Automatic speech recognition
can you understand me?

Susana Pérez Castillejo[1]

Potential impact	medium
Timescale	medium term
Keywords	automatic speech recognition, pronunciation training, speech to print, feedback

What is it?

Automatic Speech Recognition (ASR) is a digital communication method that transforms spoken discourse into written text. This rapidly evolving technology is used in email, text messaging, or live video captioning. Current ASR systems operate in conjunction with Natural Language Processing (NLP) technology to transform speech into text that people – and machines – can read. NLP refers to the methodologies and computational tools that analyze data produced in a natural language, such as English.

When users talk into an ASR-enabled application, the speech signal turns into an audio file that is first filtered for background noise and then parsed into phonemes, which are the smallest sound units in a language: the word 'push', for example, has three phonemes ('p', 'u', and 'sh'). Through statistical probability, the ASR system analyzes the phoneme sequences it 'recognizes' and deduces the words that best match those sound strings. The auto-generated text can then be 'read' by a machine to perform some other tasks.

1. University of St. Thomas, Saint Paul, Minnesota, United States; pere9775@stthomas.edu; https://orcid.org/0000-0001-7543-4506

How to cite: Pérez Castillejo, S. (2021). Automatic speech recognition: can you understand me? In T. Beaven & F. Rosell-Aguilar (Eds), *Innovative language pedagogy report* (pp. 121-126). Research-publishing.net. https://doi.org/10.14705/rpnet.2021.50.1246

Self-study is the most frequent pedagogical approach taken when integrating ASR into language education, as it usually mediates learner-device interactions instead of learner-learner exchanges.

ASR is effectively used for pronunciation training (Pennington & Rogerson-Revell, 2019), but more recent uses (Istrate, 2019; Liakin, Cardoso, & Liakina, 2015; Nickolai, 2015) show that ASR can also promote oral skills beyond pronunciation.

Examples

iSpraak.com (Nickolai, 2015), a cloud-based ASR tool, 'listens' to how a student pronounces a text provided by the teacher and returns a similarity score based on native speech patterns. The auto-scoring feature encourages independent study: learners keep practicing until they reach a certain score, but the teacher does not need to listen to every file produced.

Auto-generated transcripts from speech-to-text engines such as *Microsoft Stream* can also support independent language development (Liakin et al., 2015). As learners compare what the tool 'understood' to what they were trying to say, they improve their performance. Some of these tools pair ASR with automated translation, which can further help learners self-assess their accuracy.

An emerging ASR application is the use of Virtual Assistants (VA) such as *Alexa* or *Siri* (Istrate, 2019; see also Underwood, this volume). The communicative functions that VAs motivate include uttering commands ("Alexa, play some music!") or asking factual questions ("Siri, what is the weather like in Tokyo today?"). Successfully getting a VA to perform the desired action or to provide the needed information requires not only pronunciation accuracy, but also some knowledge of L2 vocabulary and sentence structure: the learners are not reading or repeating model sentences. If the task involves asking questions and using the information obtained, listening comprehension is an additional skill practiced.

Benefits

Using ASR for pronunciation training may encourage learner autonomy: the immediate feedback provided by the software, in the form of a transcript or an accuracy score, makes learners more aware of their progress, and the ability to carry out the exercises without the teacher gives them more control over their practice.

Speaking tasks with VAs also increase speaking opportunities beyond the classroom. VAs are not suitable for conversational practice, yet, but producing the short action-oriented or information-seeking utterances typical in these tasks is still a good proficiency-building exercise that can prepare learners for more involved oral discourses. In fact, frequent use of VAs for independent practice has been linked to significant improvements in L2 speaking proficiency (Dizon, 2020).

Potential issues

An important issue in ASR's pedagogical application is data privacy. As with other web-based interactions, exchanges with VAs produce personal data that could be commercially exploited. Thus, it is important for educators to be mindful of the data privacy policies for the technologies they use.

A second concern is robustness. ASR accuracy depends much on the acoustic conditions (performance suffers in noisy environments) and, most importantly for language educators, the speaker's experience with the language. Users often complain that the ASR tool 'detected the wrong thing', even though they know they were saying it right.

Although 'comprehension' of accented speech keeps improving, ASR performance is still not ideal when transcribing speech produced by low-proficiency learners. This issue may be resolved as more data from this type of learner becomes available. ASR accuracy with non-native speech has improved due to increased computing power and data availability from commercial sources

(telephone-based transactions, for example). These sources of data, however, do not include low-proficiency speakers: who dares to complete a phone transaction in a language they are not fluent in?

