

APRAXIA WORLD: DEPLOYING A MOBILE GAME AND AUTOMATIC SPEECH
RECOGNITION FOR INDEPENDENT CHILD SPEECH THERAPY

A Dissertation

by

ADAM HAIR

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Ricardo Gutierrez-Osuna
Committee Members,	Beena Ahmed
	Theodora Chaspari
	Ruihong Huang
	Frank Shipman
Head of Department,	Scott Schaefer

December 2020

Major Subject: Computer Science

Copyright 2020 Adam Hair

ABSTRACT

Children with speech sound disorders typically improve pronunciation quality by undergoing speech therapy, which must be delivered frequently and with high intensity to be effective. As such, clinic sessions are supplemented with home practice, often under caregiver supervision. However, traditional home practice can grow boring for children due to monotony. Furthermore, practice frequency is limited by caregiver availability, making it difficult for some children to reach therapy dosage. To address these issues, this dissertation presents a novel speech therapy game to increase engagement, and explores automatic pronunciation evaluation techniques to afford children independent practice.

The therapy game, called Apraxia World, delivers customizable, repetition-based speech therapy while children play through platformer-style levels using typical on-screen tablet controls; children complete in-game speech exercises to collect assets required to progress through the levels. Additionally, Apraxia World provides pronunciation feedback according to an automated pronunciation evaluation system running locally on the tablet. Apraxia World offers two advantages over current commercial and research speech therapy games; first, the game provides extended gameplay to support long therapy treatments; second, it affords some therapy practice independence via automatic pronunciation evaluation, allowing caregivers to lightly supervise instead of directly administer the practice. Pilot testing indicated that children enjoyed the game-based therapy much more than traditional practice and that the exercises did not interfere with gameplay. During a longitudinal study, children made clinically-significant pronunciation

improvements while playing Apraxia World at home. Furthermore, children remained engaged in the game-based therapy over the two-month testing period and some even wanted to continue playing post-study.

The second part of the dissertation explores word- and phoneme-level pronunciation verification for child speech therapy applications. Word-level pronunciation verification is accomplished using a child-specific template-matching framework, where an utterance is compared against correctly and incorrectly pronounced examples of the word. This framework identified mispronounced words better than both a standard automated baseline and co-located caregivers. Phoneme-level mispronunciation detection is investigated using a technique from the second-language learning literature: training phoneme-specific classifiers with phonetic posterior features. This method also outperformed the standard baseline, but more significantly, identified mispronunciations better than student clinicians.

DEDICATION

For Micaela.

ACKNOWLEDGEMENTS

I am indebted to my advisor, Dr. Ricardo Gutierrez-Osuna, for his guidance, support, and mentorship throughout the course of my graduate career. I have grown personally and professionally in ways I could not have anticipated when I first started this journey, and I will be forever grateful.

I am thankful for the opportunity to work with our collaborators in Australia, Dr. Beena Ahmed and Dr. Kirrie J. Ballard. I also appreciate the feedback I received from my committee members, Dr. Beena Ahmed, Dr. Theodora Chaspari, Dr. Ruihong Huang, and Dr. Frank Shipman.

My time at Texas A&M was made immeasurably better by the people I met there. Thanks to Chris, Dennis, Roger, Guanlong, Zelun, Raniero, Avinash, Jin, Sandesh, Shaojin, Nitin, Anurag, and Sudip for always being available to catch a game, grab a coffee, or just talk. Your friendship has been a true gift.

I would also like to thank my family for their love and support over the duration of my time in academia. I owe the biggest thank you to my wife Micaela for her constant encouragement and unwavering support. I could not have made it through this program without her by my side.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professors Ricardo Gutierrez-Osuna, Theodora Chaspari, Ruihong Huang, and Frank Shipman of the Department of Computer Science and Engineering and Professor Beena Ahmed the School of Electrical Engineering and Telecommunication, University of New South Wales.

The user studies described in Chapters 3 and 4 were conducted in part by Dr. Kirrie J. Ballard, Penelope Monroe, Dr. Jacqueline McKechnie, and Constantina Markoulli at the University of Sydney. They also facilitated data collection for Chapters 5 and 6. Analysis in Chapter 3 was conducted in part by Dr. Ballard and Ms. Monroe. Analysis in Chapter 4 was conducted in part by Dr. Ballard. Algorithm development in Chapter 6 was aided by Dr. Guanlong Zhao.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

This work was made possible by NPRP Grant # [8-293-2-124] from the Qatar National Research Fund (a member of Qatar Foundation). Travel to the 2019 ACM SIGACCESS Conference on Computers and Accessibility (Pittsburgh, Pennsylvania) was supported by a travel award from the Texas A&M University Graduate and Professional Student Council. The statements made herein are solely the responsibility of the author.

NOMENCLATURE

ASR	Automatic Speech Recognizer
CAS	Childhood Apraxia of Speech
DNN	Deep Neural Network
DTW	Dynamic Time Warping
GOP	Goodness of Pronunciation
GMM	Gaussian Mixture Model
GMM	Hidden Markov Model
ISTRA	Indiana Speech Training Aid
KCR	Knowledge of Correct Response
KR	Knowledge of Response
LPP	Log Posterior Probability
LPR	Log Posterior Ratio
LRC	Logistic Regression Classifier
MAP	Maximum A Posteriori
MCN	Mel-Cepstral Normalization
MFCC	Mel-Frequency Cepstral Coefficient
MLLR	Maximum Likelihood Linear Regression
NDP3	Nuffield Dyspraxia Programme
PS	PocketSphinx
SLP	Speech-Language Pathologist
SSD	Speech Sound Disorder
SVM	Support Vector Machine
TD	Typically-Developing
TM	Template Matching

TABLE OF CONTENTS

	Page
ABSTRACT	ii
DEDICATION	iv
ACKNOWLEDGEMENTS	v
CONTRIBUTORS AND FUNDING SOURCES.....	vi
NOMENCLATURE.....	vii
TABLE OF CONTENTS	viii
LIST OF FIGURES.....	xii
LIST OF TABLES	xv
1. INTRODUCTION.....	1
1.1. Specific research goals	4
1.2. Dissertation outline	6
2. BACKGROUND.....	8
2.1. Childhood apraxia of speech.....	8
2.2. Speech therapy	9
2.3. Digital speech therapy tools	10
2.4. Speech therapy games	11
2.5. Speech processing	14
2.5.1. Automatic speech recognition.....	14
2.5.2. Mispronunciation detection.....	17
2.5.3. Child speech processing obstacles	18
3. APRAXIA WORLD: A SPEECH THERAPY GAME FOR CHILDREN WITH SPEECH SOUND DISORDERS	19
3.1. Overview	19
3.2. Introduction	19
3.3. Background	23
3.3.1. Speech-driven games.....	23

3.3.2.	Game-based therapy	25
3.4.	Game design	27
3.4.1.	Game development	27
3.4.2.	Speech exercises	31
3.4.3.	Speech assessment	34
3.5.	Methods	35
3.5.1.	Participants	35
3.5.2.	Selection and participation of children	36
3.5.3.	Procedure	36
3.6.	Results	38
3.6.1.	Feedback from children	38
3.6.2.	Observations on strategy, gameplay, and engagement	42
3.7.	Discussion	45
3.8.	Conclusion	50
4.	A LONGITUDINAL EVALUATION OF TABLET-BASED CHILD SPEECH THERAPY WITH APRAXIA WORLD	51
4.1.	Overview	51
4.2.	Introduction	52
4.3.	Background and related work	56
4.3.1.	Digital speech therapy tools	56
4.3.2.	Automatic mispronunciation detection	59
4.4.	Apraxia World	62
4.4.1.	Game design	62
4.4.2.	Speech therapy program	67
4.4.3.	Pronunciation evaluation	69
4.5.	Experimental design	73
4.5.1.	Participants	73
4.5.2.	Protocol	74
4.6.	Results	78
4.6.1.	Gameplay analysis	78
4.6.2.	Therapy analysis	83
4.6.3.	Quality of audio recordings	85
4.6.4.	Manual and automatic pronunciation evaluation	87
4.7.	Discussion	90
4.7.1.	Implications for future work	94
4.8.	Conclusion	97
5.	EVALUATING AUTOMATIC SPEECH RECOGNITION FOR CHILD SPEECH THERAPY APPLICATIONS	99
5.1.	Overview	99
5.2.	Introduction	100

5.3.	Related work	102
5.4.	Automatic speech recognition	104
5.4.1.	PocketSphinx	104
5.4.2.	Template matching	106
5.5.	Experiments	107
5.5.1.	Data collection	108
5.5.2.	Experiment setup	109
5.6.	Results	110
5.7.	Discussion and conclusion	113
6.	EXPLORING CLASSIFIER-BASED MISPRONUNCIATION DETECTION FOR CHILD SPEECH THERAPY	116
6.1.	Overview	116
6.2.	Introduction	116
6.3.	Background	119
6.4.	Methods	121
6.4.1.	Acoustic modeling and posterior probabilities	121
6.4.2.	Feature generation and classification	122
6.4.3.	Goodness of pronunciation baseline	124
6.5.	Experiments	125
6.6.	Results	127
6.6.1.	Comparison against human raters	129
6.7.	Discussion and conclusion	130
7.	CONCLUSIONS FROM THIS DISSERTATION	132
7.1.	Summary	132
7.2.	Contributions	135
7.3.	Limitations	136
7.4.	Future work	138
7.4.1.	Game work	138
7.4.2.	Speech work	142
	REFERENCES	145
	APPENDIX A HUMAN RESEARCH ETHICS COMMITTEE MATERIALS	161
	APPENDIX B APRAXIA WORLD USER GUIDE	182
B.1.	Overview	182
B.2.	Installation	182
B.3.	Main screens	182
B.4.	Game settings	185
B.5.	Gameplay	188

B.6. Exercise delivery 190
B.7. Keyboard Evaluation 192
B.8. Logging..... 192

APPENDIX C APRAXIA WORLD RECORDER USER GUIDE 193

C.1. Summary..... 193
C.2. Start screen 193
C.3. Calibration 194
C.4. Probes 196

APPENDIX D CHILD SPEECH CORPORA 198

D.1. Typically-developing speech..... 198
D.2. Disordered speech from children 199

LIST OF FIGURES

	Page
Figure 1 (a) Start screen showing all of the available characters. Players start with the monkey on the far left as the default (b) On-screen information shown to players: collectibles and health in the top left, available power-ups in the top right, and a progress bar in the lower center.....	22
Figure 2 (a) The clothing store offers different pieces to fully dress up the character (b) The weapons store offers four types of weapons with increasing power (c) The power-up store offers uses of power-ups and increases to power duration	30
Figure 3 (a) Speech exercise popup in the <i>during-game</i> condition contains both a pictorial and text cue (b) The game displays a warning message when a player tries to finish the level before collecting enough stars (c) Speech exercise popup in the <i>after-game</i> condition. An awarded star count has been added to help children know how far along they are in the exercises (d) Speech exercise popup in the <i>after-game</i> condition once the minimum numbers of exercises have been completed. The message tells the player that they can either complete more exercises for a bonus or press the button to continue to the next level.....	32
Figure 4 Boxplots for survey responses from all children (some children did not answer all questions).....	41
Figure 5 Boxplot of exercises completed per finished level (unfinished and restarted levels excluded)	44
Figure 6 (a) A level from the jungle world (b) A level from the desert world (c) Speech exercise popup with both pictorial and text cues.....	63
Figure 7 (a) Various characters available for purchase (b) Costume items to dress up the character (c) Power-ups to give the character “superpowers” (d) Weapons with different attack behaviors.	67
Figure 8 Pictorial prompts for (a) pumpkin, (b) unicorn, and (c) banjo.	69
Figure 9 (a) Spectral information is extracted from an utterance, mean cepstral normalized (MCN), and trimmed (b) Template and test utterances are aligned and scored based on RMSE.	71
Figure 10 Word recording interface in AWR. Recordings are labeled as correctly (green check) or incorrectly (red x) pronounced.	73

Figure 11 Experimental protocol with two treatment blocks. Pronunciation is probed before treatment and weekly during treatment.	75
Figure 12 Minutes spent within a level per day for treatment phases one (P1) and two (P2) (** indicates $p < 0.05$, two-sample t-test).....	79
Figure 13 (a) Powerup purchases across all participants (b) Exercises completed per day.....	81
Figure 14 Absolute increase in pronunciation scores at the beginning and end of each treatment phase for caregivers (CG) and template matching (TM).....	84
Figure 15 Probability density for the number of phoneme errors per utterance.	87
Figure 16 (a) Spectral information is extracted from a trimmed utterance and then normalized (b) Template and test utterances are aligned and scored based on RMSE	107
Figure 17 Per-speaker, per-word word recognition accuracy for template matching with an increasing number of templates	111
Figure 18 Per-speaker, per-word accuracy (Using 15 utterances per word for adaptation and template matching).....	112
Figure 19 Phoneme-level feature vector feature extraction pipeline	123
Figure 20 Apraxia World start screen only appears when the app starts anew, not after the application is paused.....	183
Figure 21 Select the user profile associated with the calibration data you want to use.	183
Figure 22 World selection screen. Bottom menu has buttons for the character store, costume store, world selection screen, weapon store, powerup store, and settings screen.....	184
Figure 23 Level selection example. This is World One, which has a training level marked by the T.	185
Figure 24 Settings page with the word list selected for each level (left) and exercise-specific parameters (right).	187
Figure 25 Administrative options hidden to the right in the settings page.	187
Figure 26 Example level with heads-up display and overlaid controls.....	189

