

AI-BASED SPEAKING SKILLS ASSESSMENT AMONG ESL STUDENTS Akopiants N.

National Technical University «Kharkiv Polytechnic Institute», Kharkiv

The advancement of AI and its application to ESL students' assessment has recently become a relevant topic among educators and sparked debates about its effectiveness compared to traditional human-led assessment [1,2]. As a result, an initial experiment in the row of future prospective experiments in the area of speaking competence was conducted within the scope of Scientific and Methodological Laboratory at Business Foreign Languages and Translation Department in NTU "KhPI". The research is focused on the efficiency of the role of AI as a facilitator in assessing and improving students' speaking skills. The obtained results prompted us to conduct further comparative analysis regarding the accuracy and effectiveness of assessment performed by a human teacher versus AI-driven software.

The experiment was performed on the AI-powered platform smalltalk2me, which is designed to be a simulator to self-practice. Fifteen second- and third-year students majoring in Translation at Business Foreign Languages and Translation Department in NTU "KhPI" agreed to participate in an experiment where their English-speaking skills are assessed by artificial intelligence on this platform. The recorded respondents' answers and their AI-based feedback became the focus of this research. The obtained data were collected and subsequently analyzed. Smalltalk2me AI-driven software estimated the English speaking level based on five criteria: pronunciation, grammar, fluency, vocabulary, and interaction.

To conclude, the agreement between human and AI assessments reveals how often AI gives the same rating as a human for a particular skill. In this case, agreement was calculated as the percentage of students for whom AI assigned the exact same level as the human evaluator. Calculations show that the agreement between human and AI assessments varies across categories:

- **Interaction (37.5%)** has the highest agreement.
- **Grammar, Fluency, and Vocabulary (25%)** show moderate agreement.
- **Pronunciation (18.75%)** and **Overall Level (12.5%)** have the lowest agreement.

This suggests that AI mostly struggles with assessing **overall speaking level** and **pronunciation**, while it performs relatively better in evaluating **interaction skills**.

References:

1. M. Zhu, and Chaoran Wang, A Systematic Review of Artificial Intelligence in Language Education from 2013 to 2023: Current Status and Future Implications. Preprint, Posted: 4 Jan 2024. <http://dx.doi.org/10.2139/ssrn.4684304>. URL: <https://ssrn.com/abstract=4684304>
2. T. H. B. Nguyen, and T. D. H. Tran, Exploring the Efficacy of ChatGPT in Language Teaching, AsiaCALL Online Journal, 14(2), 2023, pp. 156–167. DOI: <https://doi.org/10.54855/acoj.2314210>