Towards a Real-time Swedish Speech Analyzer for Language Learning Games: A Hybrid AI Approach to Language Assessment

Tianyi Geng
Department of Philosophy,
Linguistics, Theory of Science
University of Gothenburg
`gusgenti@student.gu.se`

David Alfter
Gothenburg Research Infrastructure
in Digital Humanities
Department of Literature,
History of Ideas, Religion
University of Gothenburg
`david.alfter@gu.se`

Abstract

This paper presents an automatic speech assessment system designed for Swedish language learners. We introduce a novel hybrid approach that integrates Microsoft Azure speech services with open-source Large Language Models (LLMs). Our system is implemented as a web-based application that provides real-time quick assessment with a game-like experience. Through testing against COREFL English corpus data and Swedish L2 speech data, our system demonstrates effectiveness in distinguishing different language proficiencies, closely aligning with CEFR levels. This ongoing work addresses the gap in current low-resource language assessment technologies with a pilot system developed for automated speech analysis.

1 Introduction

In recent years, the integration of state-of-the-art artificial intelligence (AI) technologies—particularly large language models (LLMs)—has shown considerable promise across a range of domains, including Intelligent Computer-Assisted Language Learning (ICALL), Technology-Enhanced Language Learning (TELL), and Second Language Acquisition (SLA) (Zhang and Zou, 2022; Huang et al., 2023). A growing body of research has demonstrated the effectiveness of AI-driven language assessment tools (Daniels, 2022; Huawei and Aryadoust, 2023; Settles et al., 2020), highlighting their potential to facilitate language learning within contextually rich environments (Zou et al., 2023; Dizon, 2020; Huang et al., 2023). For instance, Brena et al. (2021) proposed supervised machine learning approaches capable of evaluating L2 English fluency and pronunciation with reported accuracy rates exceeding 90%. Despite these advancements, a recent systematic review of AI-based assessment in language learning (Chen et al., 2024) indicates a marked imbalance: 88% of the reviewed tools were developed for English learning, and only 3 out of 25 studies focused on assessing learners' speaking skills. This disparity underscores a significant gap in the current research landscape.

This paper aims to address the gap in automatic speech assessment tools, specifically for non-English languages by proposing a hybrid AI approach. We examine the adaptability of a pronunciation assessment tool optimized for English (Azure Speech Services; Microsoft 2024) to the low-resource Swedish language, then extend it by integrating large language models for content and delivery assessment, forming a detailed assessment system. In addition, the system is built as a Web App, providing real-time feedback as well as a game-like user experience. In the following sections, we will first justify the importance of building an automatic Swedish speech assessment system by reviewing recent related studies and applications around low-resource language speech assessment. We will then introduce our system design, followed by the evaluation and validation of the system with the English speech data from the COREFL corpus (Lozano et al., 2020) and an initial collection of Swedish L2 samples. Finally, we will discuss the results of the system tested for Swedish speech assessment and address the conclusions.

2 Related Work

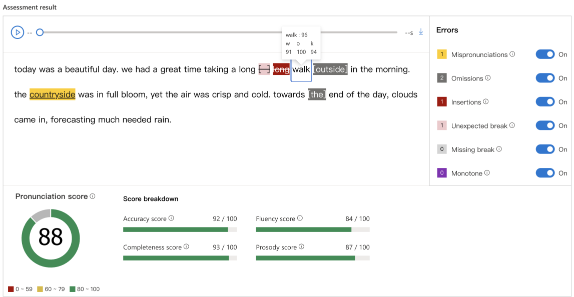
2.1 Automatic Speech Assessment Systems

Using mobile-assisted language learning (MALL) applications like Duolingo and Babbel has been a popular option for learners (Lehman et al., 2020; Loewen et al., 2020), especially for those studying low-resource languages for which accessible learning resources are scarce. Although MALL apps offer beginners a quick start, there is a lack of efficient or systematic follow-ups. Those apps mostly give a binary score ("correct or not"), or star-based assessment restricted to pronunciation practices, providing neither a comprehensive overview

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 186–201
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

of speaking ability nor detailed feedback such as pronunciation suggestions (Lehman et al., 2020; Chang et al., 2022).

For more detailed pronunciation assessment, Microsoft Azure Speech Studio (Microsoft, 2024) offers metrics related to accuracy, fluency, completeness, and prosody, as illustrated in Figure 1. While the service provides a multifaceted analysis at the phoneme, word, and sentence levels, the resulting scores remain relatively abstract and are not accompanied by pedagogically oriented feedback or actionable guidance for instructional use. In our experiments, the open-source Azure SDK was found to be primarily optimized for English language assessment, exhibiting limited capacity to accurately process Swedish phonemes. Notably, the system was unable to generate prosody scores for Swedish speech. Despite these limitations, the platform represents a promising prototype for pronunciation assessment and has the potential to be developed into a more robust tool for evaluating spoken language performance, particularly in the context of low-resource languages.



Assessment result

today was a beautiful day. we had a great time taking a long [-] ~~long~~ walk [outside] in the morning. the countryside was in full bloom, yet the air was crisp and cold. towards [the] end of the day, clouds came in, forecasting much needed rain.

walk : 96

w	ɔ	k
91	100	94

Errors

1	Mispronunciations	On
2	Omissions	On
1	Insertions	On
1	Unexpected break	On
0	Missing break	On
0	Monotone	On

Pronunciation score 88

Score breakdown

Accuracy score 92 / 100

Fluency score 84 / 100

Completeness score 93 / 100

Prosody score 87 / 100

Figure 1: Assessment interface of Azure Speech Studio

2.2 Swedish Learner Data

While the rapid advancement of artificial intelligence models has provided language education with handy tools for quick evaluations (Daniels, 2022; Löber et al., 2024), there is a lack of reliable and detailed automatic systems targeting lower-resource languages such as Swedish. Recent research has been working on filling the blank of Swedish learner data sets through building corpora of which language and proficiency levels are collected from coursebooks (COCTAILL corpus, representing learners' receptive ability) and learner essays (SweLL-pilot, representing learners' productive ability) (Volodina et al., 2019).