Figure 27 The blue anchor represents the checkpoint. After crossing this point, the player will reset here if they die.	189
Figure 28 Exercise popup with an image and text prompt.....	191
Figure 29 A prompt that the child should go collect more stars before finishing the level.....	191
Figure 30 Apraxia World Recorder start (a) and username creation (b) screens.	194
Figure 31 Apraxia World Recorder prompt calibration screens.	195
Figure 32 Apraxia World Recorder probe screens.....	197

LIST OF TABLES

	Page
Table 1 Words selected to address speaker-specific speech difficulties.	77
Table 2 Maximum progress in the game for each player.	80
Table 3 In-game purchases made by players during the study.	83
Table 4 Recorded utterances gathered during gameplay.	86
Table 5 Evaluator performance (True positive is an identified mispronunciation).	90
Table 6 Average word recognition accuracy (%) using 15 utterances per word for adaptation and template matching	113
Table 7 Top 15 phonemes in the corpus as percent of total non-silence phonemes	126
Table 8 Average combined F1 score from 5-fold cross validation (std. err.)	128
Table 9 Average combined F1 score for each set of student clinician annotations (std. err.).....	130
Table 10 Possible speech therapy game genres and speech integration methods.....	139

1. INTRODUCTION

As children begin to speak, they commonly learn some speech sounds early during development, while other sounds take longer to acquire. According to the American Speech-Language-Hearing Association, children should generally be able to produce most speech sounds by the time they are four years old [1]. Children who cannot form sounds by the expected age may have a speech sound disorder (SSD) affecting the development of accurate speech sound and prosody production [1]. Children with SSDs may also struggle with phonological representation, phonological awareness, and print awareness, which can lead to difficulties learning to read or reading disabilities [2], and negatively impact communication skills development [3]. Organic SSDs have an identifiable cause, such as motor difficulties (e.g., dysarthria), structural issues (e.g., cleft palate), or sensory problems (e.g., hearing impairments), whereas functional SSDs have no known cause [4]. Prevalence estimates for SSDs are varied; some researchers report that anywhere between 2% and 25% of children aged 5-7 may have an SSD [5], while others suggest that prevalence is closer to 1% of primary-school-aged children [6].

To improve speech production quality, children with SSDs typically undergo speech therapy with a trained speech-language pathologist (SLP) in a clinic environment. For speech therapy to be effective, treatments must be “frequent, high-intensity, individualized, and naturalistic” [7] so that children can practice new habits and skills [8]. However, scheduling appointments with SLPs can be logistically difficult [9-11]; children with SSDs constitute more than 40% of SL caseloads [12, 13] and up to 70% of SLPs

have waiting lists [13], which slows access to services. To meet high dosage requirements, clinic-based interventions must be supplemented with considerable home practice, typically directed by primary caregivers (e.g., parents, guardians). However, home practice poses its own problems. First, therapy sessions based on worksheets and flashcards can be tedious for children. Second, caregivers often have busy schedules that make it challenging to supervise the required amount of therapy, which can decrease practice [14]. *As such, this dissertation represents an effort to make speech therapy more engaging and decrease the time and skill burden on caregivers.*

A promising approach to address the tedious nature of speech therapy is to incorporate the practice into digital games. Digital therapy games can have a positive impact on child motivation and satisfaction [15], and have been shown to increase participant engagement and persistence [16, 17]. Most importantly, research has shown that computerized and tablet-based speech therapy interventions can be as effective as traditional interventions (e.g., worksheets, tabletop exercises) [18-23]. Children often enjoy using digital therapy interventions in short-term tests, and sometimes even play beyond the required time [24, 25]. However, applications often employ an arcade or casual game with simple play mechanics, which do not lend themselves to long periods of gameplay/speech practice and may quickly lose child interest [26, 27]. *Accordingly, the first part of this dissertation presents the development and evaluation of a game designed for lengthy use.*

Although work has gone into increasing therapy motivation, close caregiver supervision is still typically required during practice. For example, many game-like

applications for speech therapy have been commercially developed and are available for purchase [28] (e.g., Apraxia Farm [29], Articulation Station [30], ArtikPix [31]); however, these commercial applications do not include automated production feedback, which means that the caregiver must monitor productions and provide appropriate feedback. A handful of speech therapy games include basic production feedback through word recognition [26, 32] or monitoring vocalization volume, pitch, and duration [33, 34], but much research on field-delivered pronunciation feedback for children is still in its infancy. Although mispronunciation detection research is widespread in the language-learning community, less attention has been paid to children with speech sound disorders. This is likely due to the inherent difficulty of processing child speech caused by normally-occurring production inconsistencies [35] and the relative dearth of corpora containing error-annotated disordered speech from children. Some groups have explored processing disordered speech from children (e.g., Shahin et al. [36, 37], Dudy et al. [38, 39]), but it remains to be seen how these systems perform on field-collected disordered speech from children. This is especially important, as these systems must be robust enough to process child speech collected in children's homes, which are typically imperfect recording environments. *Therefore, the second part of this dissertation investigates child pronunciation verification with data collected under realistic home therapy conditions.*

To address the motivation and independence issues associated with home practice, this dissertation presents the development of Apraxia World, a tablet-based speech therapy game, and an investigation of automated pronunciation evaluation for speech therapy applications. This work is divided into three primary focuses: developing Apraxia World,

evaluating the game longitudinally, and examining speech recognition and mispronunciation detection performance on disordered speech from children. These developments are described across four manuscripts, which constitute Chapters 3 through 6 in this document. The research goals and contributions from this dissertation are described below.

1.1. Specific research goals

This dissertation research contains three main objectives:

1. **Game development:** Design and develop a speech therapy game for home practice based on clinician, caregiver, and child feedback from previous prototypes and update after pilot tests.
2. **Game evaluation:** Conduct initial pilot testing of the speech therapy game prototype and a longitudinal examination of the final version to study engagement and speech production improvements. This objective explores the following questions:
 - a. What do children think about the Apraxia World gameplay?
 - b. How and when should speech exercises be delivered in a platformer-style therapy game?
 - c. Do in-game speech exercises detract from the gameplay experience?
 - d. Do children remain engaged in game-based speech therapy over long periods?
 - e. Are pronunciation improvements achieved while playing Apraxia World comparable to those from traditional speech therapy?

- f. How accurately do caregivers and the in-game mispronunciation detection evaluate pronunciation?
3. **Automatic pronunciation evaluation:** Implement low-resource mispronunciation detection for use in the game and use data collected during longitudinal testing to explore additional mispronunciation detection techniques for disordered speech from children. This objective examines the following:
- a. Can limited speaker-specific audio be used to improve the word error rate on disordered speech from children?
 - b. Can a method from second-language learning mispronunciation detection work for disordered speech from children?
 - c. How does automatic mispronunciation detection performance compare against student evaluators with some training?

This dissertation research contains the following major contributions. Objective 1 yields a long-form speech therapy game and the first platformer-style speech therapy game. Objective 2 suggests that this novel therapy game increases engagement over both short- and long-term use. Critically, it also indicates that pronunciation improvements measured during the game-based therapy are comparable to those reported for traditional speech therapy practice. Objective 3 suggests that limited field-collected disordered speech data can be used to detect phoneme-level mispronunciations in child speech better than baseline methods. More importantly, it shows that automatic mispronunciation detection can mimic expert clinician labels better than student evaluators in offline testing.

1.2. Dissertation outline

The remainder of this dissertation is organized accordingly; it first presents relevant background for this work, then four manuscripts describing Apraxia World and processing disordered speech from children, and finally a summary chapter that offers discussion and directions for future work. Chapter 3 presents the prototype of Apraxia World and a pilot study to evaluate how children interact with the game and determine when to present the speech exercises to children during gameplay. This manuscript was published at the 2018 ACM Conference on Interaction Design and Children [25]. Chapter 4 describes the final version of Apraxia World, the automatic pronunciation evaluation framework used in the game, and a longitudinal study to explore long-term use and speech improvements arising from gameplay. This manuscript is under review in the ACM Transactions on Accessible Computing [40] at the time of completing this dissertation. Preliminary results from Chapter 4 were also published as late-breaking-work at the 2020 ACM CHI Conference on Human Factors in Computing Systems [41]. Chapter 5 explores the use of limited disordered speech from children to improve automatic speech recognition word error rates so that whole-word recognition can be used to verify children attempted to produce a word close to the prompted target. A portion of this chapter was presented as an extended abstract at the 2019 ACM SIGACCESS Conference on Computers and Accessibility [42]. Chapter 6 demonstrates performance of a classifier-based mispronunciation detection framework on disordered speech from children using recordings captured during the longitudinal evaluation of Apraxia World. Chapter 7 concludes this dissertation with a summary and discussion of all manuscripts, and suggests future directions for speech

therapy games targeting children. Appendix A contains the consent forms, information sheets, and questionnaires used in the Apraxia World user studies. Appendix B is the Apraxia World user guide for administering clinicians, along with additional game screenshots. Appendix C is the user guide for Apraxia World Recorder. Appendix D lists notable child speech corpora.

2. BACKGROUND

2.1. Childhood apraxia of speech

Childhood apraxia of speech is a speech disorder that affects the ability to correctly produce speech sounds and words. The American Speech-Language-Hearing Association proposed the following definition [43]: “Childhood apraxia of speech (CAS) is a neurological childhood (pediatric) speech sound disorder in which the precision and consistency of movements underlying speech are impaired in the absence of neuromuscular deficits (e.g., abnormal reflexes, abnormal tone). CAS may occur as a result of known neurological impairment, in association with complex neurobehavioral disorders of known or unknown origin, or as an idiopathic neurogenic speech sound disorder. The core impairment in planning and/or programming spatiotemporal parameters of movement sequences results in errors in speech sound production and prosody.”

Children with CAS may exhibit the following:

- Vowel errors [3, 44]
- Consonant distortions [44, 45]
- Stress/prosody errors [46, 47]
- Adding incorrect pauses between syllables [47, 48]
- Adding schwa sounds into consonant clusters [46-48]
- Speaking at the incorrect rate [47, 49]
- Difficulty with multi-syllabic words [47, 48]
- Inconsistent speech sound production [47, 50]

- Incorrect nasality [44, 47]

2.2. Speech therapy

If a caregiver or pediatrician suspects that a child may have a speech sound disorder, the child will first go through screening with a clinician before starting therapy practice. Assessments are culturally and linguistically sensitive, so a child's scores must be compared against those from a representative population [4]. These assessments examine sounds both within single words and connected speech. Disorder severity can be measured either on a continuum (e.g., mild to severe [51]) or quantitatively (e.g., percent consonants correct [52], percent vowels correct [44]). Once a speech sound disorder has been established, clinicians identify stimulative sounds, which means that the child is able to accurately imitate the problematic sound after being provided a model [4]. Stimulability testing is important to determine which target sounds are currently appropriate for therapy practice [53].

Child speech therapy consists of clinic sessions that are usually paired with caregiver-led home practice. Clinicians in the United States have reported providing 30 to 60 minutes of intervention across one or two sessions weekly [54]. However, clinicians in Australia (where the studies in this dissertation were conducted) meet with children less frequently, anywhere from once a week to once a month [55]. Regardless of clinic visit frequency, children need additional practice to meet treatment dosage. As such, caregivers are often involved in the therapy by supervising additional speech practice at home with their children [56]. This homework may include paper worksheets [57, 58] or fun activities

like games, reading books with repetitive phrases, and simply encouraging speech during regular interactions [59, 60].

2.3. Digital speech therapy tools

There has long been interest in digital speech therapy applications thanks to the opportunity to provide automatic feedback or remote therapy. Two notable speech therapy applications precede the work presented in this dissertation: the Indiana Speech Training Aid (ISTRA) and Tabby Talks. ISTRA is important for historical reasons, as it was one of the earliest speech therapy tools to offer automated pronunciation feedback. Tabby Talks is the precursor to Apraxia World and was developed as part of the same overarching project; lessons learned from Tabby Talks provided valuable insights when designing Apraxia World. Both ISTRA and Tabby Talks are described below.

