**Paul John, Walcir Cardoso, Carol Johnson**

Université du Québec à Trois-Rivières, Trois-Rivières, Canada
Concordia University, Montreal, Canada
Concordia University, Montreal, Canada

paul.john@uqtr.ca, walcir.cardoso@concordia.ca, carol.johnson@concordia.ca

# On the adequacy of L2 pronunciation feedback from automatic speech recognition: A focus on Google Translate

## Bio data

Paul John is Associate Professor of Modern Languages at the Université du Québec à Trois-Rivières. His research focuses on L2 phonological acquisition, with particular interest in variation and in using neuroimaging to investigate L2 perception. Another research interest concerns the use of automatic speech recognition, text-to-speech and grammar checkers for language learning.

Walcir Cardoso is Professor of Applied Linguistics at Concordia University. He conducts research on the L2 acquisition of phonology, morphosyntax and vocabulary, and the effects of computer technology (e.g., clickers, text-to-speech synthesizers, automatic speech recognition, intelligent personal assistants) on L2 learning.

Carol Johnson is a PhD student in Education (specialization in Applied Linguistics) at Concordia University. Her research interests include the pedagogical use of speech technologies in L2 learning, particularly the use of automatic speech recognition to improve pronunciation and writing. She currently teaches university-level ESL courses.

## Abstract

This study investigates automatic speech recognition (ASR) in Google Translate as a source for L2 pronunciation feedback. To be effective, ASR should transcribe learner errors accurately and perform equally well on male and female voices, avoiding gender bias. We assess Google Translate on three Quebec francophone (QF) segmental errors in English: th-substitution (*think* → [t]*ink*); h-deletion (*happy* → *_appy*); and h-epenthesis (*ice* → [h]*ice*). Eight QFs (4F/4M) recorded 120 sentences with and without an error on the final item (e.g., *I don't know who to *tank/thank*). Errors were equally divided between real word output (**tank*) and nonword output (e.g., *My sister is afraid of *tunder*). We anticipate real word errors, corresponding to entries in the Google Translate lexicon, will be accurately transcribed, whereas nonwords, by definition absent from the lexicon, should be erroneously matched to similar-sounding real words (i.e., the intended output "thunder"), constituting misleading feedback.

Forthcoming data analyses will determine the relative contribution of error type, real/nonword output, and gender to final-word transcription and feedback accuracy. Preliminary findings suggest a hierarchy of accuracy (h-deletion, h-epenthesis >

th-substitution) specific to real-word output. Indeed, ASR shows a clear inability to flag nonword errors. A gender bias effect is not apparent; in fact, ASR generally transcribed the sentences recorded by females more accurately. Mistranscriptions unrelated to final items have yet to be examined. Our presentation will address the implications of our findings for L2 teachers/learners and for developers seeking to design ASR specifically for L2 uses.

## Conference paper

### Introduction

Second language (L2) learners generally require help with pronunciation since L2 phonology is hard to acquire through mere exposure (Flege, Munro & Mackay, 1995). One means of promoting pronunciation accuracy in a classroom context is through corrective feedback (Lyster et al., 2013). The potential of automatic speech recognition (ASR) to supply feedback that learners can access autonomously and ubiquitously is thus appealing. ASR is a function widely available in tools such as Google Translate (GT), the system focused on here. Nonetheless, different aspects of the adequacy of ASR feedback remain to be determined.

The current study investigates ASR feedback with respect to three segmental pronunciation errors typical of Quebec francophone (QF) learners of English: th-substitution (*thank* → *tank*), h-deletion (*heat* → *_eat*) and h-epenthesis (*ice* → *hice*) (Brannen, 2011; John & Cardoso, 2009; Mah et al., 2016; White et al., 2015). These errors are usually variable. QF learners alternate between pronouncing items such as *thank*, *heat* and *ice* inaccurately and accurately: [tæŋk]~[θæŋk]; [it]~[hit]; and [hajs]~[ajs]. For pronunciation feedback purposes, ASR systems should distinguish between correct and incorrect pronunciation of L2 segments, reinforcing learners' accurate production while also flagging inaccuracies.

First, when QFs pronounce *thank*, *heat* and *ice* correctly, it is crucial that ASR confirms the targetlike output by providing an accurate transcription. Otherwise, by transcribing something else the system is sending the misleading message that learners' pronunciation is off (a 'false alarm'). Second, when QFs mispronounce *thank*, *heat* and *ice*, it is important that ASR indicates that learners have missed the mark. Importantly, we anticipated that such corrective feedback would be more accurate with mispronunciations leading to real-word output (e.g., *thank* → *tank*) than nonword output (*ice* → *hice*).

