Kuronen, M., P. Lintunen & T. Nieminen (toim.) 2017. Näkökulmia toisen kielen puheeseen – Insights into second language speech. AFinLA-e. Soveltavan kielitieteen tutkimuksia 2017 / n:o 10. 193–213.

Heini Kallio[1], Juraj Šimko[1], Ari Huhta[2], Reima Karhila[3], Martti Vainio[1], Erik Lindroos[1], Raili Hildén[1] & Mikko Kurimo[3]

[1]University of Helsinki, [2]University of Jyväskylä, [3]Aalto University

Towards the phonetic basis of spoken second language assessment: temporal features as indicators of perceived proficiency level

This study investigates whether temporal features in speech can predict the perceived proficiency level in Finnish learners of Swedish. In so doing, seven expert raters assessed speech samples produced by 60 upper secondary school students using the revised CEFR scale for phonological control. The effect of temporal features was studied with a cumulative link mixed model, and the assessments were further analyzed to study inter-rater variation. The results indicate that articulation rate and certain types of disfluencies in speech can predict the perceived proficiency level. Furthermore, assessors seem to weigh temporal features differently depending on the speech type and their individual focus.

Keywords: L2 proficiency assessment, L2 phonetics, prosody, temporal features

1 Introduction

Speaking a language fluently is often the ultimate goal of mastering a second or foreign language (L2), and fluency in some sense is frequently tested in high-stakes tests that have a tangible effect on learners' futures. In Finland, the Ministry of Education and Culture (MEC) have set a goal to include oral language skills as part of language tests at the end of upper secondary education by 2022 (Ministry of Education and Culture 2017). This increases the importance and relevance of research on spoken L2 skills and their assessment.

Despite the regular use of the term fluency in pedagogy, language testing and linguistics, its definition varies (Huhta 1993; Chambers 1997). Lennon (2000) presents two types of fluency definitions: a broad one and a narrow one. The broad sense corresponds to a higher-order, global oral proficiency, while the narrower definition of fluency refers to spoken performance, more closely to the temporal properties and "smoothness" of the speech. The present study approaches fluency from the latter, narrower perspective with the reliability and efficiency of the assessments in mind. Lennon (2000) states that an individual's fluency is acceptable as long as it engages the listener. L2 speakers' fluency is indeed very likely the primary measure that ordinary interlocutors assess in everyday interaction. It is therefore crucial to understand and study fluency and its relation to language proficiency.

1.1 Measuring fluency: an overview

While proficiency is generally the target of assessments, the assessment criteria are often based on assumptions on fluency. Reliable ways to measure spoken fluency are therefore important in language testing as well as future research in applied linguistics, but the variables behind the perception of fluency remain less studied (Koponen & Riggenbach 2000). Over the last decades, however, speech researchers have been increasingly interested in the prosodic (tonal, temporal and dynamic) features of L2 speech. In this paper we will focus on the temporal features of L2 speech fluency and their effect on language proficiency assessments. Speech fluency is generally related to temporal features that have also proven to be strong predictors of prosodic competence (see, e.g., Cheng 2011). Previous studies on L2 speech have used measures such as speech and articulation rate, mean length of run, phonation-time ratio, number of stressed words as well as number and duration of pauses to measure second language fluency (Cucchiarini et al. 2002; Derwing et al. 2004; Kormos & Dénes 2004; Trofimovich & Baker 2006; Hönig et al. 2010; Bosker et al. 2013). Derwing et al. (2004) also stated that the measure of standardized pruned syllables (self-corrections, self-repetitions, false starts and non-lexical filled pauses) was a successful predictor of fluency judgements. Bosker et al.

(2013), in turn, found that filled and silent pauses and the mean length of pause together with mean length of syllable are significant predictors of fluency. Researchers have obtained slightly mixed results, however, depending on the target language, fluency measures taken and the design of the study. The relevance of temporal variables together with the measurement methods have been under discussion. For example, researchers lack consensus on the minimum length of a pause, despite the wide range of pause studies in both native (L1) and L2 speech. This is problematic, since all temporal measurement results (except speech rate) depend on the pause threshold used. Many studies of L1 and L2 speech follow Goldman-Eislers' (1968) proposal of a pause threshold of 250 ms based on distinguishing "articulatory" (< 250 ms) and "hesitation" (> 250 ms) pauses (see, e.g., Bosker et al. 2013; De Jong & Bosker 2013). Yet many pauses shorter than 250 ms cannot be attributed to articulation (Hieke et al. 1983; Campione & Véronis 2002) but have a psychological function. For example, speech rate and pausing are involved in expressing paralinguistic information, such as confidence or emotions (Scherer et al. 1973; Scherer 1986). In the current study, we consider also the relevance of pauses shorter than 250 ms as part of perceived fluency.

Another issue in fluency measures is the types and number of disfluencies. Silent pauses – periods of vocal inactivity during speech performance – are easy to detect automatically from speech, especially if pause threshold is set high enough. There can, however, be non-silent, non-lexical pauses in speech that affect the perception of fluency. These pauses are often referred to as filled or hesitation pauses. Other temporal disfluencies include phenomena such as self-corrections and self-repetitions (Derwing et al. 2004; Bosker et al. 2013).

1.2 The context-dependent disfluencies

Many studies concentrate on several disfluencies at a time (Lennon 1990; Foster & Skehan 1996; Towell et al. 1996; Derwing et al. 2004). Filled pauses are typical for spontaneous speech and rarely occur in read speech. The use of filled and unfilled pauses, however, is highly context-dependent. For example in French filled pauses seem to be frequent and long especially in conversational speech, and the distribution of silent pauses is related to the syntactic structure of the sentence (Duez 1982). In spontaneous Finnish speech filled pauses might be less common and silent pauses more acknowledged than in French, but there is no extensive study comparing pauses in these languages. However, Toivola et al. (2009) found that native Finnish speakers tend to have longer silent pauses in read speech than non-native speakers of Finnish. Their results indicate that long silent pauses do not necessarily affect the quality of L2 speech. Campione & Véronis (2002) noted that Italians make generally

shorter pauses and Spanish, in turn, longer pauses than French, English and German. These results prompt to take into account the language-specific temporal features and individual differences in prosody, when analyzing pauses as fluency measures. Whether a certain type of pause is considered as a speech disfluency or not is indeed language-specific. It is thus relevant to study the effect of various disfluencies in the perception of fluency. However, the use of too many variables can lead to unreliable results, since many analysis methods require uncorrelating variables, but many temporal features in speech depend on each other: for example, the amount of pauses affects strongly the measurement of speech rate. This leads to a high correlation between these variables. In the current study, we avoid intercorrelating variables and use revised, systematic measurements that are less used in fluency research.