ISTRA is a digital speech therapy project introduced in the late 1980s that used commercially-available digital speech processing hardware to provide pronunciation feedback to patients [61, 62]. Automatic pronunciation feedback is provided using template matching, where a compressed feature vector extracted from a test utterance is compared against a previously-captured template for the utterance [63]. These fixed-length templates consist of the averaged samples of the best pronunciation the child could produce under clinician supervision [64]. ISTRA offers patient-specific computerized drill sessions with graphical feedback representing utterance scores (e.g., bar graphs, bull's-eye displays) and pronunciation quality reports. Some speech exercises are also delivered through game-like applications such as Baseball and Bowling, where pronunciation scores are displayed as game performance [64].

Tabby Talks [65, 66] is a speech therapy application that includes a clinician interface for configuring exercises and monitoring progress, a mobile interface for patients to complete exercises and record speech, and a server-based speech-processing engine. The speech processing is designed to provide clinicians with automated speech assessments in their reports; patients do not receive automated feedback via their interface. Speech exercises are delivered through a flashcard or memory game interface, both of which record utterances for later evaluation. In flashcard mode, children are presented with a series of prompts to record before moving to the next image with a screen tap or swipe. Starting and stopping the recording function is handled with discreet button presses. The memory game mode is a card matching game where the player must match five pairs of images hidden behind cards [26]. To flip a card over, the player taps it and then must record the word prompted by the card. Once the player has recorded the utterance, they can flip another card to look for a match. As neither mode provides automated production feedback, clinicians or caregivers can manually score the production by awarding in-app stars: gold for good and silver for fair.

2.4. Speech therapy games

To make speech therapy more engaging for children, researchers have investigated adding therapy exercises into digital games. SpokeIt [67, 68] is an example worth discussing because it is one of the few games intended for long-term use, similar to Apraxia World. SpokeIt is a storybook-style game centered around Nova, a star powered by speech energy who fell from the sky and must help the Migs return color to their world by using the player's voice. As the story progresses, the game prompts the player to produce speech

targets that will help the characters in that scene; these targets may be sounds, words, or sentences. The game is controlled only through speech, with no touchscreen interaction required for play; as such, the game will move on automatically if the player struggles for 10 seconds to produce the correct speech. SpokeIt provides feedback on what the recognizer actually heard, so players can compare against the target. Speech recognition is handled with an iOS-specific implementation of the PocketSphinx automatic speech recognizer. SpokeIt is unique in that the developers are working to include procedurally-generated content and a narrative generator to afford repeated use of the game [69], something often missing in speech therapy games.

This dissertation research is part of a larger project to explore digital speech therapy. While Apraxia World is the largest game to come from this project, many smaller therapy games have been developed by other team members. These are discussed below to demonstrate the variety of therapy game types and build context for Apraxia World. Aside from Flappy Voice, all games provide word-level pronunciation verification automatic speech recognition technology. Unless otherwise stated, the following games use PocketSphinx to detect if the produced word matches the target word.

Flappy Voice [34]: This is a Flappy Bird clone where players move the bird up and down by modulating their vocal loudness. Loudness is measured according to speech amplitude, which is normalized according the minimum and maximum amplitude observed so far in a play session. Increased volume moves the bird up and decreased volume lowers the bird. Points are awarded for every obstacle cleared. The game offers two play modes; free mode is similar to the original Flappy Bird game, where hitting an

obstacle ends the game; assisted mode keeps the bird within a defined region so the player never hits an obstacle and can play endlessly, but points are only awarded for obstacles cleared without touching the region barriers.

sPeAK-MAN [70]: This is a clone of the classic Pac-Man game. Players move the character with the standard four-directional controls to eat pellets and avoid ghosts. A level is cleared when all of the pellets have been eaten. If a ghost touches the character, a life is lost. In Pac-Man, the player can get a power-up that briefly makes all ghosts vulnerable and the character can eat them to gain points and temporarily clear them from the play-field. In sPeAK-MAN, the player has to say a target word associated with a specific ghost to make it vulnerable, instead of getting a power-up. Word recognition is handled with the Microsoft Speech SDK.

Asteroids [26]: This is an open-source clone of the retro Asteroids game. Players move a continuously-shooting spaceship with on-screen controls to shoot asteroids and avoid having any hit their ship. Large asteroids must be broken up by selecting them with a touch, which starts the recorder, and then correctly saying the displayed target word. If PocketSphinx recognizes the word, the asteroid breaks into smaller pieces that the ship can destroy by shooting them. Players earn extra lives by reaching specified point thresholds. Once all lives are lost due to asteroid collisions, the game is over.

Whack-a-Mole [26]: This game displays a set of 10 cards that flip over one at a time. If the flipped card shows a word prompt, the player must tap (“whack”) it to start the recorder and say the word before a timer runs out. If PocketSphinx correctly recognizes the word, the player earns a star. If not, the card turns back over and no stars are awarded.

Periodically, the flipped cards display a bomb instead of a word prompt. If the player taps these cards, they lose a star, if they currently have any.

WordPop [26]: This game displays a target word with letters contained in colorful bubbles. The player touches the tablet screen to start the recorder, says the word, and releases the touch when the utterance is complete. If PocketSphinx detects that the utterance matches the target, the letter bubbles break apart and float away, while making popping noises. If the word is not correctly recognized, the player can try again infinitely or request a new word. Players earn points for each letter bubble that floats away.

Speech Worm [26]: This is a word-search-style game where the letters forming a word are contiguous within a search field, but not necessarily in a single row, column, or diagonal. The target word is displayed above the search field and the player must first find the word by swiping their finger over the letters in the correct order. Once the word has been located, the player must press the “Speak” button to activate the recorder and press it again once they finish speaking. Players earn points for each word that PocketSphinx recognizes as correct. Players can say the word until PocketSphinx correctly recognizes it or request a new word.

2.5. Speech processing

2.5.1. Automatic speech recognition

Automatic speech recognizers (ASR) convert speech into a digital transcript, either for use by humans (e.g., composing an email by voice) or computers (e.g., smart assistants). For processing, the speech signal is typically converted to frequency domain features, the most common of which are Mel-Frequency Cepstrum Coefficients (MFCC)

[71]. These are derived by applying a Fourier transform to speech signal frames to convert them into a spectrogram, which represents power spectra over time. Next, the Mel filter bank energies are computed by passing the spectrogram through a series of triangular filters spaced according to the Mel scale, which mimics the frequency resolution of human perception [72]. Finally, the discrete cosine transform is applied to the log of the Mel filter bank energies to arrive at the final feature vector.

Once the speech signal has been converted into a feature vector, it can be passed through an acoustic model to determine the feature sequence phonetic probabilities. Given that feature sequences vary in length, acoustic models typically use Hidden Markov Models (HMMs) to represent transitions between phonetic states [73]. Historically, Gaussian Mixture Models (GMMs) were used to represent the phonetic states in feature space [74]; this combination of models is referred to as a GMM-HMM acoustic model. However, more recently, deep neural networks (DNNs) have surpassed GMM accuracy due to computation power increases and the availability of large-scale data [75]. As such, the DNN-HMM acoustic model has become a more popular alternative to the GMM-HMM.

Another important component of an ASR is the language model, which determines the word or symbol sequence probability. Language models are typically n-gram models with sequence statistics or finite state models that use weighted or unweighted finite state automata to represent sequences [76]. The final step in the speech recognition process is to combine the acoustic and language model probabilities to search for the most likely hypothesized word sequence, which is returned as the recognition result [74].

2.5.1.1 PocketSphinx

PocketSphinx is the mobile-ready implementation of the CMU Sphinx speech recognition platform. Sphinx was originally developed in the late 1980s to address limitations of speech recognition at the time, namely the lack of speaker-independent, large vocabulary, continuous speech recognizers [77]. The latest version, Sphinx-4, was rewritten completely in Java and introduced modular components to make the system more flexible [78]. PocketSphinx was introduced in the mid-2000s specifically to run on the limited hand-held device hardware of the day [79]. The recognizer uses a GMM-HMM acoustic model, which support adaptation through either maximum likelihood linear regression or maximum a posteriori, both of which are run with scripts provided by the project developers. The last stable version (5prealpha) was released in 2016 [80] and recent development has slowed, according to activity on the project’s GitHub repository [81]. The CMU Sphinx team addressed the lack of project updates in a blog post, saying that they have been contributing to state-of-the-art speech recognition projects, even creating a mobile port of another open-source speech recognizer, Kaldi [82]. It is unclear what the future holds for PocketSphinx, given that it is developer-friendly, but falls behind current speech recognition technology in terms of performance [83].

2.5.1.2 Kaldi

Kaldi is the most popular open-source speech recognition framework at the time of writing this dissertation. When originally introduced in 2011, Kaldi only used GMM-HMM acoustic models, although it supported two types of GMMs [84]. However, Kaldi has supported neural-network-based acoustic models for some time now, with the latest

neural network framework (nnet3) being released in 2014 [85]. The framework is written in C++ and is generally interfaced with through BASH scripts that call the various processing steps. Kaldi uses BASH script “recipes” to handle model training, which are typically written for specific corpora and distributed so that others with the speech data can train their own copy of the acoustic model. Python libraries like Pykaldi [86] have been introduced to make interacting with Kaldi easier and more recently, PyTorch-Kaldi made it simpler to train Kaldi models using the PyTorch neural network library [87]. Until recently, Kaldi was largely limited to running on personal machines for research or servers for business use, as the library was difficult to run natively on mobile devices. Although compiling Kaldi for Android had been previously documented [88], the process was highly-involved and required writing custom code to interface with all desired Kaldi functions, which put the tool out of reach of non-experts. However, in 2019, the developers behind PocketSphinx released a Kaldi port for Android [89], which enables state-of-the-art speech recognition to run completely on-device. This opens the door for future developers to use this advanced speech recognizer within mobile speech therapy games.

2.5.2. Mispronunciation detection

Mispronunciation detection aims to use speech processing techniques to identify speech segments that diverge from the expected “correct” pronunciation. In general, mispronunciation detection methods can be characterized as posterior-probability-based, classifier-based, or rule-based. Posterior-probability-based methods compare acoustic model likelihood outputs against a threshold to determine if a segment is correctly pronounced [90-93]. Classifier-based approaches extract acoustic features from samples

to train classifiers that discriminate between correct and incorrect pronunciations [94-96]. Rule-based approaches manipulate the language model according to predefined error patterns to identify mispronunciations [97-100]. Mispronunciation detection is an active research area for adult speakers, but less attention has been paid to disordered speech from children.

2.5.3. Child speech processing obstacles

In general, all child speech is more challenging to process than adult speech due to physiological differences and production inconsistency or inaccuracy during development and skill acquisition [101]. The vocal tract in children is smaller than in adults, which affects how they produce speech. For example, formants (frequency bands that define vowel sounds) extracted from child speech have been found to be roughly 50% higher than those extracted from adult speech [102]. As a child grows and the vocal tract changes, the formant production also shifts; Narayanan and Potamianos reported a close-to-linear decrease in formant frequency as age increases [103]. Furthermore, as children grow, average vowel duration and variance decreases, along with average pitch, which eventually decreases more for males than females [104]; these spectral variations make it difficult for ASR to accurately parse child speech. Although production inconsistencies and errors may arise due to typical development timelines, they can be more prominent due to SSDs, making disordered speech from children especially difficult to automatically process. As such, child speech recognition training data must be carefully selected to model the desired pronunciations from the appropriate populations, as simply adding more data when training acoustic models does not guarantee improved performance [105].

3. APRAXIA WORLD: A SPEECH THERAPY GAME FOR CHILDREN WITH SPEECH SOUND DISORDERS*

3.1. Overview

This paper presents Apraxia World, a remote therapy tool for speech sound disorders that integrates speech exercises into an engaging platformer-style game. In Apraxia World, the player controls the avatar with virtual buttons/joystick, whereas speech input is associated with assets needed to advance from one level to the next. We tested performance and child preference of two strategies for delivering speech exercises: *during* each level, and *after* it. Most children indicated that doing exercises after completing each level was less disruptive and preferable to doing exercises scattered through the level. We also found that children liked having perceived control over the game (character appearance, exercise behavior). Our results indicate that (i) a familiar style of game successfully engages children, (ii) speech exercises function well when decoupled from game control, and (iii) children are willing to complete required speech exercises while playing a game they enjoy.

