Designing Pronunciation Learning Tools: the Case for Interactivity against Over-Engineering

Sean Robertson
University of Toronto
Toronto, ON, Canada
sdrobert@cs.toronto.edu

Cosmin Munteanu
University of Toronto
Mississauga, ON, Canada
cosmin@taglab.ca

Gerald Penn
University of Toronto
Toronto, ON, Canada
gpenn@cs.toronto.edu

ABSTRACT

Paired role-play is a common collaborative activity in language learning classrooms, adding meaning and cultural context to the learning process. This is complemented by teachers' immediate and explicit feedback. Interactive tools that provide explicit feedback during collaborative learning are scarce, however. More commonly, supporting dialogue practice takes the form of computer-aided single-student read-and-record activities. This limitation is partly due to the complexity of processing language learners' speech in unconstrained tasks. In this paper, we assess the value of pronunciation error detection algorithms within a realistic, software-aided, paired role-playing task with beginning learners of French. We found that students' pronunciations improve regardless of the type of error detector employed – even for those using simple heuristics. We suggest that speech technologies for language learning have been too focused on engineering goals. Instead, new interactive designs supporting collaboration may be used to overcome engineering limitations and properly support students' engagement.

ACM Classification Keywords

H.5.2. User Interfaces: Evaluation/methodology

Author Keywords

Computer Assisted Language Learning (CALL); Collaborative Education; Computer Assisted Pronunciation Training (CAPT)

INTRODUCTION

Paired role-play is a staple of modern language learning. Paired role-play appears in task-based language learning [33]. In this pedagogy, students must work together in their new second language to complete a task. The task is structured to make use of the target language, and may explicitly or implicitly draw attention to its structures. It has been observed that language learners will employ a number of communicative strategies that can add meaning to the language learning process [22].

FULL VERSION AT: http://dx.doi.org/10.1145/3173574.3173930
CITE AS: Robertson, S., Munteanu, C., and Penn, G., Designing pronunciation learning tools: the case for interactivity against over-engineering. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, (2018), ACM, 356:1-356:13.

For advanced learners, paired role-play does not require many additional resources - it is sufficient to have learners engage in direct communication (face-to-face or through text or video chat). However, for beginner learners who have yet to develop conversational fluency, some explicit, immediate feedback is necessary. Explicit feedback is known to be beneficial to language learning [8, 34].

Computer-Assisted Language Learning (CALL) applications could take on some of the logistical responsibility of the teacher in the role-playing by providing timely, targeted feedback to language learners. But research and commercial pronunciation assessment products for language learning have been dominated by read-and-record activities that mimic the environment in which those technologies have been trained. The question of whether speech-based CALL technologies can be applied to more realistic language activities, such as paired role-playing tasks, still remains unexplored.ăThis is due in part to a lack of research on interactive design.

In this paper, we explore whether a specific component of CALL applications, the Pronunciation Error Detector (PED), can be used to provide explicit feedback in paired role-playing between beginner language learners. In the Background Section, we outline how PEDs are made and how they have been used in CALL. To judge its efficacy and suitability in supporting more collaborative tasks such as paired role-playing, we designed an experiment (described in the Method Section) that only varies the quality of the PED in order to avoid any confounding effects such as new features or teaching styles. Our results illustrate that all forms of PED can lead to significant improvements in participants' pronunciation over time when embedded in a larger learning context, *regardless of the intrinsic performance of a PED*. The results also suggest that a larger proportion of variability in performance is due to dynamics between participants, rather than between a participant and the application. We discuss how HCI practitioners can avoid the complexity of speech technologies and leverage group dynamics in the Opportunities for HCI Section.

BACKGROUND

It is well known that second language learners are often unable to hear the difference between sound groupings (*phonemes*) that are distinct to native speakers of the target second language [16, 15, 3]. Some distinctions can drastically affect the intelligibility of the learner's speech. Of those surveyed in [6], a majority of teachers agree that pronunciation should be taught to second language learners, but only a minority

have formal training. This justifies the efforts spent in areas such as Computer-Assisted Pronunciation Training (CAPT) (a subfield of CALL). In this section we briefly survey CAPT, discuss how some of its technologies support pronunciation training, and identify the gaps in that support.

CAPT research expanded in the early 1990s with the advent of better acoustic models for automatic speech recognition [14]. Initially focused on special needs groups, CAPT grew largely independently from traditional language learning. Although its achievements were difficult to incorporate into the then-dominant communicative approach to language learning [32], which strongly favoured implicit feedback, empirical evidence in favour of CALL [40], the more relaxed view on metalinguistic feedback in newer pedagogies, and more technologically savvy teachers [17] helped to ease CAPT into the mainstream. More information on the history of speech technologies in CALL can be found in [14].

There are three broad classes of research on CAPT. The first is the classifier type, common in engineering circles, which focuses on designing algorithms that, given some input set, generate an output which closely matches some possible output. In the case of PEDs, the inputs are recording segments and the outputs are scores that label how well the speaker pronounced the target segment. PEDs judge speech according to various criteria (e.g. nativeness/accentedness, comprehensibility, or intelligibility) at various segment granularities (phoneme, syllable, word, prosody, rhythm, or phrase). Overviews of this class of research can be found in be found in [38, 14]. The second class, founded in pedagogical theory, consists of reflections on past experience (e.g. [31, 9]), with the goal of presenting theories or guidelines to educators on how CALL may facilitate learning. The third class can broadly be characterized as HCI-related, mostly in the form of usability studies. While the goal of these studies can vary considerably [19], the usual approach with respect to CAPT is to present a larger CALL application and evaluate it with user satisfaction surveys, expert (teacher) critique, oral exams, or some combination thereof. Golonka et al. [18] provide examples of technologies that have been explored from pedagogical and HCI standpoints.

A number of well-known commercial CALL systems perform some kind of spoken language assessment. *Babbel*[1] appears to use template matching against native speakers' utterances in its PED[2]. *Rosetta Stone*[3] has published on using edit distances at the phoneme level for the purposes of error detection [35]. There has also been CAPT activity within *Duolingo*[4], although details on its implementation are sparse. An overview of commercial products that use PEDs can be found in [38].

A number of research projects build PEDs into their applications, too. *Ville* is talking-head software that implements a suite of pronunciation exercises [37]. It has been evaluated with post-study user surveys. *PLASER* [25], developed for children, contains a number of pronunciation activities, including read-along exercises and minimal-pair speaking exercises. *PLASER* was evaluated by a combination of user and teacher surveys. A 3-month longitudinal assessment of students' pronunciations was also perfromed. However, their improvement was judged only according to the PED's built-in metric. A prototype was developed in [39] that likewise featured a number of pronunciation activities. A usablility study of the prototype, which included eye tracking, click tracking, and success rate of activities, was performed to determine what aspects of the program participants found most useful. *DISCO* [12] has repeatedly been assessed over the years using expert (teacher) and user reviews. In *DISCO*, players engage in dialogues with the application, choosing their responses from a series of written prompts. The player must speak the prompt, at which point a number of pronunciation errors may be highlighted. A predecessor to *DISCO*, called *Dutch-CAPT*, was validated in a month-long assessment [27]. Adult participants' abilities to pronounce Dutch words were assessed by experts at the beginning and end of the experiment. In addition to performing classwork, participants were placed in one of three experimental conditions. In the control, participants received no extra pronunciation training. In another, participants were provided an abridged version of *Dutch-CAPT*, where automated pronunciation feedback was replaced with a simple record-and-playback mechanism. In the final condition, participants were provided the full *Dutch-CAPT*, including the automatic feedback provided by a PED. It should be noted that students were distributed to classrooms non-randomly and that those classrooms had significant differences in pronunciation skill prior to the experiment. The authors found that each experimental condition led to fewer mistakes over time, but with no significant difference between the conditions. The authors did, however, find a significant positive improvement of the PED condition over the record-and-playback condition on certain, "target" phonemes *post hoc*.

