

Developing a digital tool for L2 speaking assessment in low-resourced languages

Raili Hilden¹ , Mikko Kuronen² , Ekaterina Voskoboinik³ ,
Yaroslav Getman³  & Mikko Kurimo³ 

¹ University of Helsinki, Finland

² University of Jyväskylä, Finland

³ Aalto University, Finland

Previous research on training and assessment of oral skills has mainly focused on English as L2, but since languages and learning contexts vary, it is imperative to study automatic speaking assessment (ASA) in other languages with local relevance as well. This paper summarizes a project which set out to develop a prototype tool to support training and assessment of oral skills in two low-resourced languages, Finnish and Swedish. This project addressed the applicability of automated assessment to measure multiple features of monologue speech, the accuracy of human ratings used for training a system based on automatic speech recognition (ASR) and the technical conditions of providing individualized feedback to improve student learning. Encouraging results for both Finnish and Swedish were gained when adapting a big pre-trained wav2vec2.0 speech model that was fine-tuned first with a larger L1 dataset, and then with an L2 dataset collected in the project. The results suggested that the most suitable features for automatic analysis were quantifiable fluency measures and vocabulary range. Machine and human estimates were most consistent for assessing fluency, range and accuracy, while the results were more controversial for pronunciation features other than fluency. The prototype will be further developed, fine-tuned and adjusted to address the needs of adult learners preparing for the final test of integration training in L2 Finnish.

Email address for correspondence: raili.hilden@helsinki.fi

Keywords: L2 learning, oral proficiency, automatic speaking assessment, signal processing

Introduction

Learning to speak foreign languages is a fundamentally important skill in modern society. The societal and technological trends have expanded our understanding of learning environments, and the world that L2 (second language) learners of today live in is multilingual, multimodal and mobile. In the digitized world, teaching happens increasingly over the internet and by utilizing computer programs and artificial intelligence to assess speaking skills and give individual feedback. The same tools are, in principle, applicable to learning and assessment alike as soon as suitable technical solutions are developed.

In Finland in 2010, a ministry-appointed working group proposed the introduction of an optional oral course in the most commonly taught syllabi of foreign languages at the upper-secondary level (National Board of Education, 2010). The course grade was based on an oral test delivered by the National Board of Education, and a separate certification was issued upon completion of the course and the test. However, the certificate document did not have an official status comparable to the actual final school reports or the Matriculation Examination Certificate. Despite some technical and pedagogical improvements, the situation remains the same today: speaking is taught and assessed in the classroom, but it is not included in the most impactful high-stakes test at the end of general upper secondary education. The reasons for this are primarily practical: problems of implementation and the lack of assessment resources of time and rater salaries. Finnish legislation requires a double human rating of each performance to be included in the Matriculation Exam and scoring even a few minutes of speech would significantly increase the costs due to the need for a larger number of raters to ensure that the test results are delivered on time.

The digitalization of the exam towards the end of the 2010s opened new prospects for using the recent advances in machine learning and speech technology to solve these problems. A consortium of three universities combining expertise in language education and assessment (University of Helsinki), phonetics and assessment (University of Jyväskylä) and speech technology and machine learning (Aalto

University) set out to test and investigate the possibilities of reducing human workload in speaking assessment. The research group's goal was to develop a prototype tool for giving feedback on a speech sample. This feedback can be used for both formative and summative decisions about a student's oral skills. In the first phase, the formative purpose is emphasized, but in the long run, as the effectiveness and reliability of the tool increases, it could be used for summative and even high-stakes decisions, such as those made in the matriculation exam at the end of upper secondary education.

The DigiTala project, funded by the Research Council of Finland for 2019–2023, was derived from a number of recognized practical and theoretical gaps in teaching and testing oral skills in second and foreign languages in Finland. The national high-stakes exam at the end of upper secondary education (the Matriculation Examination) in Finland has for a long time been at odds with the communicative language curricula it is intended to measure, since there is no speaking section in the second or foreign language tests in this examination. There have been several attempts over the decades to increase the attention paid to oral skills. Already in the early 1990s, when Finland was approaching the European Union to become a member state, the first steps were taken by the Finnish National Board of Education (currently The Finnish National Agency for Education) to focus on speaking by producing teaching and testing materials to be used at the end of upper secondary education (Hilden, 2000, pp. 373–412). These initiatives led to active research endeavors across the country, and at the beginning of the new millennium, transparent guidelines for assessing speaking with scales adapted from the CEFR (Council of Europe, 2001) were included in the official core curricula for general basic and upper secondary education (National Board of Education, 2003, 2004).

In addition to the practical needs that gave rise to the project, it also had more purely scientific objectives. These included (1) unpacking and concretizing the construct of speaking in the Common European Framework of Reference (Council of Europe, 2020), (2) the development of automatic speaking assessment (ASA) for under-resourced languages, such as the two national languages of Finland (Finnish and Swedish), and through this also (3) supporting the teaching and learning of speaking a foreign or second language in educational settings.

Theoretical background

For assessment purposes, the validity of automated assessment is scrutinized in relation to the context of use (Xi, 2021). The DigiTala project acknowledged Finnish language education as its primary context, more specifically, general upper secondary and adult education. The former is rather strictly guided by a formal core curriculum (Finnish National Agency of Education, 2015), while the curricula for adult learners may vary depending on the instruction provider. In both settings, functional communicative competence in all skills is pursued.

The studies in this project draw on the definition of communicative competence presented in the CEFR (2001). The model consists of two major sets of competences: general and communicative language competences. They all aim at active social participation in the local and global community. In broad terms, language-independent general competences function as enablers of adequate use of language. Communicative competences, again, incorporate linguistic, pragmatic and socio-linguistic components that are deployed according to task demands. In oral language use, such as spoken production and interaction, the unique feature is phonological competence, the other components (lexical, morphological, grammatical and semantic) being exploited by all communicative activities alike. Despite remarkable differences between spoken production and interaction, they share the basic physical and cognitive implementation of speech production. In this sense, both are seen as essential parts of communicative competence, although spoken interaction is frequently the ultimate aim of teaching and learning oral communication.

