Computer Assisted Language Instruction Consortium

CALICO JOURNAL (ONLINE) ISSN 2056–9017

Article

Traditional Versus ASR-Based Pronunciation Instruction: An Empirical Study

Christina García[1], Dan Nickolai[2], and Lillian Jones[3]

Abstract

This paper presents a 15-week classroom study measuring the student outcomes of instructor-led pronunciation lessons versus entirely ASR-based pronunciation training. Seventy-six second-semester Spanish language learners were divided into two groups, one experimental (n=44) and one control (n=32). Over the course of six modules, both groups completed a pre- and post-study recording, as well as explicit pronunciation training sessions. These sessions included pre- and post-recordings, with either traditional or ASR pronunciation practice in between, which aimed attention at targeted phonemes. All student recordings were evaluated by native and near-natives for comprehensibility, nativeness, fluency, and perceived confidence. The results show that the effect of explicit and ASR instruction varies depending on the module and characteristic evaluated. ASR seems to outperform traditional instruction when targeting specific phonemes, especially in the short-term, while the explicit instruction group saw longer-term gains in regards to comprehensibility. Holistically, the data suggest that ASR-based instruction shows promise to improve certain aspects of pronunciation, but that using both techniques in tandem would be the most strategic approach to handling the development of this fundamental aspect of learner speech. The data presented here highlight the role and effectiveness of computer-assisted pronunciation training for lower-level Spanish courses.

KEYWORDS:ASR; CAPT; TTS; PRONUNCIATION.

Affiliations

[1]Saint Louis University, USA.
email: christina.garcia@slu.edu

[2]Saint Louis University, USA.
email: dan.nickolai@slu.edu

[3]University of California, Davis, USA.
email: liljones@ucdavis.edu

CALICO JOURNAL VOL 37.3 2020 213–232
https://doi.org/10.1558/cj.40379

equinox
www.equinoxpub.com
equinoxonline

1. Introduction

Educational technology platforms, textbook publishers, and commercial entities are increasingly leveraging real-time Automatic Speech Recognition (ASR) to provide instantaneous corrective feedback to language learners. As these tools grow in prominence and popularity, language educators may well wonder whether ASR-based software should supplement or supplant traditional methods of explicit pronunciation practice and instruction. A survey of prior research on Computer-Assisted Language Learning (CALL) tools points to the promise of ASR being especially impactful with regards to improving target language pronunciation (Golonka, Bowles, Frank, Richardson, & Freynik, 2014). The present study seeks to explore these claims by examining the pronunciation gains following different instructional techniques in the context of a second-semester college-level Spanish course. Results from a 15-week period indicated that there is indeed promise for ASR-based instruction to improve certain aspects of learner speech. Furthermore, data analysis revealed that learner pronunciation gains were phoneme-specific for the ASR treatment, but individual measures of pronunciation were subject to gains for both the ASR and traditional instruction groups. The research findings point strongly to the benefit of using traditional methods of pronunciation instruction in tandem with emerging ASR-based tools.

2. Background

2.1 Second Language Pronunciation

Despite its importance in facilitating communication, the explicit instruction of pronunciation had momentarily fallen out of favor with the rise of more communicative models of language teaching (Derwing & Munro, 2005). Careful and deliberate pronunciation instruction may have been too strongly associated with behaviorist pedagogies of yore to sustain a privileged place in the modern language curriculum. Whatever the reason behind its decline, teachers and researchers alike have recently pushed back against the neglect of this fundamental competence in language instruction (Thomson & Derwing, 2014). This pushback is perhaps unsurprising, given that any pronunciation that sufficiently deviates from native-like speech will ultimately become incomprehensible, or understood with only a great deal of effort by a sympathetic interlocutor. For this reason, it cannot be assumed that students will improve their pronunciation through exposure alone. Research has demonstrated that the implicit instruction of pronunciation by itself does not yield the same results as targeted training (Grant & Brinton, 2014). Recent years have seen a renewed interest in pronunciation studies amongst CALL researchers, and

this activity has fortuitously coincided with major advances in speech technologies. Because of its communicative importance, pronunciation training cannot remain neglected, and one potential path forward is to embrace new technologies if they prove to be effective.

