EUROCALL 2023: CALL for all Languages
15-18 August 2023, University of Iceland, Reykjavik
DOI: https://doi.org/10.4995/EuroCALL2023.2023.17007

Evaluating the effectiveness of Microsoft Transcribe for automating the assessment of pronunciation in language proficiency tests

Carey Nelson[a] and Walcir Cardoso[b]

[a]Department of Education, Concordia University, Montreal, , nelson.carey@uqam.ca and [b]Department of Education, Concordia University, Montreal, , walcir.cardoso@concordia.ca

How to cite: Nelson, C.; Cardoso, W. (2023). Evaluating the effectiveness of Microsoft Transcribe for automating the assessment of pronunciation in language proficiency tests. In *CALL for all Languages - EUROCALL 2023 Short Papers.* 15-18 August 2023, University of Iceland, Reykjavik. https://doi.org/10.4995/EuroCALL2023.2023.17007

Abstract

Improvements in Automatic Speech Recognition (ASR) have created opportunities for using it as a tool to facilitate second and foreign language (L2) assessment. These technical improvements have not only enabled automation of language proficiency test scoring but also reduced evaluator bias and errors, decreased processing time, and lowered costs for testing organizations. The purpose of this study was to evaluate English as a Second Language (ESL) pronunciation using the ASR feature in the Microsoft 365 product suite, Transcribe (MS-T). The study involved adult ESL learners at a Canadian university that partook in a language proficiency test. We examined the audio recordings of 56 candidates during the pronunciation portion of the test. Building on previous studies that found a strong correlation between scores from Google Voice Typing and human raters, the current study conducted a similar analysis comparing scores derived from MS-T to both human ratings and Google Voice Typing. Our findings indicate that the ASR capabilities of MS-T, similar to Google Voice Typing, can assume an important role in L2 speaking assessment by providing objectivity and reliability to the testing process, expediting scoring, and reducing costs.

Keywords: *automated evaluation, Automatic Speech Recognition (ASR), Language assessment, ESL pronunciation evaluation, Microsoft Transcribe (MS-T), placement tests.*

1. Introduction

Language proficiency testing is an important field that tries to meet the demand of evaluating candidates' proficiency in L2 for purposes such as employment, immigration, and academic admissions. These tests may be internationally recognized, e.g. TOEIC, IELTS, Duolingo, Versant, or created locally by individual institutions. In either case, language proficiency testing often touches on the assessment of a combination of language skills to determine the test-takers' abilities in that language. However, assessments with subjective components that require human raters to manually apply scoring rubrics to students' oral performances can be extremely time-consuming and labor-intensive (Coombe et al., 2020). Moreover, relying on human raters to manually score assessments leaves room for potential errors and inconsistencies based on subjective interpretations of scoring rubrics (Inbar-Lourie, 2017). There are also time and cost considerations associated with having human raters evaluate students' oral performance.

An ASR system can be particularly useful for language proficiency tests where multiple evaluators may be involved (Bernstein et al., 2010). Firstly, traditional language proficiency tests, e.g. TOEFL and IELTS, have limitations in terms of providing an accurate assessment of a learner's language proficiency. These tests rely on human evaluators who may be subject to biases and inconsistencies, and who may not be able to provide an objective and standardized evaluation of language proficiency (Xerri & Briffa, 2018). In contrast, ASR technology has the potential to provide a less biased, more consistend and standardized assessment of language proficiency, which can reduce variability in scores across different evaluators (Mroz, 2020). Secondly, due to significant technological advancements in recent years, ASR has been used for various applications, including speech recognition in smartphones, virtual assistants, and language learning platforms (McCrocklin & Edalatishams, 2020). Tools that use ASR dictation, such as MS-T, offer a promising avenue for a cost-effective solution. The possibility of implementing ASR technology could help post-secondary institutions operate efficient and effective proficiency tests.

Considering a recent study by Johnson et al. (2022), who found a strong correlation between scores assessed by Google Voice Typing (GVT) and human raters, demonstrated that GVT provided reliable and valid results in evaluating participants’ oral performance on a set of phonological criteria (e.g. phonemic accuracy, stress) and overall proficiency (e.g. comprehensibility) – on par with human raters. This raises the question: Can the same conclusions be drawn with other ASR tools? In a study from 2017, Këpuska and Bohouta found that GVT offered a superior recognition to Microsoft’s ASR. Given the rapid advancements in speech technology over recent years, does this comparison still hold true today? Our study thus aimed at building on these previous findings by examining whether a similar ASR application, e.g. Microsoft Transcribe (MS-T; found in the Microsoft 365 product suite), has the potential to improve the validity efficiency of L2 speaking assessment. While MS-T is not as accessible to the general public as GVT, this study utilized MS-T to offer an alternative to universities that have access to Microsoft products. In addition, the choice to use MS-T will allow us to gain additional insights into the impacts of ASR on assessment through the analysis of an alternative platform. Finally, this decision will allow us to compare the performance of two predominant ASR platforms, MS-T and GVT. As such, our study aimed to answer the following Research Questions (RQs):

1. What is the relationship between MS-T-rated pronunciation scores and human-rated pronunciation scores (RQ1)?
    a. Do relationships vary between MS-T-rated scores and human-rated scores across the set of evaluation criteria used by the human raters?
    b. Do relationships vary between MS-T-rated scores and human-rated scores across participant proficiency levels?
2. How does MS-T fare compared to GVT at pronunciation scoring across the set of evaluation criteria used by the human raters (RQ2)?