1.3 Assessing fluency as a part of language proficiency

Spoken language fluency studies have generally involved trained assessors (see, e.g., Wennerstrom 2000; Kormos & Dénes 2004), but also native speakers of the target language have proved to be reliable raters of L2 fluency (Derwing et al. 2004, 2006; Rossiter 2009). Studies on temporal features and fluency vary with regard to instructions and criteria given to assessors as well as speech type and sample length. Speech samples have generally been short, from less than 20 seconds to 2–3 minutes, and included either read sentences or narrative speech. The studies have commonly used Likert-type scales (see, e.g., Bosker et al. 2013; Pinget et al. 2014), but the wide descriptors of the Common European Framework of Reference for languages (CEFR, Council of Europe 2001) have also been used in fluency assessments (Préfontaine et al. 2016). The CEFR is a guideline used to describe achievements of L2 learners across Europe and in other countries. The present study is novel in that it uses updated descriptors of the CEFR scale for phonological control to investigate the perceived fluency of Finnish learners of Swedish. The revised phonological control scale consists of two descriptive subsections: prosodic features and articulation of sounds. In this study we examine the descriptor scale for prosodic features, since temporal properties of speech are considered as part of prosody. Additionally, our research reflects actual test contexts with regard to data collection methods and the assessment protocol.

1.4 The goal of this study

The goal of this study is to make spoken L2 assessment more objective and reliable by scrutinizing the quantitative temporal features in speech that affect the assessments, even when the assessors are unaware of these features. This gives us valuable information about the L2 learning and assessing process. Our

study addresses the following research questions (RQ):

a) Can objectively measured temporal features in speech be used to predict the proficiency level of the speaker?

b) How do read and semi-spontaneous speech differ with regard to temporal features and assessments?

c) How do the raters differ in their assessments?

To answer RQ(a), objective acoustic measurements of speech are related to subjective proficiency ratings of the same speech samples. A group of trained raters assessed the proficiency level of Finnish learners of Swedish. Based on the previous studies (Cucchiarini et al. 2002; Derwing et al. 2004; Kormos & Dénes 2004; Bosker et al. 2013), we expect that articulation rate and pauses have stronger effect on proficiency ratings than disfluencies related to repairing (repetitions and corrections). To answer RQ(b), we study the two speech types separately. We consider the speech type elicited from different test tasks as a meaningful variable. To answer RQ(c), we study the effect of acoustic variables separately on each assessor. The assessors are also analyzed with respect to intra- and inter-rater consistency. We expect to find some systematic differences between the assessors.

2 Material and methods

2.1 Speech data

The data used in this study is part of a larger speech corpus, which has been collected while piloting a computer-aided oral language test for large-scale purposes (the project DigiTala[1]). The piloting was done in Finnish Swedish as a second language. Finnish Swedish is a variant of Swedish spoken in Finland and differs from standard Swedish with regard to e. g. sound production as well as prosodic features like sentence and word stress. Swedish is a compulsory subject in basic education in Finland, and the national matriculation examination test of L2 Swedish is taken by approximately 8,000 upper secondary school students yearly, which makes Swedish the second most tested L2 in Finland (Finnish Matriculation Examination Board 2017).

Seven upper secondary schools from six municipalities around Finland participated in the pilot tests, and speech data from approximately 760 voluntary pupils has been recorded and stored anonymously in a database. The pilot test works as a web-based application and includes four subtasks:

[1] http://blogs.helsinki.fi/digitala-projekti/

a) A read-aloud task: newspaper headlines or a written phone message

b) Situational reacting task: reacting to situations given in written L1 or with a picture and a written clue

c) A simulated video phone call with pre-recorded replies from one native speaker of the target language

d) A live dialogue task with a peer.

These tasks cover various dimensions of speaking proficiency, differing in discourse genre, formality and complexity. The use of computer-administered tasks helps to standardize the testing conditions for all participants as well as to reduce the effect of the behavior of other individuals in the communication, enabling to study language proficiency as an individual attribute. The subtasks were split to smaller sections and every task was timed. Instructions were in written Finnish or Swedish depending on the task.

A subset of 60 speech samples from the larger pilot data was used in this study. 50 samples were produced by native Finnish and 10 by native Finnish Swedish upper secondary school students, aged 16–18 years. The native samples were selected to elicitate higher proficiency scores and thus enable investigating all levels of the CEFR scale. The speech samples included read (n = 19) and semi-spontaneus (n = 41) utterances from subtasks a, b and c.

2.2 Assessments

Seven trained evaluators, Swedish language teachers or native speakers of Finnish Swedish (see Table 1) assessed each speech sample using the updated Common European Framework of Reference for Languages (CEFR) scales (Council of Europe 2017). The assessments were part of piloting the revised descriptor scales from a proposed version of the CEFR illustrative descriptors, authorized by the Council of Europe Language Policy Section. This study focuses on the assessments of prosodic features, which is a subsection of the revised CEFR descriptor scale for phonological control and pays attention to features such as word and sentence stress, rhythm and intonation with respect to the perceived intelligibility of the speech.

The descriptor scales for prosodic features were translated to Finnish. The assessors were familiarized with the new reference scale and some speech samples were assessed and discussed together before the actual rating task. No specific instructions to focus on the temporal features were given to assessors, unlike many previous studies have done (see, e.g., Derwing et al. 2004; Bosker et al. 2013; Pinget et al. 2014). The assessments were collected using a web interface, where the assessors could listen and assess the speech samples at their own pace. Before the actual assessment task the assessors were

TABLE 1. The background of the assessors.