Aspects of adequate construct representation need to be considered in case the digital tools are used for making decisions that have an impact on students' future. Technology still limits the scope of the speaking construct (Dimova et al., 2020), since automatic assessment is strongest in addressing quantitative features of monologue speech, such as speech rate and pronunciation in read-aloud speech (Hsieh et al., 2020). Free speech and especially aspects of interactional dialogue are still weakly captured by automated assessment, although these are quite common goals of language teaching and targets of educational assessment and testing. Consequently, automatic assessment of these features is rare in high-stakes contexts. The

encouraging examples available at the time of embarking on the DigiTala project were the hybrid approach deployed by Educational Testing Service (Evanini & Zechner, 2020) and Pearson (n.d.), which focused on more widespread languages than Finnish and Swedish. However, in delivering feedback and making low-stakes teaching and learning decisions, digital companions are more frequently used for promoting learner autonomy, apparently with good results (Zou et al., 2023).

To safeguard the optimal reliability of ASA, the model needs to be trained by using human ratings, which should be consistent and show a reasonable inter-rater agreement. The strengths of human and machine scoring are complementary: while the computer can analyze test takers' speech tirelessly and systematically, the human rater can attend to aspects of speaking (e.g. sociolinguistic appropriacy and content) that are still difficult for the computer. Therefore, content is still assessed by humans in some contexts, while other domains of oral language proficiency, such as pronunciation, fluency and different aspects of vocabulary, are assessed by the ASA system (Yoon & Zechner, 2017).

Since the algorithm to support the automatic assessment is trained by human ratings, we need to consider the priorities of human raters regarding the multiple aspects of speaking and how these relate to overall level ratings. Previous research on English shows, for example, that fluency, especially speech rate, articulation rate and silent pauses, has a clear correlation with the assessment of a speaker's level of proficiency, while the relationship between prosodic features such as some tonal and temporal features, has a more complex relationship with the assessment (Iwashita 2010; Kang & Johnson 2018). However, since prosodic features are partly language-dependent, it is important to investigate their impact on assessment also in languages other than English.

Recent advancements in self-supervised learning (SSL) have enabled the development of speech models and considerably improved the recognition and classification performance in speech-related tasks, even with limited labeled target data. Large pre-trained speech models, also referred to as speech foundation models, leverage SSL to learn deep acoustic representations. Instead of relying on human-labeled data, they generate training targets from the data itself, learning from large amounts of untranscribed mono- or multilingual speech during pre-training. These learned

representations are then fine-tuned with labeled data to specific tasks and domains. For our research, we selected the wav2vec 2.0 SSL speech framework (Baevski et al., 2020).

When introducing and implementing pedagogic innovations, such as novel digital systems, it is imperative to consider the perspective of present and future users. Perceptions of automatic assessment tools and procedures have been explored to an increasing extent in recent years, primarily in English and other world languages. High-school learners were found to appreciate the convenience, motivational impact and efficiency of an AI-based speech evaluation system for autonomous oral practice, while certain issues of voice recognition and deficiencies of feedback were identified as limitations (Zou et al., 2024). Furthermore, the lack of live interaction was perceived as a drawback (Zou et al., 2023). The feedback function is highly favored by most studies, more by students than by teachers (Gu et al., 2020; Liu et al., 2025). The ASR technology is also conceived as beneficial for the enhancing of L2 pronunciation and overall speaking skills (Sun, 2023), while the crucial factor to determine its success in effectivizing the learning process might be engagement (Huang, 2025).

As Xi (2021) maintains, stakeholders' perceptions of and interactions with automated scoring systems is dependent on multiple factors, such as age, culture and technical preparedness. Therefore, investigating stakeholders in the Finnish context is timely and necessary for the development of a tool for the Finnish national languages Finnish and Swedish.

In light of these developments, the following research questions guided this project.

RQ1. (Objective 1) What features of speech in Finnish and Swedish are important for human raters and connected with specific CEFR level ratings?

RQ2. (Objective 2) How can ASR help the process of measuring and standardization of relevant speech features?

RQ3. (Objective 3) What attitudes and beliefs do students and teachers hold towards the emerging pedagogic innovation?

Methodology

Project description

In the planning and implementation of the project, we followed the multistage path depicted in Figure 1.

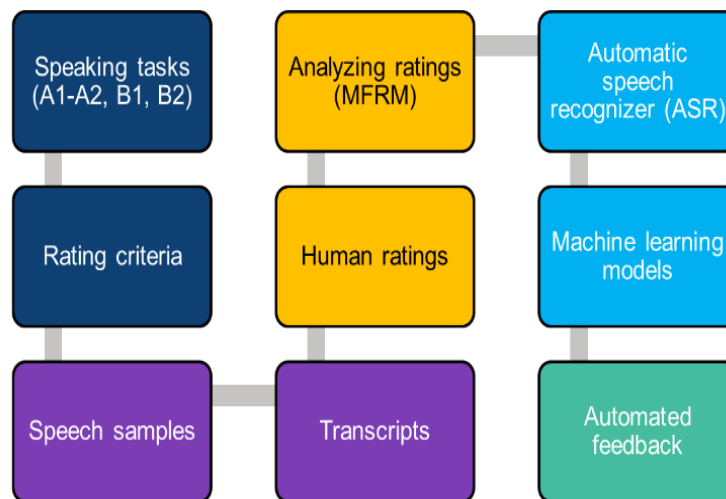


Figure 1. The multistage path of the DigiTala project in a flowchart.

Speaking tasks were designed for Common European Framework levels A1–B2. In the Finnish basic and general upper secondary education sublevels, such as A1.1, are used to enable the same scale to apply for both learning and assessment purposes. The alignment with CEFR and this scale in its original form was reported by Hilden and Takala (2007). Task content and types were informed by the core curriculum for general upper secondary education, focusing mostly on everyday situations and topics of interest and relevance for approximately 16 to 20-year-old students. We considered the findings from previous research by designing both read-aloud tasks, tasks tapping short utterances and more extended speech (Kallio et al., 2020; Kallio et al., 2023).

Participants

Speech samples in Finnish and Swedish were gathered from upper secondary schools and universities. Some of the university participants were foreign students at one of the consortium partner universities, while others were test-takers in the general language tests system provided by another partner university (see the next section for more detailed information). The students recorded their samples on a Moodle

software environment. Ultimately, the recorded samples were turned into text by the speech recognizer.

Tasks and speech samples

The dataset of speech samples comprises both read-aloud and spontaneous speech samples. The read-aloud task presented single words and short sentences to be produced by the students (e.g. *Fi. korkea kerrostalo* = a high block of flats; *Fi. Tuuli tuli kotiin* = A girl called Tuuli came home.) Guided short samples of speech were based on prompts, such as “Ask for a cup of coffee” or simulated written or spoken turn words that the student should respond to (e.g. “Where do you live?”). Extended samples were elicited by prompts such as “Tell about your day” or a picture to be described in the target language.