There are two critical concerns that arise from the above products and research. First, to the best of our knowledge, pronunciation training only occurs as a read-and-record activity. The activity is more a reflection of the way that PEDs are trained than of an underlying pedagogy. The read-and-record task is built to remove other sources of error, such as unexpected words or false starts. As a result, it fails to account for the cognitive load imposed by linguistic pedagogy, such as having to recall the correct words or form the correct sentence structure. Also, the type of pronunciation errors in read-and-record activities are strongly coloured by orthography. For example, the spelling of the French word "et" could imply to anglophone learners a pronunciation that rhymes with "bet," although the word properly sounds closer to "say." Learners make such mistakes prompted by their reading of the word as opposed to by hearing it. Further, the learner may not understand the phrase if only required to repeat it. The read-and-record task might be suitable solely for pronunciation practice, but it is not for communicative learning.

The second concern relates to the goals of their research. Engineering research tends to focus on the aforementioned PED criteria (e.g. nativeness) independently of how the PED will

[1] https://www.babbel.com/

[2] http://blog.babbel.com/tech-background-babbel-speech-recognition/

[3] http://www.rosettastone.com/

[4] https://www.duolingo.com/

used in the classroom. These are *intrinsic goals*. In contrast, research in pedagogical theory and large-scale user studies are concerned with *holistic goals*, such as long-term pronunciation improvement. The length and breadth of such interventions blur together the effects of many day-to-day activities. Neither set of goals addresses how the CAPT technology supports the task for which it was designed. We call this third set of goals a technology's *functional goals*. By focusing on how CAPT performs in the real world (such as in classrooms), research on the functional goals of an application has more ecological validity than research on the intrinsic goals of a PED. And with a limited scope, its results can be more directly attributed to the technology, rather than to an entire curriculum. This sort of task-oriented assessment is commonplace when developing applications in a myriad of domains at CHI [**?**, **?**, **?**, **?**, **?**, **?**, **?**], but has been conspicuously absent in the case of CALL.

Our experiment addresses the above two concerns by embedding a PED into a classroom role-playing activity. At the absolute beginner level, learners do not have an adequate command of the target language to role-play without explicit instruction from an expert. This and a curriculum modified by an expert in the field, defines a real-world task for a CALL application. Properly supporting role-playing also demands more than just pronunciation training from the application. Crucially, role-playing involves a social, interactive element that is often not accounted for in CAPT research, but could be extremely important when designing applications. We have found that, though even very simple PEDs were beneficial to participants, individual differences and the interactions between participants were much stronger indicators of participant success. Evidence — quantitative and qualitative — suggests that reframing CAPT as an interaction-design problem has the potential to be far more fruitful to students than merely improving classifier accuracy.

METHOD

As mentioned in the *Background* section, there has been little to no evaluation of how CAPT feedback supports classroom tasks. We designed an experiment to explore the efficacy of PED technologies in beginner role-playing tasks. A PED announces a mispronunciation by either labelling, tagging, or scoring part of a recording, so an experiment testing PEDs will be focused on what the PED is labelling. We formulated hypotheses along two axes. We hypothesized that a very simple PED could outperform a state-of-the-art PED in terms of both meeting learning goals (hypothesis 1) and in providing a simple, enjoyable user experience (hypothesis 2). To that end, our experiment uses three "PEDs" as experimental conditions: one based on state-of-the-art engineering; one PED acting as a gold standard by using labels provided by a teacher; and a set of heuristics which are not directly dependent on the audio signal.

Three driving concerns of the experimental design were technical constraints (i.e. facilitating the PED), pedagogy, and that the experiment should be ecologically valid.

PEDs are designed to evaluate recordings of fixed phrases. In order to accomodate for the PED, we needed participants to take turns recording phrases. Fortunately, we could avoid forcing participants to read the phrases by leveraging their limited vocabulary. Nonetheless, the PED defined the way participants would interact with the application.

Category	Choice(number)
Total Participants	36
First Language	English(11), Chinese(10), DualEnglish(7)
Fluency in French (1-5 asc.)	1(34), 2(2)
French Experience	None(30), Software(2), Misc(4)
Median Age	22
Gender	Male(19), Female(17)
Num. Languages Fluent	2(21), 3(8), 1(7)
Hours/Day on Mobile	5(8), 3(6), 2(3), 6(3)

Table 1. Demographic information. "Mandarin" and "Cantonese" are coded as "Chinese." "DualEnglish" refers to those who indicated more than one first language (one of which was always "English").

The specific brand of pedagogy we followed is based on an abbreviated curriculum designed by an expert in the field for an upcoming mobile French-teaching video game. The curriculum is intended to satisfy the criterion for the Common European Framework of Reference for Languages absolute beginner level [29]. Learning occurs by modelling everyday interactions with native speakers, then allowing learners to role-play the dialogues themselves. The pedagogy is mostly communicative [32], though it was modified for this experiment to allow minimal focus-on-form explicit feedback. The experimental procedure was mostly dictated by this pedagogy.

To be ecologically valid, the experiment needs to be evaluated in a realistic setting using realistic measures. The above curriculum defined part of the realistic setting: the learning goals and tasks. In order to provide more realistic support to learners, we decided on a *Wizard-of-Oz* experimental design. The wizard, an expert language teacher, would guide learners from the guise of the application. A *Wizard-of-Oz* design had the added benefit of providing teacher labels for the first experimental condition, and a realistic means by which the pronunciation of paricipants could be assessed.

Wizard

To recruit a wizard for our experiment, we advertised in our university's institute of education for a French second languge teacher with the following qualifications: a strong background of experience teaching French, especially to beginners; enough knowledge of phonetics to identify and fix the pronunciation errors of her students; and exposure to modern mainstream pedagogy. The wizard we eventuallly decided on has a Master's education in Linguistics, had taught university and high school French for 4 years, and is not part of the experimental investigation team.

Participants

Participants were primarily recruited by poster throughout University of Toronto facilities and through Facebook groups. Participants were required to be at least 18 years of age, proficient in English, and have no formal education in French. Pairs

of subjects in the experimental sessions were always booked individually. Pre-study demographic information collected can be found in Table 1.

Procedure

Each experimental session is about one hour long, though pre-, mid-, and post- study materials extend the total time to 1.5 hours. Pre-study materials include a demographic survey, consent forms, and a short practice scenario in English to teach participants how to use the application. Mid-study materials are discussed below. Post-study materials include a short user-experience survey and a debriefing.