Assessor	L1	Language teacher	Phonetic studies	Studied/taught spoken L2 skills
A1	Finnish	Swedish	Yes	Yes
A2	Finnish	Swedish	Yes	Yes
A3	Finnish	Swedish	Yes	Yes
A4	Swedish	No	Yes	No
A5	Finnish	Swedish	Yes	Yes
A6	Finnish	Swedish	Yes	Yes
A7	Swedish	Other language	Yes	No

asked to answer questions concerning their background. The background information of the assessors is presented in Table 1.

2.3 Acoustic analysis

Annotation and analysis were done using the Praat software (Boersma & Weenink 2010) and a script for large-scale systematic analysis of continuous prosodic events (Xu 2013). The acoustic variables measured for statistical analysis were articulation rate (AR), silent pause-time ratio (S), filled pause-time ratio (F), and corrections or repetitions-time ratio (CR). These four fluency variables were chosen because previous research suggests that similar measures are salient predictors of fluency. Intercorrelating variables, like speech rate and overall pause-time ratio, were avoided. Disfluency-time ratios are used as fluency measures instead of more commonly used number of disfluencies, since the sole frequency of disfluencies does not give information on the amount of time they take during speech. Disfluency-time ratio is therefore seen as more comprehensive measure, because it depends on both the number and the length of disfluencies. However, relative amount of disfluencies and distributions of disfluency durations are used to illustrate and reflect on the differences between read and semi-spontaneous speech.

All disfluencies of 50 ms or longer were marked manually and measured from every sample (see Figure 1 for annotation example). Articulatory pauses, such as plosive closure phases or prepausal lengthening, that were not part of hesitations, self-corrections or self-repetitions were excluded. Disfluency-time ratios for all disfluency types and articulation rate were calculated separately for read and semi-spontaneous speech samples. The variables were operationalized as follows:

a) Articulation rate (AR): The number of syllables produced per second excluding all pauses longer than 50 ms. Number of syllables was used instead

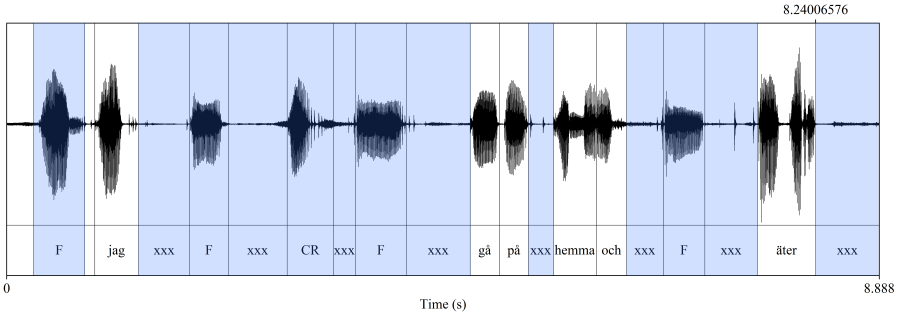

FIGURE 1. Example of pause annotation. F = filled pause, CR = self-correction or repetition, xxx = silent pause.

of phones, since the syllable is considered as the smallest speech unit that carries relevant prosodic and articulatory information (Collier 1992: 206).

b) Silent pause-time ratio (S): The total duration of silent pauses above 50 ms (no vocal activity) divided by total duration of the given speech sample. Furthermore, silent pauses were grouped by their length into three categories: short (SS, < 200 ms), medium (MS, 200–1,000 ms) and long (LS, > 1,000 ms), following the trimodal distribution presented by Campione & Véronis (2002).

c) Filled pause-time ratio (F): The total duration of filled pauses above 50 ms divided by total duration of the given speech sample. Pauses were considered as filled, when they included non-lexical vocal activity, often used as hesitation markers such as “umm”.

d) Corrections or repetitions-time ratio (CR): The total duration of self-corrections and self-repetitions divided by total duration of the given speech sample.

Silent pause-time ratio is an equivalent of more commonly used phonation-time ratio, but silent pause-time ratio is easier to compare to filled pause-time ratio and corrections or repetitions-time ratio than phonation-time ratio, because all nominators are considered as types of disfluencies. A trimodal length distribution was used for silent pauses, but the number of other disfluencies were not sufficient for such grouping.

2.4 Statistical analysis

The relation between temporal features and proficiency assessments was studied using the R program (Baayen 2008). We analyzed our data with a cumulative link (also called as ordinal regression or proportional odds) mixed

(CLM) model implemented in the clmm2 package in R and designed specifically for the analysis of responses measured on an ordinal scale (see, e.g., Christensen 2015). The CLM model gives improved estimates of regression coefficients compared to continuous models that do not take into account the ceiling and floor effects nor the possible skewness of the ordinal variable. In our analysis we treated grades given by assessors as an ordered response (categories A1-C2), and acoustic measurements as explanatory variables. Assessor was treated as a random effect, and analysis was done separately for read and semi-spontaneous speech samples. We also studied the effect of temporal measurements on each assessor with a separate cumulative link (CL) model.

The assessments were further analyzed with a multi-faceted Rasch measurement (MFRM) using the Facets-program first developed by Linacre (1989). Facets is a development of simpler Rasch models (Rasch 1960) which are psychometric models used in analyzing data from tests, questionnaires and other types of assessment. Facets is widely used in language assessment to analyze ratings of speaking and writing performances (Bachman et al. 1995; McNamara & Knoch 2012) because it can simultaneously model, and take into account, such factors as learner ability, rater severity, scale, and difficulty of the tasks and dimensions of language. The analysis produces a logit scale (an interval scale) against which all the facets (the components conceptualized to combine to produce the data, e.g., persons, items, judges, tasks) included in the analysis are directly comparable. In the present study, Facets was used to investigate the quality of the ratings (particularly consistency of the assessors) and of the new CEFR scale for phonological control.

3 Results

3.1 Temporal features

Articulation rate, silent pauses, filled pauses, and corrections and repetitions were measured from each speech sample. For silent and filled pauses, and corrections and repetitions, disfluency-time ratio was calculated from each sample.