In total, 4,405 recordings were collected and rated for Finnish, amounting to 18.6 hours, while 4,266 recordings totaling 12.8 hours were gathered for Finland Swedish². For Finnish, 1883 recordings originated from 113 university students, and 2522 recordings came from 202 upper secondary school students. For Swedish, 120 university students provided 2241 samples, and 181 upper secondary school students contributed 2025 recordings. Verbatim transcriptions preserving hesitations and mispronunciations were created for all the recordings of participants completing freeform tasks.

Table 1. Speech samples dataset statistics.

	Finnish		Swedish	
	school	university	school	university
# of recordings	2,522	1,883	2,025	2,241
duration, h	12.7	5.9	7.1	5.7
# students	202	113	181	120
# raters	26	24	16	14
# of read aloud tasks	6	7	22	11
# of freeform tasks	20	10	0	8

² Finland Swedish is the variety of Swedish spoken in Finland that differs in pronunciation from the varieties spoken in Sweden both in terms of individual sounds and prosody (see e.g. Helgason et al., 2013). In our project, Finland Swedish was the primary variety for the development of the automatic application, as it is the variety that is mainly taught and learned in Finland. However, in this article we use the term “Swedish” without specifying that we are referring to Finland Swedish. We do this so that the reader does not get the incorrect impression that Finland Swedish is a separate language.

School participants speaking Finnish were under 22 years old, mostly female, with Finnish, Swedish, or Russian as their primary L1s. University participants speaking Finnish primarily ranged from 22 to 26 years old, were mostly male, and predominantly spoke English, Russian, German, or Vietnamese as their L1s. University participants speaking Swedish were mainly above 26 years old, predominantly female, with Vietnamese, Russian, or English as their primary languages. For more details about participants' characteristics and dataset statistics, see Kurimo et al. (2023).

Raters

The 37 raters were experienced assessors of Finnish and Swedish language with a solid background in master or doctoral level of academic studies in the respective language (von Zansen et al., 2022). Furthermore, they had been working on a regular basis as raters for General language exams (YKI) or the Finnish Matriculation Examination Board. The Swedish recordings were assessed by 18 raters in 2020 and the Finnish recordings by 26 raters in 2021–2022 using Moodle Quiz module (for details concerning human ratings see von Zansen et al., 2022).

Rating scales

We used two types of rating scales in assessing the training and test data to develop the ASA system, one for holistic oral proficiency level (from A1 to C1) applied from the CEFR (2001) and its Companion volume (2020), and five analytic dimensions often used in assessing L2 speaking skills: fluency, pronunciation, vocabulary range, accuracy (in particular the impact of grammatical errors on intelligibility) and task completion. The analytic dimensions were rated on a scale of 0–3 with specifications for the different steps. These scales and dimensions were also incorporated into the final ASA system (described below). The rating criteria were piloted by four raters from the project with experience in assessing Finnish and Swedish as L2. Benchmarks were chosen from this pilot and used in the proper rating round. Raters received a two-hour training on how to assess the speech samples, during which the scales were presented, speech samples were listened to and assessed, and the assessments were discussed. The training was conducted online, recorded and made available for the raters to view after the training. Each rater rated around 200 samples and used approximately eight

hours for the task. Rater agreement was calculated with the Facets program (Linacre, 2024) as well as with Spearman's Correlation Coefficient and Quadratically Weighted Kappa using the scikit-learn library (Pedregosa et al., 2011).

Automated speech recognition (ASR) development

Traditionally, ASR is an essential component in developing ASA systems, as it allows the extraction of both fluency and content-related features. However, building effective ASR systems for L2 speakers remains challenging due to the irregularity and variability of learner speech, influenced by learners' diverse linguistic backgrounds, and the limited availability of training data to effectively model these variations (Kurimo et al., 2023). In the project, machine learning models were trained to generate scores that match human ratings as closely as possible. Hand-crafted, linguistically informed features were first extracted from learner responses, after which the models learned to map them onto assessment scores. Additionally, deep acoustic representations derived from pre-trained models were utilized. In the last phase, short verbal feedback was attached to the level ratings. In the final prototype, teachers are allowed to design their own tasks on the Moodle platform to focus their students' oral training on desired topics and levels (von Zansen & Kallio, 2024).

Results

In the following sections, we describe the main results of the project based on the path depicted in Figure 1. However, automatic feedback is not presented, as we have not yet had the possibility to develop that feature of our ASA system.

Phonetic studies (RQ1)

We investigated fluency and prosody in a number of phonetic studies. The methods used were acoustic analyses and perceptual (listener) tests, and in several cases, comparison of the findings in these two types of studies. The aim was to gain knowledge about which features of fluency and prosody (cf. Table 2) are important for human raters when assessing speakers' level of oral proficiency and fluency and should therefore be incorporated into the project's ASA system.

Table 2. The fluency and prosody features examined in the project's phonetic studies.

FLUENCY	PROSODY
articulation rate (syllables per sec. excluding pauses)	nPVI, normalized pairwise variability index (a rate-normalized mean difference between consecutive syllables)
speech rate (syllables per sec. including pauses)	n Δ S, a rate-normalized mean standard deviation of syllable duration
silent pauses (frequency and ratio)	f ₀ range (pitch range in semitones)
filled pauses (frequency and ratio)	f ₀ std (standard deviation of pitch in semitones)
pause location (between and within clauses and phrases and after interrupted words)	f ₀ slope (mean pitch slope in semitones)
repetitions and corrections	occurrence of creaky voice (vocal fry)
wrong language (frequency and ratio)	

Regarding fluency, we found that speakers at lower fluency and proficiency levels were more disfluent than speakers at higher levels due to slower articulation and speech rate, longer pause time and more silent pauses (Kallio & Kuronen, 2023; Kautonen & Kuronen, 2021). This applied to L2 speakers in both Finnish and Swedish. A study on Swedish revealed that articulation rate, speech rate and silent pause ratio (i.e., silent pauses in relation to speech time) turned out to be important features in predicting fluency in that they showed significant differences both between L2 and L1 speakers and between L2 speakers at different fluency levels (Figure 2). Articulation rate and silent pause ratio, however, did not differ significantly between L2 speakers at different proficiency levels, while a significant difference in speech rate among L2 speakers was found only between proficiency levels B and below A (Figure 3). Further, pauses within phrases and after interrupted words led to lower fluency ratings in human assessments in both Finnish and Swedish (Kallio et al., 2022; Kautonen et al., *subm.*), and pause location also affected the human assessments of proficiency level (Kallio & Kuronen, 2023). These findings indicate that integrating pause location into an ASA system would further improve the system's ability to assess proficiency and fluency in L2 speakers. However, this was not yet done in the ASA system developed in the project because automatic detection of pause location is difficult. Recent research shows promising developments in automatic detection of this aspect of fluency (Matsuura et al., 2025), and it is likely that the parameter can be included in ASA systems like ours in the near future.