The hour-long session itself is divided into three twenty minute sections, each of which is assigned an experimental condition according to a Latin square. Most of the lesson material is non-overlapping, though some of the earliest material reappears in later scenarios. As a result, sections present scenarios in fixed chronological order. After twenty minutes, plus or minus three minutes at an experiment administrator's discretion, participants stop what they are doing and are given five minutes to fill out a very brief multiple choice test, a three question experience survey, and rest. Test and survey results depend in part upon the dependent variables described in the Evaluation Section. As only the earliest material in the first section is re-used in later sections, participants are told to skip ahead to the scenarios in the next section if the previous section is not completed.

Within each section, participants complete a sequence of scenarios. Each scenario consists of a modelling phase and a practice phase. Often, subsequent scenarios will present almost the same material but force participants to switch roles.

In the modelling phase, participants watch a video depicting native speakers of French engaging in a French dialogue. Generally, participants are expected to figure out what is happening by the scenario's title and paralinguistic cues from the dialogue, though some scenarios provide additional instructions in English at the beginning of the video (when prepiloting revealed the scenario to be too difficult to be self-understood). Most of the scenarios involve props (replica fruits, coins, etc.) to better immerse themselves in their roles. Participants are forced to watch the entirety of the video once prior to the practice phase, though they can return to the video at any time afterwards to play back parts of the video.

In the practice phase, participants are expected to repeat the dialogue line-by-line into the application. Participants are further expected to adapt the dialogue slightly to their situation, e.g., by changing the gender of the salutation they use according to that of their partner, the name of the person they are addressing, or, if they are shopping, what items they are shopping for. Though the interactions are necessarily simple to accommodate beginner learners, dialogue changes are intended to obviate participants simply memorizing the dialogues. Participants record their utterances line-by-line and receive feedback from the application. Given that participants are newly exposed to the language, there are very few phrases that could be considered valid for a given turn in the dialogue at their level. Part of the feedback they receive is an "acceptance" or "rejection," which dictates whether the dialogue can proceed, or whether the participant must first fix an issue with that dialogue turn.

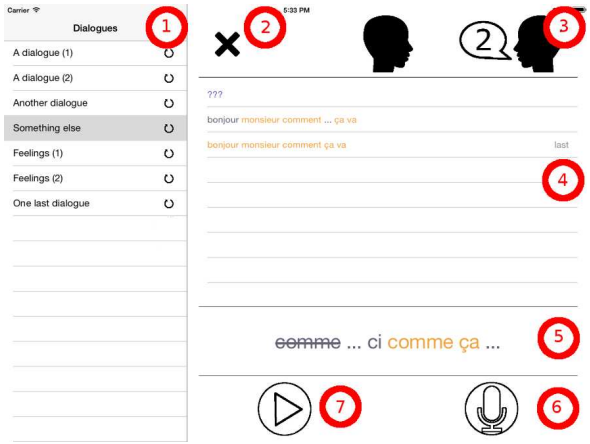

Figure 1. Client interface. See text for description.

While this occurs, the wizard decides on the feedback to present to participants in real time. She administers the feedback through a desktop interface that is connected to the participants' iPad through a local area network. During piloting, she was allowed to watch participants' reactions to her feedback. During the actual experiment, she is sequestered in another room, having access only to audio. She is informed by text message when a section is over, at which point she fills out a 1-5 ranking of each participant's pronunciation aptitude.

An administrator - one of the authors - oversees each experimental session. His goals are to ensure each section lasts approximately 20 minutes, to ensure that participants are not under undue stress, and to answer any questions relating to the application. He is not permitted to provide help regarding lesson content. Like the wizard, he is unaware in advance of the order in which experimental conditions are presented.

INSTRUMENTS

Client Interface

Figure 1 depicts the iOS interface presented to participants. It consists of: *(1)* an ordered list of all the scenarios; *(2)* a cross or checkmark indicating whether the last utterance was accepted or rejected; *(3)* whose turn it is to speak (players 1 and 2 are assigned prior to the first section); *(4)* a history of the feedback presented to participants during the active scenario; *(5)* word-level feedback for the last utterance; *(6)* a press-and-hold button to record an utterance; and *(7)* a button to open iOS' default video playback interface to play the current scenario's modelling video. Figure 2 shows how the user engages the application to complete a scenario. Each utterance recorded by a participant will result in three forms of feedback: an accept/reject icon, a sound corresponding to either the acceptance or rejection, and word-level feedback. The accept/reject and word-level feedback are determined by the wizard. The wizard may choose to display three question marks (???) instead of word-level feedback. Though the primary use of the question marks is as a catch-all for invalid

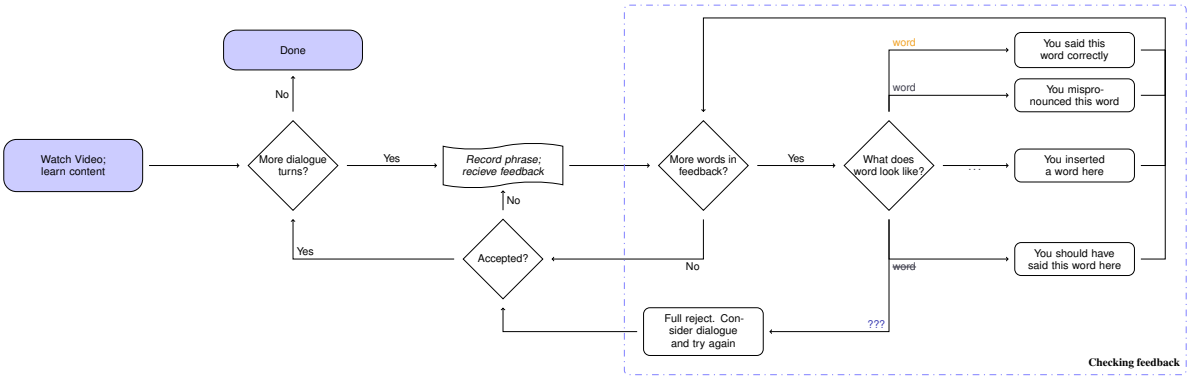

Figure 2. Scenario flow chart from the user perspective.

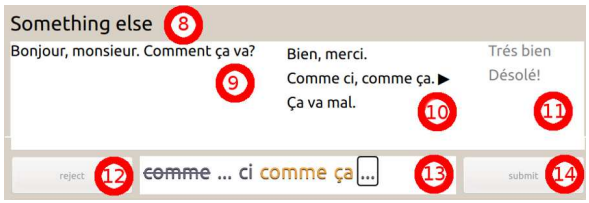

Figure 3. Wizard interface. Doctored to remove whitespace.

phrases, it can also be used to hide valid phrases from participants if they are trying to "game" the system into presenting them with a correct response. In word-level feedback, a valid phrase is highlighted as follows:

- An orange word indicates that it was pronounced correctly.
- A grey word with no strikethrough indicates that it was pronounced incorrectly.
- A grey word with a ~~strikethrough~~ indicates that the word was omitted, i.e., it should have been uttered at that point,
- Grey ellipses (. . .) indicate that one or more additional words were uttered at that point which should not have been.

Colours were chosen to be visibly distinct to those with red-green colour blindness.

In addition, a word that is part of the word-level feedback can be tapped to play a recording of a native speaker of French saying that word. This feature was added because participants must otherwise skip through the video to reach the target word. It is also a form of recast, a common type of focus-on-form feedback in communicative classrooms [28].