The reason is that ASR bases its transcriptions on items stored in its lexicon. ASR attempts to match the phonetic output to existing lexical entries. Consequently, the system should transcribe learner errors more accurately when the output corresponds to a real word, directly signaling that learners have mispronounced the target item. In the case of nonword output (*thief → tief, head → _ead, ice → hice*), ASR will likely search for the closest lexical match for the phonetic output, potentially arriving at *thief, head* and *ice.* This is particularly the case when items are embedded in sentences, such that ASR can make use of top-down prediction based on collocations, semantic associations and syntactic analysis. With correctly pronounced items, top-down processing helps the system make an accurate lexical match, thus reducing the likelihood of false alarms. With incorrectly pronounced items, however, top-down processing risks overriding phonetic cues, supplying the appropriate (i.e., target) item for the context despite the mispronunciation. To our knowledge, these issues have yet to be investigated empirically.

Our study examines ASR transcriptions for accurately and inaccurately pronounced target items that begin with /θ/, /h/ or a vowel. One objective was to establish the extent to which, in a sentence context, ASR generates false alarms when faced with accurately pronounced items. Furthermore, we used inaccurately pronounced items constituting real-word or nonword output. We anticipated the following hierarchy for transcription accuracy:

no error condition > error condition (real-word) > error condition (nonword)

Given mispronounced items, another objective was to explore whether ASR flags the three error types with comparable accuracy. In addition, the sentences were recorded by male and female QF learners of English in order to investigate the question of gender bias. Given that ASR systems are often trained on corpora with a preponderance of male voices, they may perform less well on female output (Tatman, 2017). For L2 teachers, it would be a serious concern if they were to recommend a tool that provides less accurate feedback to their female learners.

## Methodology

### ***Materials***

A total of 120 sentences with a final item starting with either /θ/, /h/ or a vowel were recorded by 4 male and 4 female QFs with and without a pronunciation error on the final item. For example:

*I don't know who to* ***thank/*tank.***
*I still need to brush my* ***hair/*air****.*
*We slipped and fell on the* ***ice/*hice****.*

In the error condition, the 120 sentences were evenly divided between 60 real-word and 60 nonword errors.

### *Data collection and analysis*

480 recordings (4 versions of each sentence: in the error and no error conditions spoken by a male and female speaker) were played into GT's ASR function. We noted whether each transcription corresponded to the *error* or *no error* condition and whether the speaker was *male* or *female*. We further noted whether the final target item began with /θ/, /h/ or a vowel and whether, given an error, the output contained a *real-word* or *nonword*. The following independent variables were thus included: pronunciation accuracy (error-no error), gender (M-F), target sound (/θ/-/h/-V), and output form in the error condition (real-nonword).

The ASR transcriptions of final items were also examined for two dependent variables: *transcription accuracy* and *accuracy of pronunciation feedback*. Transcription and feedback accuracy are partly, but not entirely, coextensive. Where a transcription captures exactly the phonetic output, clearly it provides accurate pronunciation feedback. Conversely, if target *thank* is correctly pronounced [θæŋk] but transcribed 'tank' OR if target *thank* is mispronounced [tæŋk] but transcribed 'thank', this constitutes incontrovertibly inaccurate feedback on pronunciation.

A transcription can, however, diverge from these clearcut cases (e.g., [θæŋk] may be transcribed 'talk' or 'thong' or 'sank'), and the resulting feedback can vary in how misleading it is. We classified such inaccurate transcriptions into *problematic*, *part-accurate* and *neutral feedback*.

In the no error condition, given a target with initial /θ/, /h/ or a vowel, the categories refer to:

> Problematic feedback: The transcribed item starts with /t/, a vowel or /h/, sending the misleading message that the learner has engaged in th-substitution, h-deletion or h-epenthesis.

> *Part-accurate* feedback: Although different from the target item, the transcribed item starts with /θ/, /h/ or a vowel, appropriately signaling the correct pronunciation.
>
> Neutral feedback: The transcribed item starts with another sound entirely, indicating neither a mispronunciation nor a correct pronunciation.

