European Association for
Computer-Assisted Language Learning

THE EUROCALL
REVIEW

The EuroCALL Review
Volume 31, No. 1, 2024, 16-23
ISSN: 1695-2618
http//doi.org/10.4995/eurocall.2024.20221

RESEARCH PAPER

Quantifying the impact of ASR-based instruction: What does the *iSpraak* platform learner data show?

Dan Nickolai
Saint Louis University, USA

dan.nickolai@slu.edu

How to cite this article:

Nickolai, D. (2024). Quantifying the impact of ASR-based instruction: What does the *iSpraak* platform learner data show? *The EuroCALL Review, 31(1),* 16-23. https://doi.org/10.4995/eurocall.2024.20221

Abstract

Computer-assisted Pronunciation Training (CAPT) tools have become increasingly dependent on Automatic Speech Recognition (ASR) technology to provide automated corrective pronunciation feedback to learners. The extent to which ASR-based tools measurably improve second language (L2) pronunciation is of great interest to language educators globally, and Computer-assisted Language Learning (CALL) researchers. Studies to date have largely been conducted by research practitioners with small-to-medium sized samples at single institutions. The findings and conclusions drawn from such small-scale data collection might be significantly bolstered by analysing the vast stores of learner data from large CAPT platforms. This study is informed by a sizable eight-year dataset from *iSpraak,* an open-source pronunciation tool designed to model and evaluate L2 speech. Quantitative analysis of anonymised learner interactions with this application reveals significant gains in intelligibility measures across multiple languages. Results also suggest that the extent of ASR's ability to improve learner pronunciation may be L2 dependent.

Keywords

Automatic Speech Recognition (ASR); Computer-Assisted Pronunciation Training (CAPT); Pronunciation; Corrective feedback

1. Introduction

Once a cutting-edge and experimental feature of Computer Assisted Language Learning (CALL) applications, Automatic Speech Recognition (ASR) is now integrated into virtually all modern language learning software. Commercial offerings from Babbel, DuoLingo, Mango Languages, and Rosetta Stone all tout their ability to provide immediate corrective feedback on learners' pronunciation of the target language (Babbel, 2023; DuoLingo, 2023; Mango, 2023; Rosetta Stone, 2023). Rosetta Stone, for example, champions their ASR tool named "TruAccent", which "leaves no syllable behind as learners progress" (Rosetta Stone, 2023). In the academic course material market, claims regarding pronunciation tools are not substantially different. Vista Higher Learning claims their ASR technology will "boost learner confidence" and "increase student awareness" of L2 pronunciation (Vista Higher Learning, 2023; Vista Higher Learning, 2024). The present ubiquity of ASR in CALL applications is a testament to the high level of interest in automating pronunciation evaluation and instruction, and its perceived potential. While insufficiently substantiated claims about being able to speak "with a near-native accent" (Pimsleur, 2023) still abound in marketing materials, the empirical evidence in support of ASR tools continues to accumulate year after year.

Empirical studies investigating the impact and suitability of ASR for language learning are almost exclusively carried out by research practitioners in K-16 educational settings. Regrettably, but understandably, the evaluation of technology interventions in such contexts frequently highlights institutional and practical constraints. Study populations are generally very small, and often limited to convenience sampling of researchers' own students. The interventions themselves are limited in number and duration, in order to avoid disrupting existing course curricula or risk compromising other planned instructional activities. Furthermore, the scope of any given study may be limited longitudinally due to the logistical or institutional challenges of collaborating across levels or curricula. Even when educators enthusiastically embrace emergent tools, formally conducting a study requires an additional investment of time and resources. Often it is simply not feasible for an instructor to coordinate pre- and post-tests, draft alternative assignments for a control group, or to otherwise optimally employ robust research methods. Meta-analyses and systematic reviews of existing ASR studies seek to counter the weaknesses of isolated findings (viz., Cengiz, 2023; Ngo et al., 2023; Shadiev and Lieu, 2023) but accessing larger alternative data sources on pronunciation learning may be even more edifying.

