# Using Google Voice Typing to automatically assess pronunciation

Carol Johnson[1], Walcir Cardoso[2], Beau Zuercher[3], Kathleen Brannen[4], and Suzanne Springer[5]

**Abstract**. This study examined the use of a popular Automatic Speech Recognition (ASR), Google Voice Typing (GVT), to automatically assess English as second language pronunciation. It aimed to answer the following question: What is the relationship between GVT-rated scores and human-rated scores? To answer this question, we compared audio recordings of 56 oral placement tests, rated by both human raters and GVT. Our results indicate that GVT scores strongly correlated with human-rater scores, indicating that this non-customizable ASR technology could be leveraged to increase the test usefulness of language placement tests.

**Keywords**: automatic speech recognition, automatic assessment, L2 pronunciation, Google voice typing.

## 1. Introduction

Language programs and schools rely on in-house placement tests to ensure students register in level-appropriate classes. However, these tests require extensive financial and human resources (Isaacs, 2018). This is especially true when assessing pronunciation, which often involves interviewers and multiple raters (Cox & Davis, 2012). However, there are known problems with rater reliability. For example, raters may overestimate the comprehensibility of second

1. Concordia University, Montreal, Canada; carol.johnson@concordia.ca; https://orcid.org/0000-0001-6859-6566
2. Concordia University, Montreal, Canada; walcir.cardoso@concordia.ca; https://orcid.org/0000-0001-6376-185X
3. Université du Québec à Montréal, Montreal, Canada; zuercher.beau_ryan@uqam.ca
4. Université du Québec à Montréal, Montreal, Canada; brannen.kathleen@uqam.ca
5. Université du Québec à Montréal, Montreal, Canada; springer.suzanne@uqam.ca

**How to cite this article**: Johnson, C., Cardoso, W., Zuercher, B., Brannen, K., & Springer, S. (2022). Using Google Voice Typing to automatically assess pronunciation. In B. Arnbjörnsdóttir, B. Bédi, L. Bradley, K. Friðriksdóttir, H. Garðarsdóttir, S. Thouësny, & M. J. Whelpton (Eds), *Intelligent CALL, granular systems, and learner data: short papers from EUROCALL 2022* (pp. 203-207). Research-publishing.net. https://doi.org/10.14705/rpnet.2022.61.1459

language (L2) speakers when familiar with their accents (Carey, Mannell, & Dunn, 2010). This lack of reliability can lead not only to the incorrect placement of students, but also to skepticism of test results (van der Walt, de Wet, & Niesler, 2008).

Using ASR to assess pronunciation could mitigate these problems. Large testing companies have been using custom-built ASR technology for more than two decades (e.g. Pearson's Versant – Bernstein, Van Moere, & Cheng, 2010). These organizations have access to financial and human resources beyond that of language institutions, but with advances in technology, it is now feasible for smaller organizations to take advantage of ASR to assess pronunciation. Studies of customized ASR-based assessment tools developed for language placement tests have found that ASR scores are strongly correlated with human-rater scores (Cox & Davis, 2012, van der Walt et al., 2008). Nevertheless, customizing ASR still requires a substantial budget and specific knowledge for development and maintenance (Isaacs, 2018). One way of mitigating this might be the use of free non-customizable ASR such as GVT, which has reached a high recognition rate of English L2 speech for high proficiency learners (McCrocklin & Edalatishams, 2020) and, consequently, has the potential to provide language institutions with simple low-cost solutions for pronunciation assessment.

The aim of this study was to determine if the use of GVT to assess pronunciation could increase the test usefulness of a university language placement test, based upon Bachman and Palmer's (1996) test usefulness model. This model consists of six qualities that determine the usefulness of a test: reliability, construct validity, authenticity, interactiveness, impact, and practicality. The relative importance of these qualities depends on the context of the test. As such, there is no perfect intertwining of the qualities. Instead, test developers must balance these qualities and accept that some may be negatively impacted for the sake of others, based on the purpose of the test. The research question that guided this study was:

- (1) What is the relationship between GVT-rated and human-rated pronunciation scores?
- (1a) Do relationships vary between GVT-rated and human-rated scores across a set of evaluation criteria?
- (1b) Do relationships vary between GVT-rated and human-rated scores across participant proficiency levels?

## 2. Method

Fifty-six undergraduate students of various oral proficiency levels at a university in Canada were recorded during the pronunciation portion of their placement tests for English second language (ESL) courses (Novice=2, Beginner=6, Intermediate=12, Advanced=14, and Fluent=22). Participants read aloud five increasingly difficult sentences (randomly chosen from a bank of sentences), which appeared sequentially on a screen for a period of 20 seconds each. To obtain the human-rated score, a rubric assessing five criteria (comprehensibility; phonemes; connected speech; word stress and rhythm; thought groups, sentence stress, and intonation) was used by three experienced ESL instructors who came to a consensus about each participant's score for each criterion. The same recordings were then played into GVT in Google Docs to obtain the GVT score. The output was analyzed manually with a point given for each correctly recognized word. The total number of points was divided by the total number of words in the sentences and multiplied by 100. Correlations were run to determine if a relationship existed between the human-rated and the GVT scores.

## 3. Results and discussion

The results are summarized in Table 1 below. In answer to the first research question regarding the relationship between human-rated and GVT-rated scores, a statistically significant strong correlation was found between human-rated and GVT scores. Regarding the first sub-question about relationships between GVT scores and the rubric criteria (1a), statistically significant strong correlations were also found between each criterion and the GVT scores. In regard to the second sub-question concerning the relationship between GVT scores and test-taker proficiency (1b), a significant strong correlation was found for lower-proficiency test-takers (Beginner-Intermediate), but a non-significant weak correlation was found for higher-proficiency test-takers (Advanced-Fluent).