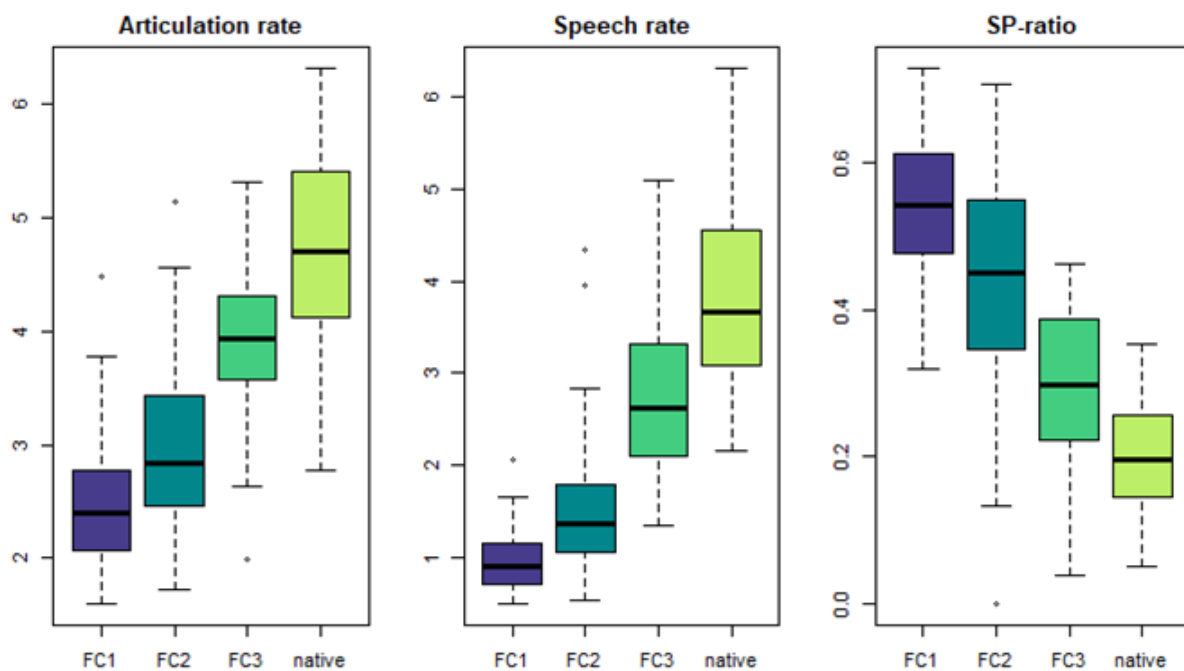


Figure 2. Articulation rate (syllables per sec. pauses excluded), speech rate (syllables per sec. pauses included), and silent pause ratio in L2 Swedish by fluency level (FC1, FC2, FC3) compared with L1 speakers (Kallio et al., 2023, p. 75).

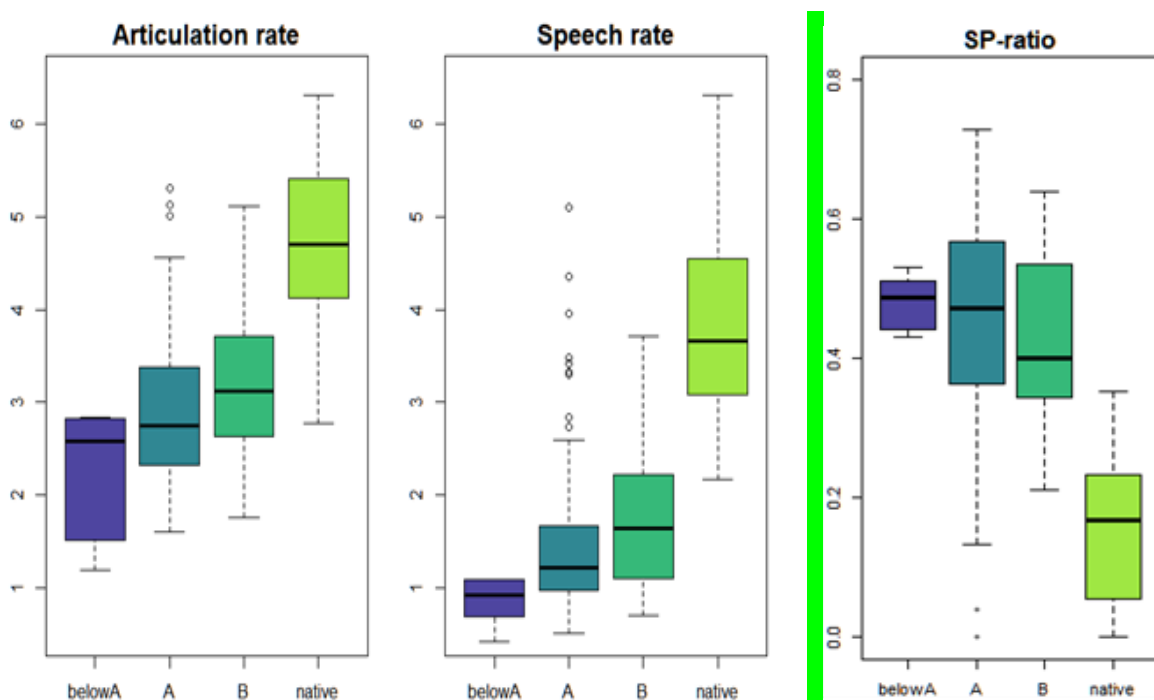


Figure 3. Articulation rate (syllables per sec. pauses excluded), speech rate (syllables per sec. pauses included), and silent pause ratio in L2 Swedish by proficiency level (below A-level, A-level, B-level) compared with L1 speakers (Kallio et al., 2023, p. 73).

Regarding prosody, we investigated whether speakers at different proficiency and fluency levels differ in speech rhythm and intonation. This was done by studying

parameters of timing (pairwise variability index nPVI and rate normalized mean standard deviation of syllable duration) and pitch (fo range, standard deviation and mean slope) (Kallio et al., 2023). We found that many of the investigated pitch parameters differ significantly between L1 and L2 speakers, while, somewhat surprisingly, this was not the case regarding timing. The mean standard deviation of syllable duration and pitch slope also showed potential in predicting fluency ratings in L2 speakers.