Some speech samples included no disfluencies, while most samples included more than one type of disfluencies. Figure 2 shows the relative amount of disfluencies in speech samples with respect to speech type. Silent pause of 200–1,000 ms was the most frequent for both speech types, and long silent pauses (> 1,000 ms) occurred almost exclusively in semi-spontaneous speech samples. Corrections and repetitions occurred more frequently in read than in semi-spontaneous speech samples, and filled pauses occurred more frequently in semi-spontaneous than in read speech samples.

Figure 3 illustrates the density of disfluencies with respect to disfluency

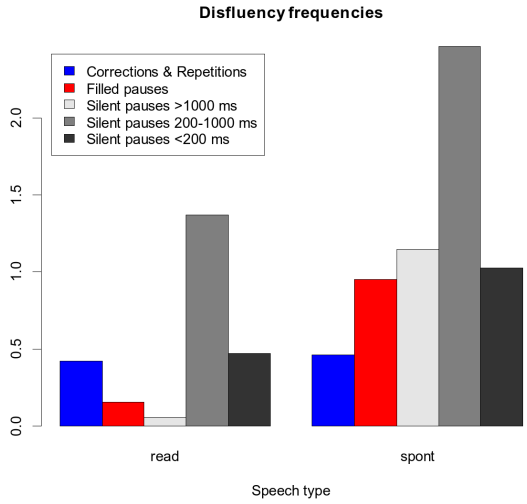

FIGURE 2. The relative amount of disfluencies in read and semi-spontaneous speech samples.

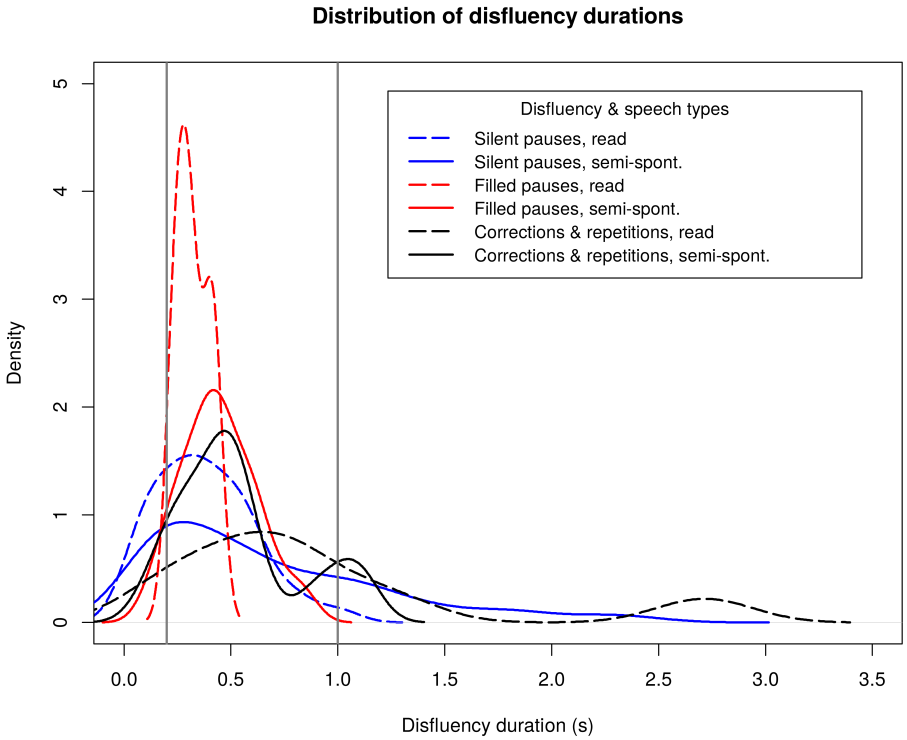

FIGURE 3. Distribution of disfluency durations. Dotted line: read speech, solid line: semi-spontaneous speech. Vertical lines are positioned at 200 ms and 1 s to illustrate the trimodal distribution of disfluency durations.

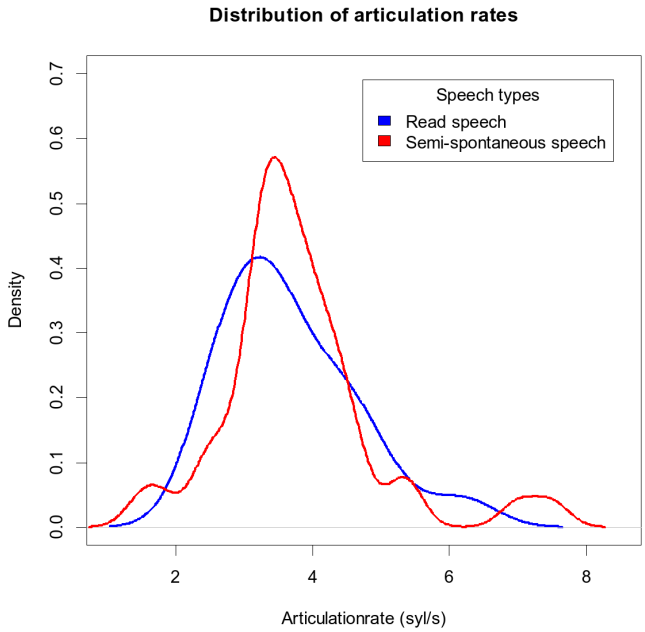

FIGURE 4. Distribution of articulation rates (syllables per second). Blue line: read speech, red line: semi-spontaneous speech.

duration, disfluency type, and speech type. The figure shows that most disfluencies are shorter than one second, but disfluencies even longer than three seconds were also detected. Filled pauses tend to be shorter than other disfluency types in both speech types, but read speech contains a considerable amount of filled pauses shorter than 500 milliseconds. Longer corrections and repetitions occurred more often in read speech than semi-spontaneous speech.

Articulation rate varied from 2.5 to 6.2 syllables per second in read speech and from 1.5 to 7.5 syllables per second in semi-spontaneous speech. The mean articulation rate was 3.7 syllables per second in read speech, and 3.8 syllables per second in semi-spontaneous speech. Figure 4 shows the distribution of articulation rates with respect to speech type.