EdTech companies offering data-based learning solutions hold the key to improve ASR's robustness: tools such as *Extempore* are using a wide range of non-native deidentified speech data in their servers for research and development (Figure 1).

Auto-generated transcripts that are still highly accurate with novice learners will be a welcome grading aid for teachers. Reading is faster than listening, particularly if the audio file is plagued with the long pauses typical in low-proficiency speech. While auto-generated fluency scores can indicate progress on the temporal aspects of speech (frequency and mean duration of pauses, percentage of speaking time), transcripts can help teachers provide feedback on lexical and syntactic accuracy faster.

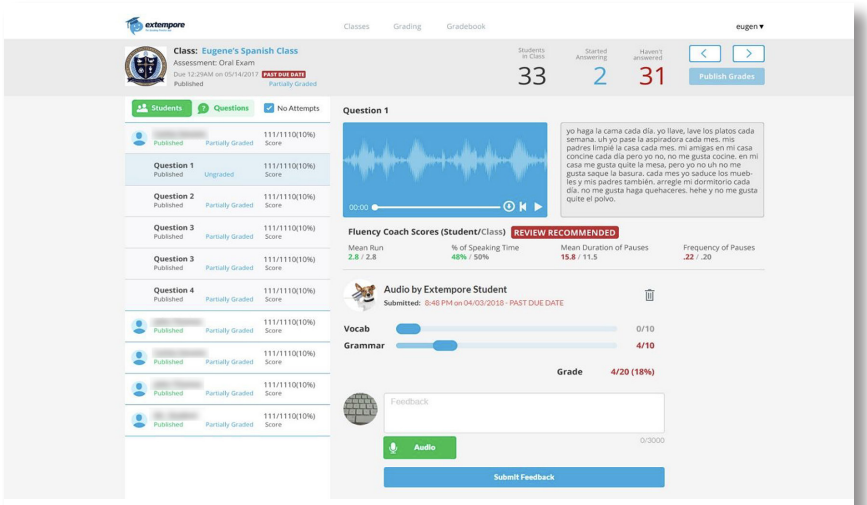

Figure 1. Prototype for Extempore's ASR-enhanced features. Metadata provided by ASR can assist language instructors when grading oral tasks

Looking to the future

The pedagogical examples described above show that ASR technology can have an important impact in language teaching and learning: automated comparison with native speech patterns encourages pronunciation accuracy, self-access speaking tasks promote learner autonomy, and independent oral practice with VAs builds proficiency.

There is a need for increased speaking practice outside the classroom targeting skills beyond pronunciation.

Through robust ASR-enabled applications, this supplemental oral practice can be completed without necessarily turning into additional grading for the teacher. Thus, as ASR with low-proficiency speakers becomes more reliable, this technology will be more widely adopted for independent and classroom-based language learning.

References

Dizon, G. (2020). Evaluating intelligent personal assistants for L2 listening and speaking development. *Language Learning & Technology, 24*(1), 16-26. https://doi.org/10125/44705

Istrate, A. M. (2019). The impact of the virtual assistant (VA) on language classes. In *Proceedings of the 15th International Scientific Conference eLearning and Software for Education* (pp. 296-301). Carol I National Defense University.

Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *Calico, 32*(1), 1-25. https://doi.org/10.1558/cj.v32i1.25962

Nickolai, D. (2015, October 30). iSpraak: automated online pronunciation feedback. *The FLTMAG*. http://www.fltmg.com

Pennington, M. C., & Rogerson-Revell, P. (2019). Using technology for pronunciation teaching, learning, and assessment. In *English Pronunciation Teaching and Research* (pp. 235-286). Palgrave Macmillan. https://doi.org/10.1057/978-1-137-47677-7_5

Underwood, J. (2021). Speaking to machines: motivating speaking through oral interaction with intelligent assistants. In T. Beaven & F. Rosell-Aguilar (Eds), *Innovative language pedagogy report* (pp. 127-132). Research-publishing.net. https://doi.org/10.14705/rpnet.2021.50.1247

Resource

For some advice on which ASR apps to try out, see: https://www.techradar.com/news/best-speech-to-text-app

Research-publishing.net

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

Innovative language pedagogy report
Edited by Tita Beaven and Fernando Rosell-Aguilar

Publication date: 2021/03/22

Typeset by Research-publishing.net
Cover layout by © 2021 Raphaël Savina (raphael@savina.net)
Photo by Digital Buggu from Pexels (CC0)

ISBN13: 978-2-490057-86-3 (Ebook, PDF, colour)
ISBN13: 978-2-490057-87-0 (Ebook, EPUB, colour)
ISBN13: 978-2-490057-85-6 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

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: mars 2021.