PressToPronounce: An output-oriented approach to mobile language learning

by

Kaiying Liao

S.M., Massachusetts Institute of Technology (2014)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Master of Engineering in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2015

© Massachusetts Institute of Technology 2015. All rights reserved.

Author .................................................................
Department of Electrical Engineering and Computer Science
January 30, 2015

Certified by ............................................................
Prof. Randall Davis
Professor of Computer Science and Electrical Engineering
Thesis Supervisor

Certified by ............................................................
Dr. Darren Edge
Microsoft Research Asia, Researcher
Thesis Supervisor

Accepted by ............................................................
Prof. Albert R. Meyer
Chairman, Masters of Engineering Thesis Committee
PressToPronounce: An output-oriented approach to mobile language learning

by

Kaiying Liao

Submitted to the Department of Electrical Engineering and Computer Science on January 30, 2015, in partial fulfillment of the requirements for the degree of
Master of Engineering in Computer Science and Engineering

Abstract

Learning a second language has become popular not only for people who wish to communicate with those who do not speak their native language, but also for professionals who are required to speak to colleagues around the globe. Unfortunately, learning a language requires time and practice that is difficult to fit into many people’s daily schedules. Currently, mobile applications are available to help language learners study “on-the-go;” however, most mobile-based exercises focus on learning to understand input, specifically on memorizing vocabulary, and fail to support users in developing their output skills, such as writing and speaking.

In this thesis, I describe the design and implementation for PressToPronounce, a mobile application that helps native Mandarin Chinese speakers practice English speaking skills. Ideas from second language acquisition studies and persuasive technology strategies guided the design of short, focused, user-centered exercises to help learners develop their competency of the target language and improve their performance through practicing pronunciation. The application also takes advantage of Computer-Assisted Pronunciation Training (CAPT), a state-of-the-art pronunciation analysis tool that provides quantitative speech and pronunciation feedback to support users in tracking their progress. Preliminary user tests for the system showed that users were enthusiastic about output-oriented exercises, and given the opportunity, would continue to use the application to practice their speech.

Thesis Supervisor: Prof. Randall Davis
Title: Professor of Computer Science and Electrical Engineering

Thesis Supervisor: Dr. Darren Edge
Title: Microsoft Research Asia, Researcher
Acknowledgments

I would like to acknowledge Hong Z. Tan and Darren Edge, my supervisors at Microsoft Research Asia. I learned a tremendous amount about research in human-computer interaction as well as finding balance between research and product development when working on research in industry.

I would also like to acknowledge Prof. Markus Zahn and Kathy Sullivan for their work in the VI-A program. In addition, I want to thank Eric Chang for supporting the VI-A program at Microsoft Research Asia, where I had the opportunity to spend 9 months in Beijing while working on my thesis project.

I sincerely thank Professor Randall Davis for being my MIT faculty supervisor in the program. I am extremely grateful that he agreed to be my supervisor despite his busy schedule.

I would also like to thank Professor Seth Teller for supporting me as my adviser, instructor, and inspiration throughout my time at MIT.

Finally, I want to thank my friends and family for all of their support.
## Contents

1 Motivation 9
1.1 Second Language Acquisition .................................. 10
   1.1.1 Definitions .................................................. 10
   1.1.2 The Output Hypothesis ................................... 10
   1.1.3 Usage-Based Grammar Theory ............................. 12
1.2 Vision ............................................................. 12

2 Computer-Assisted Language Learning 15
2.1 Mobile Microlearning in SLA .................................. 15
2.2 Persuasive Computing .......................................... 17
   2.2.1 The Seven Tactics in Persuasive Computing ............ 17
   2.2.2 Persuasive Computing in MALL Research ............. 18
2.3 Pronunciation-driven Design ................................. 19
   2.3.1 Computer-assisted Pronunciation Training ............ 20
   2.3.2 Sentence-Based System ..................................... 20

3 Mobile Application Design 23
3.1 Overview .......................................................... 23
3.2 Exercise Design .................................................. 24
   3.2.1 Target Words ................................................. 24
   3.2.2 N-Grams ....................................................... 25
   3.2.3 Sentences ...................................................... 29
3.3 Feedback for Users ............................................. 30
Chapter 1

Motivation

In the era of globalization, learning a second language has become a necessity for many people in their careers, especially if they are relocated to other countries or need to communicate with colleagues in other parts of the world. Just looking at the statistics for the number of English speakers, both native and non-native, shows how common it is for people to learn and use another language. According to Encyclopaedia Britannica, there are currently approximately 1 billion people who can speak English total, out of which 665 million people speak the language non-natively \[21\]. The British Council reports that by 2020, there would be around 2 billion people speaking and/or studying English \[4\]. China specifically had the largest number of people learning English in 2012 \[4\]. People are very welcoming of resources to help them improve their language skills, so developing technology that is accessible and easy to use could benefit many learners.

Creating resources for second language learners goes beyond developing software that caters to a range of students. The capacity of technological resources to help people effectively learn a language, particularly to speak it, needs to be considered in depth. I turn to research in Second Language Acquisition to guide my understanding of the way people learn languages and find suggestions for how their learning can be supported.
1.1 Second Language Acquisition

Second Language Acquisition (SLA) is the study of how people learn a language outside of their native language. It looks beyond memorization of lexicon and grammatical structure and instead explores more general concepts from which students and teachers of any language can glean insight.

To understand how I can create a resource to help students learn a target language, I look at some broad principles of SLA.

1.1.1 Definitions

The following list defines some basic terms in SLA. Different words may be used to describe the same concepts depending on the expert, but the following will be used throughout this thesis.

**Input**  Input is everything the learner is exposed to; what the learner hears and reads [7].

**Output**  Output is anything that the learner produces in the target language; what the learner says and writes [7].

**Competence**  Competence is the first step of output and refers to the knowledge and understanding of a language. This involves consciously understanding grammatical structures [3].

**Performance**  Performance is the second step of output and describes the “actual use of the language” [3]. Specifically for spoken output, this includes the pronunciation, the intonation, and the overall fluency of speech.

1.1.2 The Output Hypothesis

The Output Hypothesis claims that “the act of producing language (speaking or writing) constitutes, under certain circumstances, part of the process of second language
learning” [20]. The circumstances refer to a user’s realizations that gaps of knowledge may be present while trying to produce output. Becoming aware of these gaps and working to fill those gaps can help a user learn more about the target language itself. This claim emphasizes that production is a strong indicator of how fluent a learner is in the target language. Simply understanding the meaning behind input is not enough to support fluid interaction with others in the target language, which is often the goal of the learner. Thus, my main focus and motivation for this thesis project is to provide a resource for second language learners to improve their output and thereby improve their fluency in the target language, as suggested by the Output Hypothesis.