3.2 Temporal feature effect on assessments

Each speech sample was assessed by seven expert raters using the revised CEFR scale for phonological control. In our analysis, we treated grades given by assessors as an ordered response (categories A1-C2). The data was first analyzed with a cumulative link mixed (CLM) model, where assessments were treated as dependent variables and acoustic measurements as explanatory

TABLE 2. The variables used in cumulative link mixed model analysis.

Type	Variable
Dependent	Assessment
Explanatory	Articulation rate (AR)
	Short (< 200 ms) silent pause-time ratio (SS)
	Medium (200–1,000 ms) silent pause-time ratio (SM)
	Long (> 1,000 ms) silent pause-time ratio (SL)
	Filled pause-time ratio (F)
	Corrections or repetitions-time ratio (CR)
Random	Assessor

variables (see Table 2 for list of variables). Assessor was treated as a random variable, and analysis was done separately for read and semi-spontaneous speech samples. We also examined the effect of articulation rate and disfluency types on each assessor (see tables 4 and 5 in section 3.3).

Figure 5 shows the assessment distribution for all samples. The most common proficiency grades for prosodic features were B1 (n = 136) and A2 (n = 109). Standard deviation of assessments between assessors varied from 0.38 to 2.9 proficiency scales, leaving the mean SD to 0.81. The assessment distribution is skewed clearly towards lower proficiency levels. Three samples were excluded from statistical analysis, since they were marked as non-analyzable: the assessors were unable to assess prosody from these speech samples.

Table 3 shows the results of our CLM-model. Estimate values and statistical significance of articulation rate and different disfluency types on assessments are introduced with regard to speech type. The model showed a positive effect for AR and negative effect for disfluencies: that is, faster articulation rate and small disfluency-time ratio indicate higher perceived proficiency level. AR and F proved to be extremely significant predictors of the assessed proficiency level in both read and spontaneous speech. Both SS and SL as well as CR were significant predictors in read speech, while in semi-spontaneous speech only SL was somewhat significant, but SS, SM and CR remained insignificant.

As the CLM-model itself does not provide a straightforward method to estimate the quality of fit, we approximate the multinomial logistic regression with a linear mixed effect regression models using the numerical assessment as a dependent variable (with the same independent variables and random effect structure as in the original logistic regression). The marginal r2-value provides an estimate of the quality of fits in terms of variance explained by fixed effects in the linear mixed-effect models (Nakagawa & Schielzeth 2013).

The linear model yielded very similar effect estimates; in fact, the signifi-

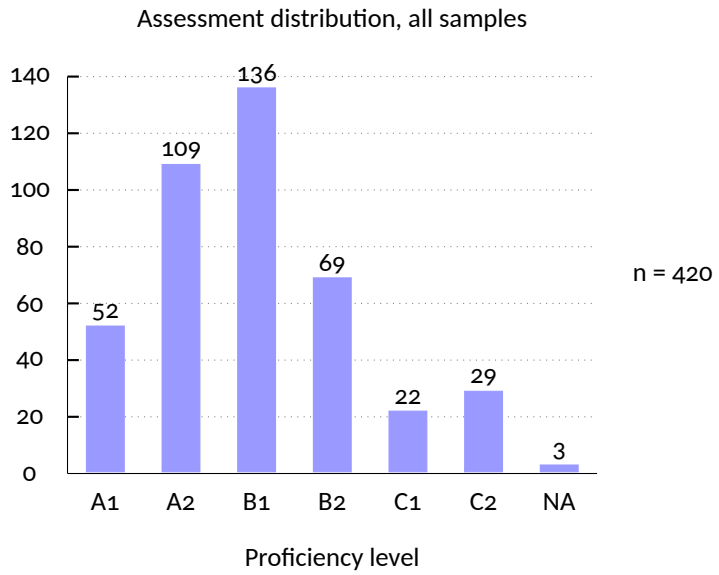

FIGURE 5. Assessment distributions computed from all assessed speech samples.

TABLE 3. The effect of articulation rate and disfluency-time ratios (estimate values and statistical significance) on prosodic features assessments. AR = articulation rate, SS = silent pauses < 200 ms, SM = silent pauses 200–1,000 ms, SL = silent pauses > 1,000 ms, F = filled pauses, CR = corrections and repetitions. p-values: 0.01–0.05*, 0.01–0.001**, < 0.001***.

Speech type	AR	SS	SM	SL	F	CR
Read	2.40***	−39.78**	−1.29	−34.96***	−41.26***	−4.48**
Spont	0.98***	−7.33	−0.65	−1.79*	−9.67***	−4.97

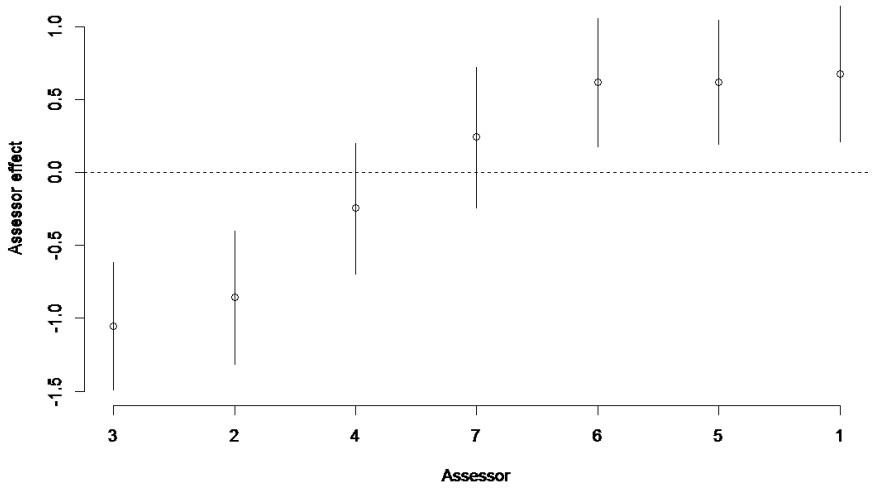

FIGURE 6. The assessor effect in prosody assessments.

cance of different effects was identical to the CLM-model as reported above. The marginal r2-value was 0.59 for read speech and 0.34 for semi-spontaneous speech, showing that the dependent variables explain a substantial amount of assessment variance; these measures provide a lower bound for the quality of fit of the more elaborate CLM-models.