2.2 ASR

One must be extremely cautious when drawing conclusions about the impact of ASR technology and its historical timeline when reviewing CALL research to date. Researchers have been investigating ASR performance for over three decades, but these technologies have improved exponentially over time. References that predate the rise of neural networks circa 2012 refer to long abandoned underlying ASR technologies such as hidden Markov or Gaussian mixture models (Pieraccini, 2012; Beaufays, 2015). When investigating the accuracy, validity, or promise of ASR systems, it is thus essential to understand that modern language modeling now relies on fundamentally different and more sophisticated systems. One year before the watershed introduction of deep and recurrent neural networks, Thomson (2011) expressed some misgivings about the potential of ASR software for pronunciation instruction. These concerns included frequently erroneous feedback to learners and a significant mismatch between how a machine and human speaker might rate intelligibility. While CALL researchers continue to echo these concerns (Lord, 2019), it is worth carefully paying attention to the publication dates of these studies. More recent studies suggest that ASR-based pronunciation training software actually outperforms traditional classroom instruction (Liakin, Cardoso, & Liakina, 2013; Elimat & AbuSeileek, 2014).

Two distinct approaches tend to emerge in CALL research when investigating the impact of ASR tools for pronunciation instruction. The first involves endeavoring to measure overall pronunciation gains, as a matter of comprehensibility for any given utterance. This can happen at the word, sentence, or dialogue level. The second approach is to target a specific phoneme that may prove to be especially problematic for a group of learners. Elimat and AbuSeileek (2014) focused on the first approach to compare regular (instructor-led) instruction to an ASR-based tool and reported a significant difference in treatments in favor of the ASR group. Liakin et al. (2013) focused their comparative analysis on the second approach by exploring the French phoneme /y/ and similarly found statistically significant gains for the ASR group, but not for a control group. Both sets of researchers conclude their studies by advocating strongly for ASR tools in pronunciation instruction. It is also worth observing that both of these studies found high levels of student appreciation for the

ASR-based approach to pronunciation instruction (Liakin et al., 2013; Elimat & AbuSeileek, 2014).

2.3 CAPT

One of the great promises of CALL is the ability to provide individualized, automated, and impactful feedback to language learners. With regards to pronunciation instruction and evaluation, we find numerous studies exploring computer-assisted pronunciation training (CAPT) software (Lord, 2019). It should be noted here that the utility of a CAPT application should be measured by the quality of the feedback it provides the learner. This feedback tends to be visual in nature, and has existed in some fashion or another for over four decades (Lord, 2019). Of course, not all visual feedback on pronunciation is helpful to the student. In the earliest days, it was common to display a simple waveform to represent both learner and model speech utterances. However, beyond duration and amplitude, there is no meaningful or actionable information represented in a waveform. Furthermore, Neri, Cucchiarini, and Strik (2002a) suggest that providing visual waveforms to learners is merely a gimmick, and done only to suggest (an absent) sophistication of the tool. In an effort to provide more meaningful visual representations of speech, some CAPT applications instead elect to display information-rich spectrograms. While this could be considered a laudable advance, Thomson (2011) reminds us that "spectrograms are uninterpretable to non-experts", and thus require extensive training in order to be useful for L2 learners; although see Olson (2014), for example, for a study showing gains with spectrographic feedback. Given the potential drawbacks of waveforms and spectrograms as visual pronunciation aids, it is of little surprise that textual feedback via ASR has long been advocated as the preferred technology underlying modern CAPT tools (Neri, Cucchiarini, Strik, & Boves, 2002b).

Neri, Cucchiarini, and Strik (2003) proposed a pedagogically-ideal design model for CAPT software just as ASR began to grab the wider attention of CALL researchers. The authors note that the most essential feature of any such program be a robust ASR engine that can reliably transcribe non-native speech. Liakin, Cardoso, and Liakina (2015) remind us that the earliest ASR systems were speaker-dependent, needed user training, and did not handle accented speech well. Clearly, the sine qua non of an ASR-based CAPT application lies in it being sufficiently flexible to handle learner speech. Transforming an incoming speech signal from language learners and reliably transcribing it is thus of paramount importance. As noted earlier, a shift in how ASR acoustic models are constructed with deep neural networks has helped to address this key feature. The second design concern advocated by Neri et al. (2003)

is that the CAPT tool evaluate the overall quality of any given utterance as a global numerical score. This score must be sufficiently granular to measure any improvement in successive attempts at repeating the same speech exercise. Further design considerations for an ideal CAPT system include the detection and diagnosis of errors at the phonemic or word level. A learner should understand where in an utterance the pronunciation deviated from the model, and what the nature of the error was. The final design suggestion from Neri et al. (2003) is providing automated corrective feedback that combines the transcription, global score, and the specific location and nature of errors. This feedback needs to be sufficiently rich and actionable to guide subsequent learner efforts.