**Output Apprehension**

A lack of both competence and performance of a target language can lead to apprehension, especially for a learner who is not confident of his speaking skills. Situations that call for speech do not afford much time to find the right vocabulary and parts of speech to make a learner’s speech both grammatically correct and true to the intended meaning. This pressure makes competence a crucial part of output, and without sufficient competence in the language, speaking is not only difficult but often nerve-wracking. Apprehension can also arise during the performance phase of speech, where a learner may be worried about his accent or pronunciation of certain words. Nuances between similar sounds that are difficult to differentiate between may occasionally lead to being misheard by others, which could result in embarrassment or miscommunication.

Apprehension in speaking is often a result of the complexity inherent in language, which tends to be noticed by a non-native speaker but often feels irrelevant to a native speaker. The Usage-Based Grammar Theory in SLA provides possible explanations and offers suggestions for how second language learners may be able look past the complexity of the target language and produce output similar to that of native speakers.
1.1.3 Usage-Based Grammar Theory

Usage-Based Grammar Theory suggests that practicing prefabricated sequences improves fluency more effectively than consciously learning all of the grammatical rules of a language because frequently used combinations of words could be stored as units in memory in addition to individual words. Thus, much of the process of finding words and fitting them into the appropriate grammatical structures is simplified to finding the proper “chunks” of lexicon to put together.

Phrases and expressions are much more efficient to piece together than individual words. As stated by Bybee, “selection of pre-formed word sequences provides not only native-like word combinations but also enables native-like fluency since words are selected more efficiently in groups rather than individually” [19]. Not only are lexical chunks shown to be more efficient, the idea that “native-like fluency” can be established through practicing chunks is supported through studies of prefabricated sequences. These studies show that during output, “55 percent of word choices [are] determined lexically by occurrence in [prefabricated phrases]” [19]. Therefore, if a speaker can store a wide variety of chunks and the contexts in which they are used, his speech will become both more efficient and fluid.

1.2 Vision

Despite the fact that learning a second language is difficult and time-consuming, suggestions from SLA provide significant evidence that exercises for studying and practicing effectively in short amounts of time can be created. I focus on developing a resource that is flexible enough for people to use for a few minutes at a time, given that many people who want to learn a second language do not have extended periods of time to study to reach their desired level of fluency. A significant amount of the free time that people find throughout the day often comes during periods of waiting, e.g. on public transportation or in line at shops. These chunks of time can be turned into opportunities for microlearning [8], and a mobile application best suits this need. Thus, I want to answer the following question:
How can we support spoken output for a second language based on suggestions from Second Language Acquisition by leveraging the qualities of mobility?

I developed PressToPronounce, a mobile application that guides the building of language from single units to progressively larger chunks through micro-exercises while providing feedback on pronunciation. It currently supports English as the target language for native Mandarin Chinese speakers, but it is designed to generalize to any target language and native language combination.

The main aspect of PressToPronounce that separates it from other English practice mobile applications is the focus on both competence and performance of spoken output. To support competence of the language, exercises are created for the user to think about lexical chunks and how they are formed before they are used in context. To support performance, state-of-the-art pronunciation analysis tools are used to provide quantitative feedback to the user. The system is also adaptive in that the combination of the results from both steps of language output sets the pace of the exercises to fit the needs of the user.
Chapter 2

Computer-Assisted Language Learning

With the ease of access to computers and smart phones, using technology as a resource to improve language skills is becoming more common. Computer-assisted language learning (CALL) has been around for many years, but has mainly focused on recall of vocabulary. During CALL exercises, user input is often limited to multiple choice or typing in translations between the user’s native language and the target language. Sometimes, the exercises do not require input at all and rely on user’s self-reported success in recall. Despite extensive training on such systems, many learners may still feel uncomfortable in settings with native speakers.

By practicing with technology that focuses on output, learners can improve and potentially change their attitude about communicating with others, which will further their skills in the target language. Incorporating mobile microlearning and persuasive computing into an output-oriented resource is one approach to supporting language learners.

2.1 Mobile Microlearning in SLA

CALL technology specifically developed for mobile devices is known as mobile-assisted language learning (MALL). MALL applications have been popular recently because
they are accessible and easy to use as immediate references, particularly in situations where people want to communicate in a language they are not fluent in. These applications are also good for practicing and studying during gaps when people have a tendency to check or use applications on their smart phones. The tasks presented through MALL applications are short and therefore support microlearning [8].

Many MALL applications are flashcard based systems. Flashcard systems require minimal amount of text on a display, which consequently give users small amounts of information to consume quickly in a single setting. Pleco is a good example of a flashcard-based mobile application for English speakers to learn Chinese [18]. Users can set both the number and the subject of flashcards to study in a given session, and sessions may be paused and restarted at the user’s convenience. Users may purchase dictionaries as well as extensive audio recordings of words through Pleco. While this application is excellent for exposing users to a variety of inputs, there is essentially no support for output as mastery of language knowledge is determined by the number of consecutive correct responses.

Outside of flashcard exercises, there are other applications that also focus on building vocabulary. Dearman and Truong created a mobile interface called Vocabulary Wallpaper that displays content through a live wallpaper, which users can view and interact with while the phone’s display is active [5]. Vocabulary Wallpaper provided many opportunities for "implicit exposure and learning" because the user did not have to open an application to see content as it continued to run in the background [5]. Like Pleco, this application centered on exposure to input, which is only part of the language acquisition process.

Even more applications exist to help users learn vocabulary in a second language, but most do not go beyond encouraging the user to memorize definitions and short phrases. For a user who is looking to improve fluency in their output of the target language, there is a clear limit to how much these applications can help them improve, as there are no writing or speaking components involved. To develop a resource appropriate for supporting output, I look beyond typical mobile language learning applications and take ideas from persuasive computing.
2.2 Persuasive Computing

Persuasive technology can be defined as “an interactive product designed to change attitudes or behaviors or both by making a desired outcome easier to achieve” [13]. Using tactics from persuasive computing would help a learner overcome their apprehension about learning and practicing a second language. For spoken output, persuasive technology can help a learner practice more so that usage of the language is increased and understanding can be further solidified. As mentioned before, second language learners are typically exposed to large amounts of input material, but the output is often limited to what the speaker is comfortable forming.

Part of the reason for this varied level of comfort and therefore potentially limited amount of output is due to personality. Gass and Selinker specifically state that “personality factors such as confidence in one’s ability to produce correct target language sentences may influence whether or not a learner produced target language material” [7]. People may be shy about practicing in front of others. However, practicing output with technology may be a way to build the confidence needed to communicate fluently in the target language. By using technology as a persuasive tool, my project aims to encourage output from a second language learner so that fluency and pronunciation in speaking may be improved.