3.3 Inter-rater variation

The assessor effect was analyzed from the CLM-model, based on the conditional distribution of the random effects (Assessor, see Table 3). Figure 6 shows that assessors A06, A05 and A01 give generally higher proficiency scores for prosody than other assessors, and differ statistically significantly from assessors A03, A02, and A04, but not from A07. Assessors A03 and A02, in turn, are stricter than the other assessors. The distance between the strictest and most permissive assessor is 1.7 logits, which corresponds to about 0.8 proficiency levels. The average proficiency raters for prosodic features are the native Finnish Swedish speaking assessors A04 and A07.

The effect of articulation rate and disfluency types on each assessor was examined separately with a cumulative link (CL) model using the R software. The data was divided into subsets by assessors, and analysis was done separately for read and semi-spontaneous speech samples. Assessments were treated as dependent variables and acoustic measurements as explanatory variables.

TABLE 4. The effect of articulation rate and disfluency types (estimate values and statistical significance) on prosodic features assessments of the read speech samples. p-values: 0.01–0.05*, 0.01–0.001**, < 0.001***. When the field is empty, the model found probabilities of 0 or 1.

Assessor	AR	SS	SM	SL	F	CR
A01	3.57**	−2.43	8.63	−1.58***	−6.68*	−2.22
A02	1.93*	1.79	2.29	−1.49***	−4.53*	2.37
A03	3.94**		2.89	−1.47***	−2.7	−2.36
A04		−6.29*	−8.54	−1.46***	−1.48	−3.49
A05	3.25**		3.02	−2.59	−61.29*	−16.04**
A06		−7.68*	−1.09	−1.46***	−1.11	−3.47
A07	2.99**	−9.7	−7.38	−1.53***	−2.79	−1.05*

TABLE 5. The effect of articulation rate and disfluency types (estimate values and statistical significance) on prosodic features assessments of the semi-spontaneous speech samples. p-values: 0.01–0.05*, 0.01–0.001**, < 0.001***. When the field is empty, the model found probabilities of 0 or 1.

Assessor	AR	SS	SM	SL	F	CR
A01	1.09**	−9.58	−4.7	−4.26	−9.35	1.9
A02	0.73*	−9.19	0.31	−1.7	−5.71	−3.38
A03	0.91*	−13.84	0.12	−1.26	−14.53	−11.12
A04	1.54***	−6.74	1.25	0.18	−11.54	−13.44
A05	1.28**		3.57		−21.63*	−16.59*
A06	1.58***	−6.97	2.16	−1.4	−21.66*	−6.18
A07	1.10**	−12.29	−1.57	−2.56	−5.42	−1.28

Tables 4 and 5 show the estimate values and statistical significance of the examined temporal features on assessments with regard to speech type and individual assessors. Table 4 shows the effects for read speech and Table 5 for semi-spontaneous speech. In read speech, the negative effect of SL was extremely significant for all assessors but A05; instead, CR had a strong negative effect on A05's assessments. The positive effect of AR was very significant or significant to all assessors, unless the probability was 0 or 1. In semi-spontaneous speech, AR was significant or very significant for all assessors, but CR was significant only for assessor A05.

The assessments were further analyzed with the Facets program in order to measure the consistency of the assessments. Table 6 shows the infit and outfit values that indicate how systematic the raters' assessments are (Linacre 1989). Outfit values are sensitive to outliers in the data whereas infit values focus on the core set of ratings. Ideal infit / outfit mean square values are

TABLE 6. The Infit and Outfit values of assessors measured from prosody assessments.

	Infit		Outfit	
Assessor	MnSq	ZStd	MnSq	ZStd
A01	0.81	-0.9	0.8	-0.9
A02	0.67	-1.7	0.6	-1.8
A03	0.85	-0.6	0.71	-1.1
A04	1.09	0.4	1.08	0.4
A05	1.46	2.0	1.78	3.2
A06	0.73	-1.4	0.7	-1.5
A07	1.21	1.0	1.15	0.7

close to 1.0. Very high values (general cut-off point 1.5) are problematic as they indicate that the rater's assessments are unsystematic and unpredictable. Very low values usually indicate that the rater is behaving more systematically than could be modeled but it may also result from the rater not using the full scale but overusing, for example, the middle part of the rating scale. Assessor A05's infit mean square value is close to 1.5, which indicates some inconsistency in this assessor's assessments.

4 Discussion

This study investigated whether objectively measured temporal features in speech can be used in predicting the proficiency level assessed by expert raters. The effect of articulation rate, pause-time ratios of silent and filled pauses as well as corrections and repetitions-time ratio was studied with a cumulative link mixed model. The assessments were further analyzed with Facets to study inter-rater variation and reliability. Below, we discuss the effect of temporal features on assessments and differences between the assessors.

4.1 Temporal features as proficiency indicators

The effect of articulation rate (AR), silent (S) and filled (F) pause-time ratio, and corrections and repetitions-time ratio (CR) was examined. Our findings suggest that different speech types have different requirements when it comes to prosodic features. AR and F proved to be extremely significant predictors in both read and semi-spontaneous speech. Higher AR implicates higher perceived proficiency level for both speech types. Higher F, in turn, implicates lower proficiency level for both speech types. Although SL occurred almost exclusively in semi-spontaneous speech samples, SS, SL and CR were all significant for read speech. Only SL was somewhat significant for semi-spontaneous