Wizard Interface

The wizard's interface, depicted in Figure 3, allows for rapid online feedback to participants. It consists of: *(8)* the scenario name; *(9)* the most recent utterance accepted in the scenario; *(10)* an inventory of valid phrases for the current dialogue turn; *(11)* the phrases that the currently selected valid phrase will lead to in the next dialogue turn; *(12)* the "reject" button; *(13)* the modification palette; and *(14)* the "submit" button. Though mouse control was provided while the wizard familiarized herself with the interface during prepiloting, the primary means of interaction is through keyboard shortcuts.

The interface is enabled once a participant begins recording an utterance. The wizard hears the ongoing utterance with less than a second of latency and can begin formulating her feedback during recording, though she cannot submit the feedback until recording has finished. She first chooses a valid phrase as a template *(10)*. If she rejects without picking a template, then it is considered a full rejection. The selected template appears in the modification palette wherein the wizard chooses to label some words as mispronounced or deleted, and then adds inserted word indicators. The wizard then chooses whether the word-level feedback should be accompanied with an accept or reject. In the special case where the wizard rejects the utterance but all words remain labelled correct, the system treats the feedback as a full rejection but shows all words to the participant. This allows the wizard to tell participants how to proceed when they are very stuck. This feedback is unlikely to be confused with actual word-level feedback since it likely follows a sequence of full rejections.

Feedback Mechanism

Figure 4 illustrates the phases of feedback generation.

In the first phase *I*, the participant records his attempt at the dialogue turn.

In the second phase *II*, the wizard is given a list of phrases appropriate to that turn. At the wizard's discretion, she may fully reject the utterance. Doing so bypasses the next phase and sends the participant the oblique "???" feedback. If the utterance is not fully rejected, the wizard modifies a valid phrase with inserted, deleted, and mispronounced labels. She also chooses whether to accept or reject the phrase. Within this restricted mechanism, the wizard is again given full discretion over what to label and whether to accept or reject.

In the third phase *III*, the PEDs calculate their feedback. The PEDs are provided with the wizard's labelled transcription and accept/reject decision from phase *II*. However, neither PED is privy to which, if any, words were labelled mispronounced. If the third phase is reached, then the feedback from all PEDs is always calculated, regardless of the actual experimental condition. This prevents the participants and wizard from detecting the conditions due to increased or decreased processing times. After calculations, the PEDs return their guesses of which

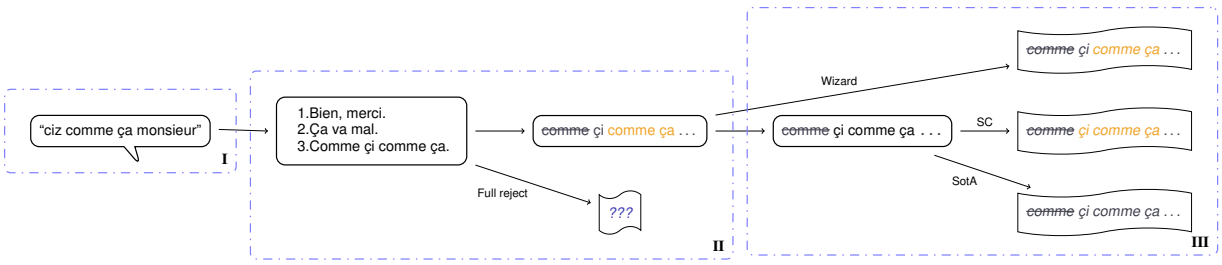

Figure 4. Phases of feedback generation. See text for details.

words were mispronounced. The feedback ultimately returned to the participants varies according to experimental condition.

One of the goals of this experimental design is to hide the experimental condition from the wizard. There are two reasons for this. First, one of the measures collected relies on the wizard's perception of participant improvement, which may be altered if she was aware that her feedback was being overridden or tampered with. Second, the wizard might have changed her teaching style to accommodate what she perceives as an unreliable feedback mechanism, which would have invalidate the comparison between conditions. Instead, the wizard controls the flow of the dialogue and reliably provides word-level feedback. By hiding the results of phase *III*, the wizard cannot reliably distinguish a mismatched label from an obstinate participant.

Pronunciation Error Detectors

As mentioned previously, two PEDs in addition to the wizard herself are tested as experimental conditions. The first is based on a state-of-the-art PED which determines from acoustic data whether words are pronounced natively or non-natively. The second is a heuristic model which uses bootstrapped scores from phonetics literature combined with assumptions about the dialogue state to guess which words are mispronounced. The PEDs determine which words to label correct and which to label mispronounced.

The State-of-the-Art (SotA) PED trains per-word Gaussian mixture models over all instances of that word in a corpus as background (prior) models. It then adapts two new models from the background model: one for native and one for non-native word instances. For a new word and its corresponding audio, SotA guesses the label by discovering the model that would sample that word with greater likelihood. In short, SotA decides whether words are mispronounced according to whether they are closer to a prototype native or prototype non-native recording. SotA was trained using pilot data and expert labels, described in [30]. Word segments are computed online through Viterbi alignment of the audio data with Pocketsphinx [21].

The heuristic model, called the Sum-Context (SC) model, decides which words to label as mispronounced based on a series of rules and a bootstrapped scoring mechanism derived from phonetics research literature. Given a sequence of words, each word is given a difficulty score by summing the scores of its phonemic transcription. Scores are derived from a few papers which specifically study French vowels often confused by anglophone learners. The other part of the model is the rules, which are based on assumptions about the dialogue state. They are, in brief:

1. There is no need to label mispronunciations in accepted utterances - participants are unlikely to search for errors in this case.
2. If there are inserted or deleted words, then do not label any words as mispronounced. Participants will focus on the insertions/deletions first.
3. Otherwise, sort the words in a list according to score, select the worst, and label all instances of that word as mispronounced. If the same phrase is rejected again immediately afterwards, select the second worst, etc. If the phrase is accepted or a new phrase is rejected, reset the list.

SC only labels words as mispronounced when there is no other possible source of mistake, then picks one word to label as mispronounced. Generally that word is the most difficult one by score, but after repeated rejections it will attempt other words. More implementation details can be found in our supplementary document.

The motivation behind SC is not to provide correct, immediate feedback to learners, as per our hypotheses. The scoring mechanism is very crude and, importantly, does not analyze any audio recordings at all. Instead, SC operates under the assumption that it is more important that learners engage in critical self-evaluation of their pronunciation than it is to provide the right feedback. Note that, even with SC, the proposed CAPT application would not be free of all audio processing: something would have to take the role of the wizard in recognizing speech and accepting or rejecting utterances. However, the remaining roles could be filled by bootstrapping an off-the-shelf speech recognizer, such as Google Speech[5], to a decision-making algorithm.

In [30] it was found that the effectiveness of the PEDs tested were highly sensitive to tuning their labelling thresholds. In order to make SotA as competitive as possible, SotA's thresholds were tuned to maximize agreement with wizard labels on a few sessions of pilot data. This courtesy, an unrealistic boon to SotA, was not extended to SC.

[5] https://cloud.google.com/speech/

Evaluation

Tests for significance and effects are conducted using Multi-level Linear Modelling (MLM). MLM is a powerful scheme of linearly modelling dependent variables with independent variables. MLM extends the notion of grouping repeated measures (such as over subjects in Repeated Measures Analysis Of VAriance (RM-ANOVA)) to an arbitrary number of grouping variables. For our experiment, we wished to capture both individual and paired effects. By using MLM's hierarchy of "levels," we can specify that paired effects (level 3) can only influence the dependent variables (level 1) indirectly by influencing the individual (level 2). Lastly, unlike RM-ANOVA, MLM can handle violations of sphericity. More information on MLM can be found in [36].