2.2.1 The Seven Tactics in Persuasive Computing

There are seven main persuasive computing tactics, all of which relate to how SLA can be supported through technology [13].

1. *Simplification* takes a complex sentence and breaks it down into its lexical chunks so that the user thoroughly understands patterns and structures present in the language.

2. *Tunneling* creates a guided experience that leads users through a “predetermined sequence of actions or events,” which provides users with adequate instructions in each exercise.
3. *Self-Monitoring* helps a user track his actions and progress, which promotes user motivation and learning.

4. *Suggestion* proposes behaviors at opportune moments to give users options for a course of action rather than force the user to perform specific tasks.

5. *Reinforcement* uses positive feedback to stress aspects of the user’s strengths by presenting tasks involving words and structures that the user is familiar with to encourage motivation and engagement.

6. *Tailoring* describes the process of identifying and analyzing a user’s needs to present new material while providing further practice and instruction in aspects that need improvement, which presents a productive experience to the user.

7. *Surveillance* makes the user aware of being monitored throughout each exercise and therefore encourages users to perform well.

### 2.2.2 Persuasive Computing in MALL Research

Persuasive tactics are not new to MALL and have been incorporated into research on mobile systems. The following are examples of these systems that have combined aspects of both persuasive technology and mobile microlearning effectively.

The first example is MicroMandarin, an application that supports microlearning of Mandarin Chinese through contextual flashcards [9]. It takes advantage of the user’s location using Foursquare so that people can learn new vocabulary in the context of their surroundings. For example, when a learner is at a cafe, MicroMandarin caters its flashcard set to words related to items seen in a coffee shop, such as “latte” or “cappuccino.” Through location as context, MicroMandarin uses suggestion as a persuasive tactic to help learners form associations and remember words and phrases more easily. In addition, the context is a way of personalizing the experience for the user and acts as a form of contextual tailoring, another one of the seven persuasive tactics.
Tip Tap Tones aims to improve tone recognition in Mandarin Chinese, an aspect of listening that is often the most difficult for non-native speakers to understand [10]. Its manifestation as a mobile game motivates learners to use the application as well as practice their listening skills more frequently. The brevity of each individual game is also a good way for users to take advantage of short breaks, making it an effective microlearning tool. The major persuasive tactics present in Tip Tap Tones include reinforcement and simplification. Reinforcement is implemented through the increased difficulty of items and tasks presented as the user improves in differentiating between tones. Tip Tap Tones uses simplification by focusing on individual syllables rather than entire phrases that are difficult to process.

MemReflex is a mobile application that uses flashcards as a method of studying Chinese characters, targeted towards an English-speaking audience [11]. This algorithm builds on prior approaches to spaced repetition [23][16][17], but explicitly targets satisfying learning experiences from short, sparse learning sessions as is characteristic of mobile interaction. The precise guidance through scheduled exercises is an implementation of both tunneling and tailoring in which the system provides tasks via constrained material that are adapted specifically to help the user.

2.3 Pronunciation-driven Design

I begin the discussion on output-oriented design by focusing on pronunciation for two main reasons. First, there are no commonly used software applications available to the public that provide quantitative feedback for pronunciation. Creating an application that provides such feedback can benefit many second language learners, especially those who are motivated to practice their pronunciation skills. The second reason is that the Speech Group at Microsoft Research Asia currently has a state-of-the-art speech analysis tool that can provide quantitative pronunciation feedback. The tool uses deep neural networks to produce accurate scores for speech according to how closely the speaker's pronunciation resembles the accent-free pronunciation of accomplished speakers. Most developers do not have access to this technology and
therefore could not give users feedback on their speech even if they designed an application for such a feature. Therefore, a significant portion of PressToPronounce is designed around pronunciation as one of its most crucial components.

2.3.1 Computer-assisted Pronunciation Training

Computer-assisted Pronunciation Training (CAPT) is state-of-the-art speech analysis technology that can accurately provide quantitative pronunciation feedback to second language learners. CAPT technology created using deep neural networks has been developed by Microsoft Research and deployed internally as part of the Bing Dictionary service.

CAPT takes a string of text along with a wave audio file as input and provides a raw score for the speech given in the wave file. The raw score between 0 and 1 gets converted to a percentile score based on how close the pronunciation of the sentence is to a standard, clear American accent. A score of 99% indicates that the speech recorded in the audio file is closer to a standard accent than 99% of non-native English speakers’ accents.

The output from CAPT is not limited to a single percentile score of the entire string of text. Each component of the text is evaluated at the word, syllable, and phoneme level. In addition, the start and end times of each component within the audio file are given, rounded to the closest microsecond.

The challenge remains for how to systematically develop a person’s language knowledge and skills using such technology. By combining CAPT with a well-designed mobile interface, second language learners can bring their speech and pronunciation skills to a higher level.

2.3.2 Sentence-Based System

Microsoft Research Asia currently uses an interface called 我爱英语, literally “I love English” and romanized as Wo Ai Ying Yu that is used to demo the CAPT system’s ability to assess a person’s pronunciation of sentences in English. The generic interface
that includes the basics of *Wo Ai Ying Yu* will be referred to as the “sentence-based system.”

The sentence-based system provides the simplest interface for using CAPT to assess a user’s pronunciation. It gives the user a sentence and asks him to read it aloud. CAPT evaluates the speech and returns feedback by showing a numerical score for the overall pronunciation of the sentence. In addition, the sentence itself is color-coded according to the raw scores of individual words compared to other non-native speakers, which red indicating below average, yellow indicating average, and green indicating above average.

Because the goal of the sentence-based system is to demonstrate the main components of the CAPT system, it is too simple to be an interface that second language learners can use effectively as a supplement for improving their pronunciation. Through brief interactions with some non-native speakers who have tried the sentence-based system, I have taken note of a few major features that are missing for this system to be a good resource.

1. The system does not keep track of users and therefore does not track user history. Users cannot compare their scores from previous uses of the system unless they track feedback on their own.

2. Some users want to improve their pronunciation of words that they perform poorly on; however, the system does not provide suggestions for how to improve.

3. Even if a user sees that he needs to improve pronunciation of certain words and has practiced speaking them, he is unable to use the system to check his score for those particular words because sentences are provided at random from a set.

In addition, this system focuses solely on pronunciation, which is only a single aspect of speech. As mentioned before, fluency can be conceived of in terms of two main components: competency and performance. Without sufficient competency and understanding of the language and its structures, simply being able to pronounce words clearly will not improve a learner’s fluency as the learner has no content to say.
Thus, a system that incorporates tasks aimed at improving both aspects would be more beneficial than a system that focuses only on one.