speech. Interestingly, SM remained non-significant for both speech types, although silent pauses of 200–1,000 ms were the most common in our data. The significance of short silences in read speech suggest that silences shorter than 200 ms should be taken into account when studying fluency and the commonly used threshold of 250 ms for silent pauses could cause misleading results. The insignificance of CR in semi-spontaneous speech, in turn, support the findings of previous studies (Cucchiarini et al. 2002; Bosker et al. 2013), and the general assumption that disfluencies such as self-corrections are more tolerated in conversational speech, which our semi-spontaneous speech samples reflect, than in read or formal speech. On the other hand, CR were on average longer in read speech than in semi-spontaneous speech, which could have caused the difference in effect. Our measures fail to distinguish between different types of self-corrections, remaining insensitive to the extent of these disfluencies, for example the length of the correction or repetition. This should be studied further, since several quick repetitions of single words or syllables may be perceived as less obstructive than long misrepresentations requiring notable backtracking. Disfluencies should be studied also by their length, but further research requires larger speech and assessment data. In this study the size of the speech data allowed scrutinizing only the length of silent pauses; the number of filled pauses and corrections and repetitions were insufficient for such grouping. Another interesting question is, which disfluencies are automatically detectable. New state-of-the-art prosody analysis methods (Suni et al. 2017) could be used to examine the possible automatization of fluency analysis. Automatic measurement of disfluencies could make the assessment procedure more efficient especially in large-scale testing, but many existing systems based on automatic speech recognition still struggle with the evaluation of spontaneous speech (see, e.g., Witt 2012).

4.2 Inter-rater variation

Many previous fluency studies have used the mean rating for each assessed item, thereby disregarding the individual differences between assessors. We studied each assessor individually, and the results revealed differences in the severity and consistency of the assessors as well as indications of the assessors' individual focuses. In our study the native assessors were ranked closer to the average assessor with respect to the severity of their assessments. The assessor A05, whose Infit Mean Square value indicated inconsistency in assessments, differed from other assessors also with respect to the significance of temporal variables: only the estimate values for articulation rate reflected the ones of other assessors. The results indicate that the assessors weigh the various disfluencies differently. This issue has not been discussed much in previous studies and should be scrutinized in more detail in order to gain better

knowledge on the assessors' foci on temporal properties of speech. Modeling the variation between assessors enables profiling the assessors and development of more relevant training procedures, which can improve the reliability of the assessments especially in large-scale tests.

5 Conclusions

We studied whether objectively measured temporal features in speech can be used in predicting the proficiency level assessed by expert raters. Our results suggest that certain temporal features do have an effect on perceived proficiency level, but speech type needs to be taken into account: corrections, repetitions and short silences are more tolerated in semi-spontaneous than in read speech. For both speech types, however, higher articulation rate implicates higher proficiency level, whereas higher filled pause-time ratio implicates lower proficiency level. Larger data is yet recommended to study different types of disfluencies more closely, but objective phonetic measurements can serve as an anchor for assessments. Precise measurements related to assessments can also be used to develop automatic methods for more effective and reliable assessment protocol.

Furthermore, we found that assessors seem to weigh temporal features differently depending on the speech type and their individual focus. This issue should be scrutinized further and methods for improving inter-rater agreement should be examined. Profiling the assessors can help to improve their assessing performance with useful feedback that will increase their phonetic awareness. Moreover, implementing human-machine hybrid scoring can significantly improve inter-rater agreement and thus the reliability and objectivity of the assessments (see, e.g., Luo et al. 2016).

When integrating oral tests to large-scale high-stakes exams, such as the Finnish Matriculation Examination, there is an abiding need for an objective and effective assessment of students' proficiency. Inconsistency between or within assessors have a negative effect on assessments and assessed L2 learners. Detecting indicators of what assessors are reacting to when subjectively assessing speech fluency helps make L2 assessment more reliable in three ways: the assessment criteria can be clarified with more specific features related to fluency, the assessors can be profiled according to their individual focus, and the assessors can be better trained to recognize these prosodic features they subconsciously use while assessing L2 proficiency.

References

Baayen, R. H. 2008. *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bachman, L., B. Lynch & M. Mason 1995. Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12* (2), 238–257.

Boersma, P. & D. Weenink 2010. PRAAT. Amsterdam: University of Amsterdam. http://www.fon.hum.uva.nl/praat/.

Bosker, H. R., A.-F. Pinget, H. Quené, T. Sanders & N. H. De Jong 2013. What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30* (2), 159–175. DOI: 10.1177/0265532212455394.

Campione, E. & J. Véronis 2002. A large-scale multilingual study of silent pause duration. In B. Bel & I. Marlien (eds) *Speech prosody 2002*. Aix-en-Provence: Laboratoire Parole et Langage, 199–202.

Chambers, F. 1997. What do we mean by fluency? *System*, *25*, 535–544. DOI: 10.1016/S0346-251X(97)00046-8.

Cheng, J. 2011. Automatic assessment of prosody in high-stakes English tests. In P. Cosi, R. De Mori, G. Di Fabbrizio & R. Pieraccini (eds) *INTERSPEECH 2011: 12th annual conference of the International Speech Communication Association, Florence, Italy, August 27–31, 2011*, 1589–1592. http://www.isca-speech.org/archive/interspeech_2011/.

Christensen, R. H. B. 2015. A tutorial on fitting cumulative link mixed models with clmm2 from the ordinal package. ftp://ftp.ussg.indiana.edu/pub/CRAN/web/packages/ordinal/vignettes/clmm2_tutorial.pdf.

Collier, R. 1992. A comment on the prediction of prosody. In G. Bailly & C. Benôit (eds) *Talking machines*. Amsterdam: North-Holland, 205–208.

Council of Europe 2001. *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

— 2017. *Common European framework of reference for languages: learning, teaching, assessment*. Council of Europe. (Companion Volume with New Descriptors).

Cucchiarini, C., H. Strik & L. Boves 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. DOI: 10.1121/1.428279.

De Jong, N. H. & H. R. Bosker 2013. Choosing a threshold for silent pauses to measure second language fluency. In Robert Eklund (ed.) *Proceedings of DiSS 2013: the 6th workshop on disfluency in spontaneous speech (DiSS), KTH Royal Institute of Technology, Stockholm, Sweden, 21–23 August 2013*. Tal, musik, hörsel – Quarterly Progress and Status Report 54(1). Stockholm: Kungliga Tekniska Högskolan, 17–20. http://liu.diva-portal.org/smash/record.jsf?pid=diva2%3A1079414%5C&dswid=647.