After every section in a session, we measured seven dependent variables related to participants' experience and performance over the last section. These variables were chosen after discussion with our pre-pilot wizard (a trained teacher) as well as our own observations. Likert-scale variables reflect user affect related to motivation [5]. Others are analogue to standard classroom assessments.

Average Rejections Per Accept (AvgRej): The wizard may halt participants' progress through the scenarios by rejecting utterances until a participant addresses some problem with them. *AvgRej* is the average number of times that utterances are rejected per dialogue turn, measured over the last section.

Number of Scenarios Completed (NumScen): The number of scenarios completed in the last section. Whereas *AvgRej* focuses on improvements at the utterance level, including pronunciation and grammar, *NumScen* captures the rate at which participants are exposed to new concepts.

Number of Quiz Questions Correct (NumQuiz): 5 questions on vocabulary and semantics are administered to participants at the end of each section pertaining to that section. The questions were written beforehand by the wizard and can be answered unambiguously. Points are not deducted for spelling.

Harder than Last (Hard): Post-section, participants are asked to rate between 1 and 5 inclusively how much they agree with the statement, "The scenarios have been harder since my last break," where 1 is full disagreement.

More Confusing than Last (Confusing): Same as *Hard*, but with the statement, "The feedback I've recieved since my last break has been more confusing."

Enjoyment: 1-5 agreement with the assertion, "I am enjoying myself." Unlike *Hard* and *Confusing*, this rating could be collected for the initial segment.

Wizard Pronunciation Score (PronScore): After every section, the wizard determines a score for each player. The score is from 1-5, a score of 1 indicating "most words are still causing him or her issues," and 5 indicating "all words pronounced well; errors very rare." This is the closest to a direct measure of pronunciation skill that we collect. The wizard participated a number of prepilot sessions in which she could hone her assessment.

Tests of the significance of fixed effects in MLM are performed with F-tests, just like Analysis Of VAriance (ANOVA). Fixed effects are the constant contribution of an independent variable. By analogy, if we measure the effect of studying for some test versus not at all, the fixed effect would be some constant improvement in test scores. In this experiment, we model the fixed effects of the choice of PED and the section number (both as level 1 predictors). The effect of studying would likely be corrupted by a number of random effects, such as the student, the time of day, etc. In this experiment, we treat the participant (level 2) and group (level 3) IDs as random effects. In RM-ANOVA, an F-test is perfomed by taking the ratio of the mean-squared error of a linear model *including* the repeated measures (fixed) variable over one that *excludes* the fixed variable. Likewise, the MLM F-test compares the error of a model with a fixed effect over a model without it.

Since we are interested in only the main effects of variables, we perform Type-II tests of significance (F_{Cond} and $F_{Section}$, resp.). We use Kenward-Roger approximations of degrees of freedom [24]. If the results are significant, we look at the fixed (constant) effects of the section variable and each factor of the experimental condition. As is always the case with a factored experimental design, one factor gets absorbed into the intercept of the model. We chose the wizard condition to model the intercept; fixed effects of experimental condition are therefore relative to the wizard's condition. The fixed effects of variables are reported as β, with a margin of error of $\widehat{\sigma}$. β can be interpreted as the difference in the measured variable relative to either the wizard condition or the first section. β and $\widehat{\sigma}$ are only calculated when fixed effects reach significance.

Though there is no significance test for random effects, we can measure how "important" the effect is by its standard deviation, σ. ρ measures the interclass correlation coefficient of a random variable, which acts similarly to ANOVA's partial η^2.

MLM analysis was performed using the Linear Mixed-Effects package `lme4` [2], Restricted Maximum Likelihood fitting, and with a power level of $\alpha < 0.05$. Prior to analysis, data were screened for univariate skewness and kurtosis at power $\alpha < .001$. *AvjRej* had highly positive kurtosis ($z = 36.2, p = .000$). *AvjRej* ($z = 12.2, p = .000$) and *NumQuiz* ($z = 5.19, p = .000$) had highly positive and negative skewness, respectively. Attempted transformations did little to improve the biases; dependent variables were analyzed untransformed for interpretability. There was only one instance of a univariate outlier at $\alpha < .001$, and only for *AvgRej*. The specific formulae for the MLM analysis can be found in our supplementary materials.

In addition, we collected the PEDs' and wizard's labels. As mentioned in the *Background* section, the standard method of evaluating PEDs is to compare generated labels to a "correct" set. This method can be applied here by considering the wizard labels the correct set. We used Cohen's Kappa (κ) to measure the degree of agreement between wizard labels and each of the PEDs' labels. κ allows us to judge whether the speech engineer's method of evaluation predicts the real-world *PronScore*.

Variable	$F_{Section}$	F_{Cond}
AvgRej	8.5**	.37
NumScen	41***	.30
NumQuiz	7.3**	2.1
Hard	2.0	1.3
Confusing	1.6	.79
Enjoyment	2.2	3.4*
PronScore	43***	2.9†

Table 2. Type II F-tests of significance against measured variables. † $p \in (.05, .06)$, * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

PED	κ
SotA	.186
SC	.344

Table 3. Cohen's κ between Wizard and PED labels.

	$\beta(\widehat{\sigma})$		
Variable	*SotA*	*SC*	*Section*
AvgRej	-	-	$-.12(.04)$**
NumScen	-	-	$-1.1(.17)$***
NumQuiz	-	-	$-.33(.12)$**
Enjoyment	$-.06(.11)$	$.22(.11)$†	-
PronScore	$-.31(.17)$	$-.39(.17)$*	$.56(.08)$***

Table 4. Fixed effects and standard errors. 2-tailed t-tests of significance are included. See warning in text about analyzing *PronScore* effects. † $p \in (.05, .06)$, * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Variable	$\sigma_{indiv}(\rho)$	$\sigma_{pair}(\rho)$
AvgRej	$.31(.42)$	$.09(.03)$
NumScen	-	$.98(.49)$
NumQuiz	$.67(.29)$	$.07(.00)$
Hard	$.58(.29)$	$.38(.12)$
Confusing	$.79(.38)$	$.54(.18)$
Enjoyment	$.64(.64)$	$0(0)$
PronScore	$.67(.38)$	$.48(.19)$

Table 5. Random effects of individuals and pairs.

RESULTS

Table 2 shows a significant contribution of experimental conditions only on *Enjoyment*. There was no significant effect of *Section* on any of the user experience variables (*Hard,Confusing*, and *Enjoyment*), but a significant effect on all the rest. We performed a fixed-effect analysis of PED on *PronScore* because: a) the F-test was marginally significant, and b) it is highly relevant to hypothesis 1. However, those fixed effects (listed in Table 4) can only be considered part of a post-hoc analysis - *PronScore* did not actually meet significance.