3.2. Introduction

Speech sound disorders (SSDs) can affect language production and speech articulation in children, leading to serious communicative disabilities [43]. Estimates for

* This chapter was published at IDC 2018. Reprinted with permission. Hair, A., Monroe, P., Ahmed, B., Ballard, K. J., & Gutierrez-Osuna, R. (2018, June). Apraxia world: A speech therapy game for children with speech sound disorders. In *Proceedings of the 17th ACM Conference on Interaction Design and Children* (pp. 119-131). <https://doi.org/10.1145/3202185.3202733>

the prevalence of SSDs in children vary; some suggest between 2% and 25% of children aged 5-7 years may be affected [5], while others estimate values closer to 1% of the primary-school-aged population [6]. Regardless of their exact prevalence, SSDs can have potentially devastating effects on a child's communication development [3]. Fortunately, children can reduce symptoms and improve speech skills by working closely with a speech language pathologist (SLP) [43]. To be effective, these treatments must be "frequent, high-intensity, individualized, and naturalistic" [7]. However, scheduling appointments with SLPs can be difficult, especially for children who live far from clinics [9-11]. Thus, clinic-based intervention typically must be supplemented with considerable home practice. Previous work indicates that remote digital sessions can be as effective as clinic-based sessions [18]. To alleviate the repetitive nature of frequent intense practice, however, these computerized therapies must be engaging.

A promising strategy to increase engagement is to deliver the speech exercises through mobile games. Accordingly, a number of game-like applications for speech therapy have been developed (e.g., Apraxiaville [29], Tiga Talk [106], Tabby Talks [16, 65], Articulation Station [30], ArtikPix [31], Pocket SLP [107]), though few provide feedback on speech productions. Among those that do, Tabby Talks [16, 65] combines (i) a mobile game that embeds speech exercises into a "memory/concentration" game where the goal is to find all pairs of identical cards in a deck, and (ii) an automatic speech recognition (ASR) engine running on a remote server that scores each individual production from the child [66]. In a pilot study [65], Tabby Talks was well received by parents, SLPs, and the children themselves, though feedback also suggested that the

intervention needed more game-like features to increase the player's interest, especially for younger children. A second area for improvement in Tabby Talks was in terms of providing real-time feedback on productions, which was not possible with the remote ASR engine due to transmission and computation delays. To address these concerns, we have developed Apraxia World, a speech-therapy game constructed on top of a full-fledged, two-dimensional platformer game, which will later be coupled with a mobile ASR engine capable of providing real-time feedback on productions. In Apraxia World, the player guides an avatar (the cheerful monkey character shown in Figure 1) through a multi-level world where the goal is to collect assets while avoiding enemies and traversing an obstacle course.

This paper describes the gaming and therapy elements of Apraxia World, with special emphasis on how to integrate speech production into the game¹. In Apraxia World, the player controls the avatar with standard inputs (virtual buttons and joystick), and speech input is tied to assets that the player must collect in order to advance from one level to the next. By associating speech production with the assets, players are able to anticipate and control when speech exercises appear, and the speech exercises do not detract from the gameplay or interrupt the player while executing complicated moves.

We validated Apraxia World through a pilot study with 14 children with SSDs (4-12 years old) and 7 typically-developing (TD) children (5-12 years old). This diverse

¹As will be discussed in the Game Design section, speech assessments in the present study were conducted by an SLP during gameplay rather than by a mobile ASR engine. This allowed us to isolate the game aspects of Apraxia World from issues pertaining to mobile ASR performance, which will be addressed in a separate publication.

population allowed us to gather feedback from children with varying exposure to speech therapy and their perception of how the speech exercises impacted gameplay. Specifically, we examined two strategies for integrating the speech exercises into the game, a *during-game* condition where the exercises were distributed throughout each level, and an *after-game* condition where the exercises were presented after finishing a level. Each child played both versions of Apraxia World and answered corresponding questionnaires on enjoyability, preference, improvements, etc. We also examined child engagement with Apraxia World based on qualitative questionnaire responses.

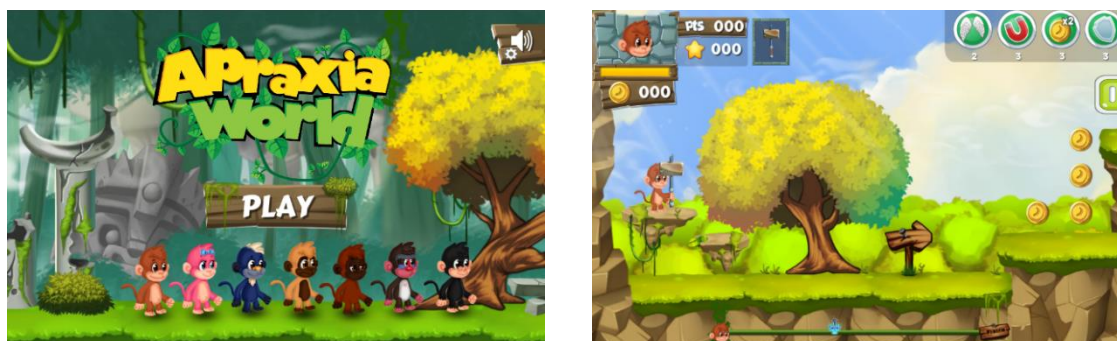


Figure 1 (a) Start screen showing all of the available characters. Players start with the monkey on the far left as the default (b) On-screen information shown to players: collectibles and health in the top left, available power-ups in the top right, and a progress bar in the lower center

The remaining parts of this paper are organized as follows. First, we provide background information on SSDs and review related work on speech-driven games and game-based therapy. Next, we describe the game, the integration of speech exercises, and

the (manual) assessment of productions. We then outline the experimental methods, including participant recruitment and study protocol, followed by the results from a pilot study with SSD children and TD children. The paper concludes with a discussion and directions for future work.

3.3. Background

The term speech sound disorder (SSD) describes a collection of difficulties with perception and/or production of individual speech sounds that affect a person's ability to produce intelligible speech [43]. SSDs that affect the production of the correct form of sounds are associated with motor-based or structural disorders (e.g., childhood apraxia of speech (CAS) or cleft palate, respectively) and are considered to be articulation disorders. SSDs that affect the functional employment of sounds (i.e., when the sounds produced are correct in form but not in usage, for example, a person may say 'dar' for 'car') are considered to be disorders involving the individual's phonological representation of sounds and/or speech segments. These speech difficulties are often overcome with regular and frequent practice [43], the repetitive nature of which makes speech therapy an excellent candidate for game integration.

3.3.1. Speech-driven games

In the context of gaming, speech input has been used to improve accessibility [108], novel interaction [109], physical therapy [110], speech therapy [34], and social skill development [111]. In previous speech-integrated games, the player's voice [108, 112, 113] or vocal features [34, 114, 115] have been used for game control. However, this model limits the choice of game to those slow enough for the player to produce the correct

voice command, and impedes gameplay if players struggle to produce command words. Furthermore, once a word-to-input mapping has been established, it is difficult to change the word without causing confusion or increasing cognitive load by making players keep track of new command words.

Cai et al. [112] took a different approach for using voice within an arcade-style game. The authors implemented a version of Tetris where voice commands unlock Tetromino (Tetris piece) rotation, rather than using the speech to directly move the piece; this allowed words to be reinforced without slow speech dramatically hindering gameplay. Researchers have also examined non-verbal features as inputs for games; common features include pitch changes or vowel sounds. Sporka et al. [115] designed a version of Tetris where players moved and rotated the Tetromino with pre-defined pitch patterns. They later extended their study of pitch as an input by comparing verbal and non-verbal commands for driving a radio-controlled car [116]. In both studies, users preferred the non-verbal commands due to ease of use and quick response.

The Vocal Joystick [117] maps pitch, power, and vowel quality to computer mouse movements. In tests, users quickly learned how to use the Vocal Joystick and found it less frustrating than using command words. In later work, Harada et al. [118] used non-verbal inputs for four different games, where game-specific commands were mapped to vowels and pitch intensities. They found that system processing time was significantly shorter for non-verbal commands, which is ideal for quick arcade-style games. House et al. [119] further expanded upon the idea of the Vocal Joystick by implementing a 5 degrees-of-

freedom control mechanism for a robotic arm moving in three-dimensional space. Vowel sounds have also been used to control retro-style games [120].

Automatic speech recognizers (ASR) are often found in speech-input systems, but they tend to struggle with children's speech. When ASR frameworks are tested with different forms of children's speech, performance decreases dramatically for continuous speech and long sentences as compared to adult speech, and the best results come from limiting the dictionary to single words and short phrases [121]. Speech patterns are typically harder to identify in children's speech due to large variations in vocal tract length, formant frequency and pronunciation quality [104, 122]. Additionally, even when ASR systems perform well with TD speech, they struggle with SSD speech [123].

3.3.2. *Game-based therapy*

Games have been evaluated for a variety of therapy applications across many disciplines. For example, in a recent IDC paper, Alessandrini et al. [124] developed a collaborative storytelling application to engage children with autism alongside their therapist, and found that the application helped fixate the child's attention on the activity. In another IDC study, Ferri et al. [125] conducted a research-through-design study of games for cognitive behavioral therapy. They prototyped three games to help children improve self-reflection and emotional analysis skills. These games were either non-competitive or gently competitive, without real loss scenarios. After surveying 18 physical therapists, Annema et al. [126] provided three implications for therapy game design: (i) configuration and setup should be simple and quick for the therapist; (ii) games should support the child and therapist by supporting on-the-fly changes and easy pausing or level

ending; and (iii) games should report and log child performance to give an overview or report across multiple therapy sessions. While simple games work well for infrequent events, such as a single clinical evaluation [127], arcade-style games may not be the most appropriate for long-term therapy, as gameplay can quickly grow stale [26, 27].

Previous applications for mobile speech therapy, such as Tabby Talks [16, 66], were developed as a proof-of-concept for remote speech therapy with a simple prompt interface. Similarly, Vocaliza [128] is a speech recognition system to help children with phonological, semantic, and syntactic therapy that shows progress over time. Research suggests that children engage better in and make fewer response errors with these types of electronic interventions than with traditional therapy [129].

Speech interventions have also been incorporated into casual games. Ganzeboom et al. [113] developed a multiplayer speech therapy game based on feedback from individuals with dysarthria. Players give each other verbal instructions through the interface – the game extracts loudness and pitch from the speech to provide feedback to help the player stay within a certain range. Umanski et al. [130] developed a game that helps children practice syllabic production rhythms. The game is a downhill slalom competition where the player makes their skier turn by producing the syllable at the correct time, with more accurate timing resulting in a tighter turn. Flappy Voice [34] is a modified clone of the popular game Flappy Bird where vocal loudness and pitch are mapped to the bird's position along the vertical axis. Players can use any verbal or non-verbal utterance to guide the bird through openings in the pipes, so long as pitch and loudness patterns can be extracted from the utterances. Lopes et al. [114] developed a game to practice sustained

vowel sounds. A bird flies from one branch to another if a vowel is produced with consistent intensity for a set duration, otherwise, it falls and the game resets. A more novel approach is demonstrated by Shtern et al. [131], where the speech articulators (i.e., tongue) are examined rather than the produced speech. In their game, the player uses tongue movements to control a flying bee.

3.4. Game design

3.4.1. Game development

We developed Apraxia World atop a full-featured, multi-world game project available for the Unity Game Engine. The game (Ekume Engine 2D) is a colorful adventure game where the player controls a monkey character. It comes with 48 levels divided into 8 worlds, multiple characters, and an in-game store for clothing and power-ups. All of the characters are shown in the start screen; see Figure 1a. Gameplay is linear – players must work their way towards the goal line at the right side of each level by navigating platforms, caverns, and other obstacles while trying to collect assets and avoid or eliminate enemies. Players control their character with a directional pad and two buttons, all overlaid on the tablet screen. Level and character information is shown in a heads-up display; see Figure 1b. The game offers two types of assets to collect: coins and stars. Coins are plentifully dispersed throughout the levels and are used to purchase items in the store. Stars originally served as a secondary challenge where a player could try to collect all stars within a level before finishing; this is similar to other games where players try to find all items of an object class. Each level contains a checkpoint (represented by a large anchor icon) around halfway through – if the player dies before reaching this point,

they lose the assets (coins and stars) collected so far in that level. However, if they die after reaching the checkpoint, they keep the assets and restart at the checkpoint.

The in-game store sells clothing/costumes, weapons, and power-ups. The store uses in-game currency, either collected in the levels or awarded for doing exercises. The prices for store items range from 50 to 6,000 coins. The clothing store is shown in Figure 2a, where the player can see how the different items look on their character. The weapons (Figure 2b) vary in power and striking distance (e.g. slingshots can shoot far but swords and hammers are close-proximity weapons)². Power-ups (Figure 2c) include coin value duplication, flight, invincibility, and coin magnets, all of which last for a short duration that can be lengthened by upgrading the power-up in the store.

We left the core gameplay unchanged, and instead modified the role of the stars. In our modified game, a player must collect a predetermined number of stars to complete a level, each star in turn requiring the player to complete a number of speech exercises. The game delivers these speech exercises either *during* or *after* gameplay; the delivery method is explained in the next section. We associated speech production with the stars so that players would be able to anticipate and control when speech exercises would appear. Additionally, we needed a “safe” time to display the exercise that would not detract from the gameplay or interrupt the player while executing complicated moves.