These findings corroborate the existing literature that found ASR scores correlate with human-rater scores (Cox & Davies, 2012; van der Walt et al., 2008). This study further contributes to the field as our findings seem to indicate that non-customizable ASR, such as GVT, is a valid option when automatizing the assessment of pronunciation.

In terms of Bachman and Palmer's (1996) test usefulness model, the use of GVT can improve the overall test usefulness of a pronunciation placement test. Validity,

reliability, and practicality are the only three elements of the model affected by changing the assessment method. Validity may be somewhat reduced as GVT assessed intelligibility (e.g. what it understands) whereas the human raters assessed multiple aspects of pronunciation (e.g. comprehensibility and prosody). However, it allows for an increase in both reliability (i.e. the consistency of machine scoring) and practicality (i.e. reduced human resource costs and time to score tests).

Table 1. Descriptive statistics and correlations: GVT scores

| Variable | M | SD | rho | p | 95% BCa Cis |
|---|---|---|---|---|---|
| GVT Score (/100) | 73.09 | 22.49 | - | - | - |
| Human-rater score (/100) | 72.00 | 26.95 | .78** | <.001 | .64,.88 |
| Comprehensibility (/5) | 4.14 | 1.20 | .85** | <.001 | .74,.90 |
| Phonemes (/5) | 3.39 | 1.47 | .78** | <.001 | .63,.88 |
| Connected speech (/5) | 3.34 | 1.51 | .72** | <.001 | .53,.84 |
| Word stress and Rhythm (/5) | 3.63 | 1.34 | .71** | <.001 | .52,.84 |
| Thought groups, sentence stress, intonation (/5) | 3.50 | 1.51 | .79** | <.001 | .65,.88 |
| Lower-proficiency – GVT score (/100) | 47.74 | 13.78 | - | - | - |
| Lower-proficiency – Human-rater score (/100) | 40.20 | 16.80 | .78** | <.001 | .56,.89 |
| Higher-proficiency – GVT score (/100) | 87.18 | 10.96 | - | - | - |
| Higher-proficiency – Human-rater score (/100) | 89.67 | 9.44 | .28 | .10 | -.07,.56 |

Note. Confidence intervals based on 1,000 bootstrap samples.
**p<.001.

## 4. Conclusions

The purpose of this study was to explore the use of GVT to score a pronunciation assessment. The findings reported indicate that GVT scores strongly correlated with human-rater scores, suggesting that non-customizable ASR could be leveraged to increase the usefulness of a placement test.

Certain limitations should be taken into consideration. With only 56 participants, the findings are not generalizable to other populations. Additionally, recordings were used rather than live speech, and at times, this may have affected the technology's ability to correctly represent what was being said, negatively impacting the scores generated by GVT.

This study has shown the potential of using non-customizable ASR technology to score pronunciation placement tests. The next step is to fully automate the scoring process. This would provide opportunities to further research its use, such as determining if GVT has any biases in terms of first language, gender, or age.

Language programs and schools rely on placement tests to ensure that the learning experiences of students are optimal. Automating the placement process for pronunciation courses by using free and readily available technology such as GVT would allow institutions to offer their students a more streamlined and reliable service. Language programs and schools should consider taking advantage of it to facilitate the placement of students while at the same time ensuring that students continue to receive the learning experiences they want and deserve.

## 5. Acknowledgments

We would like to thank Vicente Amati and Taras Kamtchatnikov (*Centre d'évaluation des compétences linguistiques*). This research was funded by *Entente Canada-Québec* (UQAM, *École de langues*) and the *Social Sciences and Humanities Research Council of Canada* (Carol Johnson).

## References

Bachman, L., & Palmer, A. (1996). *Language testing in practice.* Oxford University Press.

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

Carey, M., Mannell, R., & Dunn, P. (2010). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201-219. https://doi.org/10.1177/0265532210393704

Cox, T., & Davies, R. (2012). Using automatic speech recognition technology with elicited oral response testing. *CALICO Journal, 29*(4), 601-618.

Isaacs, T. (2018). Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly*, *15*(3), 273-293. https://doi.org/10.1080/15434303.2018.1472264

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

Van der Walt, C., de Wet, F., & Niesler, T. (2008). Oral proficiency assessment: the use of automatic speech recognition systems. *Southern African Linguistics and Applied Language Studies, 26(*1), 135-146. https://doi.org/10.2989/SALALS.2008.26.1.11.426

Published by Research-publishing.net, a not-for-profit association
Contact: info@research-publishing.net



**Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022**
**Edited by Birna Arnbjörnsdóttir, Branislav Bédi, Linda Bradley, Kolbrún Friðriksdóttir, Hólmfríður Garðarsdóttir, Sylvie Thouësny, and Matthew James Whelpton**

**Publication date: 2022/12/12**



Typeset by Research-publishing.net
Cover photo by © 2022 Kristinn Ingvarsson (photo is taken inside Veröld – House of Vigdís)
Cover layout by © 2022 Raphaël Savina (raphael@savina.net)

ISBN13: 978-2-38372-015-7 (PDF, colour)

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

**Legal deposit, France**: Bibliothèque Nationale de France - Dépôt légal: décembre 2022.