Nevertheless, the current progress has been made centering mostly texts rather than speech. The absence of a variety of publicly accessible, annotated Swedish speech data remains a significant obstacle for training robust (deep) learning models. While resources like Common Voice (Mozilla Foundation, 2020) provide raw speech data from native speakers, there is a scarcity of language learner speech samples on the spectrum of proficiency levels needed for developing language assessment or learning applications.

Getman et al. (2023) introduced an AI-assisted language learning application aimed at supporting children's second language acquisition in low-resource languages, specifically Swedish and Finnish, through the self-collection of relevant datasets. They also highlighted a significant gap in the field, noting: "To the best of our knowledge, in the context of Computer-Assisted Pronunciation Training (CAPT) for L2 Swedish and Finnish children, there are no previous work on automatic pronunciation assessment, not even for L2 Swedish and L2 Finnish adults" (Getman et al., 2023, p. 86026). In response to this gap, the present study contributes to the underexplored area of automatic speech assessment for L2 Swedish by developing a dedicated assessment system and conducting initial evaluations based on authentic speech data produced by L2 learners.

2.3 Language Proficiency Assessment Standards

The Common European Framework of Reference for Languages (CEFR; Council of Europe 2001) has been a widely recognized standard for assessing language proficiency, and recent research (Chen et al., 2024; Volodina et al., 2024) continues to use the CEFR standards and descriptors as reference-framework. While the Common European Framework of Reference for Languages (CEFR) remains a widely recognized standard, its limitations have been noted. As Alderson (2007, p. 660) observed, "the methodologies being used [to compile these descriptions] are unclear or suspect." The CEFR's abstract classification into six proficiency levels (A1 to C2) relies heavily on human evaluators—such as language instructors and linguists—which introduces concerns regarding subjectivity and scalability. Furthermore, although learners may be broadly categorized according to CEFR levels, the framework offers limited granular guidance tailored to specific proficiency levels or individual languages. This highlights a disconnect between the standardized assessment framework and the practical demands of language learning and instruction (Settles et al., 2020).

Our proposed automated speech system generates detailed analysis including:

- **Overall performance** Scores in pronunciation, content, and delivery of the speech; the corresponding CEFR level
- **Word-level pronunciation performance** demonstrating specific pronunciation strengths and weaknesses
- **Real-time feedback** with next-step learning suggestions

By combining the traditional assessment metrics and detailed, heuristic assessment analysis, we aim to build a system that generates more readable, informative results, to better serve both learners and educators.

3 System Design

Building on the automated speaking assessment framework developed by Educational Testing Service (ETS) and outlined by Zechner and Evanini (2019), the primary innovation of our system lies in the integration of complementary technologies to evaluate distinct dimensions of speech performance. The system is structured around three core modules: Pronunciation Assessment (based on two read-aloud tasks), Content and Delivery Assessment (based on a free-speech task), and CEFR Level Classification. The implementation takes the form of a web-based application featuring a gamified interface designed to enhance user engagement and learning experience.

3.1 Pronunciation Assessment Module

The system incorporates the pronunciation assessment module provided by Microsoft Azure's Speech SDK (Microsoft, 2024), which generates evaluation scores across five dimensions: Accuracy, Completeness, Fluency, Confidence, and Word-level confidence scores. Although the module does not support prosodic analysis for Swedish, our integration extends its applicability to the Swedish language and compensates for this limitation by supplementing it with two additional assessment modules.

3.2 Content-and-Delivery Assessment Module

The system utilizes a generative large language model (Llama 3.1; Touvron et al. 2023) to assess aspects of speech beyond pronunciation, specifically focusing on content relevance and language complexity in delivery. Based on predefined prompts (see details in Appendix F), the model produces quantified evaluation scores for these dimensions. Additionally, Llama 3 is prompted to generate human-like feedback in the form of constructive suggestions (see detailed examples in Appendix G), offering learners insights into how they can improve both the content and delivery of their spoken language.

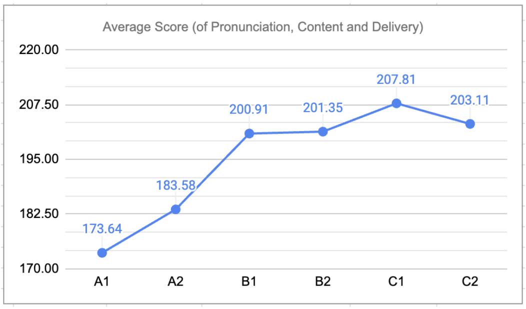


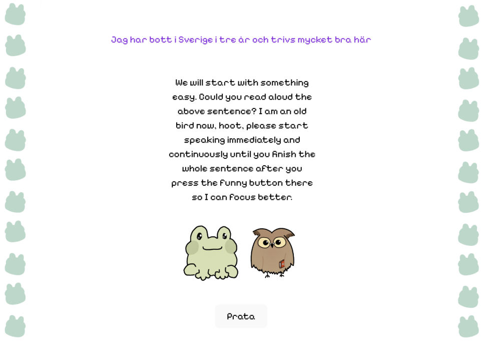
Figure 2: Average combined scores of pronunciation, content and delivery

3.3 CEFR Classification Module

Due to the lack of available Swedish data, and in order to provide an overall CEFR-based proficiency label for speech performance, we conducted a preliminary calibration of the combined scores generated by the two aforementioned AI modules. This calibration aligns the system's output with CEFR proficiency levels, using threshold values derived from test results on 55 carefully sampled English speech recordings ranging from A1 to C2, drawn from the COREFL corpus (Lozano et al., 2020) (see Figure 2). Notably, the system demonstrates strong discriminative capability at lower proficiency levels, whereas the distinction between B1 and B2 remains relatively subtle. The observed decline in scores from C1 to C2 is consistent with the known ambiguity of official CEFR descriptors at higher proficiency levels, as previously discussed by Isbell (2017) and Settles et al. (2020).