As well as adding the speech exercise, we also edited the levels to make them age-appropriate and increased the number of stars to 7-10 per level. In addition, we set stars to

²Although the game contains weapons and some combat, it is very mild in terms of violence. There are neither blood nor death animations – characters and enemies simply fall over and then disappear.

regenerate in the same place 10 seconds after being collected. We wanted a surplus of stars in different locations throughout the level to encourage players to gather extra and complete additional speech exercises if they so desired.



(a)



(b)



(c)

Figure 2 (a) The clothing store offers different pieces to fully dress up the character (b) The weapons store offers four types of weapons with increasing power (c) The power-up store offers uses of power-ups and increases to power duration

3.4.2. *Speech exercises*

The SLP can set how many exercises must be completed for each level, as well as provide a customized list of words per level, according to each child's therapy needs. In what follows, let E denote the number of speech exercises (i.e., word prompts) that must be completed per star, S denote the number of stars per level, and C denote the value of each star (in coins), all as defined by the SLP. Prompts are randomly selected from the word list such that they do not repeat until all words have been prompted.

The game delivers exercises in two ways: *during-game* or *after-game*. In the *during-game* mode, an exercise popup (see Figure 3a) appears when the player attempts to collect a star, at which point the player must complete E prompts. Correctly producing the target word triggers the game to either load the next prompt, or dismiss the popup if enough prompts have been completed. Incorrectly producing the target word causes a "Try again!" message to display briefly before the word prompt is displayed again. When the child has completed E prompts, the popup window disappears, a star is awarded, and C coins are also awarded. Players can collect as many stars as they like, each star yielding C coins. If the player attempts to complete the level before collecting S stars, a text banner prompts them to turn around and collect additional stars – see Figure 3b. Once the child collects at least S stars and crosses the goal line, the level ends.



(a)



(b)



(c)



(d)

Figure 3 (a) Speech exercise popup in the *during-game* condition contains both a pictorial and text cue (b) The game displays a warning message when a player tries to finish the level before collecting enough stars (c) Speech exercise popup in the *after-game* condition. An awarded star count has been added to help children know how far along they are in the exercises (d) Speech exercise popup in the *after-game* condition once the minimum numbers of exercises have been completed. The message tells the player that they can either complete more exercises for a bonus or press the button to continue to the next level

In contrast, the *after-game* condition allows children to play the game as normal until they attempt to cross the goal line, at which point they must complete $S \times E$ exercises – the same number as the *during-game* condition. Before attempting to cross the goal line, the player is allowed to collect as many stars dispersed through the game as they want, but

these stars do not award any bonus coins nor do they trigger speech exercises. If players so choose, they can collect no stars and go straight for the goal line. Once the player reaches the goal, the exercise popup appears; this popup (Figure 3c) is identical to the one in the *during-game* condition, except that it has a Star Counter so that the player knows their exercise progress. After each correct utterance, the game loads the next prompt. The same brief “Try again!” message as in the *during-game* condition appears if the child incorrectly produced the target word. Every E prompts, the game awards C coins and one star; this reward is reflected in the Star Counter. Once $S \times E$ exercises have been completed, two text banners and a continue button appear (Figure 3d); the banners inform the child that they can continue producing speech to gain additional coins or they can press the continue button to end the level. Once the child presses the continue button, the popup disappears and the level ends.

The speech exercises (i.e., word prompts) are based on the Nuffield Dyspraxia Programme (NDP3), an intervention program for young children with severe SSDs, including CAS [132]. NDP3 is designed to address specific effects of CAS, such as single consonant and vowel articulation, sequencing sounds together, and maintaining accurate prosody. We selected NDP3 because it comes with a 750-image set representing CV, CVC, CVCV, and multisyllabic words, which can easily be displayed in the exercise popup. Furthermore, NDP3 shows good treatment and generalization gains when delivered intensively [133]. Nonetheless, Apraxia World can be extended to other practice materials beyond (or instead of) the NDP3 set.

3.4.3. Speech assessment

Previous mobile speech therapy applications have used some form of automatic speech recognition (ASR), such as Pocketsphinx [34, 65] or custom approaches [37]. However, ASR on mobile devices either produces poor recognition rates with disordered speech or requires an internet connection such that a server can process the audio (e.g., Google Speech, Apple's Siri). Additionally, ASR performs especially poorly on speech from children [121]. Therefore, for the present study, we decided to isolate the game aspects of Apraxia World from issues pertaining to mobile ASR performance. Accordingly, we used a Wizard of Oz design where speech was evaluated manually by an SLP via a Bluetooth keyboard that allowed them to indicate (as the child plays the game) whether or not each word had been produced correctly. While ASR will be used in future iterations of the game, using the human evaluator gave us the children's best-case impression of the game and speech exercise integration, without any frustration from ASR errors.

We designed the keyboard input to mimic a binary decision: the SLP marks a speech production either as correct or as incorrect. We implemented rules to reduce the number of incorrect attempts on a single word and minimize reinforcing the wrong pronunciation; 4 consecutive incorrect pronunciations will cause a new prompt to come up (i.e., skip the problematic prompt) and 3 skipped prompts during an attempt at collecting a star (i.e., 3 prompts were skipped before 2 prompts were said correctly and a star was awarded) causes the exercise popup to disappear without awarding a star. These rules were

put in place now, so that the exercise logic will be the same between the current and future versions when ASR is enabled.

3.5. Methods

We evaluated Apraxia World in a within-subject study where children played two versions of the game, where speech exercises were delivered either *during* or *after* gameplay. In the process, we surveyed the children's impressions of this style of game in terms of enjoyability, ease of play, likes, dislikes, suggestions for improvement; we queried preference for game version; and we analyzed meta-data to identify differences across versions in amount of speech practice completed.

3.5.1. Participants

Twenty-one English speakers took part in the study. Participants included 14 children with diagnosed SSDs ranging from mild to severe (7 motor-speech and 7 phonological impairments; 13 male and 1 female; mean age: 7.4 years, range: 4-12 years old), and 7 children reported by parents to be TD (4 male and 3 female; mean age: 8.7 years; range: 5-12 years old). The children with SSDs had all been formally assessed and diagnosed as having a speech sound disorder by a qualified SLP and, at the time of participation, had no other developmental diagnosis (e.g., autism spectrum disorder or cognitive impairment). All procedures were approved by the University's Human Research Ethics Committee and all children and guardians provided written informed assent/consent, respectively, before participating in the study.

3.5.2. Selection and participation of children

Participating families self-referred in response to flyers and advertisements placed within the University's Speech Clinic, sent out by email, posted on social media, and posted in a local magazine. They were then selected for participation on the basis of SSD diagnosis occurring without other developmental diagnosis or no speech or developmental diagnosis (i.e., TD). Children and parents were asked if they would like to participate in a study looking into the development of tablet-based games to help children with their speech therapy exercises. Children were told that they would be shown two versions of the same game and asked some questions to help the research team continue to develop the game. They were told they could stop playing/discontinue participation at any time.

3.5.3. Procedure

All children were asked to test both versions of the game (*during-game* and *after-game* conditions). The order of presentation of the two game versions was randomized. Audio was recorded during the exercises for later analysis and debugging. The SLP sat beside the child and evaluated speech in real time. Exercise parameters were fixed for all children ($E = 2, S = 10, C = 25$), such that each child had to correctly produce at least 20 words before completing a level.

Two individualized word lists of approximately equal complexity were created for each child, one for each version of the game. The words were chosen by the accompanying parent and both lists contained (i) five words the parent believed the child should have no difficulty producing and (ii) five that they believed the child would have some difficulty

producing³. This was done in order to mimic a home-practice setting where some “easy” words are included to ensure some success. Each child’s ability to say the words chosen for them was checked before they began playing the game.

The children were first provided a description of the game, its aim (to collect coins and stars to buy things for the character as progression is made through the levels), and instructions on how to play. A brief demonstration of how to use the controls was also provided. The children were not explicitly told that their word productions would be judged as correct/incorrect by the SLP conducting the study. They were asked to play each version for as long as they wanted, up to a maximum of 15 minutes per version. The children were then given the game to play on a Samsung Tab A 10.1-inch tablet (Android 6.0 Marshmallow). All children started with a training level that had no exercises, no enemies, and no chance of falling off the platform. The purpose of this training level was for children to learn the game mechanics. Each child progressed from the training level into Level 1 of the full game in the same way as they transitioned between other levels of the game. Once a child had played the first version of the game (for as long as they wanted to, up to 15 minutes), they were asked to complete a questionnaire about the game before being presented with the second version. On average, the questionnaire took 5 minutes to complete. The child was then again given the game for as long as they wanted to play (up to 15 minutes). After playing the second version, they were asked a series of follow-up questions before being asked (i) which version they preferred and which version they

³e.g., for one child, “watch” and “witch” were hard words, while “rabbit” and “peach” were easy words. The same words may not work for different children.

would now like to play again, and (ii) if they would like to play again. The questionnaire focused on game enjoyability, ease of play, likes, dislikes, and suggestions for improvement. It contained a combination of 5-point Likert-scales and open-ended questions; see Figure 4. The questions were read to all children and all responses were written down by the SLP. After answering the questionnaire, the child was allowed to play their version of choice again, if desired.

During gameplay, each child's behavior was also observed to monitor for signs of reduced concentration or signs of frustration, such as fidgeting. Were such signs observed, the child was reminded that they could cease gameplay at any time. Observations on each child's approach to gameplay were also collected, including a willingness to collect additional stars in either condition; focus on collecting all the available coins; a desire to explore the levels or to try to progress through the levels as fast as possible; and use of coins collected to purchase items from the store.

We logged the number of levels completed, strategy of gameplay (e.g., focus on completing the level vs acquiring assets), and number of exercises completed (i.e. words produced) for each child per level. This allowed us to explore whether the two game versions facilitated different amounts of practice.

3.6. Results

3.6.1. Feedback from children

Figure 4 summarizes responses to questions that used numeric ratings via boxplots. Four of the children did not answer all questions on the questionnaire, but their available responses are included in the analysis.

Nineteen of the 21 children found the game enjoyable and said they would like to play it again. All 19 would have continued playing beyond the 15-minute time cap had they not been stopped by the researcher. The other two children (one SSD and one TD) requested to discontinue during the allotted testing time because they were not engaged with the game and said that they would probably not play it again. However, they did play both versions and their data are included in all analyses.

The younger children (4-5 years) conflated the question “How difficult was the game?” with ease of control manipulation; for example, some children who struggled to complete a level still rated the game as easy to play. The older children were better able to dissociate ease of game control and gameplay, and their answers as to how easy they found the game more closely reflected their game progression.

Responses to whether the *during-game* condition made the game harder were varied and depended, in part, on whether the child liked having the speech exercises during or after gameplay. Responses included: “[...] *because I liked the game and wanted to concentrate on it*” and “[*the exercises*] *keep on popping and almost killing you.*” Most children agreed that the *after-game* condition did not make the game harder.

When asked which version of the game they preferred playing, 13/21 of the children selected the *after-game* condition (eight preferred *during-game*). The reasons for this preference included: “*the words at the end of the game didn’t interrupt your game,*” “*instead of collecting stars you can just say them,*” and “*playing [the during-game condition] made the words harder.*” One of the children who liked the *during-game* condition said that they “*liked the exercises popping up.*” A child said that although they

liked the *during-game* condition, they “*would play [the after-game condition] again because of the risk of dying while doing exercises in [the during-game condition]*” (some children struggled to navigate immediately after the game un-paused following the exercises). Some children offered alternatives to the two conditions we included: one said they “*would like exercises before the level*” and another said they “*would choose neither – would like the words during the game and then again at the end of the level so that you can practice them and get extra points.*”

Other verbal responses surveyed the child’s likes, dislikes, and suggestions for improvement. When asked what they liked, children mentioned the monkey characters and fighting the enemy characters (e.g., “*bashing monkeys,*” “*the monkey and hitting the monsters,*” “*fighting the monkeys,*” the “*bashing hammer,*” and “*hitting enemies*”). One child said the game structure “*reminds me of Donkey Kong [and I] like that it was hard.*” Two other children also commented that they liked that the game increased in difficulty, saying: “*it gets harder*” and “*it takes work/skill to play.*” Other likes included: “*[there are] not a million things to remember*” and “*all the super powers.*”

