratur, Lund.

Busà, M.G., Urbani, M., 2011. A cross linguistic analysis of pitch range in English L1 and L2. In: Proceedings of the 17th International Congress of Phonetic Sciences, pp. 380–383.

Campione, E., Véronis, J., 2002. A large-scale multilingual study of silent pause duration. In: Speech Prosody 2002, International Conference.

Chambers, F., 1997. What do we mean by fluency? System 25 (4), 535–544.

Cheng, J., 2011. Automatic assessment of prosody in high-stakes English tests. In: Twelfth Annual Conference of the International Speech Communication Association. Florence, Italy.

Cucchiarini, C., Strik, H., Boves, L., 2002. Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech. Journal of the Acoustical Society of America 111 (6), 2862–2873.

Council of Europe, 2001. Common European Framework of Reference for Languages: learning, teaching, assessment. https://rm.coe.int/1680459f97.

Council of Europe, 2020. Common European Framework of Reference for Languages: Learning. Teaching, Assessment. Companion volume. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4.

Cucchiarini, C., Doremalen, J.v., Strik, H., 2010. Fluency in non-native read and spontaneous speech. In: DiSS-LPSS Joint Workshop 2010. Tokyo, Japan.

Dellwo, V., 2006. Rhythm and speech rate: a variation coefficient for deltaC. In: P. Karnowski, I. Szigeti (Eds.), Language and Language-Processing. Peter Lang, Frankfurt/Main, pp. 231–241.
Derwing, T., Munro, M., 1997. Accent, intelligibility, and comprehensibility: evidence from Four L1s. Studies in Second Language Acquisition 19 (1), 1–16.
Derwing, T.M., Rossiter, M.J., Munro, M.J., Thomson, R.I., 2004. Second language fluency: judgments on different tasks. Language Learning 54 (4), 655–679.
Engstrand, O., Krull, D., 1994. Durational correlates of quantity in Swedish, Finnish and Estonian: cross-language evidence for a theory of adaptive dispersion. Phonetica 51 (1–3), 80–91.
Eriksson, A., 1991. Aspects of Swedish speech rhythm. Gothenburg Monographs in Linguistics 9. University of Gothenburg, Department of Linguistics, Gothenburg.
Escudero Mancebo, D., González Ferreras, C., Aguilar Cuevas, L., Estebas Vilaplana, E., 2017. Automatic assessment of non-native prosody by measuring distances on prosodic label sequences. In: Proceedings of Interspeech 2017, pp. 1442–1446. https://doi.org/10.21437/Interspeech.2017-366.
Finnish National Agency for Education, 2019. National Core Curriculum for General Upper Secondary Education 2019.
Finnish National Agency for Education, 2003. National Core Curriculum for Upper Secondary Schools 2003: National Core Curriculum for General Upper Secondary Education Intended for Young People.
Gamer, M., Lemon, J., Fellows, I., Singh, P., 2012. Package irr: Various coefficients of interrater reliability and agreement (Version 0.84). R Archive Network. https://cran.r-project.Org/web/packages/irr/irr.Pdf.
Golonka, E., Bowles, A., Frank, V., Richardson, D., Freynik, S., 2014. Technologies for foreign language learning: a review of technology types and their effectiveness. Computer Assisted Language Learning 27 (1), 70–105. https://doi.org/10.1080/09588221.2012.700315.
Grabe, E., Low, E.L., 2002. Durational variability in speech and the rhythm class hypothesis. In: Gussenhoven, C., Warner, N. (Eds.), Papers in Laboratory Phonology, VII. Mouton de Gruyter, Berlin, pp. 515–546.
Grosser, W., 1993. Aspects of intonation L2 acquisition. Current issues in European second language acquisition research. 81–94.
Grosser, W., James, A., Leather, J., 1997. On the acquisition of tonal and accentual features of English by Austrian learners. In: James, A., Leather, J. (Eds.), Second language speech: Structure and process. Studies on language acquisition 13. Mouton de Gruyter, Berlin, pp. 211–228.
Gut, U., 2009. Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German. Peter Lang, Frankfurt.
Götz, S., 2013. Fluency in native and nonnative English speech. John Benjamins Publishing, Amsterdam.
Heinonen, H., 2019. Durationsförhållandena i finskspråkiga gymnasisters uttal av L2-svenska: hur relaterar de till begripligheten? In: Bianchi, M., Håkansson, D., Melander, B., Pfister, L., Westman, M., Östman, C. (Eds.), Svenskans Beskrivning 36. University of Uppsala, Uppsala, pp. 95–106. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-378180.