3.4 Web Implementation and User Experience Design

In our system, the player assumes the role of *Frog*, a character motivated to learn Swedish, and engages with Professowl, a fictional language professor who provides feedback and evaluations of the player's spoken Swedish. This narrative framing is intended to enhance learner engagement by embedding assessment within an interactive and playful context.



Jag har bott i Sverige i tre år och trivs mycket bra här

We will start with something easy. Could you read aloud the above sentence? I am an old bird now, hoot, please start speaking immediately and continuously until you finish the whole sentence after you press the funny button there so I can focus better.

Figure 3: Professowl guiding Frog through the pronunciation assessment tasks

The dialogue flow begins with Professowl guiding Frog through reading two Swedish sentences of different CEFR proficiency levels and then a free speech on the topic of "self introduction". Professowl gives corresponding feedback including scores and suggestions in an encouraging way.

4 Preliminary Results and Discussion

Given the limited availability of Swedish L2 speech data, we collected five original sets of preliminary speech samples from L2 learners at varying proficiency levels (see detailed results in Appendix C and D). These samples were manually evaluated by an experienced Swedish language instructor using the same scoring metrics employed by the automated system, enabling a direct comparison between human and machine assessments. While the dataset remains modest relative to high-resource languages such as English, it establishes an essential foundation and provides a baseline for subsequent analyses.

Due to the scale difference between Azure assessment metrics (0 to 100%) and our assessment metrics (1 to 5 Likert Scale) for the human rating, the system assessment scores were proportionally converted to 1 to 5 point scale based on thresholds at 20%, 40%, 60%, 80%.

As illustrated in Figure 4, a general alignment can be observed between the system-generated assessments and those provided by the human evaluator. However, the system is currently unable to assess prosody in Swedish, resulting in missing scores for this dimension. Furthermore, limitations in handling Swedish phonological characteristics lead to a rigid, word-by-word evaluation approach. For instance, commonly (phonologically) reduced function words in Swedish—such as *att* 'to' and *i* 'in/at' were frequently misclassified as "weak words" even when produced fluently. This issue is highlighted in the comparison of strong and weak word assessments between the system and the human evaluator (Appendix E).

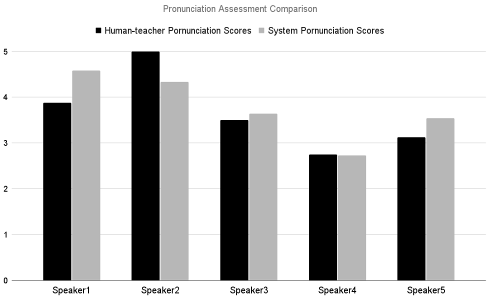


Figure 4: Average Pronunciation Scores Comparison

5 Conclusion and Future Work

In this paper, we present an initial prototype of a speech assessment system designed for Swedish. Our speech analyzer generates meaningful evaluation scores, provides reference word lists based on word-level pronunciation performance, and delivers both general feedback and personalized suggestions to support language learning.

The system combines Microsoft Azure's speech services with large language models to divide the assessment process into distinct tasks, each handled by separate tools. The game-like user experience design intends to promote learners' engagement (Hung, 2017; Hung et al., 2018). This approach demonstrates the potential of digital language learning tools in low-resource settings.

For future work, we plan to focus on several key aspects to improve the effectiveness and reliability of our system. First, we aim to achieve greater integration stability by stabilizing the speech services and embedding appropriate transition cues. This will reduce unintended delays during gameplay and ensure a smoother user experience throughout the learning process.

Second, we intend to enhance our phonological analysis capabilities by improving the system's ability to recognize and analyze phonological patterns in naturally spoken Swedish. This further development will enable more precise assessment of learners' pronunciation and speaking skills, particularly the nuances of Swedish phonology that are crucial for assessing language proficiency.

Third, we plan to significantly expand our data by collecting a larger and more comprehensive dataset covering learners at all proficiency levels from A1 to C2. This expanded dataset will better represent the full spectrum of Swedish learners and enable more robust training and reliable evaluation of our assessment algorithms.

Finally, we are focusing on improved validation procedures. To do this, we will engage additional teachers and annotators to rate language samples, thus confirming the accuracy of our automated assessments through inter-rater reliability measures. Furthermore, we plan to calibrate our CEFR classification system using authentic data from Swedish second language learners. This should help ensure that our proficiency level assignments conform to established CEFR standards and reflect the specific characteristics of Swedish language acquisition.

Limitations

This study presents a prototype system for automatic speech assessment in Swedish as a second language, but several limitations should be acknowledged. First, the evaluation relies on a small and preliminary dataset consisting of only five learner speech samples, which restricts the generalizability and statistical robustness of the findings. Second, the calibration of CEFR levels was based on English L2 data due to the lack of sufficient annotated Swedish learner corpora, which may have introduced cross-linguistic biases in proficiency classification. Third, the Azure speech assessment module lacks support for prosodic features in Swedish, limiting the system's ability to fully capture suprasegmental aspects of pronunciation. Additionally, the rigid word-by-word evaluation method often misinterprets function word reductions common in fluent speech, potentially penalizing natural speaking patterns. Furthermore, despite the robustness of the Microsoft Azure speech assessment analysis, the reliance limits replicability of this work. Other open-source alternatives such as Whisper-based assessment will be considered in future research to maximize the accessibility of the system.