2. Method

2.1. Context and participants

The sample of participants comprised 56 adults (n = 56; 21 males, 35 females; mean age: 28.1), with the following distribution of native languages: French (n=39, 68.4%), Spanish (n=4; 7.0%), Arabic (n=3; 5.6%), and others (5; 8.92%). They were undergraduate students who had taken a proficiency test either to enroll in proficiency-appropriate ESL classes or to fulfill linguistic requirements for their academic programs. They had oral production levels ranging from A1 to C2 according to the Common European Framework of Reference (CEFR) for languages (Council of Europe, 2001). The sample for this study comprised 75 recordings drawn from over 20,000 proficiency tests administered between 2015 and 2020. Only the pronunciation portion of these tests were considered, which were randomly selected across six proficiency levels, based on the candidates' pronunciation scores obtained on the tests. Nineteen sound files were eliminated from the sample because they were not clearly audible.

2.2. Procedure

This study made use of secondary data, which were originally collected by a modern language department at a French-speaking university in Canada. It was part of a larger project to implement new rubrics for scoring pronunciation on the ESL proficiency test. With respect to the pronunciation section, the participants were given two practice sentences to be read aloud. This ensured that they understood the evaluation activity procedure and that the computer system was functioning correctly. The test then gave the participants five sentences that appeared one after the other, after 20 seconds, with increasing levels of speaking and pronunciation difficulty. The students read these five sentences where the first one was a baseline sentence (same for all students) followed by four sentences that were randomly chosen from a pool organized by proficiency level. The test saved each sentence as an individual recording to be consulted by the raters. Due to an agreement to protect the integrity of the test, the baseline question cannot be shown here. Table 1 is an example of sentences to be read aloud by students and is taken from Johnson et al. (2022). The human raters used a rubric assessing five phonological components: *Comprehensibility, Phonemes, Connected speech, Word stress and rhythm,* and *Thought groups, sentence stress, and intonation*.

Table 1. Sample sentences.

Level	Sentence
1	[Baseline sentence]
2	*A trio sings to the audience as it streams onto the busy street in the cold rain.*
3	*These are more sophisticated pictures, aimed at a particular kind of filmgoer. Is she sure that this audience understands them?*
4	*After the stems are cut off the mushrooms, they are then going to be sautéed with a small onion, a clove of garlic, and an eighth of a cup of breadcrumbs.*
5	*Even though the trailer has been cleaned, there are still lingering traces of acetones and other toxic amalgams either in the gaskets or in the valve assembly.*

2.3. Procedure

For data analysis, the MS-T score, the GVT score, the final human-rater score, as well as the scores for each criterion were entered into SPSS 29. These results help to answer the research questions of whether a relationship exists between MS-T rated pronunciation and human raters' evaluations.

3. Results

Looking at the relationship between MS-T and the human-rated scores, the results show a statistically significant strong correlation between the two variables, $r_s(54) = .79$, $p < .001$. For the first sub-question pertaining to the relationships between the MS-T scores and the rubric criterion (RQ1a), the results indicate that there are statistically significant strong correlations between the MS-T score and each of the criteria considered for the assessment of pronunciation (Tables 2 and 3 show the summary of the correlations). With respect to the question about the relationship between MS-T scores and test-taker proficiency (RQ1b), results indicate a significant correlation between the MS-T and human-rated scores for lower-proficiency test takers, $r_s(54) = .59$, $p < .006$, but a non-significant weak correlation between the MS-T and human-rated scores for higher-proficiency test takers, $r_s(54) = .29$, $p = .89$.

Table 2. Correlations between MS-T score and human-rated scores by criteria.

Rubric criteria	r_s	95% BCa Cis
Comprehensibility	.83**	.72, .90
Phonemes	.76**	.62, .85
Connected speech	.78**	.65, .87
Word stress and rhythm	.73**	.57, .84
Thought groups, sentence stress, and intonation	.76**	.62, .86

Note. Confidence intervals based on 1000 bootstrap samples. ***p* < .001.

Table 3. Correlations between MS-T score and human-rated scores by proficiency level.

Rubric criteria	r_s	95% BCa Cis
Lower-level proficiency	.59**	.19, .82
Higher-level proficiency	.29	-.56, .57

Note. Confidence intervals based on 1000 bootstrap samples. ***p* < .006.