Many of the phonetic results we found in our studies are in line with previous findings on English. This applies to the relevance of articulation rate, speech rate, and SP ratio in the human and automatic assessment of both fluency and proficiency (cf. Cucchiarini et al. 2002; Kang & Johnson, 2018). However, some results, for example on pause location, timing and fo parameters, are probably at least to some extent language-dependent. Therefore, the results of the project are important in the development of reliable ASA systems for L2 speech in Swedish and Finnish.

Automatic speech recognizer and machine learning models (RQ2)

Data scarcity has been a significant bottleneck in the DigiTala project. However, in Al-Ghezi et al. (2021), we explored whether wav2vec 2.0 models can be applied to ASR for L2 Swedish speech. We fine-tuned publicly available mono- and multilingual foundation models on our target L2 data and compared their performance with conventional Deep Neural Network Hidden Markov Models (DNN-HMMs). Our best model achieved a 7.3% relative improvement in word error rate (WER) and a 35.3% relative improvement in character error rate (CER), reducing WER from 17.7% to 16.4% and CER from 10.6% to 6.9%. While the self-supervised learning (SSL) models produced recognition results considerably closer to the reference than the DNN-HMMs, further analysis revealed that the best model also improved generalizability by correctly recognizing words that were not part of the training data.

In Al-Ghezi et al. (2023a), we also investigated whether and how the deep acoustic representations of wav2vec 2.0 can be applied for holistic ASA of L2 Finnish and Swedish. While these representations encode acoustic information about speech, they are not directly interpretable. Our aim was to compare the representations produced by wav2vec 2.0 with manually crafted, human-interpretable features, and to

investigate whether they could complement each other. Specifically, the hand-crafted features included pronunciation (ratio of voiced frames, acoustic model score, etc.), fluency (average syllable duration, speech rate, etc.), and lexical (lexical diversity, root type-token ratio, number of unique words in the ASR transcript, etc.) features. We trained six-layer DNN classifiers using hand-crafted features, wav2vec 2.0 representations, and a combination of both.

In L2 Swedish experiments, wav2vec 2.0 representations outperformed hand-crafted features, while their combination further improved the performance. Additionally, we found that the choice of wav2vec 2.0 layer for representation extraction affected the results: intermediate representations extracted from the middle layer (layer 12 out of 24) yielded the best performance, indicating that the last layer of wav2vec 2.0 may not always be the optimal one for ASA. Finally, we applied an end-to-end training strategy, optimizing a single system to directly perform the speech classification task using raw input audio utterances. We trained an end-to-end ASA system by further fine-tuning the wav2vec 2.0 ASR model for direct speech classification, which achieved the best results for L2 Swedish in this study. Similarly, for L2 Finnish, wav2vec 2.0 representations outperformed hand-crafted features. However, neither combining the two feature types nor training a wav2vec 2.0 ASA model led to a clear additional improvement. This may be due to the use of a multilingual foundation model pre-trained on parliamentary speeches in three Uralic languages (Wang et al., 2021), in contrast to the monolingual Swedish model pre-trained on speech from various domains.

Al-Ghezi et al. (2023b) summarizes the components to be deployed in the final ASA systems providing automated feedback for L2 Finnish and Swedish, comprising ASR and five DNN-based classifiers each. The holistic scoring systems were six-layer DNNs trained on the combination of hand-crafted and wav2vec 2.0 features, as in Al-Ghezi et al. (2023a), while the analytic dimensions, including pronunciation, fluency, and lexico-grammatical features, were assessed by separate classifiers, each following the same six-layer DNN model architecture but trained on distinct corresponding manually crafted feature sets. Both the holistic and analytic scales – except for the lexico-grammatical one (potentially due to inaccuracies in automatic ASR transcripts) – demonstrated higher machine-human agreements than corresponding inter-human

agreements in both L2 Finnish and L2 Swedish. For Finnish, machine and human estimates were most consistent in assessing the holistic (Spearman's correlation 0.8) and pronunciation (Spearman's correlation 0.6) scores, while for Swedish, fluency showed the highest consistency among these scales, with the Spearman's correlation of 0.5.

For languages that are less commonly learned than English, such as Finnish and Swedish, not only data scarcity but also data imbalance poses significant challenges. In the collected datasets, different proficiency levels are unevenly represented. Specifically, intermediate proficiency levels are overrepresented, while beginners are underrepresented for both languages, and advanced learners have a limited number of samples in all data subsets other than the Finnish L2 school data (Kurimo et al., 2023). This imbalance often causes machine learning algorithms to overfit more represented classes. To mitigate these issues, we explored several approaches, including data augmentation, oversampling combined with augmentation techniques and curriculum learning (Lun et al., 2024). These methods have proven effective in addressing data imbalance, enhancing the models' performance without requiring additional data collection. Additionally, in another study focused on scoring task completion (Voskoboinik et al., 2023), we employed self-supervised pre-trained text-based models like the audio models previously discussed. In this research, we investigated whether training these models using similarity learning, rather than a traditional classification head, could better handle class imbalance, and we found that similarity learning indeed improved performance for underrepresented classes. Furthermore, we discovered that using raw, unrounded mean scores from multiple raters provided superior results compared to rounded scores. Specifically, for each speech recording, we first averaged the holistic scores provided by the human raters without rounding, and these continuous values were then used as training targets.

Another promising direction for addressing data scarcity and imbalance involves leveraging large language models (LLMs), which have demonstrated the ability to adapt to new tasks with minimal or no task-specific training data. In our initial exploration of this approach (Phan et al., 2024), we utilized GPT-4 (OpenAI, 2023) for scoring task completion in a few-shot learning setup for a picture description task. By incorporating carefully selected examples and explicitly defining the scoring criteria

within prompts, we observed substantial improvements in task completion and the assessment of task performance. Additionally, this approach enabled a novel generation of explanations and detailed feedback, enhancing the educational value of the automated assessments. We further showed that Llama 3 LLMs (Dubey et al., 2024) are effective for holistic proficiency scoring of Finnish L2 speech transcripts, especially when using in-context learning and soft labels, which aggregate the model's confidence across all proficiency levels instead of selecting only the most likely prediction (Voskoboinik et al., 2025).