Ethical Concerns

The development and deployment of automated language assessment tools raise several ethical considerations. Firstly, the system's reliance on proprietary and opaque evaluation mechanisms—such as Azure's speech scoring—may reinforce biases that are not easily observable or correctable by developers or users. Secondly, collecting and processing learner speech data involves privacy risks and must comply with ethical data handling standards, including informed consent and secure data storage. In this study, all participants were aged 18 or over and provided express consent for their speech data to be used for research purposes. Special care should be taken if the system is later extended to include minors or vulnerable populations, particularly in educational game-based settings. Lastly, while large language models can offer helpful feedback, they may inadvertently reinforce normative language ideologies or reflect implicit biases. To ensure fairness, pedagogical relevance, and user well-being, ongoing evaluation and human oversight are essential throughout system development and deployment.

Acknowledgements

This work was financially supported by the Royal Society of Arts and Sciences in Gothenburg (Kungliga Vetenskaps- och Vitterhets-Samhället i Göteborg).

References

J Charles Alderson. 2007. The CEFR and the need for more research. *The Modern Language Journal*, 91(4):659–663.

Ramon F Brena, Evelyn Zuvirie, Alan Preciado, Aristh Valdiviezo, Miguel Gonzalez-Mendoza, and Carlos Zozaya-Gorostiza. 2021. Automated evaluation of foreign language speaking performance with machine learning. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 15(2):317–331.

Younghoon Chang, Seongyong Lee, Siew Fan Wong, and Seon-phil Jeong. 2022. AI-powered learning application use and gratification: an integrative model. *Information Technology & People*, 35(7):2115–2139.

Angxuan Chen, Yuyue Zhang, Jiyou Jia, Min Liang, Yingying Cha, and Cher Ping Lim. 2024. A systematic review and meta-analysis of AI-enabled assessment in language learning: Design, implementation, and effectiveness. *Journal of Computer Assisted Learning*.

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

Paul Daniels. 2022. Auto-Scoring of Student Speech: Proprietary vs. Open-Source Solutions. *TESL-EJ*, 26(3):n3.

Gilbert Dizon. 2020. Evaluating intelligent personal assistants for L2 listening and speaking development. *Language Learning & Technology*, 24(1):16–26.

Yaroslav Getman, Nhan Phan, Ragheb Al-Ghezi, Ekaterina Voskoboinik, Mittul Singh, Tamas Grosz, Mikko Kurimo, Giampiero Salvi, Torbjørn Svendsen, Sofia Strömbergsson, et al. 2023. Developing an AI-assisted low-resource spoken language learning app for children. *IEEE Access*.

Xinyi Huang, Di Zou, Gary Cheng, Xieling Chen, and Haoran Xie. 2023. Trends, research issues and applications of artificial intelligence in language education. *Educational Technology & Society*, 26(1):112–131.

Shi Huawei and Vahid Aryadoust. 2023. A systematic review of automated writing evaluation systems. *Education and Information Technologies*, 28(1):771–795.

Hsiu-Ting Hung. 2017. Clickers in the flipped classroom: Bring your own device (byod) to promote student learning. *Interactive Learning Environments*, 25(8):983–995.

Hsiu-Ting Hung, Jie Chi Yang, Gwo-Jen Hwang, Hui-Chun Chu, and Chun-Chieh Wang. 2018. A scoping review of research on digital game-based language learning. *Computers & Education*, 126:89–104.

Daniel R Isbell. 2017. Assessing C2 writing ability on the Certificate of English Language Proficiency: Rater and examinee age effects. *Assessing Writing*, 34:37–49.

Blair Lehman, Lin Gu, Jing Zhao, Eugene Tsuprun, Christopher Kurzum, Michael Schiano, Yulin Liu, and G Tanner Jackson. 2020. Use of adaptive feedback in an app for English language spontaneous speech. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21*, pages 309–320. Springer.

Sarah Löber, Björn Rudzewitz, Daniela Verratti Souto, Luisa Ribeiro-Flucht, and Xiaobin Chen. 2024. Developing a Web-Based Intelligent Language Assessment Platform Powered by Natural Language Processing Technologies. In *Swedish Language Technology Conference and NLP4CALL*, pages 126–136.

Shawn Loewen, Daniel R Isbell, and Zachary Sporn. 2020. The effectiveness of app-based language instruction for developing receptive linguistic knowledge and oral communicative ability. *Foreign Language Annals*, 53(2):209–233.

Cristóbal Lozano, Ana Díaz-Negrillo, and Marcus Callies. 2020. Designing and compiling a learner corpus of written and spoken narratives: COREFL. *What's in a Narrative*, pages 21–46.

Microsoft. 2024. Azure Speech Studio. `https://speech.microsoft.com/portal`. Accessed: 2024-01-28.

Mozilla Foundation. 2020. Common Voice Dataset. `https://commonvoice.mozilla.org/`. Accessed: 2025-04-15.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for computational Linguistics*, 8:247–263.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Elena Volodina, David Alfter, and Therese Lindström Tiedemann. 2024. Profiles for Swedish as a second language: lexis, grammar, morphology. In *Huminfra Conference*, pages 10–19.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, et al. 2019. The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology (NEJLT)*, 6:67–104.