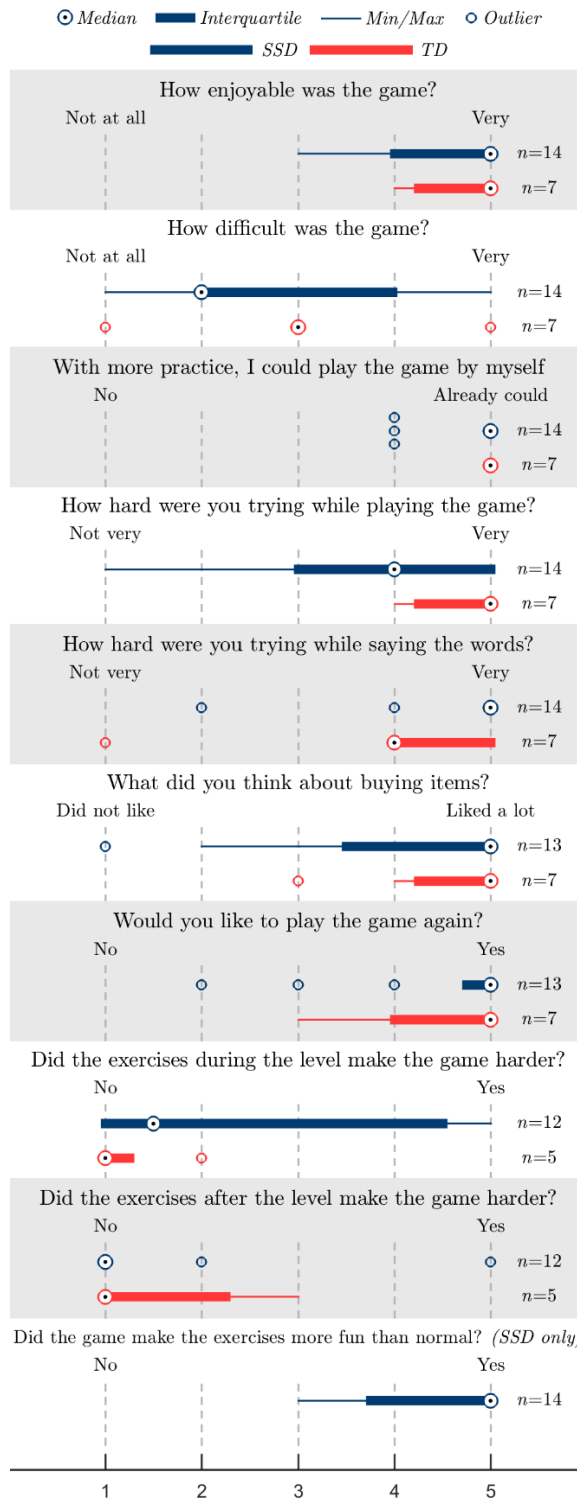


Figure 4 Boxplots for survey responses from all children (some children did not answer all questions)

The children were also asked what aspects of the game they liked the least. The most common comments were about dying, restarting, and losing stars collected (if they died before the checkpoint) (e.g., “*keeping dying,*” “*restarting when you die,*” and “*losing stars when I die before the checkpoint*”). Although some children enjoyed that difficulty level increased quickly, others cited it as an issue (e.g., “*it got hard pretty quick*”).

Suggestions for changes were varied and reflected that the children had engaged well enough with the game to imagine modifications for both individualization and development. Some suggested ensuring that the items for purchase were more varied and matched the characters, or combined with the superpowers (e.g., boots that allow you to fly). One child said they “*would rather princesses and unicorns*” than monkeys. Three children commented that they would like the game more if it had a storyline (i.e., a reason for their character’s progression through the islands). For example, one said that they would like the island to have villages so that they could then be the hero who has to save their village. Other comments reflected the same idea of fleshing out the virtual world: “*collect[ing] an army to kill the bad guys,*” having “*different types of bad guys,*” and “*buy[ing] pets to help you survive.*”

3.6.2. Observations on strategy, gameplay, and engagement

Gameplay data were available for all 14 SSD children. Data for two of the seven TD children were lost due to software malfunction.

All of the children, except the two who asked to discontinue play, were observed to concentrate well during both gameplay and exercise completion. Minor frustration was observed solely in relation to the child’s character dying and/or loss of stars collected. This

was, however, accepted by all children as a negative, but unavoidable, part of the game. The smaller children were observed to have difficulty holding the tablet and those less familiar with tablet-based games appeared to have difficulty managing the two-handed controls. One child's suggestion for easing these difficulties was to include an option for an external joystick. The double jump maneuver proved difficult for some children, who struggled with the button timing.

Approach to gameplay appeared to be linked to interest in asset collection. Sixteen children rated buying items for their characters highly (“*it made it like a quest to earn cash and buy your accessories*”). They were observed to spend more time collecting coins than the remaining five children, who said that buying items for their character did not interest them. The older children demonstrated a clear understanding of the relationship between completing exercises and asset collection, whereas the younger children did not. For example, three older children (10–12 years) purposefully undertook more than the minimum required exercises per level, with the express intent of purchasing items from the store.

Figure 5 shows the total number of speech exercises completed by the children per finished level. Speech exercises completed in unfinished or restarted levels are not included. Regardless of order of delivery, 14/19 of the children for which we have gameplay data finished more levels in the *after-game* condition; three children finished the same number of levels in both conditions and one child finished more levels in the *during-game* condition. This imbalance is due to two primary causes, (i) levels take longer to finish in the *during-game* condition because the player must spend time looking for stars

or waiting for them to regenerate, and (ii) the gameplay data include levels completed in the brief free-play portion after the test. These data were left in because the free-play more closely approximates home-practice (less evaluative pressure on the child).

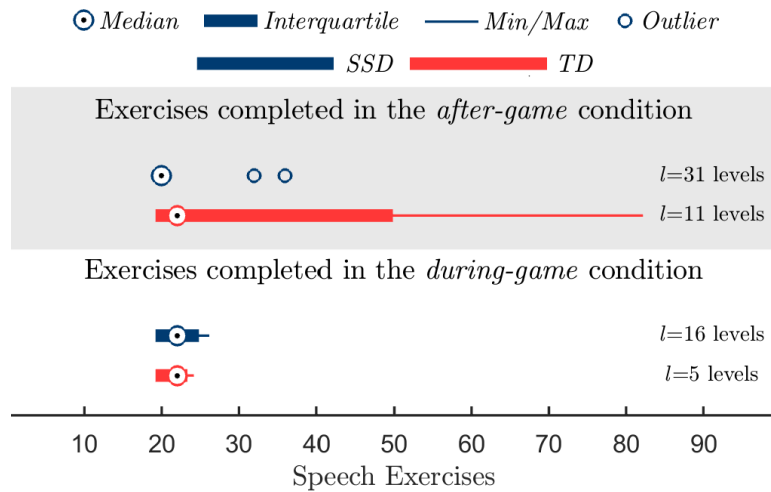


Figure 5 Boxplot of exercises completed per finished level (unfinished and restarted levels excluded)

SSD children in total finished $l=31$ levels in the *after-game* condition (median exercises per level: 20, range: 20-36) and $l=16$ levels in the *during-game* condition (median exercises per level: 22, range: 20-26). TD children in total finished $l=11$ levels in the *after-game* condition (median exercises per level: 22, range: 20-82) and $l=5$ levels in the *during-game* condition (median exercises per level: 22, range: 20-24). In general, SSD children completed the minimum number (20) of speech exercises per level in the *after-game* condition; TD children completed more exercises in the *after-game* condition due to

a lack of perceived risk (any attempted exercise essentially guarantees a reward). All children completed close to the minimum number of exercises in the *during-game* condition. Across all participants, the median exercises completed in each condition (20 vs 22 – SSD; 22 vs 22 – TD) indicate that children are unlikely to complete large quantities of speech exercises beyond a specified minimum. As such, the choice of exercise delivery method may be more important as a per-player customizable element rather than a way to ensure maximal exercise completion; we further expand upon customization below.

3.7. Discussion

This paper presents a novel approach for providing intensive and often tedious speech exercises to children with SSDs in a more engaging manner. We have developed two versions of a platformer game in which speech exercises are integrated and linked to asset collection, wherein the exercises can be presented either during or after gameplay. We surveyed children’s impressions of the overall approach and version preference, and also examined meta-data for potential influence of version on the amount of speech practice undertaken.

Overall, the children (13/21) preferred the *after-game* condition for two main reasons: (i) they did not like having their gameplay disrupted, preferring to do the exercises separately and (ii) they did not like losing collected stars in the *during-game* condition when they died before reaching the checkpoint. Although the stars were placed in locations that should have been minimally disruptive to gameplay, the children still reported worrying about controlling their character immediately after the game un-paused when in a potentially difficult position (e.g., if an enemy is close by, if they are jumping over a

platform gap). Losing stars upon dying was more discouraging to players than had been anticipated. One child compensated by strategizing: they prioritized reaching the checkpoint before collecting any stars. Stars collected before the player reached a checkpoint were intentionally not saved to encourage additional speech production. However, even though this led to all children completing many more exercises in the *during-game condition* than they did in the *after-game condition*, it also proved to reduce their motivation. Losing stars was judged as being more frustrating than repeatedly dying – the children completed an average of 25 exercises while playing in the *during-game* condition that were not saved due to restarting a level. In future versions of Apraxia World, this could be remedied by, for example, allowing players to keep all coins collected from the exercises (but not stars, still ensuring extra speech production) if their character dies before the checkpoint, or by allowing them to keep all stars and coins collected.

The eight children who preferred the *during-game condition* demonstrate that the preference for one version over the other was not unanimous. These children enjoyed having their speech exercises distributed during gameplay, with one stating that “*it seemed like I had to earn less stars [in the during-game condition].*” It could therefore be argued that providing future players access to both game versions would ensure that individual preferences will be met.

One important consideration, regardless of version, is the ratio of speech exercises to gameplay. Although the *after-game* condition was preferred by the majority, if a player struggles to make progress in the game, it becomes non-optimal in terms of number of exercises completed during gameplay, which undermines the major goal of the

intervention tool. Most children reached a point, for the younger players (4-5 yrs.) in the first level, where they had to make multiple attempts to reach the end of the level. In the *after-game* condition, this resulted in a lot of gameplay without speech exercises. Similarly, some children seemed to like exploring the level and were in no hurry to move onto the next one, which again increased playtime without speech exercises. This could be remedied with a *before-game* condition, in which players would have to complete exercises whenever starting or restarting a level (from the beginning or checkpoint). Exercises could alternatively be presented at certain time intervals throughout the level. This would ensure that children could still experience uninterrupted gameplay time, while also ensuring that the necessary ratio of gameplay to speech exercises to maintain therapeutic utility would be upheld. An alternative solution may be to add an “energy” level that decays over time and must be replenished by completing exercises; in this fashion, players would be required to complete exercises regularly, but at a time of their choosing.

Providing tiers of game difficulty to cross a broader range of age, physical ability, or SSD severity may be beneficial in future versions. Child age and prior gaming experience were observed to affect player success. Similarly, the children had varying success with the game controls. Even though the controls used are standard for tablet games, some children had trouble with button combinations that required more careful timing. Again, the children who had limited prior experience with tablet-based games were observed to find the dual-handed controls difficult. A subset of children with movement-based speech disorders, such as CAS, have limb coordination difficulties; some children

during the study were observed to have difficulty with game controls, extraneous limb movements, and rapidly timed double clicks. Compensatory strategies for these factors, such as an external joystick, need to be addressed in subsequent versions of the game.

The current study highlighted that built-in flexibility in a speech therapy tool is necessary. The subtle complexities in creating and presenting such a tool lie in matching both child and SLP expectations by balancing gameplay and child engagement against the provision of therapeutic levels of speech practice. Providing the user (SLP/child) with the ability to modify parameters such as exercises *before, during, or after* gameplay will help ensure the functionality and utility of the game as a therapeutic tool; this aligns with the implications for design put forth by Annema et al. [126] for therapy games. One of the aims of Apraxia World is to provide the child with a sense of autonomy during speech practice. Negotiation with their parent or SLP as to when they do exercises during gameplay would provide the child with a sense of control over their speech practice. However, to ensure that this negotiation does not lead to exercise avoidance, all game conditions need adjusting to ensure the ratio of gameplay to speech exercises is carefully balanced.

Similar to traditional gameplay, children undertaking gamified speech therapy want customizability in their game experience. The children generally liked the concept of buying items for their character. They purchased costumes, new weapons, and extra character power-ups. Children were motivated by a desire to customize their game character, and having items to purchase inspired them to collect coins and stars. Character customization is another method to help the child create an individualized gameplay

experience, potentially helping them further engage with Apraxia World as a therapeutic tool. Maintaining a child's motivation to use the game and engagement in speech practice over the long-term is vital for the success of Apraxia World. Both character customization and choice over when the speech exercises appear are flexible elements of Apraxia World aimed at supporting this. However, limitations in the inventory of items available were highlighted during the current study. One child commented that the costume items available did not match well and another highlighted that there were no girl clothes. The suggestion of being able to pay to change the name of their character was also made. Developing the range of items available for purchase in subsequent versions of Apraxia World would ensure a rich gameplay experience for the child, helping to maintain motivation and engagement.