Perceptions of students and raters (RQ3)

Perceptions of human raters

Perceptions of human raters towards the digital assessment process they were involved in, towards digital assessment at large, as well as task features and assessment criteria, were investigated by von Zansen et al. (2022). The raters (n=37) primarily considered the guidelines as functional and easy to follow, and the use of the Moodle platform did not pose any substantial problems.

The majority of raters (n=34) took a positive or a neutral stance on automatic assessment in general. They emphasized the major assets regarding automated assessment, such as high reliability, economy and freedom from human physiological limitations³.

The machine does not tire or make random mistakes and treats all performances in the same way. A well-suited solution for assessing large numbers of students up to level B1.

In support of human assessment, machine assessment is a very good thing, as it can increase the reliability of assessment in terms of, for example, pronunciation, the scope of vocabulary, fluency, etc.

This is important and I am strongly in favor of oral assessment, and I understand very well that, due to the lack of resources, it is only possible to a large extent by automated means.

³ The machine translation software "Kontra" has been used in translation of the citations of responses to open-ended survey questions.

On the other hand, raters also paid attention to potential drawbacks of automatic assessment. They expressed in various percentages their awareness of the lack of human sensitivity and the risk of construct underrepresentation if brief monologue speech is taken as an indication of the entire concept of speaking.

This is certainly an inevitable development in the future, but I can't help thinking that I'm making myself and my language teacher colleagues unemployed (a little more so :-)). What's more, I don't think the machine will ever be able to fully evaluate spontaneous speech?

The task selection presented to L2 Swedish speakers on the upper-secondary level consisted of a read-aloud task, short statements (reactions to prompts like how to congratulate or ask for help) and an extended monologue (talking about a favorite place). The criteria encompassed an overall task completion, fluency, pronunciation, range and accuracy. In addition, a level assignment was given on the CEFR scale. Fluency was regarded as the easiest criterion to assess, while the raters were struggling with linguistic range. The most likely reason for this was that some of the shortest statements (for example, "How are you today?") gave insufficient evidence and left the rater puzzled about how far the speaker's skill would suffice beyond the particular utterance.

For some students or performances, the criteria were impossible to use. They did not fit the situation at all. For example, pronunciation, accuracy, if the pupil said only tack (Thank you).

The challenge in this dataset was the assessment of very short sentences (read aloud), the concrete nature of some assignments (> how to get to the top level), very concise samples (a few seconds when the instructions asked for 30 sec. I will also consider how to assess task completion

In the L2 Finnish data from adult speakers, the speech samples were longer, but linguistic range and determining the overall level were considered the most complicated dimensions (von Zansen et al., 2022).

Perceptions of students

Adult L2 Finnish learners' perceptions of a computer-based speaking test developed for automated assessment purposes were investigated by von Zansen and Hilden (2022). Finnish learners (n=115) participated in the data collection of the project as part of an elementary Finnish course using Moodle and Zoom. After completing

speaking tasks, the students answered a questionnaire that included opinion-scaled and open-ended questions. The questionnaire responses were analyzed using descriptive statistics and content analysis.

The general impression of the computer-based test was mostly positive (80% of the responses). Some students expressed excitement towards speaking to a machine, and others were not used to recording their speech.

It was a new way taking a test and it felt good.

- - Having the chance to re-record them allows you to work on your pronunciation

I think it is very difficult to speak to the computer rather than to a real person. It felt unnatural and more difficult

I found it stressful and I was not used to talking like that by myself.

The Moodle environment was familiar to participants, but some comments were made about technical challenges related to the local network or recording device.

Problem with microphone - Had to switch to another browser

The instructions and assignment materials were perceived as clear and informative, and the students felt confident that they could complete what was expected of them. They also liked the topics and the variety of tasks.

They cover a good range of situations where you could find yourself in

The tasks themselves were ok, but not very conversational. - - I would have preferred roleplay - - or back and forth conversation about getting to know someone - -

The tasks made sense and were at an appropriate level

Overall, the designed speaking tasks were found to work well for elementary (A1–A2) learners, with the exception of the integrative task involving colloquial language (von Zansen & Hilden, 2022) The student views are paramount in designing tasks that they perceive as meaningful and inspiring enough to motivate them for self-regulated training and taking automated tests in the future.

Conclusions

The goals of the DigiTala project reported above addressed the development of automatic speaking assessment for under-resourced languages on one hand and the concretization of the construct of spoken production emerging in the level system of the Common European Framework of Reference on the other. The third aim was to enhance the teaching and learning of speaking as a long-term pursuit.

Regarding the first goal, the automatic speech recognizer transformed the monologue speech samples into text, and the audio samples were evaluated by experts according to the CEFR levels. In this way, the automated scoring system was taught by the algorithm to score new performances. The project made considerable progress towards this goal, although the duration of the project and the funding granted did not allow us to go much further than producing a prototype of the intended device. The prototype provides feedback to the speaker and allows independent training for learners with different language backgrounds.

Pertaining to the second goal, reading tasks and short description tasks were best suited for automatic scoring. This project focused only on monologic audio data and verbal features of students' speech, but it inspired the researchers to plan and launch a new project covering the full range of spoken interaction. The first stages of a subsequent project on automatic assessment of spoken interaction in a second language) are described in another article in this publication (see Ullakonoja et al., this issue).

The third long-term aim of the endeavor to contribute to improved teaching and learning of L2 speaking was approached via user experience that turned out to be predominantly positive. All these results were well in line with the initial hypotheses based on previous literature on the few efforts reported by the embarkment of this project. Furthermore, the project applied this knowledge to two under-resourced languages, far beyond the global languages present in previous automatic scoring developments.

The capacity of this project to improve automatic speech recognition of L2 Finnish and Swedish and to develop a digital tool for the assessment of L2 speech was considered

promising enough by the Research Council of Finland to justify a proof-of-concept funding for the years 2025–2026. This project targets Finnish as L2 for immigrants and has as a main goal to develop a mobile application for assessing L2 speaking in Finnish and for providing feedback to the learner. The increased speech corpus and new knowledge provided by the completed and ongoing projects should allow us to improve and fine-tune the ASA system for monologic speech to yield more stable and reliable results.

Acknowledgements

We would like to thank the anonymous reviewers and the editors of the journal and the special issue for their comments, which helped to improve our article. We would also like to thank Ilona Lähteenmäki for reviewing the language in the article.

Author disclosures

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


The project DigiTala, Digital support for training and assessing second language speaking, has been funded by the Research Council of Finland (grant numbers 322619, 322625 and 322965).