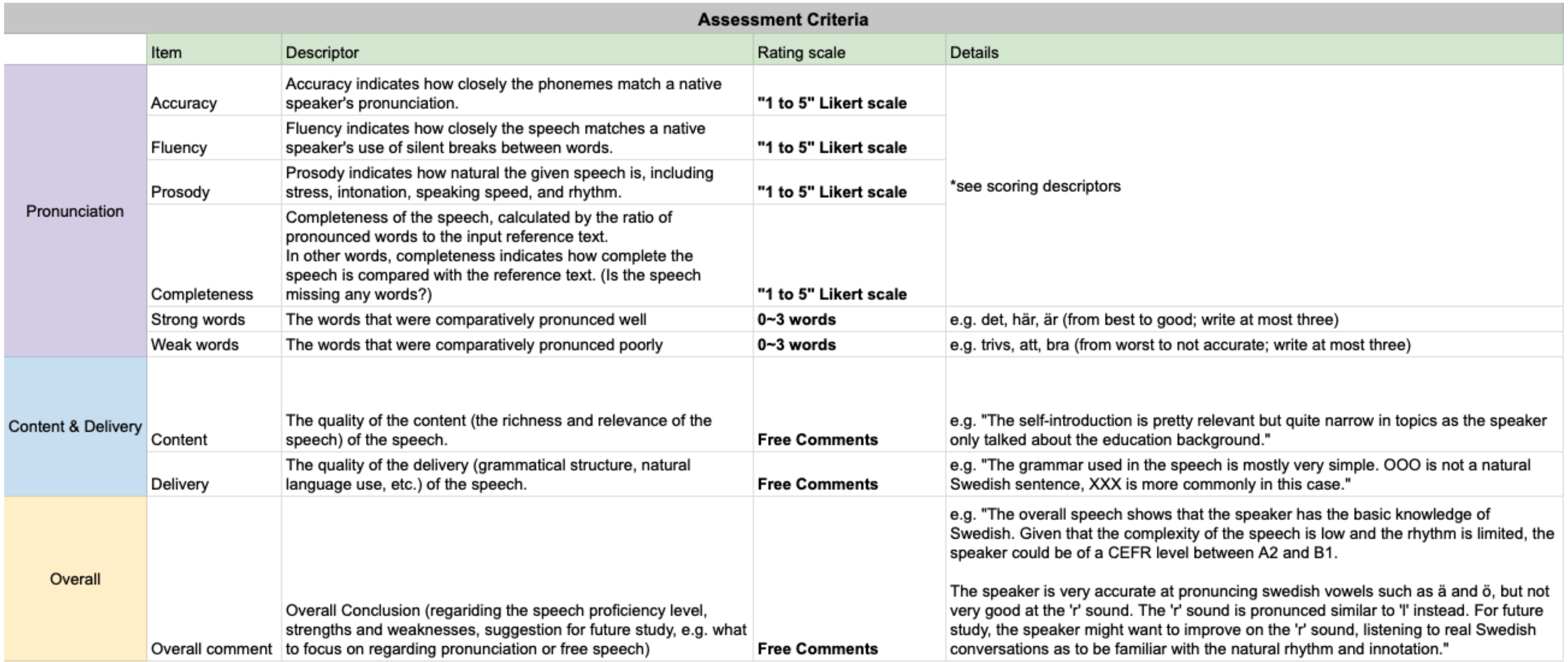
Klaus Zechner and Keelan Evanini. 2019. *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge.

Ruofei Zhang and Di Zou. 2022. Types, purposes, and effectiveness of state-of-the-art technologies for second and foreign language learning. *Computer Assisted Language Learning*, 35(4):696–742.

Bin Zou, Yiran Du, Zhimai Wang, Jinxian Chen, and Weilei Zhang. 2023. An investigation into artificial intelligence speech evaluation programs with automatic feedback for developing EFL learners' speaking skills. *Sage Open*, 13(3).

A Assessment Criteria

Figure 5 provides detailed descriptors for the pronunciation metrics used in our assessment system.



Assessment Criteria

	Item	Descriptor	Rating scale	Details
Pronunciation	Accuracy	Accuracy indicates how closely the phonemes match a native speaker's pronunciation.	**"1 to 5" Likert scale**	*see scoring descriptors
	Fluency	Fluency indicates how closely the speech matches a native speaker's use of silent breaks between words.	**"1 to 5" Likert scale**	
	Prosody	Prosody indicates how natural the given speech is, including stress, intonation, speaking speed, and rhythm.	**"1 to 5" Likert scale**	
	Completeness	Completeness of the speech, calculated by the ratio of pronounced words to the input reference text. In other words, completeness indicates how complete the speech is compared with the reference text. (Is the speech missing any words?)	**"1 to 5" Likert scale**	
	Strong words	The words that were comparatively pronunced well	**0~3 words**	e.g. det, här, är (from best to good; write at most three)
	Weak words	The words that were comparatively pronunced poorly	**0~3 words**	e.g. trivs, att, bra (from worst to not accurate; write at most three)
Content & Delivery	Content	The quality of the content (the richness and relevance of the speech) of the speech.	**Free Comments**	e.g. "The self-introduction is pretty relevant but quite narrow in topics as the speaker only talked about the education background."
	Delivery	The quality of the delivery (grammatical structure, natural language use, etc.) of the speech.	**Free Comments**	e.g. "The grammar used in the speech is mostly very simple. OOO is not a natural Swedish sentence, XXX is more commonly in this case."
Overall	Overall comment	Overall Conclusion (regariding the speech proficiency level, strengths and weaknesses, suggestion for future study, e.g. what to focus on regarding pronunciation or free speech)	**Free Comments**	e.g. "The overall speech shows that the speaker has the basic knowledge of Swedish. Given that the complexity of the speech is low and the rhythm is limited, the speaker could be of a CEFR level between A2 and B1. The speaker is very accurate at pronuncing swedish vowels such as ä and ö, but not very good at the 'r' sound. The 'r' sound is pronunced similar to 'l' instead. For future study, the speaker might want to improve on the 'r' sound, listening to real Swedish conversations as to be familiar with the natural rhythm and innotation."