In summary, the sentence-based system is clearly not designed to support the user through the process of developing fluency in speech. However, it is a good baseline for pronunciation-driven design to start with and to build upon to develop a resource that can effectively help a second language learner practice and improve their output in the target language.
Chapter 3

Mobile Application Design

3.1 Overview

PressToPronounce guides users through exercises that can help solidify their competence of English and improve their pronunciation. The exercises come in sets, which include tasks for both studying individual words and practicing phrases and sentences.

Each set centers around a target word that is shown initially in a flashcard-like warm-up exercise. Afterwards, phrases and sentences containing this target word appear in word sequence and speaking exercises to enforce the concept of learning language through chunks. At the end of each set, the application displays a summary of the user’s performance with options to browse more detailed scores and feedback. The entire process of going through a single exercise set can take as little as a minute for a veteran user or up to four minutes for a new user, which is the typical amount of time spent to complete mobile microlearning tasks.

This application was designed for users who have already had some background in English and is meant to bolster a student’s learning, not to replace instructors. Moreover, it is a practice tool for spoken output, similar to how audio recordings supplement listening comprehension and stories supplement reading comprehension. Details on the exercises presented throughout the application show how PressToPronounce incorporates an output-oriented approach to reinforce speaking skills.
3.2 Exercise Design

To address the missing components from the sentence-based system as described in the previous chapter, I design exercises around word sequence ordering, listening, and speaking. The specific exercises depend on the length of the chunks, specifically if the chunk is a word, an n-gram, or a sentence.

3.2.1 Target Words

As stated in the overview of the design, exercises are put together into sets. Each set consists of a target word and three phrases or sentences. The target word is the focus of the set — it shows up in each of the phrases or sentences so that the user can understand how to use the word in context and practice saying it fluidly with other words.

At the beginning of each set, the target word is shown to the user in a flashcard-like exercise, with two buttons that the user can interact with. The first is a play button that provides text-to-speech audio for the user to hear how the word sounds, the second displays the definition for the word.

When the user presses the “Show Definition” button, the target word’s definition will appear in English. Because there is a likelihood that the user does not understand the word or the English definition, the user may toggle between English and Chinese definitions.

I call the target word exercises flashcard-like because it gives the opportunity for the user to test his memory of the word’s meaning without context. However, it is not a strict flashcard in which the user is required to report to the system whether he remembered the word and its definition correctly. The exercise can either be used briefly to jog the user’s memory or be studied for a longer length of time depending on the user’s style of learning. At the very least, it indicates that the current word will be the focus of the next few exercises in the set.
3.2.2 N-Grams

N-grams are combinations of \( n \) words that often appear together in a specific order in both spoken and written language. For example, in the English language, "hot coffee" is a common 2-gram while "on my way" is a common 3-gram. In character-based languages such as Chinese, the n-grams are created from character-based units rather than word-based units. Examples of a 3-gram and 4-gram in Chinese are “发短信” (fa duan xin) and “从此以后” (cong ci yi hou) which mean “send a text message” and “from then on” respectively. N-grams generally have a specific meaning attributed to them when used in context and therefore provide a good source of lexical chunks for a language learner to study.

To help users solidify their use of English chunks through PressToPronounce, two tasks are given to the user for each n-gram in a set.
Task 1: Word Sequence Ordering

The first task is to provide the correct order of a given set of words. The words in an n-gram are scrambled and displayed as buttons, and the user must recreate the appropriate n-gram by pressing the correct buttons in order before moving on to the next task.

The interface for the exercise helps the user so that even if the user has never encountered the n-gram or other n-grams with a similar pattern previously, the user can use brute force to guess the correct order of the words. Words are selected one at a time, and feedback is given immediately to avoid overwhelming the user. The user starts by pressing the button containing the word he thinks should go first. If he has selected correctly, the button flashes green and the word is revealed in the n-gram at the top of the page. The button is disabled so that the user sees it has already been added to the n-gram. If the selection is incorrect, the button flashes red, indicating
Figure 3-3: An example of the word sequence exercise for a 3-gram.

to the user that he should choose another word. This goes on until all of the words in the n-gram appear together in the correct order at the top of the page.

**Task 2: Speech and Pronunciation**

After the word sequence task is complete, the application displays a study page for the n-gram. The n-gram is color-coded by word, similar to the sentence-based system. Green, yellow, and orange indicate above, near, and below threshold pronunciation scores respectively. However, the colors are shown according to the user’s previous pronunciation of the words rather than the performance of the current speech exercise that the user is asked to perform. This feedback gives the user awareness of which words he has pronounced well and which words need improvement, so that he can work towards producing better results on the current exercise.

Before beginning the speech exercise, the user can review the definitions of each word by tapping on them and also listen to the words pronounced independently of
Figure 3-4: An example of the word sequence exercise for a 3-gram as the user correctly selects the second word in the sequence. The green background of the button for the word “any” shows that it is the correct choice. The first word of the 3-gram “by” was already selected correctly by the user and shown on the top half of the screen. Its corresponding word button has been disabled so that the user does not choose it again for the remainder of the exercise.

one another. If the n-gram is new and the user wants to understand the meaning behind it or see how it is used, there is a “Show Examples” button that takes the user to a web search page with results for the n-gram. This feature is added to help users understand the meaning and use behind the n-gram as it often does not match the aggregate definitions of the independent words.

The speech exercise itself is simple; the user records himself saying the sentence while pressing down the “Hold + Speak Phrase” button. This function mimics the “Hold to Talk” button for sending audio messages through mobile chat applications like WeChat. After recording, the user can listen to the recording of his speech. If he has made a pronunciation error or wants to try saying the phrase again, he can record himself again until he is ready to submit the recording for feedback. The exercise is

28
completed when the user presses the continue button to move on to the next n-gram.

3.2.3 Sentences

Word ordering exercises work well with n-grams, but are not compatible with most sentences. It is very difficult to judge the correctness of a particular sequence of words in a sentence because meaning and grammar both have to be taken into consideration. To keep the tasks within the scope of microlearning exercises, the system skips the word sequence task and brings the user directly to the pronunciation task. Like the n-gram study page, the sentence is shown at the top of the page where each word is color-coded according to the user’s pronunciation from previous exercises. Definitions can be reviewed by tapping on the words, and a “Show Examples” button is available
for the user to see how the sentence may be used in context. The speaking exercises is completed after the user records himself saying the sentence aloud and presses the "Continue" button at the bottom of the page.

### 3.3 Feedback for Users

One of the main drawbacks of the sentence-based system was the lack of clear feedback despite the fact that CAPT is able to produce precise quantitative numbers for pronunciation. To support users’ learning and encourage improvement, PressToPronounce provides both feedback that can be understood quickly within sessions of microlearning and detailed numbers that can show progress and improvement over time.