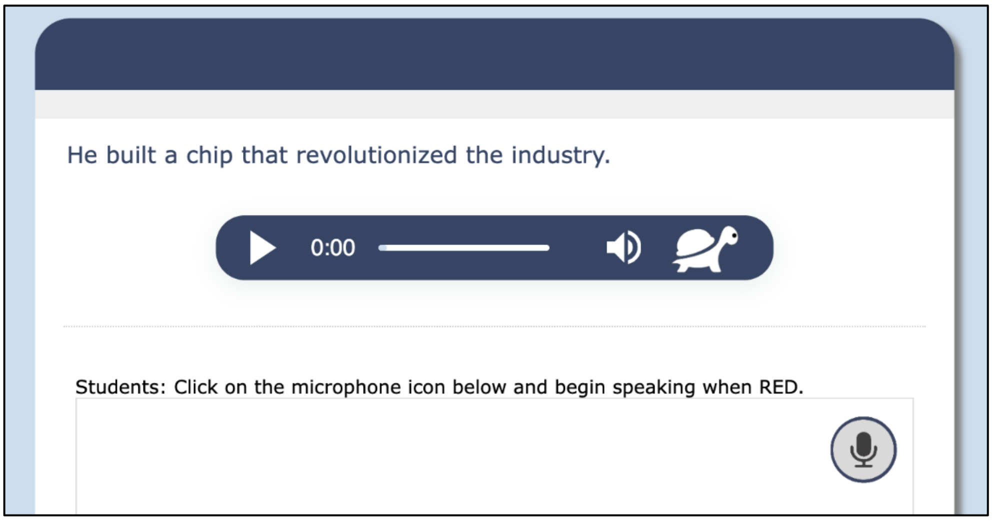
One underexplored and underutilised source of data comes from the publishers and purveyors of ASR tools. Vast quantities of learner data are systematically collected by these platforms across a large number of activities. These data are used for record keeping, tracking student progress, and calculating and storing grades. Analysing these data can provide valuable insight into the impact and efficacy of any given pronunciation exercise. There is also the potential to draw data-driven conclusions about automated feedback across many activities and languages. Of course, accessing these proprietary data invites its own challenges. Researchers must either create these tools (as is the present case) or establish a working relationship with a third-party entity and their institution that avoids conflicts of interest, protects personally identifiable information, and whose terms are mutually vetted and agreed upon. Regarding student records, these data need to be appropriately anonymised by the provider so that they don't violate of their own privacy policies. The anonymised data set informing the present study comes from the open-source pronunciation platform *iSpraak;* it is being analysed in strict accordance with the tool's end user license agreement.

2. Understanding data from *iSpraak*

The aggregate data collected by *iSpraak* over an eight-year period consists of millions of records across dozens of languages. This study focuses on a small subset of these data that represents learners who have made multiple attempts at a given activity. An example of such an activity can be seen in Figure 1, where a student can listen to and read the instructor-provided prompt: "*He built a chip that revolutionized the industry*." Once the learner completes an attempt, a score is immediately calculated based on the similarity of the transcription and model text (Figure 2). Students are subsequently invited to review the missed words and to make another submission. By measuring the number of attempts they make, and their subsequent improvement in scores, we can begin to quantitatively evaluate the impact of repeated interactions with the tool. The accuracy of this measurement does, of course, presuppose the overall accuracy of the ASR transcription. While this has historically been a contentious and dubious presupposition (Golonka et al., 2014), recent research supports the strong statistical correlation between machine and human transcriptions (Acosta & Ocasio, 2023).

Figure 1

Sample activity for iSpraak.

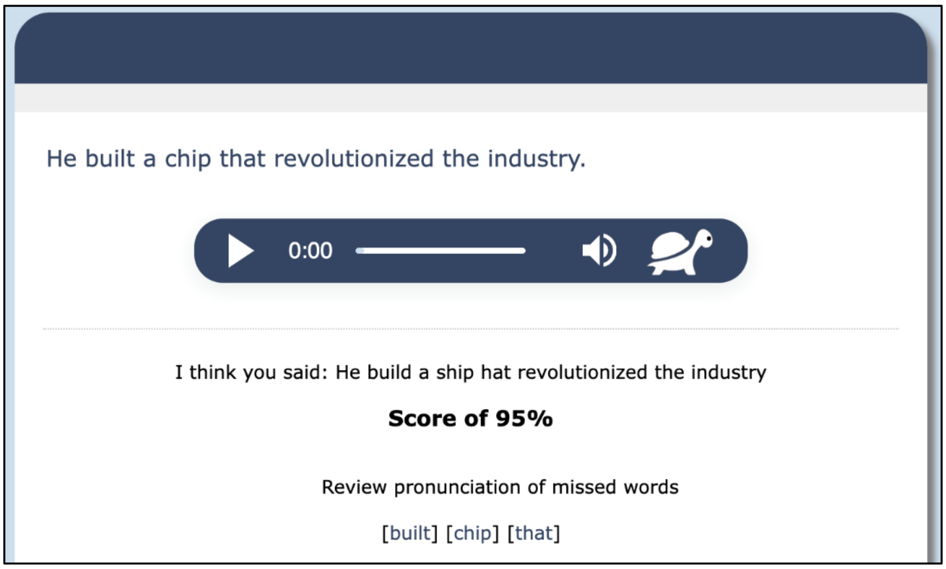


He built a chip that revolutionized the industry.