Figure 5: Assessment criteria for human-teacher assessment

B Detailed Scoring Descriptors

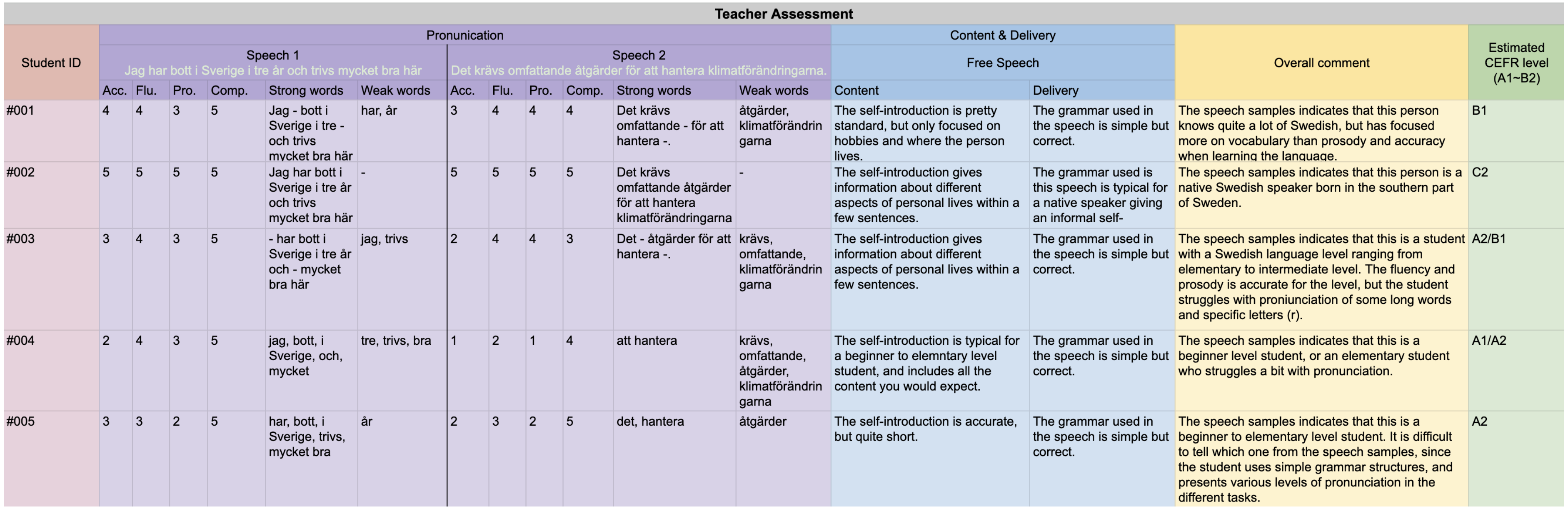
Table 1 provides detailed descriptors for the pronunciation metrics used in our assessment system.

Score	Accuracy	Completeness	Fluency	Prosody
1	Incomprehensible speech with almost no sounds that are accurate	Missing many important words (< 60%)	Very snatchy speech with frequent unnatural breaks	No variation in stress or intonation, or the rhythm is completely off
2	Many obvious errors in pronunciation, difficult to understand	Several missing words (60–75%)	Frequent hesitations and stops	Unnatural rhythm, intonation and stress patterns
3	Some noticeable errors but generally accurate and understandable	Most words included with some minor omission (75–85%)	Generally fluent flow with some unnatural stops	Some natural stress and intonation patterns
4	High accuracy with minor errors that don't affect comprehension	Nearly complete (85–95% coverage)	Generally smooth speech with occasional pauses	Generally appropriate stress, rhythm and intonation
5	Most sounds are perfectly correct, native-like speaking	Complete (95–100% coverage)	Natural, native-like speech flow with appropriate pauses	Native-like rhythm, stress, and intonation

Table 1: Detailed scoring descriptors for pronunciation metrics

C Preliminary Test Results (Human Assessment)

Figure 6 provides the human teacher's assessment results on the test speech samples.