2.4 iSpraak

The web application iSpraak is an example of an ASR-based CAPT tool that aligns well with the previously mentioned vision advocated by Neri et al. (2003). Instructors create activities by providing the platform with a short text in the L2 and an optional MP3 file to serve as the audio model. Alternatively, and as is the case in this study, the application can generate text-to-speech (TTS) audio files if a model recording does not yet exist. Both the instructor-provided text and the audio files are available to the learner to review while interacting with the platform. The synthesized audio support consists of three playback options: a male voice, a female voice, and a reduced speed (75%) female voice, as seen in Figure 1. While TTS voices may be criticized for lacking some suprasegmental features of natural speech, their pedagogical value in pronunciation exercises

Figure 1. Screenshot of iSpraak interface.

has been demonstrated in recent research (Liakin, Cardoso, & Liakina, 2017). In iSpraak, learners are instructed to listen to the model audio file(s) and then are prompted to click a microphone icon to begin the speech exercise. As the student speaks, the ASR engine displays a real-time transcription of any recognized L2 text. When the activity is submitted by the student, a numerical score (0–100%) is presented; any mispronounced words are listed in isolation for further review. These review words are also directly linked to www.Forvo.com, a crowdsourced pronunciation database of selected words and phrases.

As Lord (2019) has noted, the advantage of using a tool like iSpraak is that no phonological expertise is required by either the learner or instructor. The feedback provided by the tool, though at times imperfect, is eminently interpretable as text and audio. In addition to giving a numerical score and identifying words for review, the application also provides a final transcript preceded with the explanation: "*I think you said*: ". Providing this transcript can serve to draw learners' attention to pronunciation mistakes made at the word or phoneme level. A student trying to say *perro* "dog" but presented with *pero* "but" in the transcript may appreciate this more granular and actionable level of feedback. Whereas some CAPT tools limit feedback to a binary correct/incorrect notification, the advantage of ASR engines is that a transcription of a non-targeted word can serve to provide additional pronunciation information to the learner. Furthermore, learners may find a live transcription of their L2 speech more validating than a simple "correct" or "thumbs up" type of alert. It is also important to highlight that the iSpraak score is calculated by the similarity between the transcribed text and the model text. In the above example of *pero* versus *perro,* the platform would indicate a score of 80%, in which four out of five letters were present in the transcription.

2.5 Motivation for Present Study

An early pilot study of iSpraak was carried out by the second author for a third-semester French course. While the findings were positive in terms of pronunciation improvement, there was no control group and the class size was not sufficiently large (n=8) to draw any strong conclusions. It was subsequently determined that a more robust and carefully designed study would be required to confidently integrate and rely on the tool for future pronunciation instruction. Indeed, it was this observation that led to the genesis of the present study. The research team assembled to determine, empirically, the impact of using iSpraak as an ASR-based CAPT tool in a lower-level Spanish course. This student population was chosen primarily due to large enrollment numbers and the intuition that potential pronunciation gains would be more significant earlier in an L2 curriculum. The research team was especially motivated

to see how this emergent technology compares to traditional (instructor-led) pronunciation instruction.

When preparing a study that evaluates the impact of ASR, it is instructive to highlight prior calls for research design in this domain. Thomson and Derwing (2014) have argued that pronunciation studies should allow for replication, be comprised of sufficiently large samples, and contrast an experimental group with a control group. More recently, O'Brien et al. (2018) echoed these criteria, and further called for more variety (beyond English) in the languages being used. With these suggestions serving as a framework, our research team set out to conduct a study on the impact of ASR-based pronunciation training as it compares to traditional classroom instruction of Spanish phonology. The fundamental question being explored was whether or not the ASR treatment would result in greater or fewer gains, and what types of gains, in a set of commonly identified pronunciation metrics.

3. Methodology

3.1 Student Modules and Recordings

For the purposes of this study, traditional instruction should be understood as an instructor-led session that explicitly draws learners' attention to specific phonemes through targeted examples, minimal pairs, and choral repetition. In contrast, ASR-based pronunciation training consists of student interaction with iSpraak. As mentioned in section 2.4, this tool provides real-time automated transcriptions of student speech, while highlighting mispronounced words for subsequent review.

The study was carried out over 15 weeks in the context of a second-semester Spanish course at a mid-sized, private university in the Midwest. There were 76 student participants divided by lab section into two groups, one experimental (n=44) and one control (n=32). The only information available for the participants is their gender and class rank, which can serve as a rough proxy for their relative age given that non-traditional students are not common at this university. As seen in Table 1, there are almost twice as many female as male participants, which is typical of language courses at this institution. Their class

Table 1
Student Participant Demographics

	Participants' Information
Male:female, *n*	28:48
Class rank, *n*	Freshman: 21, sophomore: 22, junior: 24, senior: 9

rank varies quite a bit, with almost equal numbers of freshmen, sophomores, and juniors, and fewer seniors. All students are L1 English speakers.

