EUROCALL 2023: CALL for all Languages
15-18 August 2023, University of Iceland, Reykjavik
DOI: https://doi.org/10.4995/EuroCALL2023.2023.16987

Assessing Google Translate ASR for feedback on L2 pronunciation errors in unpredictable sentence contexts

Paul John[a], Carol Johnson[b] and Walcir Cardoso[c]

[a]Université du Québec à Trois-Rivières, CogNAC Research Group, Trois-Rivières, Canada, , paul.john@uqtr.ca; [b]Concordia University, Centre for the Study of Learning and Performance, Montreal, Canada, , carol.johnson@concordia.ca and [c]Concordia University, Centre for the Study of Learning and Performance, Montreal, Canada, , walcir.cardoso@concordia.ca

How to cite: John, P.; Johnson, C.; Cardoso, W. (2023). Assessing Google Translate ASR for feedback on L2 pronunciation errors in unpredictable sentence contexts. In *CALL for all Languages - EUROCALL 2023 Short Papers.* 15-18 August 2023, University of Iceland, Reykjavik. https://doi.org/10.4995/EuroCALL2023.2023.16987

Abstract

Following previous research into predictable sentence contexts, this study assesses the pronunciation feedback provided by Google Translate's (GT) Automatic Speech Recognition (ASR) in unpredictable contexts. We examined the accuracy of GT transcriptions for target items recorded by male and female Quebec Francophones (QFs). The items occurred in neutral carrier sentences such that no contextual cues help ASR identify the targets. Th-initial vs t-initial (thank-tank) *and h-initial vs vowel-initial* (heat-eat) *items were used to investigate the potential for feedback on the QF errors of th-substitution, h-deletion, and h-epenthesis, comparing real-word* (thank→tank) *vs nonword output* (thief→tief). *As with predictable contexts in our previous research, we observed high transcription accuracy for real words only. Without contextual cues, accuracy rates were lower than in predictable contexts for correctly pronounced items but higher than for incorrect pronunciations constituting real words. Unpredictable contexts are thus inferior at confirming correct pronunciation (confirmative feedback) but superior at flagging real-word errors (corrective feedback). Contrary to the anticipated ASR gender bias, female recordings showed higher transcription accuracy than male recordings. Our findings both confirm the usefulness of GT's ASR for generating pronunciation feedback and highlight the importance of context (predictable vs unpredictable) and lexical status (real vs nonword).*

Keywords: *automatic speech recognition, Google Translate, L2 pronunciation, corrective vs confirmative feedback, predictable vs unpredictable contexts, gender bias.*

1. Introduction

The current study expands on an earlier investigation into using Google Translate's (GT) Automatic Speech Recognition (ASR) for *corrective* and *confirmative feedback* on second language (L2) pronunciation errors. Although corrective feedback can help learners improve pronunciation (Saito, 2021), questions remain regarding the accuracy of ASR feedback (Inceoglu et al., 2022). L2 pronunciation errors are generally variable, meaning learners alternate between targetlike and erroneous realizations of L2 sounds. Consequently, ASR transcriptions should not only flag incorrect pronunciation but also confirm correct pronunciation, hence our introduction of the term *confirmative feedback* as a complement to *corrective feedback*. While our previous work investigated GT ASR transcription accuracy for items in predictable sentence contexts (John et al., in press), the current study assesses the technology in unpredictable contexts. The purpose is to establish the impact of presence vs absence of contextual cues on GT ASR's ability to provide feedback on correct and incorrect pronunciations.

We focus on three Quebec Francophone (QF) pronunciation errors in English: th-substitution (*thank→tank, thief→tief*), h-deletion (*heat→_eat, help→_elp*), and h-epenthesis (*old→hold, ice→hice*) (John & Frasnelli, 2022). Crucially, we investigate the role of lexical status (real vs nonword) and gender (male vs female voices) on transcription accuracy. While correct pronunciations always constitute real words, pronunciation errors can generate real or nonwords (*tank, _eat, hold* vs *tief, _elp, hice* above). Since nonwords are by definition absent from the GT lexicon, we anticipated high transcription accuracy for real words only. Nonetheless, unpredictable contexts should generate lower accuracy than correctly produced items in predictable contexts, where contextual (syntactic-semantic-collocational) cues conspire with phonetic cues to ensure correct identification of the target item (Ashwell & Elam, 2017). With respect to gender, ASR systems are often trained on datasets with more male speech samples, potentially leading to poorer performance on recognizing female voices (Garnerin et al., 2019). Gender bias in ASR would undermine the appropriacy of its use for L2 learning purposes.