This study was limited by the population demographics – only 4 of the 21 participants were female, and only one of them was in the SSD group. Although up to 2.85 times more males than females have a SSD [6], our sex ratio approaches neither that of SSD nor general populations. Seeking a better demographic balance in future studies will help to make sure Apraxia World appeals to a wide audience.

For this study, we focused on the engagement and usability aspects of Apraxia World, which serve as the foundation for ambulatory studies we plan to conduct later in 2018. Direct SLP input will not be available during gameplay; as such, we will automate the speech evaluation through ASR. While mobile ASR engines (e.g., PocketSphinx) lack the capabilities of server-based solutions [66], recent findings [134] suggest that running the ASR engine in “forced-alignment” mode can be used to assess pronunciation. While

this is generally not an option for general applications of speech recognition, in the context of speech therapy, the (target) spoken word is known in advance. Alongside the ASR, we plan to develop a therapist portal for managing the remote therapy application. Given that some children found Apraxia World too difficult, future versions will include more graduated level difficulty and adaptive difficulty, such that the game stays at an engaging level of difficulty as players' skills improve. Additionally, we will evaluate a *before-game* condition in the next version of Apraxia World, as the *during-game* and *after-game* conditions both had their own drawbacks.

3.8. Conclusion

In this paper, we presented Apraxia World, a mobile speech therapy tool built atop a full-fledged, multi-world platformer game. Apraxia World decouples speech production and game control to avoid limiting the type and variety of speech input; players complete speech exercises to make progress, but speech does not control character movement, which requires fine motor control. We conducted a user study to validate game functionality and evaluate how enjoyable children found gameplay alongside speech exercises. Overall, the children showed enthusiasm and engagement with Apraxia World and the novel mode of speech exercise delivery. Most of the children preferred to do exercises in the *after-game* condition, however, this was not unanimous; this indicates that future versions of the game should continue to offer flexibility in how players can do their speech exercises. The results of the study support the feasibility of Apraxia World as an augmentation to traditional clinic-based speech therapy.

4. A LONGITUDINAL EVALUATION OF TABLET-BASED CHILD SPEECH THERAPY WITH APRAXIA WORLD*

4.1. Overview

Digital games can make speech therapy exercises more enjoyable for children and increase their motivation during therapy. However, many such games developed to date have not been designed for long-term use. To address this issue, we developed Apraxia World, a speech therapy game specifically intended to be played over extended periods. In this study, we examined pronunciation improvements, child engagement over time, and caregiver and automated pronunciation evaluation accuracy while using our game over a multi-month period. Ten children played Apraxia World at home during two counterbalanced four-week treatment blocks separated by a two-week break. In one treatment phase, children received pronunciation feedback from caregivers and in the other treatment phase, utterances were evaluated with an automated framework built into the game. We found that children made therapeutically significant speech improvements while using Apraxia World, and that the game successfully increased engagement during speech therapy practice. Additionally, in offline mispronunciation detection tests, our automated pronunciation evaluation framework outperformed a traditional method based on

* This chapter has been accepted for publication in the ACM Transactions on Accessible Computing. Reprinted with permission. Hair, A., Ballard, K. J., Markoulli, C., Monroe, P., McKechnie, J., Ahmed, B., & Gutierrez-Osuna, R. A Longitudinal Evaluation of Tablet-Based Child Speech Therapy with Apraxia World. *ACM Transactions on Accessible Computing*. Forthcoming.

goodness-of-pronunciation scoring. Our results suggest that this type of speech therapy game is a valid complement to traditional home practice.

4.2. Introduction

The term speech sound disorder (SSD) refers to a group of disorders affecting the development of accurate speech sound and prosody production that are diagnosed in childhood [1]. Children with SSDs struggle with phonological representation, phonological awareness, and print awareness, which can lead to difficulties learning to read or reading disabilities [2], and negatively impact communication skills development [3]. Fortunately, children with SSDs often reduce symptoms and improve speech skills by working closely with speech-language pathologists (SLP) to undergo speech therapy [4]. For speech therapy to be effective, treatments must be “frequent, high-intensity, individualized, and naturalistic” [5] so that children can practice new habits and skills [6]. However, scheduling appointments with SLPs can be logistically difficult [7-9], and up to 70% of SLPs have waiting lists [10], which slows access to services. To meet high dosage requirements, clinic-based interventions must be supplemented with considerable home practice, typically directed by primary caregivers (e.g., parents, guardians). However, home practice sessions can be tedious for both caregivers and children, and busy caregiver schedules can decrease the amount of practice a child receives [11]. As such, there is a need for speech therapy systems that follow best practice principles, place less burden on the time and skill of caregivers, and make the therapy itself more engaging.

A promising approach to address barriers to frequent child speech therapy is to incorporate the therapy into digital games. Digital therapy games can have a positive

impact on child motivation and satisfaction [12], and have been shown to increase participant engagement and persistence [13, 14]. Most importantly, research has shown that computerized and tablet-based speech therapy interventions can be as effective as traditional interventions [15-21], although not all digital applications out-perform traditional methods [22] or produce clinically-significant results [23]. A number of game-like applications for speech therapy have been commercially developed and are available for purchase [24] (e.g., Apraxia Farm [25], Articulation Station [26], ArtikPix [27], Tiga Talk [28]). Children often enjoy using digital therapy interventions in short-term tests, and sometimes even play beyond the required time [29, 30]. However, applications often employ an arcade or casual game with simple play mechanics, which do not lend themselves to long periods of gameplay/speech practice and can quickly become tedious [31, 32]. Furthermore, many games do not include production feedback, which means that the therapy practice must still be closely supervised by caregivers. A handful of speech therapy games include pronunciation feedback [31, 33, 34], but much of this work is still preliminary.

To address the motivation and independence issues associated with home practice, we have designed a mobile game for speech therapy called Apraxia World that delivers repetition-based therapy to address childhood apraxia of speech (CAS). CAS is a neurological SSD that affects speech movements and can slow learning appropriate intensity, duration, and pitch for speech sounds [43]. Apraxia World was developed based on child feedback from early prototypes, and is intended for extended use to accommodate lengthy therapy treatments; we employed a participatory design approach [135] where

children, caregivers, and clinicians acted as informants and testers as the game progressed from prototype to the version presented here. Children play Apraxia World like a traditional mobile game with an on-screen joystick and buttons, but must complete short speech exercises to collect specific in-game assets that are needed to progress through the levels. In a pilot study [25], we evaluated a prototype version of Apraxia World to simulate a single therapy session conducted in an SLP office setting. In general, children were enthusiastic about playing the game and reported that the game made their speech exercises more fun than normal. However, that study did not assess long-term engagement and usage, or possible therapeutic benefits (i.e., pronunciation improvements).

In this article, we present the full-fledged version of Apraxia World and a longitudinal study to explore system usage, therapeutic benefit of home therapy with the game, and speech evaluation accuracy. In contrast to the prototype used for pilot testing, Apraxia World now includes automatic pronunciation evaluation to afford more child independence during practice. With this version of the game, we set out to answer the following research questions:

- RQ1: Do children remain engaged in the game-based therapy practice over a long period of play?
- RQ2: What level of pronunciation improvement do children achieve while playing Apraxia World?
- RQ3: How accurately do caregivers and our automated system evaluate pronunciation?

To answer these questions, we designed a longitudinal study that allowed us to examine child engagement and interest in the game over time, and compare therapeutic improvements to those reported for traditional practice. The study consisted of two four-week treatment phases with a two-week break in between. In one phase, children received pronunciation feedback from their caregivers in a Wizard-of-Oz manner (the system appeared automated, but actually had a human operator). In the other phase, children received feedback from the template matching framework. From our investigation, we found that:

- Children enjoyed the game, even over the long treatment period
- Game personalization was a popular aspect of Apraxia World
- Children made pronunciation gains with Apraxia World comparable to those reported for traditional clinician plus home-based speech therapy of similar intensity
- Caregivers tended to be lenient pronunciation evaluators, and
- Template matching outperformed goodness of pronunciation scoring in offline mispronunciation detection tests

The rest of this article is organized as follows. In Section 2 we present relevant background for digital speech therapy tools and automatic mispronunciation detection. Section 3 describes Apraxia World, the speech therapy program it delivers, and the mispronunciation detection framework. Section 4 details the experimental design of our longitudinal study, and the remaining sections present our results, discussion of findings, and concluding remarks. This article expands upon preliminary results that will be

presented as late-breaking work at the 2020 ACM CHI Conference on Human Factors in Computing Systems [35].

4.3. Background and related work

4.3.1. Digital speech therapy tools

Child speech therapy approaches can be grouped into two categories: linguistic- or articulation-based practice. Linguistic-based approaches address difficulties in using the correct sound to convey meaning [136]. As such, these therapy plans focus on organizing a child's sound system so they produce sounds in the appropriate context. Articulation-based approaches focus on the movement of articulators (e.g., tongue, lips) to produce speech sounds correctly [136]. A child will first learn the correct phoneme pronunciation by itself or in a simple word before practicing the sound in longer words or sentences. Both therapy approaches focus on drills and repetition. Previous work suggests that children receive the most benefit from frequent short sessions with randomly presented prompts, instead of repeated practice of one prompt [137]. The repetitive nature of these short sessions makes them excellent candidates for delivery via digital methods.

A variety of digital speech therapy interventions have been developed over the last 30 years. The Indiana Speech Training Aid (ISTRA) is a foundational project introduced in the late 1980s that used digital speech processing technology to provide speech therapy feedback to patients [61, 62]. ISTRA offered patient-specific computerized drill sessions with graphical feedback representing utterance scores (e.g., bar graphs, bull's-eye displays) and pronunciation quality reports. Some speech exercises were also delivered through game-like applications such as Baseball and Bowling, where pronunciation scores

were displayed as game performance [64]. Some 10-15 years later, researchers presented the Articulation Tutor (ARTUR), another computer-based speech training aid that provided specific feedback on how to remedy incorrect articulations and showed a graphical model of the correct articulator positioning [138]. Their evaluations revealed that feedback delivered through the system helped children improve articulator positioning. The Comunica Project is a digital speech therapy system from the mid-to-late 2000s for Spanish speakers [123] with three distinct components: PreLingua (basic phonation skills), Vocaliza [128] (articulation skills), and Cuéntame (language understanding). PreLingua contained a game-like child interface, Vocaliza mimicked flashcards, and Cuéntame presented simple open-ended responses or commands. Both Vocaliza and Cuéntame contained automatic pronunciation verification that allowed an SLP to track progress over time. Tabby Talks [65, 66] is a more recent therapy application that included a mobile interface for patients, a clinician interface with progress reports, and a speech-processing engine. Speech exercises were delivered through a flashcard or memory game interface, both of which recorded utterances for later evaluation. The system processed audio on a remote server and included pronunciation progress in the clinician reports, but did not provide real-time feedback to the child. Results from a pilot test [16] indicated that this type of application is a viable complement to traditional clinic-based sessions, but that additional engaging features are needed to make the application more interesting for children. These previous projects illustrate the rich history of working to improve digital speech therapy and provide a strong foundation for future speech therapy tools.

To address the issue of low motivation due to the repetitive and boring nature of home therapy practice, researchers have also worked to deliver speech therapy exercises through standalone digital games. Lan et al. [34] developed Flappy Voice, a game where players fly a bird through obstacles by modulating their vocal loudness and pitch to change altitude. Following this concept, Lopes et al. [139] presented a game where the player helps the main character reach objects by producing a constant-intensity sustained vowel sound while the character moves. Feedback is provided by moving the character up or down to represent intensity changes. While these two games focused on modulating or maintaining specific sounds, the majority of speech therapy games have focused on keyword repetitions. For example, Navarro-Newball et al. [32] designed Talking to Teo, a story-driven game in which the player must correctly complete a series of utterance repetitions to complete actions for the main character. Utterances are evaluated with a custom speech recognizer and the success of in-game actions depends on the quality of production. Cler et al. [140] proposed a ninja versus robot fighting game for velopharyngeal dysfunction therapy where the player must repeat nasal keywords correctly to attack the enemy character. Nasality was measured with an accelerometer worn on the player's nostril. Duval et al. [68, 141] introduced SpokeIt, a storybook-based game designed for cleft palate therapy, where the player helps voiceless characters navigate an unfamiliar world by producing target words associated with actions. This game provides pronunciation feedback using built-in speech recognition and is designed to afford long-term play by procedurally generating level content. Ahmed et al. [26] evaluated five speech-driven arcade-style therapy games with stakeholders and typically-developing

children. Children preferred games with rewards, challenges, and multiple difficulty levels, indicating that overly simple games may not be suitable for speech therapy. These studies demonstrate the variety of methods available to integrate speech exercises into digital games and the diversity of genres that can facilitate gamified speech therapy.