After each set of exercises, users can see a summary of their scores from the tasks they just completed. Scores are shown out of 100 as an average of the word sequence ordering exercise scores and an average of the pronunciation score percentile of the n-grams shown in that set. The average scores for the previous set are juxtaposed with both the average scores of the past 10 sets and the average scores over their entire history. Displaying the scores as a bar graph emphasizes the comparison of scores over time so that the user can measure his improvement at a glance.

Users can also choose to see a detailed history of their pronunciation scores, which give the percentile score of each of the past twenty n-grams and sentences. Despite the fact that the percentile scores on the Pronunciation History page already show far more concrete feedback to the user compared to the feedback given in the sentence-based system, the user can choose to look at even more specific pieces of information regarding their past performances. If a user taps on any of the words shown in the n-grams, the application displays a graph that shows how the user’s average pronunciation score for that specific word has changed over each of the user’s spoken exercises containing that word.

After the user has looked at his detailed pronunciation history, he can go back to the score summary page of his previous exercise set and either conclude his output
Figure 3-6: An example of the score summary page for a set of exercises. The user can compare the mean scores for pronunciation and word sequence of the past set with the mean scores of the past 10 exercises as well as the mean scores overall. PressToPronounce provides just enough concrete detail that the user can analyze his performance according to concrete numbers, but also summarizes much of the information compactly so that the user can understand his progress quickly in between microlearning sessions.

3.4 Sources for Target Language Content

Target language content used in PressToPronounce is stored as words, n-grams, and sentences. I describe the source of the content in the following sections.
3.4.1 Words

Words come from the British National Corpus and are ranked by frequency of appearance in both written and spoken language [15]. The most frequently seen 5000 words were taken from the corpus and then filtered down to 4890 words after duplicate and potentially offensive words were removed. In PressToPronounce, we assume that words encountered more frequently in the corpus are “easier,” while more “difficult” words are those seen less frequently. Both English and Chinese definitions taken from Bing dictionary are stored with each word.

Out of the words kept in the database, only the words ranked at 1000 and above are potential target words. While it is possible to make any word a target word in the PressToPronounce system, using an article or preposition as a target word does not support vocabulary exercises well as their use is better understood when combined in an n-gram or sentence. Moreover, PressToPronounce is designed for users who
already have some background in the language as it does not teach basic grammatical
rules that are needed to complete the easiest word sequence exercises.

3.4.2 N-Grams

Data for n-grams \((n = 3, 4, 5)\) and their frequencies are gathered from the Google
Research n-gram corpus [2]. To supplement the frequency data, I also calculate the
mutual information score for each n-gram, the logarithm of the probability of encoun-
tering the specific n-gram divided by the product of the probability of encountering
the words alone. A higher mutual information score implies that the words are more
likely to appear in combination than individually.

To fit n-grams into the system, we filter them by two criteria. The first filter
removes all of the n-grams that contain words outside of the 4890 word set, thereby
limiting the scope of the language presented. The second filter removes n-grams that
are permutations of one another so that users do not encounter chunks that have
multiple correct word sequence orderings. After applying both filters to the set, the
top 100 n-grams by highest frequency are added to the set of n-grams stored in the
database.

There is clearly a disadvantage to filtering the n-grams that have permutations
because it removes an entire subset of n-grams that should be learned. I go into more
depth on this topic in my discussion of further work on the system.

3.4.3 Sentences

Sentences are taken from a text simplification data set from research on lexical simpli-
fication [14]. They do not have frequency or mutual information data like the n-grams
do, but the simple sentence set provides short sentences that are easy to understand
out of context and are appropriate for second language learners to practice.
Chapter 4

Adaptivity of System

In order to help users learn and improve through these short, condensed interactions, the exercises and their contents need to be specific to the user. As such, there is no fixed order of words or n-gram. The system incorporates various levels of adaptivity to determine the content and construction of each exercise set.

4.1 User Adaptivity

For a system to be adaptive, it must be aware of the user’s actions and performance. PressToPronounce constantly records and updates user information to present exercises with the most appropriate content that supports the user’s learning experience.

4.1.1 Word Level

Users begin with an initial word level that is close to the number of words a user is comfortable using in the target language. A user can start at level 1000 and go up to level 4890 for the current set of words stored in the PressToPronounce system. These levels map directly onto the rank of the words in the system, such that if the user begins at level 3000, the system assumes the user is competent enough to include any of the 3000 most frequent words in his English output.

As the user practices with the system more often, he is given words above his
word level to expand his vocabulary and improve his comfort level with words that are encountered less often in the English language. How quickly he can move through exercises to become exposed to new words in the system depends on his performance, which is evaluated relative to thresholds set for his output level.

4.1.2 Thresholds

To adapt to the user, thresholds are determined by the user’s performance history. The exercises that get evaluated against a quantitative threshold are the word sequence exercises and the pronunciation thresholds. Word sequence exercises are scored according to the following equation.

\[
\text{Word Sequence Score} = \left( 1 - \frac{\text{number of mistakes made}}{\text{total number of possible mistakes}} \right) \times 100
\]

The \textit{number of mistakes} is the number of incorrect guesses that the user can make while trying to construct the n-gram, and the \textit{total number of possible mistakes} is the number of times a user can make a mistake in the worst possible scenario.

Pronunciation scores are based on the percentile scores returned from the CAPT system after analysis of the user’s speech. Both word sequence and pronunciation scores are taken out of 100.

A sliding window of the pronunciation scores and word sequence scores contribute to determining the mean score and standard deviation of the user’s recent performances in each exercise. While setting the thresholds to be equal to the mean scores would be simple and easy, the system actually sets the threshold to be lower. Specifically, the threshold is set to 0.674 standard deviations below the mean, which effectively gives the user a 75% chance of success if success is defined as scoring above threshold. By giving the user a greater chance to succeed, the user has more opportunity to progress. Setting a slightly lower threshold also provides a buffer for mistakes, which is necessary when each exercise set only provides three opportunities to score well in each task and takes the average score.
4.2 Progression

There are two types of progression that the user experiences through PressToPronounce, both of which adjust according to the user’s performance. The first is the micro-progression or exercise progression, the movement through tasks over a course of a single exercise set that dictates the user experience. The second is the macro-progression or chunking progression, the progression of difficulty in the exercises the user encounters over time and how difficulty levels are adapted to the user.

4.2.1 Micro-Progression: Exercise Progression

As noted, a single typical exercise set consists of a flashcard-like vocabulary exercise for the target word followed by three n-gram exercises or three sentence exercises. An n-gram exercise includes both a word sequence ordering task and a speaking task while a sentence exercise consists of only the speaking task.