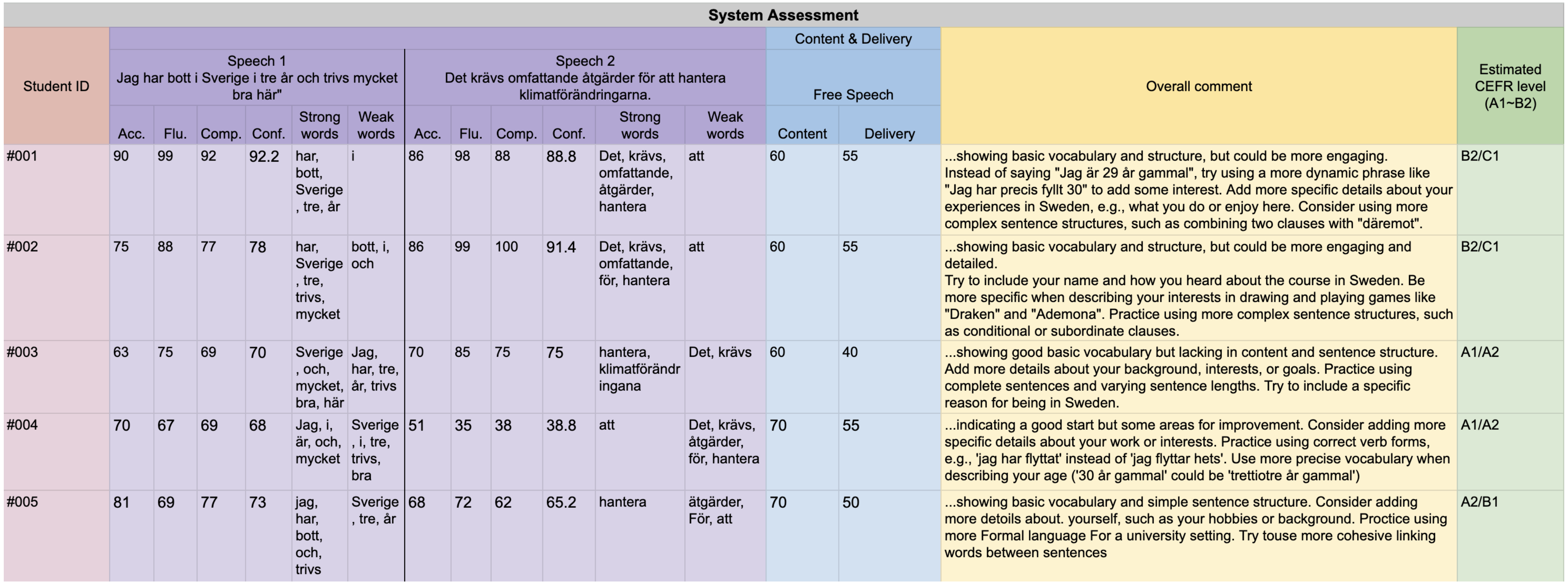
Teacher Assessment

Student ID	Pronunciation												Content & Delivery		Overall comment	Estimated CEFR level (A1~B2)
	Speech 1 Jag har bott i Sverige i tre år och trivs mycket bra här						Speech 2 Det krävs omfattande åtgärder för att hantera klimatförändringarna.						Free Speech			
	Acc.	Flu.	Pro.	Comp.	Strong words	Weak words	Acc.	Flu.	Pro.	Comp.	Strong words	Weak words	Content	Delivery		
#001	4	4	3	5	Jag - bott i Sverige i tre - och trivs mycket bra här	har, år	3	4	4	4	Det krävs omfattande - för att hantera -.	åtgärder, klimatförändringarna	The self-introduction is pretty standard, but only focused on hobbies and where the person lives.	The grammar used in the speech is simple but correct.	The speech samples indicates that this person knows quite a lot of Swedish, but has focused more on vocabulary than prosody and accuracy when learning the language.	B1
#002	5	5	5	5	Jag har bott i Sverige i tre år och trivs mycket bra här	-	5	5	5	5	Det krävs omfattande åtgärder för att hantera klimatförändringarna	-	The self-introduction gives information about different aspects of personal lives within a few sentences.	The grammar used is this speech is typical for a native speaker giving an informal self-	The speech samples indicates that this person is a native Swedish speaker born in the southern part of Sweden.	C2
#003	3	4	3	5	- har bott i Sverige i tre år och - mycket bra här	jag, trivs	2	4	4	3	Det - åtgärder för att hantera -.	krävs, omfattande, klimatförändringarna	The self-introduction gives information about different aspects of personal lives within a few sentences.	The grammar used in the speech is simple but correct.	The speech samples indicates that this is a student with a Swedish language level ranging from elementary to intermediate level. The fluency and prosody is accurate for the level, but the student struggles with proniunciation of some long words and specific letters (r).	A2/B1
#004	2	4	3	5	jag, bott, i Sverige, och, mycket	tre, trivs, bra	1	2	1	4	att hantera	krävs, omfattande, åtgärder, klimatförändringarna	The self-introduction is typical for a beginner to elemntary level student, and includes all the content you would expect.	The grammar used in the speech is simple but correct.	The speech samples indicates that this is a beginner level student, or an elementary student who struggles a bit with pronunciation.	A1/A2
#005	3	3	2	5	har, bott, i Sverige, trivs, mycket bra	år	2	3	2	5	det, hantera	åtgärder	The self-introduction is accurate, but quite short.	The grammar used in the speech is simple but correct.	The speech samples indicates that this is a beginner to elementary level student. It is difficult to tell which one from the speech samples, since the student uses simple grammar structures, and presents various levels of pronunciation in the different tasks.	A2

Figure 6: Preliminary results of human-teacher assessment

D Preliminary Test Results (System Assessment)

Figure 7 provides the system’s assessment results on the test speech samples.