In the error condition, *problematic* and *part-accurate feedback* are the reverse of above; *neutral feedback* is the same:

> Problematic feedback: The transcribed item starts with /θ/, /h/ or a vowel, thus masking the mispronunciation of the initial segment.
>
> *Part-accurate* feedback: Although different from the target item, the transcribed item starts with /t/, a vowel or/h/, appropriately signaling the mispronunciation.
>
> Neutral feedback: The transcribed item starts with another sound entirely, neither masking nor correctly signaling the mispronunciation.

## Results

### *No error condition*

Accuracy rates for transcriptions of correctly pronounced final items provide an indication of the extent to which the ASR system confirms correct pronunciation, thus providing accurate feedback. Table 1 shows the transcription accuracy rates for final items that, were they mispronounced, would result in a) real-word or b) nonword output, for M and F speakers both separately and combined. The overall mean for both sets of sentences (a + b) is also indicated.

**Table 1.** *No error condition: final word transcriptions*

**a. Accuracy rates (real-word output sentences in error condition)**

| **Target items** | **M** | **F** | **M + F** |
|---|---|---|---|
| *th-initial* | .70 | .80 | .75 |
| *h-initial* | .85 | .95 | .90 |
| *V-initial* | .90 | .95 | .93 |
| **Mean:** | **.82** | **.90** | **.86** |

**b. Accuracy rates (nonword output sentences in error condition)**

| **Target items** | **M** | **F** | **M + F** |
|---|---|---|---|
| *th-initial* | .75 | .90 | .83 |
| *h-initial* | .95 | 1.0 | .98 |
| *V-initial* | .85 | 1.0 | .93 |
| **Mean:** | **.85** | **.95** | **.90** |
| **Overall mean:** | **.83** | **.93** | **.88** |

Across the board, transcription accuracy rates for F voices are higher than for M voices (Table 1), the opposite to the predicted pattern of gender bias. In terms of the target items, h-initial and V-initial items have consistently higher accuracy rates than th-initial items. Accuracy rates were incidentally higher in the set of sentences that would lead to nonword output in the error condition (except vowel-initial items spoken by M voices). Conceivably, the final items in this second set of sentences are marginally more predictable.

Given the overall mean of .88, these results suggest learners should expect ASR to correctly transcribe approximately 9 in 10 correctly pronounced content words in a predictable sentence context. Promisingly, none of the inaccurate transcriptions supply the item corresponding to the typical QF mispronunciation (i.e., no instances of output [θæŋk], [hit] and [ɛr] being transcribed 'tank', 'eat' and 'hair'). That is, none of the inaccurate transcriptions constituted clearly inaccurate pronunciation feedback (Table 2).

**Table 2.** *No error condition: rates of inaccurate transcriptions/feedback types*

| | **Pronunciation feedback** | | | |
|---|---|---|---|---|
| **Target items** | **Inaccurate** (*thank → tank,* etc.) | **Problematic** (*theft → test*) | **Part-accurate** (*thrifty → thirsty*) | **Neutral** (*thud → fun*) |
| *th-initial* | .00 | .03 | .05 | .14 |
| *h-initial* | .00 | .00 | .06 | .00 |
| *V-initial* | .00 | .00 | .06 | .01 |

Likewise, very few of the mistranscriptions indirectly send the message that the learner has produced a typical pronunciation error. In fact, the only examples of such problematic feedback involve 2 sentences with target items starting with /θ/ (e.g., *I wonder who committed the* ***theft*** *→* ***test***). The mistranscriptions for 4 sentences with target items starting with /θ/ in fact substitute items that include /θ/, thus constituting part-accurate feedback (e.g., *Danielle has always been very* ***thrifty*** *→* ***thirsty****)*. While strictly speaking inaccurate, these mistranscriptions nonetheless indirectly suggest learners have correctly realized the difficult target sound.

All of the mistranscriptions of sentences with target items starting with /h/ and all but one with target items starting with a vowel were of this same part-accurate feedback type (Table 2). No examples of problematic or neutral feedback were observed for h-initial forms; no examples of problematic feedback were observed for vowel-initial forms. The item transcribed in each case started with /h/ or a vowel (e.g., *Their wedding was in a lovely* ***hall*** *→* ***whole***; *You need to use your* ***elbows*** *→* ***albums***). With such part-accurate feedback, learners will not be inclined to conclude they have deleted /h/ or needlessly inserted it.

The remaining 11 mistranscriptions of items starting with /θ/ constitute neutral feedback (i.e., the transcription contained another sound than /θ/ or /t/: *The book hit the floor with a loud* ***thud*** *→* ***fun***). One mistranscription of a target item beginning with a vowel also constitutes neutral feedback: *Chicken pox makes you* ***itch*** *→* ***pitch***.