The set always begins with the target word in a vocabulary exercise as single words represent the lowest level of the chunking spectrum. By starting with the target word with a mostly passive exercise, the system can refresh the user’s memory or teach a definition quickly before beginning other tasks that require more thought and action.

For a user who has been exposed to the majority of the words in the system, the progression within a single exercise set may look like vocabulary → word sequence → speaking → word sequence → speaking → word sequence → speaking. However, there is a chance that the n-grams or sentences in the exercise set may contain words that the user is not familiar with, specifically words above the user’s word level that the user has not been exposed to yet within the system. Users who are just starting with the system or have had less experience with the target language are more likely to encounter these types of chunks. While the system attempts to minimize the appearance of such a scenario, it is impossible to eliminate these types of n-grams completely.

Rather than leaving the user to guess n-gram constructions and sentence meanings blindly when new words are present, the system adapts the application to include
extra vocabulary exercises to prepare the user and keep the microlearning session productive. The vocabulary exercises are placed just before they appear in the word sequence or speaking task for the user to receive some basic idea of what the word means. Seeing the next task immediately after looking at the definition can also help the user understand the word better when it is in context.

An example exercise progression might be \textit{vocabulary (target word)} \rightarrow \textit{word sequence} \rightarrow \textit{speaking} \rightarrow \textit{vocabulary (new word)} \rightarrow \textit{word sequence} \rightarrow \textit{speaking} \rightarrow \textit{word sequence} \rightarrow \textit{speaking}. The system caters to the user at the micro-progression level to make sure that he can follow every exercise and not feel left behind. Beyond the exercise progression, the system continues to adapt to the user in the macro-progression.

### 4.2.2 Macro-Progression: Chunking Progression

The system must be able to track the user’s progress over time and increase the user’s exposure to vocabulary and the lexical chunks that contain the vocabulary. After a user has mastered words that he has encountered through the system, the system should continue to adapt to the user’s needs and provide more difficult vocabulary along with more difficult tasks for the user to complete to continue supporting his learning and practice. How quickly the progression happens depends on the user, and \textit{chunking progression} refers to the process of how the system monitors and guides the user at the appropriate pace.

#### Chunking Levels

Within a single exercise set, a target word is used to bind vocabulary and n-gram exercises together into a coherent set. For each target word in the system, the user can progress through four levels of difficulty, which is set according to the length of the chunks provided in the n-gram or sentence exercises. N-grams in the system are limited to 3-grams, 4-grams, and 5-grams, with 3-grams being the easiest and 5-grams being the most difficult. In addition, sentences are treated as a subset of lexical chunks and represent the highest level of “chunking” in the system after 5-
grams. Thus, the four levels of difficulty consist of 3-grams, 4-grams, 5-grams, and sentences.

The idea of matching the level of difficulty with the length of the chunks follows the suggestion from SLA that chunks should be developed from smaller units and then built on to form longer phrases that get stored in memory. Furthermore, the word sequence ordering task for n-grams is shown to be more overwhelming with longer phrases due to the greater number of choices of words and potential combinations the user needs to consider. Longer phrases are also more difficult to say fluidly because there are more sounds to think about in the process of performance. Therefore, when a new user begins practicing with the system, all of the target words will start at the 3-gram level. Users can progress to higher or lower levels for a target word through *chunking up* or *chunking down*.

**Mastering a Target Word**

If a user scores above the threshold for both the competence and the performance tasks given set of exercises, he chunks up to a higher level of exercises for that specific target word. If a user scores below the thresholds, he chunks down to a lower level, and if the scores are not both below or both above the threshold, the level for the target word remains the same. Levels cannot be skipped while chunking up or down, so the user receives practice at all four levels of difficulty. After a user performs above the pronunciation score threshold at the sentence level, the system considers the target word “mastered.”

While moving up or down only one level at a time may seem tedious, there are only three n-grams covered in a single exercise set. At the minimum, the user needs to cover only twelve n-gram and sentence exercises before “mastery” of the target word. Therefore, a user who has already seen a target word from studying the language outside of the system can get through the sets of exercises quickly and move on to other target words.

Although high level users may be able to get through each level of a target word quickly, they should not feel bored by the system if the target word shows up in
consecutive exercise sets. At the same time, lower level users who see the same target word appear may get stuck at a level and feel frustrated with the word, which can result in discouragement and lack of motivation to practice the target language. In order to maintain a balance between progressing quickly and keeping the user from getting stuck, all while exposing the user to a breadth of vocabulary, the system adapts to the user’s ability so that the user feels like he is progressing. How the system adapts to the user’s needs is based on the content selected for each exercise set, which is described in the next section.

4.3 Exercise Content Selection

Each exercise set is created such that it supports the learner both by reinforcing the target words that the user found success in and by drilling words that the user encountered but was less familiar with. Within a set, there is a single target word that is systematically chosen. Once the target word is chosen, n-grams and sentences are selected according to specific criteria.

4.3.1 Selecting the Target Word

When the user logs into the application, the system must decide which target word to focus on to use for the current set of exercises. Criteria are checked in “passes” where a single “pass” contains multiple conditions. In the case that no word satisfies the conditions in the current pass, the system moves onto the following pass. Regardless of the pass, the first criteria for choosing a word is that the word was not a target word for any exercise set completed within the past five minutes of the current selection process. There are a total of five passes for finding a target word, listed in order of priority below.

1. Pass 1: To encourage reinforcement and allow the user to get through easier words quickly, the word must a) have a pronunciation score above the user’s pronunciation threshold, b) not have passed the sentence level of chunking, or in
other words, has not yet been “mastered,” and c) have the highest pronunciation score of all words that satisfy the two previous criteria.

2. Pass 2: To cover a greater breadth of words without overwhelming the user, the word must a) be ranked below the user’s word level and b) not have been encountered yet either as a target word or as a part of a phrase or sentence of another target word.

3. Pass 3: To expand a user’s vocabulary, the word must a) have the minimum rank above the current user word level and b) not have been mastered.

4. Pass 4: To further encourage reinforcement regardless of the word’s pronunciation score relative to the user’s passing pronunciation threshold, the word must a) not have been mastered and b) have the highest pronunciation score given the first criteria.

5. Pass 5: To provide practice for words that need improvement, the word must have the lowest pronunciation score out of all target words, regardless of whether it has been mastered.

4.3.2 Selecting the N-Grams and Sentences

Similar to how target words are chosen, there are three passes that contribute to determining which n-grams can help the user the most. As many n-grams as needed are chosen from the first pass, but if the number of n-grams that satisfy the conditions from that pass is insufficient, the system moves to the second pass to get the remaining n-grams, and so on. Sentences are included in the process and are generalized to be included in the category of n-grams when describing each of the passes.