System Assessment																
													Content & Delivery			
Student ID	Speech 1 Jag har bott i Sverige i tre år och trivs mycket bra här"						Speech 2 Det krävs omfattande åtgärder för att hantera klimatförändringarna.						Free Speech		Overall comment	Estimated CEFR level (A1~B2)
	Acc.	Flu.	Comp.	Conf.	Strong words	Weak words	Acc.	Flu.	Comp.	Conf.	Strong words	Weak words	Content	Delivery		
#001	90	99	92	92.2	har, bott, Sverige, tre, år	i	86	98	88	88.8	Det, krävs, omfattande, åtgärder, hantera	att	60	55	...showing basic vocabulary and structure, but could be more engaging. Instead of saying "Jag är 29 år gammal", try using a more dynamic phrase like "Jag har precis fyllt 30" to add some interest. Add more specific details about your experiences in Sweden, e.g., what you do or enjoy here. Consider using more complex sentence structures, such as combining two clauses with "däremot".	B2/C1
#002	75	88	77	78	har, Sverige, tre, trivs, mycket	bott, i, och	86	99	100	91.4	Det, krävs, omfattande, för, hantera	att	60	55	...showing basic vocabulary and structure, but could be more engaging and detailed. Try to include your name and how you heard about the course in Sweden. Be more specific when describing your interests in drawing and playing games like "Draken" and "Ademona". Practice using more complex sentence structures, such as conditional or subordinate clauses.	B2/C1
#003	63	75	69	70	Sverige, och, mycket, bra, här	Jag, har, tre, år, trivs	70	85	75	75	hantera, klimatförändringana	Det, krävs	60	40	...showing good basic vocabulary but lacking in content and sentence structure. Add more details about your background, interests, or goals. Practice using complete sentences and varying sentence lengths. Try to include a specific reason for being in Sweden.	A1/A2
#004	70	67	69	68	Jag, i, är, och, mycket	Sverige, i, tre, trivs, bra	51	35	38	38.8	att	Det, krävs, åtgärder, för, hantera	70	55	...indicating a good start but some areas for improvement. Consider adding more specific details about your work or interests. Practice using correct verb forms, e.g., 'jag har flyttat' instead of 'jag flyttar hets'. Use more precise vocabulary when describing your age ('30 år gammal' could be 'trettiotre år gammal')	A1/A2
#005	81	69	77	73	jag, har, bott, och, trivs	Sverige, tre, år	68	72	62	65.2	hantera	åtgärder, För, att	70	50	...showing basic vocabulary and simple sentence structure. Consider adding more detoils about. yourself, such as your hobbies or background. Proctice using more Formal language For a university setting. Try touse more cohesive linking words between sentences	A2/B1

Figure 7: Preliminary results of system assessment

E Preliminary Test Results (Strong/Weak Words Comparison)

Table 2 shows the assessment results comparison between the system and the teacher as for strong/weak words pronunciation.

Sentence	Student ID	Evaluator	Strong words	Weak words
S1	#001	System	här, bott, Sverige, tre, är	i
		Teacher	Jag, bott, i, Sverige, tre, och trivs, mycket, bra, här	här, är
	#002	System	här, Sverige, tre, trivs, mycket	bott, i, och
		Teacher	Jag, har, bott, i, Sverige, tre, år, och trivs, mycket, bra, här	–
	#003	System	Sverige, och, mycket, bra, här	Jag, har, tre, år, trivs
		Teacher	har, bott, i, Sverige, tre, år, och, mycket, bra, här	jag, trivs
	#004	System	Jag, i, är, och, mycket	Sverige, i, tre, trivs, bra
		Teacher	jag, bott, i Sverige, och, mycket	tre, trivs, bra
	#005	System	jag, har, bott, och, trivs	Sverige, tre, år
		Teacher	har, bott, i Sverige, trivs, mycket, bra	år
S2	#001	System	Det, krävs, omfattande, åtgärder, hantera	att
		Teacher	Det, krävs, omfattande, för, att, hantera	åtgärder, klimatförändringar
	#002	System	Det, krävs, omfattande, för, hantera	att
		Teacher	Det, krävs, omfattande, åtgärder, för, att, hantera, klimatförändringar	–
	#003	System	hantera, klimatförändringarna	Det, krävs
		Teacher	Det, åtgärder, för, att, hantera	krävs, omfattande, klimatförändringar
	#004	System	att	Det, krävs, åtgärder, för, hantera
		Teacher	att, hantera	krävs, omfattande, åtgärder, klimatförändringarna
	#005	System	hantera	ätgärder, För, att
		Teacher	det, hantera	åtgärder

Table 2: Comparison of strong and weak word analysis between system and teacher

F LLaMA Model Prompting Details

The "content and delivery assessment module" employs Llama 3.1 with carefully designed and tested prompts to ensure consistent feedback. The prompt details and model configuration are as follows:

F.1 Detailed Prompt

The following prompt template is passed with relevant values for evaluating the content and delivery of every speech input:

```
You are Professowl, a Swedish language teacher.  Analyze the following student's
self-introduction in Swedish: "$input"

Provide feedback in this JSON format:
{
"analysis": {
"relevance": number,    // 0-100, how well the content works as a self-introduction
"complexity": number,    // 0-100, language complexity level
"feedback": string,      // One sentence including both relevance and complexity scores
"suggestions": string[]    // 2-3 short, specific suggestions in English
}
}

Focus on:
- Content and relevance for self-introduction
- Sentence structure and vocabulary use
- Grammar and language flow

Example response:
{
"analysis": {
"relevance": 80,
"complexity": 65,
"feedback": "Your self-introduction is quite relevant (80/100) with moderate language
complexity (65/100), showing good basic vocabulary and structure.",
"suggestions": [
"Try using more descriptive adjectives when talking about your interests",
"Add information about your future plans in Sweden",
"Practice using more complex sentence connections"
]
}
}

Always include both scores in the feedback sentence.
```

F.2 Model Configuration

To ensure consistent and reliable outputs, the following configuration parameters are used:

- **Model**: Llama 3.1
- **Temperature**: 0.3
- **Output format**: JSON
- **Deployment**: Local Ollama instance

G System Feedback Examples

The following examples demonstrate system feedback generated by the LLaMA model for speech samples. The feedback structure is designed to be encouraging and provides score-based assessment as well as actionable suggestions.

G.1 Example 1

"Your self-introduction is somewhat relevant (70/100) with relatively low language complexity (55/100), indicating a good start but some areas for improvement. Consider adding more specific details about your work or interests. Practice using correct verb forms, e.g.,'Jag har flyttat' instead of 'Jag flyttar'. Use more precise vocabulary when describing your age 30 år gammal' could be 'trettiotre år gammal)"

G.2 Example 2

"Your self-introduction is somewhat relevant (70/100) with relatively low language complexity (50/100), showing basic vocabulary and simple sentence structure. Consider adding more details about yourself, such as your hobbies or background. Practice using more formal language for a university setting. Try to use more cohesive linking words between sentences."