*Error condition*

Accuracy rates for transcriptions of mispronounced final items provide an indication of the extent to which ASR flags pronunciation errors, thus providing accurate feedback. Table 3 shows the rates of accurate transcription for mispronounced items corresponding to a) real-word or b) nonword output.

**Table 3. Error condition: final word transcriptions**

**a. Accuracy rates (real-word output)**

| **Target items** | **M** | **F** | **M + F** |
|---|---|---|---|
| *th-initial* | .20 | .30 | .25 |
| *h-initial* | .35 | .60 | .48 |
| *V-initial* | .30 | .85 | .58 |
| **Mean:** | **.27** | **.58** | **.43** |

**b. Accuracy rates (nonword output)**

| **Target items** | **M** | **F** | **M + F** |
|---|---|---|---|
| *th-initial* | .00 | .00 | .00 |
| *h-initial* | .05 | .15 | .10 |
| *V-initial* | .05 | .20 | .13 |
| **Mean:** | **.03** | **.12** | **.08** |
| **Overall mean:** | **.15** | **.35** | **.26** |

Transcription accuracy rates for F voices are, as in the error condition, almost always higher than for M voices (Table 3). This F advantage is the opposite to the predicted pattern of gender bias. H-initial and vowel-initial items once again show consistently higher accuracy rates than th-initial items. Accuracy rates are also decidedly lower in the set of sentences with nonword output. Indeed, as expected, none of the mispronounced th-initial targets resulting in nonwords were correctly transcribed. Surprisingly, some h-initial targets (1M and 3F) and some vowel-initial targets (1M and 4F) resulting in nonwords were correctly transcribed. In some cases, GT found a proper noun that corresponded to the supposed nonword (e.g., for *oil* → *hoil,* the transcription was 'Hoyle'). In one sentence, *empty* → *hempty* was transcribed 'hemp tea'.

The overall mean of .26 for M and F voices suggests learners can expect ASR to mistranscribe 7.4 of 10 content words that they mispronounce in a sentence context. Nonetheless, transcription accuracy varies considerably: while accuracy rates for nonword output are exceedingly low (as low as .00), rates for real-word output can be promisingly high (as high as .85). Correspondingly, the degree of truly accurate pronunciation feedback also varies.

Incorrect transcriptions that provide clearly inaccurate pronunciation feedback are those that supply the mispronounced target item (e.g., *thank* → *tank* is transcribed 'thank'). As reported above, no examples of the reverse phenomenon (*thank* → *thank* being transcribed 'tank') appear in the no error condition. In the error condition, however, such clearly inaccurate feedback is widespread. Table 4 shows the rates of target item transcription for mispronounced final items corresponding to a) real-word or b) nonword output.

**Table 4.** *Error condition: target item transcriptions (inaccurate feedback)*

**a. Target item rates (real-word output sentences)**

| Target items | M | F | M + F |
|---|---|---|---|
| *th-initial* | .50 | .45 | .48 |
| *h-initial* | .40 | .30 | .35 |
| *V-initial* | .40 | .15 | .28 |
| **Mean:** | **.43** | **.30** | **.37** |

**b. Target item rates (nonword output sentences)**

| Target items | M | F | M + F |
|---|---|---|---|
| *th-initial* | .65 | .70 | .68 |
| *h-initial* | .65 | .70 | .68 |
| *V-initial* | .65 | .55 | .60 |
| **Mean:** | **.65** | **.65** | **.65** |
| **Overall mean:** | **.54** | **.47** | **.51** |

While target items are erroneously transcribed just over a third of the time with real-word output, this rises to almost two thirds with nonword output. This is precisely the pattern of transcription inaccuracy – and concomitant feedback inaccuracy – according to the output type that we anticipated.

Nonetheless, not all of the mistranscriptions supplied the target item, constituting inaccurate feedback. Instead, some mistranscriptions supplied items other than the target, constituting problematic, part-accurate or neutral feedback (Table 5).

**Table 5.** *Error condition: rates of inaccurate transcriptions/feedback types*

| | **Pronunciation feedback** | | | |
|---|---|---|---|---|
| **Target items** | **Inaccurate** (*tank* → *thank,* etc.) | **Problematic** (*tumb* → *thong*) | **Part-accurate** (*teft* → *test*) | **Neutral** (*tud* → *pop*) |
| *th-initial* | .58 | .03 | .20 | .08 |
| *h-initial* | .51 | .00 | .14 | .05 |
| *V-initial* | .44 | .00 | .20 | .01 |

Problematic feedback comes from ASR transcribing other items that start with /θ/, /h/ or a vowel, given target items with these sounds. Only two such inaccurate transcriptions occurred. Both involve th-initial targets (e.g., *Unfortunately, she twisted her* ***tumb*** *→* ***thong***). Inaccurate transcriptions that provide this problematic form of pronunciation feedback are thus extremely rare.