Students: Click on the microphone icon below and begin speaking when RED.

Figure 2

Sample activity with student scoring following ASR transcription.



He built a chip that revolutionized the industry.

I think you said: He build a ship hat revolutionized the industry

Score of 95%

Review pronunciation of missed words

[built] [chip] [that]

The *iSpraak* platform provides all instructors by default with some reporting and analytics to better visualise pronunciation progress for a group of learners. For any given activity, two reports are automatically generated. The first report provides a bar chart of the top five missed words and an error frequency count of all missed words, sorted alphabetically. The second report tracks student progress when multiple attempts of an activity are recorded. This report displays a chart of each student's first, best, and final score, along with a numerical count of attempts made. These score improvements are averaged across all students to provide the instructor with an average class improvement percentage for that particular activity. Regardless of the number of attempts, this calculation is based on the delta between the first attempt and the best attempt.

While individual instructors can independently review learning data from their own activities, the aggregate data required for the present study consist of tens of thousands of anonymised records across thousands of learners. All sensitive data were coded with irreversible student and instructor hashes before being exported for this analysis. In addition to these two anonymised variables, individual records showed a unique activity ID, a student score ranging from 0-100%, the number of missed words, and a sequential record ID (with the higher numbers representing more recently completed activities). Table 1 shows a subset of this information, which totalled 138,930 records. These data could also be sorted and queried according to the language of instruction (Arabic, Dutch, English, French, German, Italian, Japanese, Korean, Mandarin, Portuguese, Russian, and Spanish).

Table 1

Sample of anonymised iSpraak *data.*

Anonymised Student ID	Anonymised Instructor ID	Activity ID	Score	Missed Words	Language	Record ID
0015	0003	13450	85%	4	Spanish	19507
0015	0003	13450	92%	2	Spanish	19508
0016	0004	13901	65%	11	English	19509
0016	0004	13901	75%	7	English	19510
0017	0005	12651	99%	1	Japanese	19511

3. Research questions

Given the variables made available for analysis, our research questions are necessarily limited to those exploring score improvement, the number of attempts, and the language of instruction. For the purposes of this analysis, we are also only looking at the improvement of a learner on a given activity, not their improvement over time across multiple activities. While such longitudinal questions might be answerable from the dataset, an analysis of this type would require an objective measure of difficulty for each targeted activity. Without this measure, we risk comparing apples with oranges and could easily misinterpret pronunciation gains or losses made over time; it could very well be that the most recently assigned activities are also the most difficult. All things considered, there are two research questions that the available *iSpraak* data can help us answer:

> **RQ1:** To what degree does L2 pronunciation improve on an ASR activity following repeated interaction?

RQ2: Are there cross-linguistic differences in L2 pronunciation improvement through ASR activities?

Having worked with ASR-based CAPT tools for some time, we have an expectation that a learner's repeated interactions with a given activity will result in some measurable improvement of their score. Expecting an error rate to decrease as a consequence of practice is the core tenet of Skill Acquisition Theory (DeKeyser, 2020). Furthermore, decades of empirical research have shown that practicing pronunciation almost always leads to its improvement (Levis, 2022). We also hypothesise that score gains will vary across languages for two reasons. The first explanation for this variation is that not all acoustic models powering ASR engines are created equally. Google's advertised transcription word error rates (WER) vary according to language, even for native speakers (Google, 2023). Secondly, some languages, such as French, invite special transcription challenges. Examples abound of homophonically ambiguous phrases, such as "ils marchent / il marche" (they walk / he walks) or "elle mange / elles mangent" (she eats / they eat). Instructors do not necessarily consider the ambiguity of oral speech when designing activities that rely on machine transcription.