Table 4 shows that participants tended to enjoy SC feedback around .2 to .3 Likert scale points more than Wizard or SotA feedback. SC and SotA feedback had similarly negative effects on *PronScore*, leading to scores around .3 to .4 points lower than the wizard's. However, only the negative effect of SC reaches significance. This relationship is not predicted by measuring the agreement between wizard and PED labels; according to Table 3, SC is in far greater agreement than SotA with the wizard. All significant *Section* effects from Table 2 are also significant in Table 4. As the section number increases, participants complete fewer scenarios and get fewer quiz questions correct. However, their responses are accepted faster and their pronunciation score improves over time as well. Table 5 shows modelled variation due to participants and pairs. Most measures have considerable variation between participants – a standard deviation between participants is often close to a full point on the variable's scale. The model also shows considerable, though smaller, variation due to group dynamics.

DISCUSSION

Improving pronunciation is one of the core goals of second language learning, and as such, of CAPT research. While our study was designed to investigate the role of PED quality in supporting pronunciation training within an interactive context, we first discuss here some observations and analyses related to our participants' short-term improvements to pronunciation performance.

Clearly, from Table 4, there is a significant positive effect of section on *PronScore*. This is easily interpreted to be an overall improvement in pronunciation over time. We did not find a significant effect of the experimental condition on *PronScore*, which means we do not have enough evidence to accept (or reject) the first hypothesis. Here, we explore the patterns in the experimental data to uncover other factors influencing the relationship between PEDs and *PronScore*.

As mentioned in *Results*, we calculated the fixed effects of experimental conditions on *PronScore*, despite not reaching significance. Table 4 suggest SC and SotA had a similarly negative effect on *PronScore*, though only the SC-Wizard relationship reached significance with a 2-tailed t-test. Though the absolute difference in effect size is very small (.08 points), small differences in performance may accumulate over time. A post-hoc F-test checking for significant interactions between section and condition gave none ($F(2,83) = .67$). Nonetheless, we plot *PronScore* over condition and time to search for indications of future performance.

The left-hand plot of Figure 5 shows the means and 95% confidence intervals of *PronScore*, distributed over section and condition. The positive effect of time on instruction is clearly seen in this graph: subsequent sections per condition have mostly non-overlapping confidence intervals. However, the raw measures do not compensate for individual or pair effects. The right-hand plot of Figure 5 normalizes the means and variances of *PronScore*s by participants. The method, described in [10, 26, 1], emulates the sort of within-subjects normalization that occurs in RM-ANOVA.

The most noticeable difference between PEDs and the clearest benefit of wizard labels is observed in the second section. This could explain the significant negative effect of SC listed in Table 4. Each activity in the second section began by reading aloud a unique place name (the only reading in the experiment). The wizard remarked that these place names were often too difficult for beginner learners to say. Given that each place name occurs once, SC's strategy of moving through possible mistakes one at a time is likely detrimental. However, by the third section, scores in each condition have almost converged.

Participants tended to enjoy the SC condition the most. As neither *PronScore* nor *NumQuiz* reached significance over experimental conditions, it is unlikely that participant enjoyment was tied to participants' real performances. Indeed, no post-study survey questions related to user satisfaction were

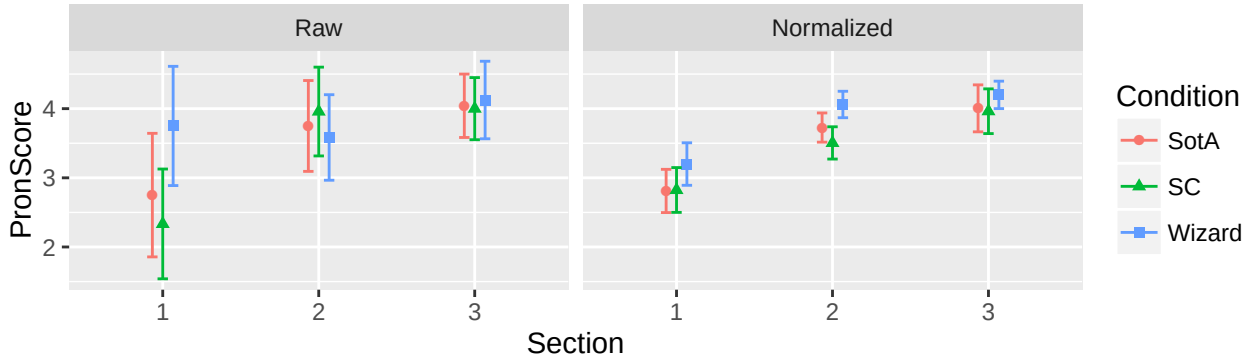

Figure 5. ***PronScore*** **over time and experimental condition. Error bars are 95% confidence intervals. See text for a description.**

significantly correlated to *PronScore*. It is also unlikely that participants were more frustrated by the quality of feedback in the wizard and state-of-the-art conditions since *Confusing* was so far from significance. There was no decrease in (and even a slight increase in) the number of rejections in the SC condition, which makes it unlikely that participants percieved faster progress in the SC condition. The implementation of SC would seem to suggest that only a select few words would be labelled mispronounced. However, the distribution of words labelled mispronounced was almost identical to the wizard condition. This means that participants were unlikely to have benefitted from attending to a smaller set of words (which should have impacted *PronScore* anyway). The most likely answer to the increased enjoyment we have found is that, in total, there were significantly fewer words labelled mispronounced in SC conditions ($\mu_{SotA} = 57.8$, $\mu_{Wiz} = 35.9$, $\mu_{SC} = 23.05$). If true, this result points to the importance of judicious application of feedback, perhaps to only target phonemes [27], and only some of the time. This further reinforces the need for HCI research into the functional goals of CALL.

While our short-term study does not establish the superiority of one PED over another, significant improvements over time *do* show the value of the computer-assisted role-playing task in itself. This post-hoc analysis shows that pronunciation scores are improving regardless of the quality of the PED. In other words, there is evidence that an HCI practitioner can design a CALL application that supports pronunciation feedback by only providing coarse, utterance-level analysis of audio. There is even some evidence suggesting that participants enjoy these systems more, in line with our second hypothesis (albeit of marginal size). We also note that the measure used in engineering PEDs, the κ, predicts the entirely opposite relationship than the one suggested by MLM. It is therefore not obvious how simple, offline evaluations such as a Cohen's Kappa could be of use in predicting future support for this collaborative task. This is of particular relevance to HCI practitioners and interaction designers who are attempting to build pronunciation support tools but face the barriers of laboriously constructing and fine-tuning PEDs or other aspects of the underlying speech technology.

Though our experimental results help to break down perceived technological barriers that keep HCI practitioners out

Assertion	Mean	Median	Stddev
I received help from my partner	4.0	4	1.2
I would have preferred learning alone	2.3	2	1.1
I was comfortable around my partner	4.2	4	0.97

Table 6. All post-study survey results relating to partners. Participants indicate between 1 and 5 how much they agree with an assertion, 1 being strongly disagree.

of CAPT, the values in Table 5 point to a much more potent source of variability (and opportunity) in the experiment: the pair of students (learners). Our experiment was specifically designed to investigate PEDs within the context of paired role-playing. The post-study survey results (Table 6) indicate that participants appeared to have generally preferred having partners. However, determining whether the paired aspect of the interaction contributed positively or negatively to learner goals would require a separate investigation and a different experimental setup (outside the scope of the present study). What we can analyze, though, is how important the collaborative aspect of a learning support interface is when compared to the choice of PED.