Previously, in Phase 1 of our research, we examined GT transcription accuracy for correctly and incorrectly pronounced items in *predictable* sentence-final contexts (e.g. *I don't know who to thank*$^{\surd}$-*tank*^{X}) (John et al., in press). Correctly pronounced items showed high transcription accuracy rates (88.33%). Real-word output in the error condition showed lower accuracy (47.50%), but considerably higher than nonword output in the error condition (8.33%). GT's ASR is thus particularly good at confirming accurate pronunciation in predictable contexts, especially given that no false alarms were observed (e.g. instances where a correctly pronounced *thank* was transcribed as *tank,* erroneously indicating an error). It also flags pronunciation errors almost half the time, as long as these lead to real words (i.e. *thank→tank* but not *thief→tief*). As summarized in Table 1 below, we thus observed more ('>') confirmative than corrective feedback; and within corrective feedback, more feedback on real than nonwords. Where it failed to transcribe errors accurately, GT's ASR usually produced false negatives (36.66% for real words; 65.00% for nonwords). False negatives are transcriptions that reflect the target item despite incorrect pronunciation (e.g. *thank* mispronounced as *tank* or *thief* mispronounced as *tief* being nonetheless transcribed as *thank* and *thief*). These occur partly because ASR can recover the target item from contextual cues despite incorrect pronunciation. Such cues are, however, exclusively available in predictable contexts, hence the importance of investigating unpredictable contexts. Interestingly, transcription accuracy for female speakers was consistently higher, so the concern that female learners might receive less accurate feedback due to gender bias appears unfounded. Tentatively, we attributed the female advantage to women's generally more targetlike L2 production and careful articulation (Moyer, 2016). This female advantage should be less evident in unpredictable contexts, since clear articulation of the carrier sentence in no way aids identification of the target; indeed, the usual pattern for gender bias, advantaging male speakers, could conceivably emerge in unpredictable contexts.

The current study, constituting Phase 2 of the research, retested GT's ASR for real and nonwords produced in an *unpredictable* carrier sentence. Without contextual cues, the distinction between real words corresponding to correct vs incorrect pronunciations no longer applies. Thus, confirmative and corrective feedback were conflated, and only the *real-word* > *nonword* advantage was investigated ('>' = 'higher transcription accuracy than'). Being identified solely via phonetic cues, decontextualized real words should show lower transcription accuracy than correctly pronounced items (< 88.33%) in predictable contexts (i.e. as observed in our previous study), but higher than incorrectly pronounced items (> 47.50%) in predictable contexts (again, from the previous study). Put differently, we expected GT's ASR to be worse in unpredictable contexts at confirming correct pronunciation, but better at flagging real-word errors (see Table 1 for a summary of these hypotheses). That is, for confirmative feedback, we anticipated a *predictable* > *unpredictable* advantage; whereas for corrective feedback, we anticipated the reverse *unpredictable* > *predictable* advantage. We likewise investigated whether gender bias emerges in unpredictable contexts or whether the *female* > *male* advantage persists, and we gathered information on false alarms/negatives (which, like confirmative and corrective feedback are necessarily conflated in unpredictable contexts).

Table 1. Summary of Phase 1 findings (predictable contexts) vs Phase 2 hypotheses (predictable contexts)

PHASE 1 (predictable contexts)	PHASE 2 (unpredictable contexts)
Real words: *confirmative* > *corrective feedback* 88.33% vs 47.50%	Real words: *confirmative*/*corrective feedback* < 88.33% *confirmative*/*corrective feedback* > 47.50%
real words > *nonwords* 88.33% / 47.50% vs 8.33%	*real words* > *nonwords*
F > M	M > F or F > M
False alarms: 0%	False alarms: no hypothesis formulated
False negatives: 36.66% (real words); 65.00% (nonwords)	False negatives: no hypothesis formulated

2. Method

Ten Male (M) and 10 Female (F) QF adults were used to record 200 items in a carrier sentence (*This is what I would like to say*, "______"). The recordings were not based on spontaneous speech with naturally occurring errors and correct pronunciations. Instead, we asked speakers to produce th-initial, t-initial, h-initial, and vowel-initial real and nonwords as presented in a written prompt, and any recordings containing genuine mispronunciations were eliminated. That is, the speakers should be viewed as L2 voice actors used to generate stimuli rather than as participants. The true participant in this research is Google Translate itself.