To minimize the perceived interruption of the study, each pronunciation session featured topic-relevant material for the existing curriculum. Additionally, researchers only intervened for a 20-minute period every other week during a regularly scheduled lab hour, which resulted in six pronunciation modules over the course of the semester. During the first and last modules, the students recorded the same short text ("*Dialectos*"), which can be found in the Appendix. In modules 2–5, both the traditional instruction and ASR practice targeted some of the most problematic sounds for Spanish learners, including vowels, <g, j, h>, <y, ll>, and voiceless stops. During these modules, students first recorded two test sentences (see Appendix), then participated in traditional or ASR pronunciation practice, and finally re-recorded the same two sentences. For the purposes of this article, we only report on the beginning versus end semester comparison, and the vowel and <g, j, h> modules, which were chosen since they had the highest student participation. Including both beginning/end and targeted modules allows us to tease out short-term and long-term effects of both training types.

3.2 Evaluation of Recordings and Analysis of Ratings

After the exclusion of approximately 20 samples for incomplete or inaudible recordings, there were 578 usable student recordings, which were first anonymized and subsequently coded for the speaker, group, module number, and whether or not the recording was pre- or post-training. The research team then devised an evaluation instrument and invited a professional network of native and near-native speakers of Spanish (mostly fellow instructors and graduate students) to provide feedback. All instructions on the evaluation task were in Spanish and all but 21 of the 118 raters self-reported their Spanish level as native or near-native. Additionally, the vast majority of raters reported being moderately or very comfortable speaking with L2 Spanish speakers, indicating they do so every other day or every day. Evaluators were asked to assess speech samples on separate six-point Likert scales for comprehensibility, nativeness, fluency, and perceived confidence. Additionally, there was a 10-point scale to indicate what percentage of the sample was not comprehensible due to pronunciation errors. These measures were chosen as they are commonly used in studies that evaluate the reliability of CAPT, as well as second language pronunciation research more broadly (see, for example, Levis, 2007, and Lord, 2019 for a review of such studies).

The ratings data were ultimately represented in a ten-column spreadsheet with columns for anonymized student code, group number, module number,

pre or post variable, comprehensibility, nativeness, fluency, perceived confidence, percentage of words that impeded comprehension, and an anonymized rater ID number. A sample of the first few rows can be seen in Table 2.

Table 2
Sample of Ratings Data

Student	Group	Module #	Pre/post	C.	N.	F.	S.C.	%	Rater ID
47	A	16296	post	4	1	2	2	60	3
53	A	16286	pre	2	1	1	1	90	3
20	B	16296	post	4	2	2	3	40	3

In total, there were 1806 ratings from 118 unique raters, with each rater evaluating 1–20 student samples. As seen in Table 2, the evaluators' responses were transformed into scalar, numerical values, where 1 corresponds to the lower end of the scale (less native, fluent, etc.) and 6 corresponds to the higher end of the scale (more native, fluent etc.). Since it seemed possible that some of the five characteristics rated may be correlated, we first performed factor analysis using the *factanal* function in R (R Core Team, 2014) with *varimax* rotation. The results of the factor analysis suggested two joint factors: the first combining comprehensibility and percentage of words in which comprehension is impeded, and the second combining nativeness and fluency. Upon further examination, however, the results for nativeness and fluency as separate factors were more informative than a joint factor combining the two. Thus, separate linear mixed effects models were created for each of the three singular characteristics (nativeness, fluency, perceived confidence) and the one joint factor (henceforth *comprehensibility*) using the *lmer* function (Bates, Mächler, Bolker, & Walker, 2014). Here the dependent variable was the evaluators' ratings, while the independent variables included the time of recording (pre/post) and group (A/B). Individual speaker and evaluator were included as random effects and the alpha p-value was set at 0.05. Post-hoc analysis (with Tukey correction) was done to determine which individual comparisons were significant in each model.

4. Results

4.1 Overall Results

The analysis of the ratings data revealed statistically significant correlations for certain characteristics being measured, but not all, as evidenced by Table 3. The control group (A), which received traditional pronunciation instruction, saw

significant gains over the course of the semester vis-à-vis increased *comprehensibility*. The experimental ASR group (B) improved in terms of *fluency* over the course of the semester, and neither group improved in terms of *nativeness* or *perceived confidence* when beginning and end of semester recordings are compared. On a much shorter time period, the individual pronunciation modules also indicated some significant gains. The vowel module showed increased *fluency* and *comprehensibility* for the ASR group (B), and increased levels of *perceived confidence* for both student groups.