4. Methodology

An initial analysis of the data indicated that 120,869 of the 138,930 records were not usable for our posited research questions. This is because only 13% of records showed repeated attempts by the same student at the same activity. As we are trying to measure improvement, we need at least two attempts per pronunciation exercise in order to calculate a delta in the scores. This reduced our usable dataset to 18,061 records. Using the programming language Python, we scripted a simple iterative function to produce a new table from the anonymised records. This table collapses all student activity on an individual assignment to the following variables: first score, last score, best score, and number of attempts. A sample from this new table is below (Table 2).

Table 2

Sample of iSpraak *score improvement data.*

Anonymised Student ID	Anonymised Instructor ID	Activity ID	First Score	Last Score	Best Score	Attempts	Improvement
0015	0003	13450	85%	97%	97%	3	12%
0014	0003	13450	91%	94%	94%	5	3%

To calculate improvement for a given exercise, the first score is subtracted from the best score. Table 2 shows one student (0015) improving by 12% and another student (0014) improving by only 3% at the same activity (13450). It should be noted we are using "best score" and not the "last score". While these scores generally mirror each other (in 71% of cases), many instances of the most recent (last) score revealed targeted efforts focused on single words, rather than the entire phrase. This resulted in extremely low final scores in numerous cases. We speculate that students would attempt to correctly pronounce only the words they missed, rather than re-reading the entire phrase in the L2. This issue may be best addressed by modifications to the application's interface or clearer directives from the instructor.

Another consideration was to focus only on languages that had the highest number of records. Of the twelve languages in the dataset, only five languages comprised more than 1000 records. This reduced our analysis to English (n=4,335), French (n=1,029), Japanese (n=3,901), Korean (n=1,137), and Spanish (n=6,971). This final filtering of our data reduced the final number of records by 688 to arrive at 17,373 usable data points.

5. Results

As hypothesised, the results indicate that repeated engagement on individual activities does improve the learner score. We also confirmed that the size of the improvement varies by language. These average gains range from 11.78% (Korean) to 13.09% (English). Table 3 provides a breakdown of improvement by language. The average number of attempts by learners was remarkably uniform across the target languages, ranging from 3.31 (Korean) to 3.61 (French). The consistency of attempts by language suggests that the variation of improvement was not due to disproportionate effort by learners across languages. Some other factor is likely at play (such as Google's WER across acoustic models) regarding differences in score improvement for different L2s. When the five languages are taken together, there was a mean improvement of 12.67% (SD=16.94%) with an average 3.44 (SD=3.93) attempts per activity.

Table 3

Results across languages ordered by average improvement percentage.

Language	Improvement	Attempts
English	M=13.09%, SD=17.00%	M=3.50, SD=4.96
Japanese	M=12.78%, SD=16.56%	M=3.47, SD=3.18
French	M=12.72%, SD=16.63%	M=3.61, SD=4.48
Spanish	M=12.49%, SD=17.71%	M=3.39, SD=3.89
Korean	M=11.78%, SD=13.48%	M=3.31, SD=4.68

6. Discussion

The present study shows measurable learner improvement of L2 pronunciation following repeated interaction with an ASR-based pronunciation platform. These findings lend credence to the hypothesis that speech recognition technologies support pronunciation instruction. Analysis of the data shows that these gains vary slightly according to language. Unfortunately, the scope of the available data sheds no additional light on the nature of the gains made by those using the platform. For example, we cannot determine which aspects of L2 speech are improving and which are not. Previous research suggests that ASR may not be effective for all phonemes (Bashori et al., 2022; Chen et al., 2020; Garcia et al., 2020; Guskaroska, 2020; Inceoglu et al., 2020). Another unanswered question of interest is whether some pronunciation improvement might be attributed to learner familiarity with the tool. Some remarkable improvements in the data (10% first score to 100% best score, for example) could easily be attributed to inadvertent or premature first submissions by students learning to navigate the platform. These outlier data, while limited, provide even more reason to restrict our analysis to the languages with the most abundant records.

Another important limitation is that we cannot say with any certainty what precisely led to the improvement in scores. In some cases (French), it may have been the text-to-speech modelling, while in others (Japanese) it may have been the transliteration support offered by the tool. *iSpraak* also directs learners to Forvo.com to review human recordings of mispronounced words. None of these interactions with the platform's features are measured nor do we currently have survey data asking learners how they used the tool to improve their pronunciation. Better metrics could paint a more comprehensive picture of what exactly is driving the improvement in scores. While the nature of the analysed data alone does not provide much explanatory power, the present study does serve to quantitatively support the notion that ASR can have a measurable positive effect on pronunciation instruction.