Most of the mistranscriptions resulting in items other than the targets generated part-accurate feedback that indirectly captures the pronunciation error. Indeed, 16 of the

23 mistranscriptions of this sort given a th-initial target provide t-initial items, reflecting the substitution error (e.g., *I wonder who committed the* ***teft → test***). This is likewise the case for 11 of 15 mistranscriptions given h-initial targets and for 16 of 17 mistranscriptions given vowel-initial targets: the items are vowel-initial and h-initial respectively, reflecting the h-deletion and h-epenthesis errors.

The few remaining mistranscriptions involving items other than the target item constitute neutral feedback. That is, the transcribed item contains another sound entirely (e.g., *The book hit the floor with a loud* ***tud → pop***). This is the case for 6 of 23 mistranscriptions of th-initial targets, for 4 of 15 mistranscriptions of h-initial targets, and for 1 of 17 mistranscriptions of vowel-initial targets.

Mistranscriptions of inaccurately realized items are thus often of the false negative type: they erroneously indicate that the speaker has correctly realized the target item. This is particularly the case when the phonetic output constitutes a nonword. Mistranscriptions that supply items other than the target item, however, in most cases supply part-accurate feedback: the transcription provides an item that captures the mispronunciation of the difficult sound, thus indirectly signaling the pronunciation error.

### Discussion/Conclusion

To recap, given accurately pronounced items (no error condition) in a sentence context, ASR in GT is highly unlikely to generate false alarms that would send L2 learners the misleading message that they have made a pronunciation error. Indeed, across 240 correctly realized sentences, only two mistranscribed target items created this false impression. This is highly reassuring from a corrective feedback perspective. Nonetheless, for feedback purposes, it is even more important how ASR transcribes inaccurately pronounced items. Unfortunately, ASR struggled to transcribe pronunciation errors accurately, particularly given nonword (.08) vs real-word (.43) output. The tendency was to transcribe the contextually appropriate target item (overall: .51) rather than the phonetically accurate item, especially given nonword output (.65). Nonetheless, the higher transcription accuracy for some items (notably .85 for F vowel-initial forms) suggests that, even given a sentence context, ASR can at times provide effective feedback. In addition, when ASR transcribed an item other than the output or target item, in most cases the initial sound was captured in the transcription, constituting part-accurate feedback. Teachers and learners who want to target th-substitution errors should be forewarned that ASR experienced greater difficulty correctly transcribing th-initial than h-initial or vowel-initial items, whether in the no error or error condition. The concern that ASR might show a gender bias, however, is not supported: in fact, ASR generally showed higher accuracy when transcribing the female recordings.

In sum, our impression is that ASR probably provides better pronunciation feedback on items spoken in isolation that avoid the influence of contextual cues and especially on minimal pairs that avoid the nonword transcription problem. Future research will investigate this issue further.

## References

Brannen, K. (2011). The perception and production of interdental fricatives in second language acquisition [Doctoral dissertation, McGill University].

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Effects of age of second-language learning on the production of English consonants. *Speech Communication, 16*(1), 1–26.

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching, 46*(1), 1-40.

John, P., & Cardoso, W. (2009). Francophone ESL learners' difficulties with English /h/. In B. Baptista, A. Rauber & M. Watkins (Eds.), *Recent research in second language phonetics/phonology: Perception and production* (pp. 118-140). Cambridge.

Mah, J., Goad, H., & Steinhauer, K. (2016). Using event-related brain potentials to assess perceptibility: The case of French speakers and English [h]. *Frontiers in Psychology, 7*, 1-14.

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. In D. Hovy, S. Spruit, M. Mitchell, E. M. Bender, M. Strube & H. Wallach (Eds.), *Proceedings of the First Workshop on Ethics in Natural Language Processing* (pp. 53-59). Association for Computational Linguistics.

White, E. J., Titone, D., Genesee, F., & Steinhauer, K. (2015). Phonological processing in late second language learners: The effects of proficiency and task. *Bilingualism: Language and Cognition*, 1-22.