The authors had the following roles in the research and writing:


Raili Hilden: conceptualization, funding acquisition, project administration, supervision, writing; Mikko Kuronen: conceptualization, formal analysis, funding acquisition, investigation, methodology, writing; Yaroslav Getman: data curation, formal analysis, investigation, software, writing; Ekaterina Voskoboinik: data curation, formal analysis, investigation, software, writing; Mikko Kurimo: conceptualization, funding acquisition, project administration, supervision, writing, methodology, data curation

ORCID iDs

Raili Hilden  <https://orcid.org/0000-0002-5114-5600>

Mikko Kuronen  <https://orcid.org/0000-0001-5971-7063>

Yaroslav Getman  <https://orcid.org/0000-0003-4680-8294>

Ekaterina Voskoboinik  <https://orcid.org/0009-0007-2691-5793>

Mikko Kurimo  <https://orcid.org/0000-0001-5278-7974>

References

- Al-Ghezi, R., Getman, Y., Rouhe, A., Hilden, R., & Kurimo, M. (2021). Self-Supervised End-to-End ASR for Low Resource L2 Swedish. In the *Proceedings of Interspeech 2021*, 1429–1433. <https://doi.org/10.21437/Interspeech.2021-1710>
- Al-Ghezi, R., Getman, Y., Voskoboinik, E., Singh, M., & Kurimo, M. (2023a). Automatic Rating of Spontaneous Speech for Low-Resource Languages. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, 339–345. <https://doi.org/10.1109/SLT54892.2023.10022381>
- Al-Ghezi, R., Voskoboinik, K., Getman, Y., Von Zansen, A., Kallio, H., Kurimo, M., Huhta, A., & Hilden, R. (2023b). Automatic Speaking Assessment of Spontaneous L2 Finnish and Swedish. *Language Assessment Quarterly*, 20(4–5), 421–444. <https://doi.org/10.1080/15434303.2023.2292265>
- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 12449–12460. <https://doi.org/10.48550/arXiv.2006.11477>
- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press. <https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions>
- Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. <http://www.coe.int/lang-cefr>
- Cucchiaroni, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous

- speech. *The Journal of the Acoustical Society of America*, 111(6), 2862–2873.
<https://doi.org/10.1121/1.1471894>
- Dimova, S., Yan, X., & Ginther, A. (2020). *Local language testing: Design, implementation, and development*. Routledge.
<https://doi.org/10.4324/9780429492242>
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. (2024). The llama 3 herd of models, arXiv preprint arXiv: 2407.21783 DOI: 10.48550/arXiv.2407.21783
- Evanini, K., & Zechner, K. (2020). Overview of automated speech scoring. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 3–20). Routledge.
<https://doi.org/10.4324/9781315165103-1>
- Finnish National Board of Education. (2003). Lukion opetussuunnitelman perusteet 2003 [National Core Curriculum for General Upper Secondary Schools 2003].
https://www.oph.fi/sites/default/files/documents/47345_lukion_opetussuunnitelman_perusteet_2003.pdf
- Finnish National Board of Education. (2004). Perusopetuksen opetussuunnitelman perusteet 2004 [National Core Curricula for Compulsory Basic Education 2004]. https://www.oph.fi/sites/default/files/documents/perusopetuksen-opetussuunnitelman-perusteet_2004.pdf
- Finnish National Agency for Education. (2015). Lukion Opetussuunnitelman Perusteet 2015 [National Core Curriculum for General Upper Secondary Schools 2015].
https://www.oph.fi/sites/default/files/documents/172124_lukion_opetussuunnitelman_perusteet_2015.pdf
- Gu, L., Davis, L., Tao, J., & Zechner, K. (2020). Using spoken language technology for generating feedback to prepare for the TOEFL iBT® test: A user perception study. *Assessment in Education: Principles, Policy & Practice*, 28(1), 58–76. <https://doi-org/10.1080/0969594X.2020.1735995>
- 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>

- Hilden, R. (2000). *Att tala bra, bättre och bäst. Suomenkielisten abiturienttien ruotsin kielen suullinen taito testisuoritusten valossa*. [Doctoral dissertation, University of Helsinki]. Helsingin yliopiston opettajankoulutuslaitos, Tutkimuksia 217.
- Hilden, R., & Takala, S. (2007). Relating descriptors of the Finnish school scale to the CEF overall scales for communicative activities. In A. Koskensalo, J. Smeds, P. Kaikkonen & V. Kohonen (Eds.), *Foreign languages and multicultural perspectives in the European context* (pp. 291–300). Dichtung – Wahrheit – Sprache, 9–10. Lit Verlag.
- Hsieh, C.-N., Zechner, K., & Xi, X. (2020). Features measuring fluency and pronunciation. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 101–122). Routledge. <https://doi.org/10.4324/9781315165103-7>
- Huang, M. (2025). Student engagement and speaking performance in AI-assisted learning environments: A mixed-methods study from Chinese middle schools. *Education and Information Technologies*, 30(6), 7143–7165. <https://doi.org/10.1007/s10639-024-12989-1>
- Iwashita, N. (2010). Features of oral proficiency in task performance by EFL and JFL learners. In M. T. Prior, Y. Watanabe, & S-K Lee (Eds.), *Selected proceedings of the 2008 Second Language Research Forum: Exploring SLA perspectives, positions, and practices* (pp. 32–47). Cascadilla Proceedings Project. <https://www.lingref.com/cpp/slrf/2008/>
- Kallio, H., Kuronen, M., & Koivusalo, L. (2022). The role of pause location in perceived fluency and proficiency in L2 Finnish. In the *Proceedings of the 4th International Symposium on Applied Phonetics (ISAPh 2022)*, 22–27. International Speech Communication Association. <https://doi.org/10.21437/ISAPh.2022-5>
- Kallio, H., Kautonen, M., & Kuronen, M. (2023). Prosody and fluency of Finland Swedish as a second language: investigating global parameters for automated speaking assessment. *Speech Communication*, 148, 66–80. <https://doi.org/10.1016/j.specom.2023.02.003>