Table 3
Summary of Statistically Significant Correlations in Ratings Data

	Comprehensibility		**Nativeness**		**Fluency**		**Confidence**	
Time	**Control (A)**	**Exp. (B)**	**Control (A)**	**Exp. (B)**	**Control (A)**	**Exp. (B)**	**Control (A)**	**Exp. (B)**
Beginning to end of semester	↑	X	X	X	X	↑	X	X
Vowel module	X	↑	X	X	X	↑	↑	↑
<g, j, h> module	X	↑	X	↑	X	↑	↑	↑

↑ = significant gain observed
X = no statistical difference between pre/post

One of the more striking findings of the data was the impact of ASR when working with the <g, j, h> module. The ASR group (B) made significant advances in all of the characteristics being evaluated. This contrasts with the traditional pronunciation group (A), which only made advances in measures of *perceived confidence* for the same module. In the following sections, each of the five characteristics rated is examined in more detail.

4.2 Comprehensibility

Figure 2 shows the ratings for the joint comprehensibility factor, which combines ratings of comprehensibility and percentage of words in which comprehension is impeded. In this figure and the boxplots that follow the diamond points denote the means for each group examined, while the line through each box represents the median. The box itself shows the middle 50% of all ratings, and the lines that extend above and below the box represent the highest and lowest 25% of all ratings, respectively. On the left panel of Figure 2 are the

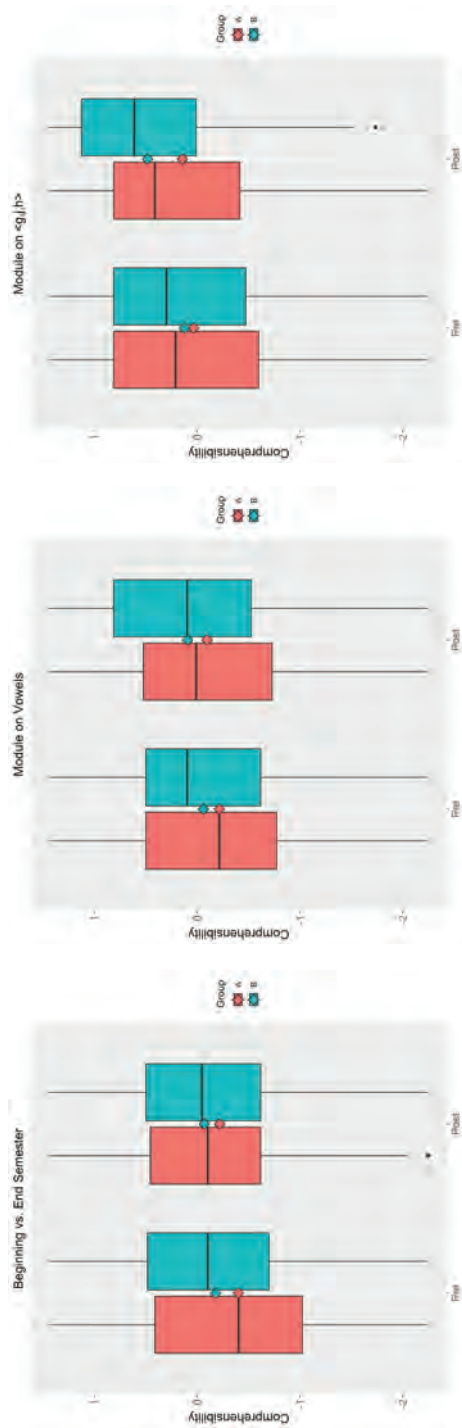

Figure 2. Boxplot of comprehensibility ratings by group comparing beginning and end of semester (left panel), vowel module (middle panel), and <g, j, h> module (right panel).

results for the beginning versus the end of semester, the middle panel shows the results for the vowel module, and the right panel displays the results for the <g, j, h> module. The scale in this figure goes from -2 to 1 instead of 1 to 6 as the two individual characteristics had to be scaled and centered before making the joint factor. In the beginning versus end of semester comparison, there is no significant difference for pre and post for the ASR group (B), but the control group (A) shows a significant increase in comprehensibility over the course of the semester ($p < 0.001$). For the vowel and <g, j, h> modules, on the other hand, the control group (A) does not exhibit any significant differences in comprehensibility, while the ASR group's (B) comprehensibility increases significantly in the vowel ($p < 0.05$) and <g, j, h> ($p < 0.001$) modules.

4.3 Nativeness

Figure 3 displays the nativeness ratings for the <g, j, h> module, where "Pre" corresponds to the recording immediately before the module and "Post" corresponds to the recording after the module. Overall the ratings for nativeness are highly skewed to the lower end of the scale (*poco nativo* "not very native"), which is to be expected given that the participants are second semester students. While there is no significant change in nativeness ratings for the control group (A), the ASR group (B) was rated as significantly more native-like in the post recording following the <g, j, h> module ($p < 0.01$). There are no significant differences in nativeness ratings for either group for the vowel module or the beginning versus end of semester recordings.