Derwing, T. M., M. J. Rossiter, M. J. Munro & R. I. Thomson 2004. Second language fluency: judgments on different tasks. *Language Learning*, *54* (4), 655–679.

Derwing, T. M., R. I. Thomson & M. J. Munro 2006. English pronunciation and fluency development in Mandarin and Slavic speakers. *System*, *34* (2), 183–193. DOI: 10.1016/j.system.2006.01.005.

Duez, D. 1982. Silent and non-silent pauses in three speech styles. *Language and Speech*, *25*, 11–28.

Finnish Matriculation Examination Board 2017. Kevään 2016 ja 2017 ylioppilastutkintoon ilmoittautuneiden määrät kokeittain. https://www.ylioppilastutkinto.fi/images/sivuston_tiedostot/stat/FB2017KT2001.pdf.

Foster, P. & P. Skehan 1996. The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, *18* (3), 299–323.

Goldman-Eisler, F. 1968. *Psycholinguistics: experiments in spontaneous speech*. London: Academic Press.

Hieke, A. E., S. Kowal & D. C. O'Connell 1983. The trouble with "articulatory" pauses. *Language and Speech*, *26* (3), 203–214.

Hönig, F., A. Batliner, K. Weilhammer & E. Nöth 2010. Automatic assessment of non-native prosody for English as L2. In *Speech prosody 2010 conference proceedings*. Vol. 100973. Chicago (Ill.): 1–4.

Huhta, A. 1993. Suullisen kielitaidon arviointi. In S. Takala (ed.) *Suullinen kielitaito ja sen arviointi*. Kasvatustieteiden tutkimuslaitoksen julkaisusarja B: Teoriaa ja käytäntöä 77. Jyväskylä: Kasvatustieteiden tutkimuslaitos, 143–225. https://journal.fi/afinla/article/view/60844/22613.

Koponen, M. & H. Riggenbach 2000. Overview: varying perspectives on fluency. In H. Riggenbach (ed.) *Perspectives on fluency*. Ann Arbor (Mich.): University of Michigan Press, 5–24.

Kormos, J. & M. Dénes 2004. Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32* (2), 145–164. DOI: 10.1016/j.system.2004.01.001.

Lennon, P. 1990. Investigating fluency in EFL: a quantitative approach. *Language Learning*, *40*, 387–417. DOI: 10.1111/j.1467-1770.1990.tb00669.x.

— 2000. The lexical element in spoken second language fluency. In H. Riggenbach (ed.) *Perspectives on fluency*. Ann Arbor (Mich.): University of Michigan Press, 25–42.

Linacre, J. 1989. *Many-facet Rasch measurement*. Chicago (Ill.): MESA Press.

Luo, D., W. Gu, R. Luo & L. Wang 2016. Investigation of the effects of automatic scoring technology on human raters' performances in L2 speech proficiency assessment. In T. Lee, L. Xei, J. Dang, H.-M. Wang, J. Wei, H. Weng, Q. Hou & Y. Wei (eds) *Chinese spoken language processing (ISCSLP): 2016 10th international symposium*. IEEE, 1–5.

McNamara, T. & U. Knoch 2012. The rasch wars: the emergence of Rasch measurement in language testing. *Language Testing*, *29* (4), 555–576.

Ministry of Education and Culture 2017. *Gaudeamus igitur: ylioppilastutkinnon kehittäminen*. Opetus- ja kulttuuriministeriön julkaisuja 2017:16. Helsinki: Opetus- ja kulttuuriministeriö.

Nakagawa, S. & H. Schielzeth 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4* (2), 133–142.

Pinget, A.-F., H. R. Bosker, H. Quené & N. H. De Jong 2014. Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing*, *31* (3), 349–365. DOI: 0.1177/0265532214526177.

Préfontaine, Y., J. Kormos & D. E. Johnson 2016. How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing*, *33* (1), 53–73. DOI: 10.1177/0265532215579530.

Rasch, G. 1960. *Probabilistic models for some intelligence and achievement tests*. Chicago (Ill.): University of Chicago Press.

Rossiter, M. J. 2009. Perceptions of l2 fluency by native and non-native speakers of English. *The Canadian Modern Language Review*, *65*, 395–412. DOI: 10.3138/cmlr.65.3.395.

Scherer, K. R. 1986. Vocal affect expression: a review and a model for future research. *Psychological Bulletin*, *99* (2), 143.

Scherer, K. R., H. London & J. J. Wolf 1973. The voice of confidence: paralinguistic cues and audience evaluation. *Journal of Research in Personality*, *7* (1), 31–44.

Suni, A., J. Šimko, D. Aalto & M. Vainio 2017. Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, *45*, 123–136. DOI: 10.1016/j.csl.2016.11.001.

Toivola, M., M. Lennes & E. Aho 2009. Speech rate and pauses in non-native Finnish. In M. Uther, R. Moore & S. Cox (eds) *Interspeech-2009*. Brighton: ISCA, 1707–1710.

Towell, R., R. Hawkins & N. Bazergui 1996. The development of fluency in advanced learners of French. *Applied Linguistics*, *17*, 84–119. DOI: 10.1093/applin/17.1.84.

Trofimovich, P. & W. Baker 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. DOI: 10.1017/S0272263106060013P.

Wennerstrom, A. 2000. The role of intonation in second language fluency. In H. Riggenbach (ed.) *Perspectives on fluency*. Ann Arbor (Mich.): University of Michigan Press, 102–127.

Witt, S. M. 2012. Automatic error detection in pronunciation training: where we are and where we need to go. In O. Engwall (ed.) *Proceedings of the international symposium on automatic detection of errors in pronunciation training: Stockholm, Sweden*. Stockholm: KTH Royal Institute of Technology, 1–8.

Xu, Y. 2013. ProsodyPro: a tool for large-scale systematic prosody analysis. In B. Bigi & D. Hirst (eds) *TRASP 2013: tools and resources for the analysis of speech prosody*. Aix-en-Provence: Laboratoire Parole et Langage, 7–10.