- Kallio, H., & Kuronen, M. (2023). Revising parameters for predicting L2 speech fluency and proficiency. In the *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS 2023)*. Guarant International.
https://drive.google.com/file/d/15U2l2y4_-9lyZAgmiccQYXYj9zBi_CAu/
- Kautonen, M. & Kuronen, M. (2021). Kvantitativa perspektiv på L2-tal på olika färdighetsnivåer. *Folkmålsstudier*, 59, 11–39.
<https://journal.fi/folkmalsstudier/article/view/112545>
- Kallio, H., Suni, A., Simko, J., & Vainio, M. (2020). Analyzing second language proficiency using wavelet-based prominence estimates. *Journal of Phonetics*, 80, 100966. <https://doi.org/10.1016/j.wocn.2020.100966>.
- Kautonen, M., Kuronen, M. & Kallio, H. (submitted). Automatisk bedömning av tal på L2-svenska med fokus på flyt.
- 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. <https://doi.org/10.1080/15434303.2018.1451531>
- Kurimo, M., Getman, Y., Voskoboinik, E., Al-Ghezi, R., Kallio, H., Kuronen, M., von Zansen, A., Hilden, R., Kronholm, S., Huhta, A., & Lindén, K. (2023). New data, benchmark and baseline for L2 speaking assessment for low-resource languages. In the *Proceedings of 9th Workshop on Speech and Language Technology in Education (SLaTE)*, 166–170. International Speech Communication Association (ISCA). <https://doi.org/10.21437/SLaTE.2023-32>
- Linacre, J. M. (2024). *Facets computer program for many-facet Rasch measurement* (Version 3.87.0) [Computer software].
<https://www.winsteps.com>
- Liu, X. J., Wang, J., & Zou, B. (2025). Evaluating an AI speaking assessment tool: Score accuracy, perceived validity, and oral peer feedback as feedback enhancement. *Journal of English for Academic Purposes*, 75, 101505.
<https://doi.org/10.1016/j.jeap.2025.101505>
- Lun, T. M., Voskoboinik, E., Al-Ghezi, R., Grosz, T., & Kurimo, M. (2024). Oversampling, Augmentation and Curriculum Learning for Speaking Assessment with Limited Training Data. In the *Proceedings of Interspeech*

- 2024, 4019–4023. International Society for Computers and Their Applications (ISCA). https://www.isca-archive.org/interspeech_2024/lun24_interspeech.html#
- Matsuura, R., Suzuki, S., Takizawa, K., Saeki, M. & Matsuyama, Y. (2025). Gauging the validity of machine learning-based temporal feature annotation to measure fluency in speech automatically. *Research Methods in Applied Linguistics*, 4(1), 100177. <https://doi.org/10.1016/j.rmal.2024.100177>
- National Board of Education (2010). Vieraiden kielten ja toisen kotimaisen kielen suullisen kielitaidon arviointi lukiossa [Assessment of oral proficiency in second national and foreign languages in upper secondary school]. Information release 49/2010. https://www.oph.fi/sites/default/files/documents/kielten_suullinen_kielitaito_arviointi.pdf
- Pearson. (n.d.). Part 1: English Speaking & Writing Test Introduction. <https://www.pearsonpte.com/articles/part-1-english-speaking-and-writing-test-introduction>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://doi.org/10.48550/arXiv.1201.0490>
- Phan, N., von Zansen, A., Kautonen, M., Voskoboinik, E., Grosz, T., Hilden, R., & Kurimo, M. (2024). Automated content assessment and feedback for Finnish L2 learners in a picture description speaking task. In the *Proceedings of Interspeech 2024*, 317–321. ISCA-International Speech Communication Association. <https://doi.org/10.21437/Interspeech.2024-1166>
- Sun, W. (2023). The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: A mixed methods investigation. *Frontiers in Psychology*, 14, 1210187. <https://doi.org/10.3389/fpsyg.2023.1210187>
- Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., & Dupoux, E. (2021). VoxPopuli: A Large-Scale Multilingual Speech Corpus

- for Representation Learning, Semi-Supervised Learning and Interpretation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 993–1003, Association for Computational Linguistics.
<https://doi.org/10.48550/arXiv.2101.00390>
- von Zansen, A., & Hilden, R. (2022). "It was cool and comfortable!" Akateemisten alkeistason S2-opiskelijoiden kokemuksia tietokoneella suoritettavasta puhumisen kokeesta. *Suomen Ainedidaktinen tutkimusseura. Ainedidaktisia tutkimuksia*, 22, 72–90. <http://hdl.handle.net/10138/353562>
- von Zansen, A., & Kallio, H. (2024). DigiTala – Moodle-sovellus suullisen kielitaidon automaattiseen arviointiin. *AFinLA-teema*, 17, 91–116.
<https://doi.org/10.30660/afinla.131465>
- von Zansen, A., Kallio, H., Sneck, M., Kuronen, M., Huhta, A., & Hilden, R. (2022). Ihmisarvioijien näkemyksiä suullisen kielitaidon automaattisesta arvioinnista, digitaalisesta arviointiprosessista sekä puhesuorituksista arvioitavista ulottuvuuksista. *AFinLa vuosikirja 2022. Suomen soveltavan kielitieteen yhdistyksen julkaisuja*, 79, 370–394.
<https://doi.org/10.30661/afinlavk.114821>
- Voskoboinik, E., Getman, Y., Al-Ghezi, R., Kurimo, M., & Grósz, T. (2023). Automated assessment of task completion in spontaneous speech for Finnish and Finland Swedish language learners. In the *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, 102–110. LiU Electronic Press. <https://aclanthology.org/2023.nlp4call-1.12/>
- Voskoboinik, E., Phan, N., Grósz, T., & Kurimo, M. (2025). Leveraging Uncertainty for Finnish L2 Speech Scoring with LLMs. In the *Proceedings of the Workshop on Automatic Assessment of Atypical Speech (AAAS-2025)*. University of Tartu Library. <https://hdl.handle.net/10062/107137>
- Xi, X. (2021). Validity and the automated scoring of performance tests. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed., pp. 513–529). Routledge. <https://doi.org/10.4324/9781003220756-40>

- Yoon, S.-Y., & Zechner, K. (2017). Combining human and automated scores for the improved assessment of non-native speech. *Speech Communication*, 93, 43–52. <https://doi.org/10.1016/j.specom.2017.08.001>
- Zou, B., Du, Y., Wang, Z., Chen, J., & Zang, W. (2023). An Investigation Into Artificial Intelligence Speech Evaluation Programs With Automatic Feedback for Developing EFL Learners' Speaking Skills. *SAGE Open* 13(3), 1–8. <https://doi.org/10.1177/21582440231193818>
- Zou, B., Liviero, S., Ma, Q., Zhang, W., Du, Y., & Xing, P. (2024). Exploring EFL learners' perceived promise and limitations of using an artificial intelligence speech evaluation system for speaking practice. *System (Linköping)*, 126, Article 103497. <https://doi.org/10.1016/j.system.2024.103497>