1. Pass 1: To provide opportunities for users to practice words that need improvement, the n-gram should contain words that have low pronunciation score such that the average score of all words in the n-gram outside of the target word is below the passing pronunciation threshold.
2. Pass 2: To expand the user’s vocabulary, the n-gram must a) contain at least one new word and b) have the minimum rank when comparing the rank of its highest ranked word.

3. Pass 3: To expose the user to the most common patterns and constructions used, the n-gram must have the highest combined frequency and mutual information value.
Chapter 5

System Evaluation

To evaluate PressToPronounce, I began a user study to gather data on pronunciation scores. While the resulting data set was incomplete, qualitative reactions and comments were found overall to be very positive. However, a long term study still needs to be completed to determine if PressToPronounce can significantly improve a second language learner’s fluency and measure how it compares to other existing systems.

5.1 User Testing and Feedback

5.1.1 Study Design

I recruited 14 native Mandarin Chinese speakers currently working at Microsoft Research Asia who wanted to improve their English speaking skills for a four-week study. The participants were divided into two groups of 7. The first group would start the study using PressToPronounce for two weeks and then switch to using the sentence-based system for the remainder of the time. The second group would begin with the sentence-based system and then switch to PressToPronounce. To measure improvement of pronunciation over each of the two week periods, three checkpoints were used to evaluate the participants’ progress: 1) a pre-study evaluation before the first half of the study, 2) a mid-study evaluation immediately before the second half, and 3) a post-study evaluation at the end with a brief interview to get feedback from the
participants. The evaluations would consist of pronunciation-based exercises scored by the CAPT system.

During the study, participants were asked to use their assigned system for approximately 20 to 30 minutes each day, though they were permitted to use either system for longer if they wished. They were also informed that they could use the application whenever they wanted for as many sessions as they liked, as long as they spent a total of 20 to 30 minutes on the exercises each day (e.g. 10 minutes in the morning and 15 minutes at night). Devices were available to users to borrow at their convenience during the day, though this limited users to only practicing with the system while they were in the Microsoft buildings, which is generally during their work hours.

5.1.2 User Feedback

Unfortunately, the study was not completed as planned due to unforeseen circumstances. After the pre-study evaluation was conducted and participants began using their first assigned system, the server supporting the communication between the mobile applications and the CAPT service showed signs of being unable to handle the necessary load for smooth completion of the study. Both systems slowed down and ultimately crashed, leading to the discontinuation of the study. Finding and fixing the bug that led to the crash required more time than expected, so there was not enough time to finish the study and gather data for the mid-study or post-study evaluation. As a result, preliminary evaluation of PressToPronounce came from user behavior and interaction with the systems in addition to qualitative reactions and feedback.

From looking at the sessions in which participants were able to use PressToPronounce consistently, we can see that users have varied ideas of when and for what duration to use the system. Some participants preferred to focus on using the application for 30 minutes or longer in a single sitting. In contrast, one participant scheduled breaks throughout his work day to use the application for approximately 10 minutes per sitting. PressToPronounce was shown to be flexible enough for users to fit into their desired study patterns, regardless of whether they consist of a single, focused, extended learning session or many brief training sessions throughout the day.
Three participants who were motivated to practice their English due to encountering many work scenarios where communication needed to be entirely in English chose to spend more than 30 minutes on the application each day. They did so to take advantage of the user study to practice their English more and expressed interest in continuing to use the application after the user test period. Other participants also noted that the application seemed helpful and could see it improve their speaking skills in the long term. This initial feedback shows that PressToPronounce is a promising tool for learners to become better English speakers.

5.2 Future Evaluation

In order to claim that PressToPronounce is an effective and preferred resource for practicing English output compared to other applications, future studies will be conducted by the Human Computer Interaction Group at Microsoft Research Asia as researchers continue to work with output-oriented language learning resources. A potential study designed is outlined in this section along with the challenges of evaluating a system like PressToPronounce.

5.2.1 Potential Study Design

For a full user study, I would compare PressToPronounce to the sentence-based system and look for differences between pronunciation scores and other factors throughout the time that participants are using each system. I would recruit at least 12 subjects who are native Mandarin Chinese speakers looking to improve their English speaking skills.

Similar to the original study, participants would be divided evenly into two groups. For the first half of the study, the first group would use PressToPronounce and the second group would use the sentence-based system. For the second half, the two groups would switch systems. The entire study would run for at minimum six weeks with three checkpoints to evaluate the user’s progress: a pre-study evaluation, a mid-study evaluation, and a post-study evaluation. The evaluations would also consist of
pronunciation-based exercises.

The main difference between this newly proposed study and the original study is the flexibility provided to participants. Participants would not be required to spend any amount of time per day on either system. They would be asked to use their assigned system at least once per day, but they would choose when, how often, and how long they use it. In addition, participants would either have the applications on their own devices or be given devices to bring outside of Microsoft buildings. By giving users the freedom to use the application with fewer constraints, we may be able to understand more about the types of scenarios in which users are most comfortable completing the exercises — particularly speaking and listening exercises — through their location and duration of use. In addition, we may be able to understand the participants’ attitudes towards each system and compare the features and scenarios that may encourage or discourage use.

5.2.2 Challenges

Learning a second language is difficult and requires a lot of time before output at near fluent levels are attained. The biggest challenge in evaluation is that a six-week study may not sufficient to show significant results in improvement, especially when considering the limitations of the metrics available to measure improvement in a second language.

Despite the advantage of having CAPT to give feedback on speech, pronunciation scores are far from sufficient to be the main quantitative metric for fluency. Fluency requires both competency and performance, and pronunciation only contributes in part to the performance aspect of language output. Moreover, speech and pronunciation vary immensely between each instance of the same utterance for both native and non-native speakers. CAPT is a precise tool that can pick up on these variations, so a single user can receive a wide range of scores, which results in greater difficulty for evaluating a user’s improvement over time in a quantitative manner.

There is also no particularly good measure of competence that the system can use to determine output fluency. While the word sequence tasks are meant to push
learners to think about construction of the target language, users are not actively using these constructions through the application. The scores for the tasks also may not contribute enough information about the user’s knowledge of the language, though it could present a glimpse of how well the user can recognize common patterns in frequently used n-grams.