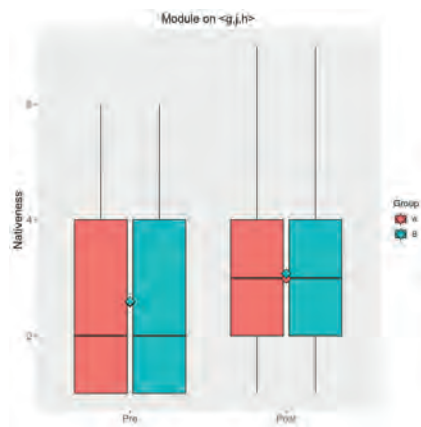

Figure 3. Boxplot of Nativeness Ratings by Group for the <g, j, h> Module.

4.4 Fluency

The ratings for fluency can be seen in Figure 4, with the beginning versus end of semester in the left panel, the vowel module in the center, and the <g, j, h> module in the right panel. In all three cases, the ASR group (B) shows a significant increase in fluency ratings, while the control group (A) does not change significantly in fluency between pre and post recordings. The increase in fluency for the ASR group (B) is smallest for the beginning versus end comparison ($p < 0.05$), somewhat larger for the vowel module ($p < 0.01$), and most notable for the <g, j, h> module ($p < 0.001$). In the case of the vowel and <g, j, h> modules, this is perhaps in part a task effect since the iSpraak interaction had the group B students repeat the same sentence several times before their final recording, whereas group A practiced the sounds in the recording sentence without practicing the sentence itself. Having had more practice with the specific recording sentence, the ASR group (B) thus sounded more fluent in their post recordings.

4.5 Perceived Confidence

Figure 5 shows the perceived confidence ratings for the vowel module (left panel) and the <g, j, h> module (right panel). For both participant groups, there are significant gains in perceived confidence when comparing pre and post recordings in these two modules. For the vowel module, the increase in perceived confidence is somewhat larger for the ASR group ($p < 0.001$), but is still significant for the control group ($p < 0.05$). Similarly, for the <g, j, h> module, the ASR group exhibits a larger gain in perceived confidence ($p < 0.001$) than the control group ($p < 0.04$). The difference observed between the two groups may in part be attributed to a task effect, as was discussed for the fluency ratings. The ASR group had more opportunities to practice the recording sentence itself, which would cause them to sound more confident the final time they said this sentence (in the post recording). Nevertheless, we see that both traditional and ASR interventions lead to increased confidence ratings. It is important to note that these are ratings of how confident (*seguro de sí mismo*) the students sounded and not how confident they actually felt. There are no significant differences in either group in confidence ratings for the beginning versus end of semester comparison.

Figure 4. Boxplot of fluency ratings by group comparing beginning and end of semester (left panel), vowel module (middle panel), and <g, j, h> module (right panel).

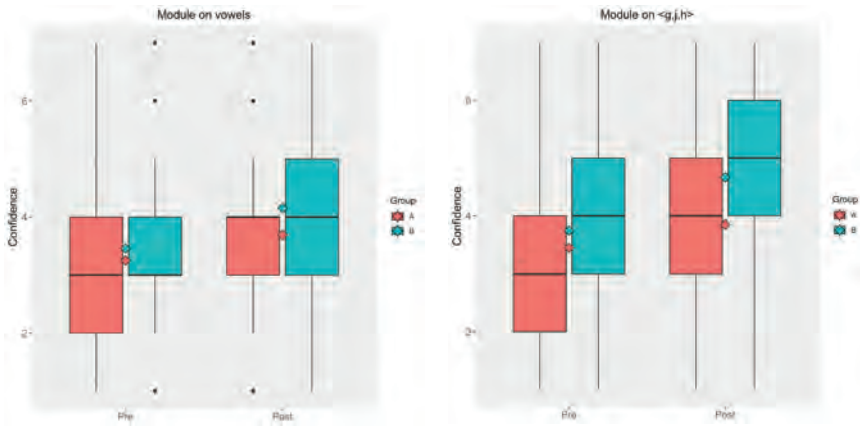

Figure 5. Boxplot of confidence ratings by group for the vowel module (left panel) and <g, j, h> module (right panel).

4.6 Summary

As was seen in the previous results sections, the effect of explicit and ASR pronunciation instruction varies based on the module considered, which allows us to see the difference between long-term and short-terms effects. As evident in Table 2, the ASR group (B) outperforms the explicit instruction group (A) for the short-term gains seen in the vowel and <g, j, h> modules. The ASR group is rated as significantly more comprehensible, fluent, and confident-sounding after the ASR intervention in both of these targeted modules, and also rated significantly more native-like in the <g, j, h> module. This could be, in part, as mentioned before, a task effect since the ASR group had more opportunities to practice with the test sentence that they produced for the pre and post recordings. The explicit instruction group only makes short-term gains for the measure of perceived confidence: they are rated as significantly more confident-sounding in the post recording for the vowel and <g, j, h> modules. On the other hand, the explicit instruction group (A) marginally outperforms the ASR group (B) for long-term gains. The explicit instruction group is rated as significantly more comprehensible when comparing the beginning and end of semester recordings. Given that "comprehensibility" is a joint factor that combines ratings of comprehensibility and percentage of words in which comprehension is impeded, the change seen here in the explicit instruction group represents a gain in two of the five measures rated. In comparison, the ASR group only shows a long-term gain for one measure: they are rated as significantly more fluent in the end of semester recording.