7. Conclusion

The present study demonstrates how anonymised learner data from large ASR tools can provide important learning metrics and insight for pronunciation researchers and educators. Conventional practitioner research often only yields small datasets due to limited populations and other institutional or practical constraints. While the challenges of accessing proprietary datasets can be substantial, being able to capitalise on these data is a boon to understanding the potential of emergent technologies such as ASR. The available data from *iSpraak* show that learners make significant improvements following repeated engagement with the platform. These data also indicate that improvement varies slightly according to language. As CALL researchers continue to explore the perils and promises of new applications, they would be well-served by pursuing novel sources of learner data as a component of their investigations.

Acknowledgements

The author would like to thank Emma Schaefer, Paula Figueroa, and Vijay Varshini Lakshmi Narayanan for their support in preparing this article. The National Endowment for the Humanities has generously funded the ongoing development of *iSpraak* and the research efforts behind this work.

References

Acosta, K., & Ocasio, M. (2023). Transparent Language: Learners' perceptions, successes, and challenges of using a speech recognition tool for molding beginner Spanish pronunciation in online courses. In *Technological Resources for Second Language Pronunciation Learning and Teaching* (pp. 127-146). Lexington Books.

Babbel (2023). Retrieved from https://support.babbel.com.

Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2024). 'Look, I can speak correctly': learning vocabulary and pronunciation through websites equipped with automatic speech recognition technology. *Computer Assisted Language Learning*, *37(*5-6), 1335-1363.

Cengiz, B. C. (2023). Computer-assisted pronunciation teaching: An analysis of empirical research. *Participatory Educational Research*, *10(*3), 72-88.

Chen, W. H., & Lim, H. (2020). Using ASR to improve Taiwanese EFL learners' pronunciation: Learning outcomes and learners' perceptions. In O. Kang, S. Staples, K. Yaw, & K. Hirschi (Eds.), *Proceedings of the 11th Pronunciation in Second Language Learning and Teaching conference*, Northern Arizona University, September 2019 (pp. 37-48). Ames, IA: Iowa State University.

DeKeyser, R. (2020). Skill acquisition theory. In *Theories in second language acquisition*. Routledge.

DuoLingo (2023). Retrieved from https://en.duolingo.com/efficacy.

Garcia, C., Nickolai, D., & Jones, L. (2020). Traditional versus ASR-based pronunciation instruction: An empirical study. *CALICO Journal*, *37(*3), 213-232.

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, *27(*1), 70-105.

Google (2023). Speech-to-text https://cloud.google.com/speech-to-text.

Guskaroska, A. (2020). ASR-dictation on smartphones for vowel pronunciation practice. *Journal of Contemporary Philology*, *3(*2), 45-61.

Inceoglu, S., Lim, H., & Chen, W. H. (2020). ASR for EFL pronunciation practice: Segmental development and learners' beliefs. *Journal of Asia TEFL*, *17(*3), 824.

iSpraak (2023). https://www.iSpraak.net/about.html.

Levis, J. M. (2022). 2. Teaching pronunciation: Truths and lies. In *Language and Literature in Education 2*, pp. 39-72.

Mango Languages (2023). Retrieved from https://mangolanguages.com/how-it-works

Ngo, T. T. N., Chen, H. H. J., & Lai, K. K. W. (2024). The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis. *ReCALL*, *36(*1), 4-21.

Pimsleur (2023). https://offers.pimsleur.com.

Rosetta Stone (2023). https://www.rosettastone.com/enterprise/resources/content/rosetta-stone-truaccent-snapshot. Accessed 5/1/23.

Shadiev, R., & Liu, J. (2023). Review of research on applications of speech recognition technology to assist language learning. *ReCALL*, *35(*1), 74-88.

Vista Higher Learning (2023). https://learn.vistahigherlearning.com/espaces/whats-new.html. Accessed 5/1/23.

Vista Higher Learning (2024). https://vhlblog.vistahigherlearning.com/speech-recognition-a-game-changer-in-language-education.html. Accessed 2/12/24.

Ethical statement

The author of this article is the Principal Investigator on a Digital Humanities Advancement Grant from the National Endowment for the Humanities. This grant funds the research and ongoing development of the platform *iSpraak*. No commercial interest exists. This investigation was carried out in the United States in accordance with professional and institutional standards of research. All user data were irreversibly anonymised before conducting quantitative analysis on learner records. All user data were processed in strict accordance with *iSpraak's* end user license agreement.