RQ2 asked about the relationship between MS-T and GVT at pronunciation scoring. The results in Table 4 show that GVT is marginally stronger, but not significant, at the criteria of *comprehensibility*, *phonemes*, and *thought groups, sentence stress, and intonation*, whereas MS-T is a little stronger at the criterion *connected speech* and *word stress and rhythm*. After running a two-tailed t-test, we observe that the mean for MS-T is significantly higher than for GVT for pronunciation scoring total *t*(54) = -3.7, *p* = .001.

Table 4. Comparison (r_s) between MS-T scores and GVT scores.

Rubric criteria	MS-T	GV-T
Comprehensibility	.83**	85**
Phonemes	.76**	.78**
Connected speech	.78**	.72**
Word stress and rhythm	.73**	.71**
Thought groups, sentence stress, and intonation	.76**	.79**
Total	.79**	.78**

Note. Confidence intervals based on 1000 bootstrap samples. ***p* < .001.

4. Discussion and conclusions

This study evaluated the effectiveness of MS-T as an ASR for scoring L2 pronunciation assessment, and compared it to another popular ASR platform, GVT. The first research question looked at how MS-T compared to human raters when evaluating L2 pronunciation samples. Our findings for RQ1 indicate that MS-T displays strong, significant correlation across each of the different phonological criteria used in the evaluation. These results indicate, like what Johnson et al. (2022) found with GVT, that there exists a strong association when evaluating L2 pronunciation between MS-T and human-rated scores. This leads us to believe that the ASR MS-T gives similar results for evaluating L2 pronunciation samples.

With regards to RQ2, we looked at how MS-T compared to GVT for evaluating pronunciation samples. Overall, considering all phonological elements together, MS-T had significantly stronger correlations than GVT. However, the results were mitigated when we delved into each of the phonological criteria. The GVT and MS-T results showed varying degrees of correlation with human rater results across different aspects of pronunciation. Specifically, GVT had higher correlation for *comprehensibility, phonemes,* and *thought groups, sentence stress, and intonation*, while MS-T correlated more closely for *connected speech* and *word stress and rhythm*. This study examined whether a relationship existed between MS-T and human-rated scores with the goal of determining if this ASR platform could provide valid and reliable automated scoring of pronunciation that aligns with human judgements. The MS-T scores have been shown to be reliable and valid as the correlations between MS-T and human scores fare quite strong. As a result, this study demonstrates there is a possibility of MS-T implementation as an automated pronunciation scoring feature for proficiency tests, thereby improving the practicality of these tests, possibly leading to a reduction in costs. In addition, the results show that MS-T can potentially be deployed to automatically score pronunciation assessments, such as low-stakes language tests (proficiency placement tests). Further research is needed to confirm this perspective. In future research, it could be of interest to find out why these differences exist between the two ASR systems. Finally, the study shows that leveraging MS-T's ASR capabilities for L2 testing has the potential to provide a more efficient and cost-effective assessment tool compared to traditional language proficiency tests. It also suggests that MS-T can serve as a viable alternative to GVT for L2 pronunciation scoring.

Acknowledgements

We would like to express our gratitude to Carol Johnson for her guidance throughout this study. We also wish to thank Kathleen Brannen, Susan Jackson, Suzanne Springer, and Beau Zuercher, all from UQAM, for their assistance with the pronunciation samples; and Vicente Amati and Taras Kamtchatnikov, also from UQAM, for their technical support.

References

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355-377. https://doi.org/10.1177%2F0265532210364404

Berry, V., Sheehan, S., & Munro, S. (2019). What does language assessment literacy mean to teachers? *ELT Journal, 73*(2), 113-123. https://doi.org/10.1093/elt/ccy055

Coombe, C., Vafadar, H., & Mohebbi, H. (2020). Language assessment literacy: What do we need to learn, unlearn, and relearn? *Language Testing in Asia*, *10*(3), 1-16. https://doi.org/10.1186/s40468-020-00101-6

Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Inbar-Lourie, O. (2017). Language assessment literacy. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment: Encyclopedia of language and education* (3rd ed., pp. 257-268). Springer.

Johnson, C., Cardoso, W., Zuercher, B., Brannen, K. et Springer, S. (2022). Using Google Voice Typing to automatically assess pronunciation. In *Intelligent CALL, granular systems, and learner data: Short papers from EUROCALL 2022* (p. 203–207). Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2022.61.1459

Këpuska, V., & Bohouta, G. (2017). Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). *Journal of Engineering Research and Applications*, *7*(03), 20-24.

McCrocklin, S., & Edalatishams, I. (2020). Revisiting popular speech recognition software for ESL speech. *TESOL Quarterly*, *54*(4), 1086-1097. https://doi.org/10.1002/tesq.3006

Mroz, A. (2020). Aiming for advanced intelligibility and proficiency using mobile ASR. *Journal of Second Language Pronunciation*, *6*(1), 12–38. https://doi.org/10.1075/jslp.18030.mro

Xerri, D., & Briffa, P. (Eds.). (2018). *Teacher involvement in high-stakes language testing*. Springer.