5. Conclusion

The present study found that both instructor-led and ASR-based instruction techniques yielded statistically significant gains in pronunciation ratings. These gains, which were often non-overlapping, were limited both in scope and nature. Given a very narrow interpretation of the data, one might conclude that ASR outperforms traditional instruction when targeting specific phonemes, especially in the short term. Zooming out to the semester level, however, we see more gains in *comprehensibility* with the explicit instruction group. When examined holistically, the data suggest that neither approach acts as a silver bullet for teaching pronunciation.

This study has important limits, of course, that should be considered from both research and pedagogical perspectives. First, it is problematic to extrapolate the findings to a different student population, with a different L1, or to different levels of proficiency, or indeed to second languages other than Spanish. Second, the pronunciation treatments were limited to brief bi-weekly visits by the research team. It is possible that more frequent or more prolonged instruction would yield different findings. Indeed, some of the targeted phoneme modules showed no significant gains at all, for either the control or experimental group. It could be that certain phonemes, generally speaking, are not easily improved through limited instruction or susceptible to any type of quick treatment. Additionally, the results rely on evaluator ratings and future studies should confirm whether these ratings correlate with measurable gains via acoustic analysis. Finally, this study used a specific CAPT tool, iSpraak, and it is unknown whether or not a different ASR-based tool would yield the same findings. Regrettably, there are no exportable data for student interaction with some features of the tool. It is unknown, for example, how many students used the TTS function or how many reviewed suggested words on *Forvo.com*. Understanding the role and impact of these functions might help paint a clearer picture of pronunciation gains as measured against interaction with the application.

The current state and sophistication of ASR technologies deserve the careful attention of L2 educators and researchers alike. Given the trends in the commercial language learning market, it is reasonable to suspect that ASR will be increasingly integrated into educational technology platforms. As pronunciation evaluation becomes more and more automated, it is vital to understand the instructional limits and pedagogical potential of ASR-powered CAPT tools. Exploring this question through further research studies will shed much-needed light on selecting optimal strategies for L2 pronunciation instruction. In particular, future studies should target a variety of sounds and proficiency levels in diverse L2s, and also with other CAPT tools, in order to

glean a more global picture of the efficacy of ASR in pronunciation instruction. The findings from the present study suggest that different interventions can yield different types of non-overlapping gains. To respond to the overarching research question of whether to supplant or supplement traditional instruction with an ASR-based tool, there appears to be compelling reasons for the latter, and it seems beneficial to maintain a two-pronged approach to pronunciation instruction for the foreseeable future.

About the Authors

Christina García, Assistant Professor of Spanish and Linguistics at Saint Louis University, is a sociolinguist and phonetician interested in phonetic variation, sociophonetic perception, and L2 pronunciation acquisition. She has done fieldwork in Argentina and Ecuador, examining how sounds are socially meaningful and contribute to the formation of regional identities, and her research on L2 pronunciation harnesses technological tools to provide diverse types of pronunciation feedback to learners. Her work has been published in journals such as Language Variation and Change and Studies in Hispanic and Lusophone Linguistics, and brings cutting-edge techniques used in sociophonetics to the forefront of Hispanic Linguistics.

Dan Nickolai is an Assistant Professor of French and the Director of the Language Resource Center at Saint Louis University. He has educational and professional backgrounds in the fields of Computer Science and Second Language Acquisition. His current research and development efforts are focused on designing software platforms that automate the evaluation and instruction of second languages. He is a familiar face on the CALL conference circuit, and his tools have been used in over 50 countries by tens of thousands of language students and educators.

Lillian Jones is a doctoral student at the University of California, Davis, studying Hispanic Linguistics and Second Language Acquisition. Her research interests include the pedagogical applications of text messaging and social media, computer-mediated communication, computer-assisted language learning, and online and hybrid teaching. Lillian's MA research paper explored the effect of text messaging on adult linguistic production. She has also published work regarding approaches to integrating emoji into L2 lessons, and is currently involved in providing curriculum and user experience support in the development of an open-source, digital vocabulary program.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Beaufays, F. (2015). The neural networks behind Google Voice transcription [web log]. Retrieved from https://ai.googleblog.com/2015/08/the-neural-networks-behind-google-voice.html

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379–397.