Helgason, P., Ringen, C., Suomi, K., 2013. Swedish quantity: Central Standard Swedish and Fenno-Swedish. Journal of Phonetics 41 (6), 534–545. https://doi.org/10.1016/j.wocn.2013.09.005.
Hieke, A.E., Kowal, S., O'Connell, D.C., 1983. The trouble with "articulatory" pauses. Language and Speech 26 (3), 203–214.
Hirst, D., Di Cristo, A., 1998. A survey of intonation systems. Intonation systems: A survey of twenty languages. 1–44.
Huhta, A., Kallio, H., Ohranen, S., Ullakonoja, R., 2019. Fluency in language assessment. In: P. Lintunen, M. Mutta, P. Peltonen (Eds.), Fluency in L2 learning and use. Multilingual Matters, Bristol, pp. 129–145.
Hönig, F., Batliner, A., Weilhammer, K., Nöth, E., 2010. Automatic assessment of non-native prosody for English as L2. In: Proceedings of Speech Prosody, 2010. Chicago.
Isaacs, T., 2018a. Shifting sands in second language pronunciation teaching and assessment research and practice. Language Assessment Quarterly 15 (3), 273–293.
Isaacs, T., 2018b. Fully automated speaking assessment: changes to proficiency testing and the role of pronunciation. In: O. Kang, R. I. Thomson & J. M. Murphy (Eds.). The Routledge Handbook of Contemporary English pronunciation. Routledge, Abingdon, UK, pp. 570–584. https://doi.org/10.4324/9781315145006-36.
Ivars, A.-M., 2015. Dialekter och småstadsspråk. Svenskan i Finland – i dag och i går. Skrifter utgivna av Svenska litteratursällskapet i Finland (SLS) Nr 798. Helsinki: SLS. http://urn.fi/URN:NBN:fi:sls-978-951-583-496-6.
Jang, T.Y., 2008. Speech rhythm metrics for automatic scoring of English speech by Korean EFL learners. Malsori 66, 41–59.
Kahng, J., 2014. Exploring utterance and cognitive fluency of L1 and L2 English speakers: temporal measures and stimulated recall. Language Learning 64 (4), 809–854.
Kallio, H., Šimko, J., Huhta, A., Karhila, R., Vainio, M., Lindroos, E., Hildén, R., Kurimo, M., 2017. Towards the phonetic basis of spoken second language assessment: temporal features as indicators of perceived proficiency level. In: M. Kuronen, P. Lintunen, T. Nieminen (Eds.),. In: AFinLA-e: Soveltavan kielitieteen tutkimuksia, 10, pp. 193–213.
Kallio, H., Suni, A., Šimko, J., Vainio, M., 2020. Analyzing second language proficiency using wavelet-based prominence estimates. Journal of Phonetics, Volume 80, 100966, ISSN 0095-4470, https://doi.org/10.1016/j.wocn.2020.100966.
Kallio, H., Suni, A., Šimko, J., 2021. Fluency-related temporal features and syllable prominence as prosodic proficiency predictors for learners of English with different language backgrounds. Language and Speech,Vol 65 (3), 571–597.
Kang, O., Rubin, D., Pickering, L., 2010. Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. Modern Language Journal 94 (4), 554–566.
Kang, O., Johnson, D., 2018. The roles of suprasegmental features in predicting English oral proficiency with an automated system. Language Assessment Quarterly, 15 (2), 150–168.
Karhila, R., Rouhe, A., Smit, P., Mansikkaniemi, A., Kallio, H., Lindroos, E., Hildén, R., Vainio, M., Kurimo, M., 2016. Digitala: an augmented test and review process prototype for high-stakes spoken foreign language examination. In: Interspeech 2016, pp. 784–785.
Karlsson, F., 1983. Suomen kielen äänne- ja muotorakenne. WSOY, Porvoo/Helsinki/Juva.
Kautonen, M., 2017. Finskspråkiga talares intonation av finlandssvenska i påståendeyttranden i fritt tal. Folkmålstudier 55, 31–60.
Kautonen, M., 2019. Finskspråkiga inlärares uttal av finlandssvenska i fritt tal på olika färdighetsnivåer. JYU Dissertations 90. University of Jyväskylä, Jyväskylä. http://urn.fi/URN:ISBN:978-951-39-7778-8.
Kautonen, M., Kuronen, M., 2021. Uttalsinlärning med fokus på svenska. Skrifter utgivna av Svenska litteratursällskapet i Finland (SLS). Nr 860. http://urn.fi/URN:ISBN:978-951-583-552-9.
Kautonen, M., von Zansen, A., 2020. DigiTala research project: automatic speech recognition in assessing L2 speaking. Kieli, koulutus ja yhteiskunta 11 (4). https://www.kieliverkosto.fi/fi/journals/kieli-koulutus-ja-yhteiskunta-kesakuu-2020/digitala-research-project-automatic-speech-recognition-in-assessing-l2-speaking.
Kisler, T., Reichel, U., Schiel, F., 2017. Multilingual processing of speech via web services. Computer Speech & Language 45, 326–347.
Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine 15 (2), 155–163.
Kormos, J., Dénes, M., 2004. Exploring measures and perceptions of fluency in the speech of second language learners. System 32 (2), 145–164.
Krull, D., Engstrand, O., 2003. Speech rhythm – intention or consequence? Cross-language observations on the hyper-hypo dimension. Phonum 9, 133–136.
Kuronen, M., 2000. Vokaluttalets akustik i sverigesvenska, finska och finlandssvenska. Studia philologica Jyväskyläensia 49. University of Jyväskylä, Jyväskylä. http://urn.fi/URN:ISBN:978-951-39-4093-5.
Kuronen, M., Tergujeff, E., 2018. Second language prosody and its development: connection between different aspects. The Language Learning Journal 48:6, 685-699. doi:10.1080/09571736.2018.1434228.
Kuronen, M., Ullakonoja, R., & Kautonen, M. (2016). Inlärningen av de svenska tonaccenterna hos finska S2-talare – automatiseras uttalet? Språk och Stil 26, 161–194. urn:nbn:se:uu:diva-314627.
Ladefoged, P., Johnson, K., 2014. A Course in Phonetics. Nelson Education, Toronto.
Leinonen, K., 2004. Finlandssvenskt sje-, tje- och s-ljud i kontrastiv belysning. Jyväskylä Studies in Humanities 17. University of Jyväskylä, Jyväskylä. http://urn.fi/URN:ISBN:951-39-1828-9.
Lennon, P., 2000. The lexical element in spoken second language fluency. In. H. Riggenbach (Ed.). Perspectives On Fluency. University of Michigan Press, pp. 25–42.
Li, K., Wu, X., Meng, H., 2017. Intonation classification for L2 English speech using multi-distribution deep neural networks. Computer Speech & Language 43, 18–33.
Lindström, J., Norrby, C., Wide, C., Nilsson, J., 2017. Intersubjectivity at the counter: artefacts and multimodal interaction in theatre box office encounters. Journal of Pragmatics 108, 81–97. https://doi.org/10.1016/j.pragma.2016.11.009.
Liss, J.M., White, L., Mattys, S.L., Lansford, K., Lotto, A.J., Spitzer, S.M., Caviness, J.N., 2009. Quantifying speech rhythm abnormalities in the dysarthrias. Journal of Speech & Hearing Research 52 (5), 1334–1352.
Low, E.L., Grabe, E., 1995. Prosodic patterns in Singapore English. In: Proceedings of the International Congress of Phonetic Sciences. Stockholm, 3, pp. 636–639.
Low, E.L., Grabe, E., Nolan, F., 2000. Quantitative characterisations of speech rhythm: 'Syllable-timing'. Language and Speech 43, 377–401.
Luo, D., Gu, W., Luo, R., Wang, L., 2016. Investigation of the effects of automatic scoring technology on human raters' performances in L2 speech proficiency assessment. In: 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2016. IEEE, pp. 1–5.
McAllister, R., Flege, J.E., Piske, T., 1999. The acquisition of Swedish long vs. short vowel contrasts by native speakers of English, Spanish and Estonian. In: Proceedings of the 14th International Congress of Phonetic Sciences, pp. 751–754.
Mennen, I., 2007. Phonological and phonetic influences in non-native intonation. In: J. Trouvain & U. Gut (Eds.), Non-Native Prosody: Phonetic Descriptions and Teaching Practice. Mouton de Gruyter, Berlin, pp. 53–76.
Ministry of Education and Culture, 2017. Gaudeamus igitur: Ylioppilastutkinnon kehittäminen. Opetus- ja kulttuuriministeriön Julkaisuja 2017:16. Opetus- ja kulttuuriministeriö, Helsinki.
Munro, M.J., Derwing, T.M., 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Language Learning 45 (1), 73–97.
Nolan, F., Asu, E.L., 2009. The pairwise variability index and coexisting rhythms in language. Phonetica 66 (1–2), 64–77.
Norrby, C., Wide, C., Lindström, J., Nilsson, J., 2011. Finland Swedish as a non-dominant variety of Swedish - extending the scope to pragmatic and interactional aspects. In R. Muhr (Ed.). Non-dominant Varieties of Pluricentric Languages. Getting the Picture. In memory of Prof. Michael Clyne. Peter Lang Verlag, pp. 11–25. Wien et Al.
O'Dell, M., Lennes, M., Werner, S., Nieminen, T., 2007. Looking for rhythms in conversational speech. In: Proceedings of the 16th International Congress of Phonetic Sciences, pp. 1201–1204.
Peltonen, P., Lintunen, P., 2016. Integrating quantitative and qualitative approaches in L2 fluency analysis: a study of Finnish-speaking and Swedish-speaking learners of English at two school levels. Europian Journal of Applied Linguistics 4 (2), 209–238.