Based on minimal pairs (e.g. *thank-tank, hate-ate*), the 140 real-word items we used were th- vs t-initial and h- vs vowel-initial. The 60 nonword targets comprised t-initial, vowel-initial, and h-initial forms (e.g. *tief, _appy, hice*). The items thus covered all of the output forms under QF correct or incorrect production of the English 'th' and 'h' sounds. Fewer nonwords were tested than real words mainly because we were confident, based on our previous findings during Phase 1 (John et al., in press), that GT would be unable to transcribe these accurately. Of the 20 recordings of each item in the carrier sentence, we selected ten (5M/5F) to play into GT's ASR. In determining which recordings to retain, those with unclear or erroneous articulation of the target items were eliminated, such that only optimal recordings for our research aims remained. These 2000 recordings were coded for final-item transcription accuracy with the aim of comparing real vs nonword output and M vs F speakers.

Inaccurate transcriptions were further investigated for 'false alarms/negatives'. False alarms/negatives involve real words being transcribed as the minimal pair opposite, such as a *thank* recording being transcribed as *tank* or vice versa. This misleadingly suggests learners have substituted 't' for 'th' (false alarm) or correctly realized 'th' when 't' was in fact substituted (false negative). A nonword transcribed as its real-word counterpart (*tief* transcribed as *thief*) also constitutes a false negative.

3. Results & discussion

Table 2 presents accuracy rates for transcriptions of real words produced by male and female speakers, both separately (M/F) and combined (M + F).

Table 2. Transcription accuracy: real words (e.g. *thank*, *tick*, *hold*, *eat*) in unpredictable contexts (%)

Target items	M	F	M + F
th-initial	60.00	64.50	62.25
t-initial	35.00	34.00	34.50
h-initial	70.00	76.00	73.00
V-initial	74.00	84.50	79.25
Mean	**59.75**	**64.75**	**62.25**

As expected, the overall accuracy rate for male and female voices combined (62.25%) is lower than observed in our previous study for correctly realized items (88.33%) in predictable contexts. Conversely (again, as expected), the rate of 62.25% is higher than our previously observed rate for incorrectly realized items leading to real-word output (47.50%) in predictable contexts. One anomaly is that t-initial real words in our current study inexplicably show lower accuracy (34.50%) than incorrectly realized items, leading to real-word output (47.50%) in predictable contexts. GT’s ASR transcription accuracy in unpredictable contexts performs equally well for both correct and incorrect pronunciations constituting real words: without contextual information, only phonetic cues participate in item identification, leading to lower accuracy in confirming correct pronunciation but higher in flagging incorrect pronunciation. Corrective feedback on error is thus more reliable in unpredictable contexts, whereas confirmative feedback on correct pronunciation is less reliable.

Furthermore, we can report that false alarms/negatives are rare among real words (2.25-4.25%), with the minor exception again of t-initial items (14.00%). GT’s ASR thus tends not to signal that learners, upon producing a real word, have either mispronounced a correctly realized sound or correctly realized a mispronounced sound. Indeed, many inaccurate transcriptions could be designated ‘near accurate’ (13.75-27.00%), meaning the transcription, while diverging from the actually realized item, nonetheless accurately reflects the quality of the initial sound. For example, output *thank*, *hold,* and *tank* transcribed as *think*, *home,* and *take;* while strictly speaking this is inaccurate, they are ‘near-accurate’ insofar as they correctly indicate how the crucial initial sound was produced.

Interestingly, as observed previously in predictable contexts, transcription accuracy for female recordings of real words is higher than for male recordings across nearly all conditions (Table 2). We anticipated that the female advantage might disappear in unpredictable contexts, since careful pronunciation of the neutral carrier sentence (expected in female L2 speech; Moyer, 2016) provides ASR with no advantage in identifying the final item. Nonetheless, clearer female articulation of just the target itself apparently aids item identification. Table 3 presents accuracy rates for transcriptions of nonwords.

Table 3. Transcription accuracy: nonwords (e.g. *tief, hice, elp*) in unpredictable contexts (%)

Target items	M	F	M + F
t-initial	0.00	0.00	0.00
h-initial	8.00	9.00	8.50
V-initial	0.00	2.00	1.00
Mean	**2.67**	**3.67**	**3.17**

The overall mean for nonword output for male and female voices is glaringly low (3.17%), but this is not surprising given that GT cannot match a nonword to an entry in its lexicon. This finding confirms that GT is essentially incapable of providing corrective feedback on nonword mispronunciations. Indeed, the few instances where transcriptions actually captured phonetic output presumed to constitute nonwords, involved instances where GT was able to identify a proper noun (e.g. *hivy* transcribed as *Hy-Vee*, a grocery store) or to segment the input into smaller units (e.g. *hegos* transcribed as *he goes*). These findings suggest that, to be effective, pronunciation activities should focus on target items resulting in real words if mispronounced. We also observed high rates of false alarms/negatives among nonwords (26.00-40.50%), which only reinforces this implication.