Elimat, A. K., & AbuSeileek, A. F. (2014). Automatic speech recognition technology as an effective means for teaching pronunciation. *The JALT CALL Journal*, *10*(1), 21–47.

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.

Grant, L., & Brinton, D. (2014). *Pronunciation myths: Applying second language research to classroom teaching.* Ann Arbor, MI: University of Michigan Press.

Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics, 27*, 184–202.

Liakin, D., Cardoso, W., & Liakina, N. (2013). Mobile speech recognition software: A tool for teaching second language pronunciation. *Cahiers De L'ilob*, *5*, 85–99.

Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal*, *32*(1), 1–25.

Liakin, D., Cardoso, W., & Liakina, N. (2017). The pedagogical use of mobile speech synthesis (TTS): Focus on French liaison. *Computer Assisted Language Learning*, *30*(3–4), 325–342.

Lord, G. (2019). Incorporating technology into the teaching of Spanish pronunciation. In R. Rao (Ed.), *Key issues in the teaching of Spanish pronunciation: From description to pedagogy* (218–236). New York, NY: Routledge.

Morgan, Terrell A. (2010). *Sonidos en contexto: una introducción a la fonética del español con especial referencia a la vida real*. New Haven, CT: Yale University Press.

Neri, A., Cucchiarini, C., & Strik, H. (2002a). Feedback in computer assisted pronunciation training: Technology push or demand pull? In Z. Tan & P. Dalsgaard (Eds.), *Proceedings of the International Conference on Spoken Language Processing* (pp. 1209–1212). Denver, CO, ICSLP.

Neri, A., Cucchiarini, C., & Strik, H. (2003). Automatic speech recognition for second language learning: How and why it actually works. In M. J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings from the 15th International Congress of Phonetic Sciences* (pp. 1157–1160). Rundle Mall, Australia: Causal Productions Pty Ltd.

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002b). The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning*, *15*(5), 441–467.

O'Brien, M. G., Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. I., Levis, G.M. (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation*, *4*(2), 182–207.

Olson, D. J. (2014). Benefits of visual feedback on segmental production in the L2 classroom. *Language Learning & Technology*, *18*(3), 173–192. Retrieved from https://www.lltjournal.org/item/2875

Pieraccini, R. (2012). *The voice in the machine: Building computers that understand speech*.

Cambridge, MA: MIT Press.
R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. www.R-project.org
Thomson, R. I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *CALICO Journal*, *28*(3), 744–765.
Thomson, R. I., & Derwing, T. M. (2014). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*(3), 326–344.

Appendix

1. Beginning/End of Semester Text Recorded in First and Last Modules

Dialectos (taken from Morgan, 2010)

Hay más de trescientos millones de personas que hablan español, principalmente en España y Latinoamérica. Por razones históricas y geográficas, han divergido los varios dialectos de la lengua. No sólo existen diferentes acentos, sino también diferentes léxicos: se dice coche, piso y maíz en España; auto, apartamento y choclo en Chile; carro, departamento y elote en México. Sin embargo, las manifestaciones culturales del mundo hispanohablante – arte, cine, deporte, literatura, música y televisión – sirven para compensar la diversidad lingüística.

[There are more than 300 million people who speak Spanish, principally in Spain and Latin America. For historical and geographic reasons, the various dialects of the language have diverged. There are not only different accents, but different lexicons: car, apartment, and corn are said in Spain; car, apartment, and corn in Chile; car, apartment, and corn in Mexico. However, the cultural manifestations of the Spanish-speaking world—arte, cinema, sports, literature, music and television—serve to compensate the linguistic diversity.]

2. Recording Sentences for Vowel and <g ,j, h> Modules

Vowels

1. *Durante el show de la moda, el modelo llevó un suéter amarillo, unos pantalones morados, y unas sandalias de playa.* [During the fashion show, the model wore a yellow sweater, purple pants, and beach sandals.]

2. *El diseñador presentó un estilo nuevo de Bogotá con unas botas modernas, una blusa típica, y un vestido tradicional.* [The designer presented a new style from Bogotá with some modern boots, a typical blouse, and a traditional dress.]

<g, j, h>

1. *En junio y julio en las Grandes Ligas de Béisbol, los jugadores juegan muchos partidos usando guantes, bates y pelotas.* [In June and July in Major League Baseball, the players play many games using gloves, bats, and balls.]

2. *José Hernández y Hugo Cabrera son dos dominicanos que juegan al béisbol en los Estados Unidos y son hombres de segunda base.* [José Hernández and Hugo Cabrera are two Dominicans who play baseball in the United States and they are second base men.]