Table 5 shows how much variability in the measured variable was attributed by the model to either individual or pair effects. While the random effects are by definition unpredictable, they give some idea of how the choice of one group or another impacts the variable. Interestingly, the variable with the greatest magnitude random pair effects (*NumScen*, *Hard*, *Confusing*, and *PronScore*) can all be related to the shared responsibility of completing scenarios. That is, they can reflect the effectiveness of the dyad's collaborative problem solving. In contrast, *NumQuiz* and *Enjoyment* relate to the the understanding and experience of the individual. The above interpretation meshes well with our informal observations that participants treated the task more as a puzzle to solve than a legitimate role-play. Unfortunately, if true, this means the application went beyond the role of providing feedback to driving the task.

Interpretation aside, comparing the fixed effects of PED in Table 4 to the random effects of pairs in Table 5, there is a clear possibility that leveraging group dynamics is more important than employing accurate speech technologies in improving participants' pronunciation. This is encouraging

for human-computer interaction: getting the interaction design right with respect to supporting collaboration between all parties involved is critical, and designers do not need to wait for perfectly-accurate PEDs. We discuss the opportunities presented by these findings in the following section.

OPPORTUNITIES FOR HCI

As this experiment has shown, the goals of CAPT speech technologies can be out of touch with realistic expectations. We have seen that, within the limited scope and length of this study, that a baseline technology with only coarse-grained processing can still lead to improved pronunciation in learners. The authors of [11] also found that recognition accuracy did not matter much in their task. They argued that doing the activity was more important to learning than the content of the activity. In other words, the speech technology is secondary to the role it takes in the learning process.

We believe that the fundamental question of CAPT is not how to use a specific technology to help teach pronunciation. The greater question is of *how* technology *could* be used to help teach pronunciation. If speech technologies are tools for CAPT (an analogy proposed in [13]), then the difference is between designing tools to help with a goal, and finding a use for a tool that one already has. Design is the principal concern, which means, contrary to the ambivalence exhibited in the past, CAPT is in the domain of HCI.

Consider our experimental setup. In order to facilitate the technological constraints of the PED, we forced learners to record each utterance individually. Instead, had we started fresh from the question "How can we facilitate explicit pronunciation feedback in classroom-based paired role-playing tasks?" then we might have tried to leverage the fact that an expert source of feedback, the teacher, would be present in the classroom. Instead of trying to accurately analyze speech and present feedback directly, we could focus on sampling real-time audio clips of different pairs to the teacher so that she may provide explicit feedback to many pairs at one time. In such a design, randomly sampling students (or giving priority based on teacher discretion) may be sufficiently useful to teachers, bypassing the need for complicated speech technologies altogether. This may or may not be enhanced by output from speech technologies that could aid the teacher (e.g. by more effectively making timely decisions about which students to sample based on pronunciation rankings determined by the speech technologies running in the background).

Of particular interest to CHI is the dynamism of interaction between pairs of participants that was not exploited in our experiment. Since CAPT is often framed as an independent study tool [4], group dynamics have been broadly ignored. One standard deviation of the random effects of pairs in Table 5 is greater than the fixed effect of PEDs in Table 4. We informally observed a strong social pressure to perform when a participant's partner was highly motivated. We found significant pairwise differences in post-study reports of participants' comfort with and the helpfulness of their partners based on their age ($p < .05$). There are also well-known cultural barriers to overcome that could impact the efficacy of group work [7, 20]. A potentially great asset in making CAPT a collaborative venture is that partners can help return some of the depth and complexity of spoken interaction when a CAPT system inevitably simplifies it.

This is not to say that other types of research are not necessary to the development of CAPT systems. A PED must be trained on some criterion that is immediately measurable. Also, as the *Dutch-CAPT* example shows us [27], they may be very effective in the read-and-record tasks. Wholistic, long-term studies can show us the long-term effects of applications when they interact with other language learning activities. Nonetheless, defining the role of a CAPT technology is integral to its evaluation. By communicating the roles of a technology to teachers, they are in better positions not only to assess their long-term feasibility, but step in when the technology is clearly failing in its task. This, in turn, eases the burden of the designer from building a "good" system to one that is "good enough" [23].

CONCLUSIONS

While speech technologies have improved dramatically in recent years, their use in language learning applications has largely been limited to "read-and-record" activities. Teaching pronunciation continues (justifiably) to be more effectively conducted in the context of classroom instruction, through activities such as paired role-playing. Speech-enabled Computer-Assisted Pronunciation Training (CAPT) applications do not yet provide adequate support to such environments despite a need for them, particularly in larger-size classrooms. We investigated the suitability of CAPT speech technologies, specifically Pronunciation Error Detectors (PEDs), for providing feedback in role-playing tasks for beginner learners of French. Our ecologically valid experiment provides evidence that sentence-level processing with no fine-grained audio processing can support the language learning process. We propose that the design of CAPT systems to perform a specific task within an interaction between two learners is far more important to learning outcomes than technological goals, making the field perfectly ripe for HCI intervention.

ACKNOWLEDGEMENTS

This research was funded by an Ontario Centres of Excellence Technical Problem Solving grant in partnernship with Speax Inc., and a Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council.

REFERENCES

1. Thom Baguley. 2012. Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods* 44, 1 (2012), 158–175. DOI: http://dx.doi.org/10.3758/s13428-011-0123-7
2. Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. DOI:http://dx.doi.org/10.18637/jss.v067.i01
3. David Birdsong. 2007. Nativelike pronunciation among late learners of French as a second language. In *Language experience in second language speech learning: in honor of James Emil Flege*. John Benjamins Publishing, 99–116.

4. Stephen Bodnar, Catia Cucchiarini, Bart Penning de Vries, Helmer Strik, and Roeland van Hout. 2017. Learner affect in computerised L2 oral grammar practice with corrective feedback. *Computer Assisted Language Learning* 30, 3-4 (2017), 223–246. DOI: http://dx.doi.org/10.1080/09588221.2017.1302964

5. Stephen Bodnar, Catia Cucchiarini, Helmer Strik, and Roeland van Hout. 2016. Evaluating the motivational impact of CALL systems: current practices and future directions. *Computer Assisted Language Learning* 29, 1 (2016), 186–212. DOI: http://dx.doi.org/10.1080/09588221.2014.927365

6. Judy Breitkreutz, Tracey Derwing, and Marian Rossiter. 2001. Pronunciation Teaching Practices in Canada. *TESL Canada Journal* 19, 1 (2001), 51–61. DOI: http://dx.doi.org/10.18806/tesl.v19i1.919

7. Barbara Burnaby and Yilin Sun. 1989. Chinese Teachers' Views of Western Language Teaching: Context Informs Paradigms. *TESOL Quarterly* 23, 2 (1989), 219–238. DOI:http://dx.doi.org/10.2307/3587334

8. Susanne Carroll and Merrill Swain. 1993. Explicit and Implicit Negative Feedback. *Studies in Second Language Acquisition* 15, 03 (1993), 357–386. DOI: http://dx.doi.org/10.1017/S0272263100012158

9. Ray Clifford. 1998. Mirror, Mirror, on the Wall: Reflections on Computer Assisted Language Learning. *CALICO Journal* 16, 1 (1998), 1. http://search.proquest.com/docview/750443820?accountid=14771

10. Denis Cousineau. 2005. Confidence intervals in within-subject designs: A simpler solution to Loftus and Massons method. *Tutorials in Quantitative Methods for Psychology* 1 (2005), 42–45. http://www.tqmp.org/Content/vol01-1/p042/p042.pdf

11. Gabriel Culbertson, Solace Shen, Erik Andersen, and Malte Jung. 2017. Have Your Cake and Eat It Too: Foreign Language Learning with a Crowdsourced Video Captioning System. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 286–296. DOI: http://dx.doi.org/10.1145/2998181.2998268

12. Joost van Doremalen, Lou Boves, Jozef Colpaert, Catia Cucchiarini, and Helmer Strik. 2016. Evaluating automatic speech recognition-based language learning systems: a case study. *Computer Assisted Language Learning* 29, 4 (2016), 833–851. DOI: http://dx.doi.org/10.1080/09588221.2016.1167090

13. Farzad Ehsani and Eva Knodt. 1998. Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm. *Language Learning & Technology* 2, 1 (1998), 45–60.