Nonetheless, we should point out that many of the inaccurate transcriptions reassuringly constitute ‘near accurate’ transcriptions (37.50-59.00%). That is, the realization of the initial sound was frequently reflected in the transcription (e.g. the realization *tief* for target *thief* was transcribed as *teeth,* accurately signaling that ‘t’ was substituted for ‘th’). Thus, while real-word output in controlled activities is ideal for generating GT’s ASR pronunciation feedback, the technology can still generate partial (‘near accurate’) corrective feedback on nonwords produced in more open activities such as those involving spontaneous speech.

4. Conclusions

GT’s ASR can provide beneficial L2 pronunciation feedback. However, our investigation of QF th-substitution, h-deletion, and h-epenthesis reveals that the accuracy of the feedback is affected by the context in which these pronunciation issues occur. For flagging pronunciation errors (corrective feedback), unpredictable contexts are better; for confirming correct pronunciation (confirmative feedback), predictable contexts are. Moreover, regardless of context, GT’s ASR is markedly better at flagging incorrect pronunciations that generate real words (*thank→tank*) than nonwords (*thief→tief*). Pronunciation activities should thus take into consideration both the presence/absence of contextual cues and the lexical status of mispronounced items. We suggest teachers create practice sentences in which mispronunciation of the target sounds results in real words, thus increasing the amount of ASR corrective feedback learners receive. Target words could initially be placed in decontextualized carrier sentences or word lists to generate more corrective feedback on pronunciation errors. As students accurately produce the sound more frequently, teachers could employ sentences that use the word in context. Doing so will provide students with more confirmative feedback from ASR, showing students that they can successfully produce the sound and increasing their self-efficacy in their pronunciation skills. It is in our plans to develop and trial activities based on these findings and suggestions. Reassuringly, the anticipated gender bias failed to emerge: even for items in unpredictable contexts, female speakers showed higher transcription accuracy than males. In sum, GT’s ASR has considerable potential to generate invaluable feedback on pronunciation, but its ability to provide both corrective and confirmative feedback is influenced crucially by the nature of the sentence context (predictable vs unpredictable) and by the lexical status of output (real word vs nonword).

Acknowledgements

We are grateful to Andrew Conroy for splicing and editing of the sound files. This project received funding from the *Luc Maurice Foundation* and the *Social Sciences and Humanities Research Council* (430-2022-00512).

References

Ashwell, T., & Elam, J. (2017). How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners’ oral production? *The JALT CALL Journal, 13*(1), 59-76. https://doi.org/10.29140/jaltcall.v13n1.212

Garnerin, M., Rossato, S, & Besacier, L. (2019). Gender representation in French broadcast corpora and its impact on ASR performance. In *AI4TV '19: Proceedings of the 1st international workshop on AI for smart TV content production* (pp. 3-9). Association for Computing Machinery. https://doi.org/10.1145/3347449.3357480

Inceoglu, S., Chen, W., & Lim, H. (2022). Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition. *ReCALL, 35*(1), 89-104. https://doi.org/10.1017/S0958344022000192

John, P., Cardoso, W., & Johnson, C. (in press). Automatic speech recognition as a source of corrective feedback on L2 pronunciation. In M. Peterson & N. Jabbari (Eds.), *Frontiers in computer assisted language learning* (pp. 1-19). Routledge.

John, P., & Frasnelli, J. (2022). On the lexical source of variable L2 phoneme production. *The Mental Lexicon, 17*(2), 239-276. https://doi.org/10.1075/ml.22002.joh

Moyer, A. (2016). The puzzle of gender effects in L2 phonology. *Journal of Second Language Pronunciation, 2*(1), 8-28. https://doi.org/10.1075/jslp.2.1.01moy

Saito, K. (2021). Effects of corrective feedback on second language pronunciation development. In H. Nassaji & E. Kartchava (Eds.), *The Cambridge handbook of corrective feedback in second language learning and teaching* (pp. 407-428). Cambridge University Press. https://doi.org/10.1017/9781108589789.020