Préfontaine, Y., Kormos, J., Johnson, D.E., 2016. How do utterance measures predict raters' perceptions of fluency in French as a second language? Language Testing 33 (1), 53–73.

Qian, Y., Lange, P., Evanini, K., 2020. Automatic speech recognition for automated speech scoring. In: Zechner, K., Evanini, K. (Eds.), Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech. Routledge, New York, pp. 61–74. https://doi.org/10.4324/9781315165103-4.

Ramus, F., Nespor, M., Mehler, J., 1999. Correlates of linguistic rhythm in the speech signal. Cognition 73 (3), 265–292.

Rasier, L., Hiligsmann, P., 2007. Prosodic Transfer from L1 to L2. Theoretical and Methodological Issues. Nouveaux Cahiers De Linguistique Francaise 28. Universite de Geneve, Genève, Schweiz, pp. 41–66.

Riad, T., 2014. The Phonology of Swedish. Oxford University Press, Oxford.

Ringen, C., Suomi, K., 2009. Fenno-Swedish VOT: influence from Finnish? In: Branderud, P., Traunmüller, H. (Eds.), Proceedings of FONETIK 2009. Department of Linguistics, University of Stockholm, Stockholm, pp. 60–65.

Ringen, C., Suomi, K., 2012. The voicing contrast in Fenno-Swedish stops. Journal of Phonetics 40 (3), 419–429. https://doi.org/10.1016/j.wocn.2012.02.010.

Pearson, 2017. PTE Academic Score Guide. https://pearsonpte.com/wp-content/uploads/2017/08/Score-Guide.pdf. (Accessed: 2021-02-20).

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D., Ripley, M.B., 2013. Package 'mass'. Cran r, 538, 113–120.

Strangert, E., 1985. Swedish Speech Rhythm in a Cross-Language Perspective. Almqvist & Wiksell International, Stockholm.

Suni, A., Šimko, J., Aalto, D., Vainio, M., 2017. Hierarchical representation and estimation of prosody using continuous wavelet transform. Computer Speech & Language 45, 123–136.

Suni, A., Kallio, H., Benus, S., Šimko, J., 2019. Characterizing second language fluency with global wavelet spectrum. In: Proceedings of the 19th International Congress of Phonetic Sciences. Australasian Speech Science and Technology Association Inc, Melbourne, Australia.

Tavakoli, P., Skehan, P., 2005. Strategic planning, task structure and performance testing. In Ellis, R. (Ed.), Planning and Task Performance in a Second Language. John Benjamins Publishing, Amsterdam, pp. 239–273.

Tevajärvi, K., 1982. Intonation in Finland-Swedish: word and sentence stress in the Helsinki dialect. In: Working papers, 22. Lund: Lund University, Department of Linguistics, pp. 175–180.

The Finnish Matriculation Examination Board, 2021. Ilmoittautumiset eri kokeisiin tutkintokerroittain 2012–2021. https://www.ylioppilastutkinto.fi/ext/stat/FS2021A2012T2010.pdf.

Thomas, E.R., Carter, P.M., 2006. Prosodic rhythm and African American English. English World-Wide 27 (3), 331–355.

Trofimovich, P., Baker, W., 2006. Learning second language suprasegmentals: effect of L2 experience on prosody and fluency characteristics of L2 speech. Studies in Second Language Acquisition 28 (1), 1–30.

Ullakonoja, R., 2007. Comparison of pitch range in Finnish (L1) and Russian (L2). In: Proceedings of the 16th International Congress of Phonetic Sciences. Saarbrücken, Germany. Universität des Saarlandes.

Vihanta, V., Leinonen, K., Pitkänen, A., 1990. On rhytmic features in Finland-Swedish and Sweden-Swedish. In: K. Wiik & I. Raimo (Eds.), Nordic Prosody V. University of Turku, Turku, pp. 325–350.

Wennerström, A., 2000. The role of intonation in second language fluency. In H. Riggenbach (Ed.), Perspectives On Fluency. University of Michigan Press, pp. 102–127.

White, L., Mattys, S., 2007. Calibrating rhythm: first language and second language studies. Journal of Phonetics 35, 501–522.

White, L., Malisz, Z., 2020. Speech Rhythm and Timing. In: C. Gussenhoven, A. Chen (Eds.), The Oxford Handbook of Language Prosody. Oxford University Press, USA, pp. 166–179.

Wik, P., 2011. The Virtual Language Teacher: Models and Applications For Language Learning Using Embodied Conversational Agents. Royal Institute of Technology, Stockholm. Doctoral Thesis. urn:nbn:se:kth:diva-33579.

Zechner, K., Higgins, D., Xi, X., Williamson, D.M., 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Communication 51 (10), 883–895. https://doi.org/10.1016/j.specom.2009.04.009.