Thus, conducting the proposed study could provide the evidence to show that PressToPronounce helps non-native English speakers improve their speaking, either directly through pronunciation scores or other data that may be collected throughout the study. More detailed feedback can also be gathered from participants so that we can understand their attitudes towards both applications and better cater to their needs in future iterations of PressToPronounce.
Chapter 6

Discussion and Conclusion

6.1 Incorporation of Persuasive Tactics Within the Design

While the ideas from SLA using the output hypothesis and concept of lexical chunks have a significant presence and influence in this application, tactics from persuasive computing have also contributed to pushing the design in the direction of making the resource more appealing to users while encouraging them to increase their practice of language output. This section discusses how well the PressToPronounce system was able to incorporate each aspect of persuasive design.

1. Simplification

Simplification is found as a part of the chunking progression. While I often describe chunking as a process of “building up,” chunks are a reduction of longer phrases and sentences. Sentences from speech and writing are broken first into n-grams and single words before they are built up again and shown in the context of more complex language structures for the purpose of learning and practice. Thus, simplification is clearly a crucial persuasive tactic incorporated into PressToPronounce.
2. Tunneling

By guiding the user through vocabulary exercises, word sequence tasks, and speaking tasks in a calculated progression, tunneling is clearly one of the main features of PressToPronounce. Providing specific activities for the user to accomplish helps the user focus and practice effectively for the short time that an exercise size set lasts.

3. Self-Monitoring

The feedback that is shown through the exercise summary allows users to monitor their own progress over time in a very efficient manner. The additional option of viewing detailed pronunciation scores from previous performances furthers the user’s ability to self-monitor their skills quickly and easily.

4. Suggestion

Color-coding the words in n-grams and sentences to help the user score well in the speech task encourages and supports user success through suggestion at the opportune moment in an exercise set. In addition, PressToPronounce provides the option for users to see definitions of vocabulary for each word as well as view examples of how n-grams and sentences are used in context, which helps further user understanding of the language.

5. Reinforcement

Reinforcement is most clearly seen in the chunking progression where a buildup of longer phrases and sentences provides more challenging tasks. These tasks indicate to the user that he is doing well enough to move on to higher difficulty exercises and are also typically based on target words that the user has pronounced well. By primarily focusing on words that the user is doing well with and implicitly drilling the words that the user needs improvement on through n-gram and sentence selection, the user’s successes are reinforced effectively without ignoring other aspects of his performance.
6. Tailoring

PressToPronounce has a heavy focus on user-centered design, as described in Chapter 4: Adaptivity of the System. It clearly tailors to the user so that the user can improve his output efficiently through the given microlearning tasks.

7. Surveillance

The user can see his past performances and how they influence his progression in chunking level through exercise sets. The clarity of the progression should indicate to the user that the system is monitoring his performance to determine when the chunking happens. Aware that progression only occurs as a result of good performance, the user should be persuaded to take the exercises seriously and try to score well in each task.

6.2 Improving PressToPronounce

6.2.1 Competency

At the moment, the chunking process treats sentences as a subset of n-grams with the progression being somewhat desegregated as two advancements: \( \text{word} \rightarrow \text{n-gram} \) and \( \text{word} \rightarrow \text{sentence} \). However, the chunking progression ideally should be \( \text{word} \rightarrow \text{n-gram} \rightarrow \text{sentence} \). At the moment, this discontinuity prevents the user from developing competency as efficiently as he can, and further work can be done to complete the progression.

One possibility for completing the progression is to give the user the opportunity to form his own sentences using given n-grams that he has seen through previous exercise sets. It is possible for the system to use speech recognition to detect what words are used within the sentence, though there are obvious caveats to this approach. First, a non-native speaker who is studying to improve his pronunciation may not be clear enough when speaking for the speech recognition to detect the correct words consistently. Even if the system prompts for the sentence using other methods of input
such as keyboard input, determining the grammatical correctness of the sentence is too difficult to incorporate into the system and requires far more work outside of the scope of this system.

6.2.2 Performance

While the sentence-based system has no way to track improvement or allow the user to test himself on specific words or sentences, PressToPronounce supports the user by offering this information. It gives the user the opportunity to make multiple attempts at a better performance and helps the user identify areas of pronunciation that need improvement. After these areas are determined, the user may also want to know how to improve and look to the system for concrete suggestions. Currently, PressToPronounce still lacks the ability to guide the user in this stage of learning to pronounce words more clearly and accurately.

Offering suggestions to the user about how to improve pronunciation is difficult because nuances in sounds often involve minor changes in the position of the mouth, lips, and tongue. Unfortunately, CAPT technology is not built to determine how the user produces the sounds that result in the scores, and even for a human instructor, it can be difficult to describe how the sounds should be formed to make pronunciation clearer.

Diagrams for how to produce specific sounds do exist, and PressToPronounce can include these diagrams in the application. However, as mentioned before, very small changes produce nuances, so speakers may already be mimicking the diagrams very closely. How to pinpoint and make the adjustments for pronunciation correction and then communicate this to the user is still an aspect that needs to be explored.

6.2.3 Fluency Evaluation

As discussed in the evaluation of the system, pronunciation and word sequence ordering scores are not sufficient for evaluating fluency. To improve the evaluation, more factors need to be included as metrics. One factor that can be considered for provid-
ing more information to both the user and the system is the duration of syllables and words spoken in a sentence. This takes advantage of the start and end times within an audio clip that the CAPT system is already able to determine.

While I do not believe that speed of a person’s speech necessarily correlates with fluency, it may be possible to compare the difference between the duration of syllables in a sentence spoken by a native English speaker with those of a non-native speaker who is attempting to produce that accent. Rhythm is often a good indicator of a speaker’s grasp of the language construction and potentially has a significant influence on the intonation and pronunciation of the spoken sentence. By taking these factors into account, a better system could provide more information for the user and suggest various relevant methods for improving and developing fluency.

6.3 Contributions

To summarize, this project made the following contributions.

1. Created a mobile application that supports study of English as a second language to native Mandarin Chinese speakers

2. Designed and implemented an adaptive, user-centered application that leverages existing state-of-the-art CAPT system at Microsoft for speech and pronunciation analysis

3. Designed a persuasive system based on n-grams that can generalize to multiple languages

4. Showed that a resource for pronunciation-driven language learning is desirable from the feedback received in preliminary user tests

A second language learner needs to master both input and output in order to become fluent in the language. While most technological resources available to user focus on listening and reading comprehension, not enough support reading and writing.
By taking concepts from SLA research and persuasive computing, I created user-centered exercises that encourage practice of output, filling in much of the resource gap that formed with the abundance of input-oriented applications. In addition, I designed PressToPronounce to take advantage of Computer-Assisted Pronunciation Training (CAPT), a state-of-the-art pronunciation analysis tool that provides quantitative speech and pronunciation feedback to support users in tracking their progress. Combining all of these aspects resulted in an application that received enthusiastic feedback from users who are trying to improve their speaking skills in a non-native language.
Bibliography