14. Maxine Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication* 51, 10 (2009), 832 – 844. DOI: http://dx.doi.org/10.1016/j.specom.2009.04.005

15. James Emil Flege. 1987. The production of new and similar phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of phonetics* 15, 1 (1987), 47–65.

16. James Emil Flege, Ocke-Schwen Bohn, and Sunyoung Jang. 1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics* 25, 4 (1997), 437 – 470. DOI: http://dx.doi.org/10.1006/jpho.1997.0052

17. Nina Garrett. 2009. Computer-Assisted Language Learning Trends and Issues Revisited: Integrating Innovation. *The Modern Language Journal* 93 (2009), 719–740. DOI: http://dx.doi.org/10.1111/j.1540-4781.2009.00969.x

18. Ewa M. Golonka, Anita R. Bowles, Victor M. Frank, Dorna L. Richardson, and Suzanne Freynik. 2014. Technologies for foreign language learning: a review of technology types and their effectiveness. *Computer Assisted Language Learning* 27, 1 (2014), 70–105. DOI: http://dx.doi.org/10.1080/09588221.2012.700315

19. Saul Greenberg and Bill Buxton. 2008. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Florence, Italy, 111–120.

20. Guangwei Hu. 2002. Potential Cultural Resistance to Pedagogical Imports: The Case of Communicative Language Teaching in China. *Language, Culture and Curriculum* 15, 2 (2002), 93–105. DOI: http://dx.doi.org/10.1080/07908310208666636

21. D. Huggins-Daines, M. Kumar, A. Chan, A.W. Black, M. Ravishankar, and A.I. Rudnicky. 2006. Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, Vol. 1. I–I. DOI: http://dx.doi.org/10.1109/ICASSP.2006.1659988

22. Yu-Wan Hung and Steve Higgins. 2016. Learners use of communication strategies in text-based and video-based synchronous computer-mediated communication environments: opportunities for language learning. *Computer Assisted Language Learning* 29, 5 (2016), 901–924. DOI: http://dx.doi.org/10.1080/09588221.2015.1074589

23. Matthew Kay, Shwetak N. Patel, and Julie A. Kientz. 2015. How Good is 85%?: A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 347–356. DOI: http://dx.doi.org/10.1145/2702123.2702603

24. Michael G. Kenward and James H. Roger. 1997. Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics* 53, 3 (1997), pp. 983–997. http://www.jstor.org/stable/2533558

25. Brian Mak, Manhung Siu, Mimi Ng, Yik-Cheung Tam, Yu-Chung Chan, Kin-Wah Chan, Ka-Yee Leung, Simon Ho, Fong-Ho Chong, Jimmy Wong, and Jacqueline Lo. 2003. PLASER: Pronunciation Learning via Automatic Speech Recognition. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing - Volume 2 (HLT-NAACL-EDUC '03)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 23–29. DOI:http://dx.doi.org/10.3115/1118894.1118898
26. Richard Morey. 2008. Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology* 4, 2 (2008), 61–64. http://www.tqmp.org/Content/vol04-2/p061/p061.pdf
27. Ambra Neri, Catia Cucchiarini, and Helmer Strik. 2006. ASR-based corrective feedback on pronunciation: does it really work?. In *Interspeech 2006*. 1982–1985. http://www.isca-speech.org/archive/archive_papers/interspeech_2006/i06_1372.pdf
28. Howard Nicholas, Patsy M. Lightbown, and Nina Spada. 2001. Recasts as Feedback to Language Learners. *Language Learning* 51, 4 (2001), 719–758. DOI: http://dx.doi.org/10.1111/0023-8333.00172
29. Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press. https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680459f97
30. Sean Robertson, Cosmin Munteanu, and Gerald Penn. 2016. Pronunciation Error Detection for New Language Learners. In *Interspeech 2016*. 2691–2695. DOI: http://dx.doi.org/10.21437/Interspeech.2016-539
31. M. R. Salaberry. 1996. A Theoretical Foundation for the Development of Pedagogical Tasks in Computer Mediated Communication. *CALICO Journal* 14, 1 (1996), 5. http://search.proquest.com/docview/750317215?accountid=14771
32. Sandra J. Savignon. 1987. Communicative Language Teaching. *Theory into Practice* 26, 4 (1987), pp. 235–242. http://www.jstor.org/stable/1476834
33. Peter Skehan. 2003. Task-based instruction. *Language Teaching* 36, 1 (2003), 1–14. DOI: http://dx.doi.org/10.1017/S026144480200188X
34. Nina Spada and Yasuyo Tomita. 2010. Interactions Between Type of Instruction and Type of Language Feature: A Meta-Analysis: Type of Instruction and Language Feature. *Language Learning* 60, 2 (2010), 263–308. DOI: http://dx.doi.org/10.1111/j.1467-9922.2010.00562.x
35. Theban Stanley, Kadri Hacioglu, and Brian Pellom. 2011. Statistical Machine Translation Framework for Modeling Phonological Errors in Computer Assisted Pronunciation Training System. In *ISCA Workshop on Speech and Language Technology in Education*. Venice, Italy. http://project.cgm.unive.it/events/SLaTE2011/papers/Stanley-mt_for_phonological_error_modeling.pdf
36. B.G. Tabachnick and L.S. Fidell. 2012. *Using Multivariate Statistics*. Pearson Education, Limited. http://books.google.ca/books?id=ucj1ygAACAAJ
37. Preben Wik, Rebecca Hincks, and Julia Hirschberg. 2009. Responses to Ville: A virtual language teacher for Swedish. (2009). http://academiccommons.columbia.edu/catalog/ac:160205
38. Silke M. Witt. 2012. Automatic error detection in pronunciation training: Where we are and where we need to go. In *Proc. of the International Symposium on Automatic Detection of Errors in Pronunciation Training (ISADEPT)*, Vol. 6. 1–8.
39. Ping Yu, Yingxin Pan, Chen Li, Zengxiu Zhang, Qin Shi, Wenpei Chu, Mingzhuo Liu, and Zhiting Zhu. 2016. User-centred design for Chinese-oriented spoken english learning system. *Computer Assisted Language Learning* 29, 5 (2016), 984–1000. DOI: http://dx.doi.org/10.1080/09588221.2015.1121877
40. Yong Zhao. 2003. Recent developments in technology and language learning: A literature review and meta-analysis. *CALICO journal* 21, 1 (2003), 7–27. DOI: http://dx.doi.org/10.1558/cj.v21i1.7-27