4.3.2. Automatic mispronunciation detection

Techniques based on automatic speech recognition (ASR) show the potential to improve child pronunciation skills by enabling automatic mispronunciation detection within speech therapy applications [142]. The standard method for detecting mispronunciations is the goodness of pronunciation (GOP) proposed by Witt and Young [90]. The GOP method scores phoneme segments based on a probability ratio between the segment containing the target phoneme and the most probable phoneme. Although the GOP method was originally developed for second language learning, it has also been adapted to process speech from children with SSDs [38, 39]. In addition to GOP, researchers have presented various methods to evaluate child speech for pronunciation training and speech therapy applications. For example, Saz et al. [143] deployed speaker normalization techniques to reduce the effects of signal variance so that their pronunciation verifier could better detect variance in phoneme productions. Specifically, the authors examined score normalization and maximum a posteriori model adaptation to increase separation in the log likelihood outputs of a Hidden Markov Model (HMM) pronunciation verifier. Their approaches reached 21.6% and 15.6% equal error rates, respectively. Shahin et al. [37] proposed a phoneme-based search lattice to model possible mispronunciations during speech decoding. Their system identified incorrectly pronounced phonemes with

over 85% accuracy. In later work [144], the authors developed a mispronunciation detection approach using one-class Support Vector Machines (SVM). Their method used a deep neural network (multilayer perceptron) to extract 26 speech attribute features before training an SVM per phoneme using correctly pronounced samples. This method outperformed GOP for both typically-developing and disordered speech from children. In contrast to the above methods that only examine phoneme correctness, Parnandi et al. [65] presented a series of speech recognition modules to identify errors associated with CAS. These included an energy-based voice activity detector, a multilayer perceptron with energy, pitch, and duration features to identify lexical stress patterns, and an HMM to detect error phonemes. They achieved 96% accuracy detecting voice delay, 78% accuracy classifying lexical stress, and 89% accuracy identifying incorrect phonemes. Although the described methods demonstrate performance close to or above the clinically-acceptable threshold of 80% accuracy [142], they require phonetically-annotated data. This means researchers often must annotate custom corpora or rely on forced alignment, which can yield inaccurate segment times on mispronounced or child speech.

Detecting child mispronunciations is made even more challenging by the inherent difficulty of processing child speech due to inconsistencies in speech features. For example, Lee, Potamianos, and Narayanan [35] reported that children, specifically those under 10 years of age, exhibit “wider dynamic range of vowel duration, longer segmental and suprasegmental durations, higher pitch and formant values, and larger within-subject variability.” Compounding these issues is the limited number of appropriate child speech corpora; for example, the OGI Kids’ Speech Corpus [145] and PF-STAR [146] only

contain typically-developing speech, the PhonBank [147] collection contains corpora of disordered speech from children [148-150], but without ready-to-use recording annotations, and the recently released BioVisualSpeech corpus only contains European Portuguese speech [151]. As a result, acoustic models tend to be built using adult speech corpora, which severely limits system accuracy. In situations where speaker data are limited, template matching [152] may be an appropriate method to provide speaker-specific pronunciation feedback. Template matching is a well-established speech recognition technique that uses dynamic time warping to compare a test utterance to previously collected examples of target words (“templates”). These templates can also be used to model the correct pronunciation of words. For example, this method has been used within a pronunciation practice application for second-language learners [64]. Template matching has also been successfully incorporated into child speech therapy systems as a pronunciation evaluator [61, 63, 153]. Template matching evaluations have been shown to correlate with human evaluations when using high-quality productions from the speaker as pronunciation templates [63]. This method successfully takes advantage of small amounts of child speech and can lower the burden of collecting calibration utterances for SLPs, caregivers, and children. Additionally, template matching does not require phonetic transcriptions, as words are evaluated holistically, which makes curating speech recordings even simpler for end users.

4.4. Apraxia World

4.4.1. Game design

Apraxia World is a brightly-themed 2D platformer game built by customizing and expanding an existing game demo (Ekume Engine 2D) using the Unity Game Engine. We explored building a game from scratch, but due to cost and time constraints, we instead opted to modify an available game. The Ekume Engine 2D was selected for its rich collection of pre-made assets, age-appropriate theming, and familiar gameplay mechanics. Players control a monkey-like avatar to navigate platforms, collect items, and fight enemies as they work to get across the finish line. Apraxia World includes 40 levels (eight levels for each of the five worlds), seven different characters, and an in-game store. These features align with recommendations that digital speech therapy systems include more game-like elements [26]. Figure 6 (a) and (b) show the level design from two different worlds (jungle and desert).

From pilot testing, we found that children enjoyed the gameplay, speech exercises did not impede gameplay, and the game made the exercises more fun, although children generally completed the minimum number of exercises, even when offered in-game rewards [25]. Since these initial tests, we modified the game as follows: we count all utterance attempts towards the session goal, similar to traditional practice; we added an “energy” timer that encourages regular star collection; we implemented an exercise progress save mechanism so children can take a break; and we added automatic speech processing (technical details in Section 4.4.3). The game mechanics are described below.



(a)



(b)



(c)

Figure 6 (a) A level from the jungle world (b) A level from the desert world (c) Speech exercise popup with both pictorial and text cues.

There are a handful of popular strategies for controlling speech therapy games: producing sustained sounds [34, 114, 139], speaking target words corresponding to actions [33, 68], or controlling specific aspects of speech [24]. While these strategies have the benefit of providing implicit feedback (progress in the game means the speech sounds are being correctly produced), they can be problematic if the player struggles to form the target sounds. Additionally, it can be difficult to navigate a character through a two-dimensional world using only speech to control complex movements or simultaneous commands (i.e., running and jumping). As such, Apraxia World incorporates speech as a secondary input used to collect in-game assets, specifically, yellow stars spread throughout the levels; see Figure 6 (a).

When the player attempts to collect the star by touching it with their character, the game pauses and a themed speech exercise popup appears; see Figure 6 (c). Within the exercise, the player is prompted to capture pronunciation attempts using separate button presses to start and stop an audio recorder. As the player follows the exercise prompts, a human listener or automated system evaluates their utterances and the game displays the appropriate feedback (e.g., “Good job!” or “Not quite!”). Once the player attempts the specified number of utterances (either correctly or incorrectly pronounced), the popup disappears and the star is added to their inventory. Collecting the exercise stars is mandatory, as the game requires a certain number of stars to complete the level; the required number of stars per level and utterances per star can be configured by clinicians. Levels have between 7 and 12 stars scattered throughout, which reappear after a short delay to encourage the player to continue to explore.

Apraxia World displays a timer showing how long until the avatar's "energy" runs out. This timer depletes continuously and must be replenished by doing speech exercises. When the character runs out of "energy," it starts to move slowly, which makes the game more challenging. This encourages players to complete speech exercises regularly during gameplay. When players complete speech exercises, they earn 10 seconds for a correct pronunciation and 5 seconds for an incorrect pronunciation. In this way, players are rewarded for all pronunciation attempts, but correct attempts are more strongly rewarded to motivate them to maintain practice effort.

Apraxia World provides players the option to purchase six additional characters and buy items in the store to encourage personalization. Players buy these items using coins (in-game currency) that they collected throughout the levels or that were awarded for doing speech exercises. The store sells costume items (pants, shirts, hats, and accessories) to dress up the characters, different weapons, and power-ups that give the characters "superpowers." Some of the items available for purchase are displayed in Figure 7. The power-ups last only briefly and provide the player a protector shield (invincibility), allow them to fly, attract coins "magnetically," or increase gathered points by a multiplier. Power-up duration can be extended via purchase, but is always temporary. The different characters and costume items are purely for cosmetic personalization; they have no effect on how the game plays. The different weapons and power-ups do impact gameplay, in order to accommodate different play strategies.

Apraxia World saves exercise progress when a player leaves the level, so they can take a break from their exercises and come back without losing their work. Once the player

comes back to the level, their character starts back at the beginning, but the previous therapy progress is reloaded so that they do not have to repeat exercise attempts. After the player completes the required number of speech exercises, the game does not allow them to do additional exercises. At this point, the player can continue until they finish the level or lose, whichever comes first. The game then locks the levels until the next day, as players are only allowed to complete one level per day to limit therapy exposure and avoid game fatigue.

Even though the controls employed in *Apraxia World* are standard for tablet games, they may not be completely accessible for populations undergoing speech therapy. For example, some children with movement-based speech disorders, such as CAS, have motor impairments [154]. Other groups going through speech therapy may also experience difficulties with specific movements (e.g., children with Autism Spectrum Disorder [155]). Although not implemented in this study, the controls could easily be mapped to an external joystick or adaptive controller to make the game more accessible to those who want to use it.

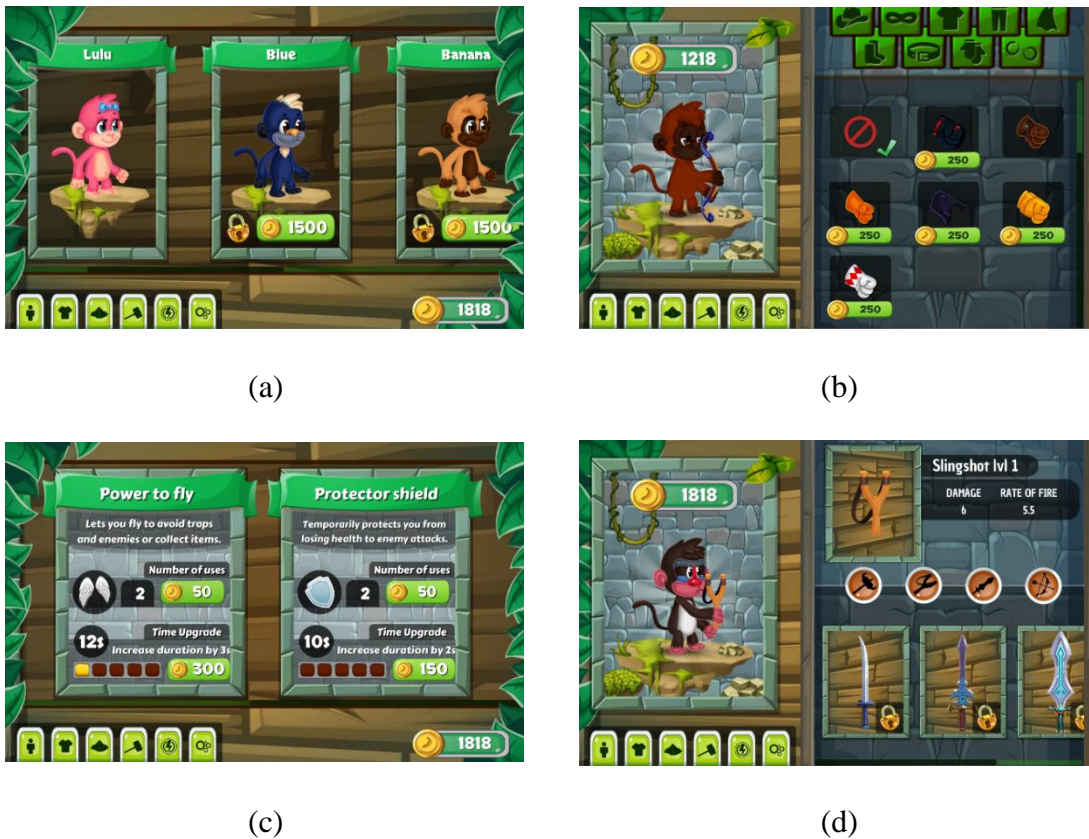


Figure 7 (a) Various characters available for purchase (b) Costume items to dress up the character (c) Power-ups to give the character “superpowers” (d) Weapons with different attack behaviors.

4.4.2. *Speech therapy program*

Apraxia World offers two types of feedback: knowledge of response (KR) and knowledge of correct response (KCR). KR informs the learner of the correctness of their response, whereas KCR informs the learner of the correct response, so that they can judge the correctness of their response themselves [156]. KR has been shown to help people using digital speech therapy systems make improvements comparable to those from

traditional speech therapy [157], although it is up to system designers to decide what granularity of feedback to deliver. Apraxia World provides word-level KR feedback alongside the speech exercises by telling the child if an utterance was correct (“Great job!”) or incorrect (“Try again!” “Not quite!”), i.e., the correctness of the response. The game also offers KCR by providing the child with an example of the correct pronunciation whenever they need help, thereby informing them of the “correct response;” the child can hear the pronunciation sample by pressing a button displayed on the speech exercise popup. These example pronunciations were generated in advance using the Google Text-to-Speech service [158].

The speech exercises in Apraxia World are based on a Principles of Motor Learning approach [137, 159], which prescribes a structure of practice and feedback to stimulate long-term learning. This means that Apraxia World can accommodate both linguistic- or articulation-based practice, depending on the target words selected by the SLP. First, an SLP assessed each child to determine problematic speech sounds and stimulability for correct production of the problematic sounds in real words. For our purposes, a sound was stimuable if the child could accurately imitate it multiple times and produce it without a model on at least 5 attempts within a 30-minute session. The SLP then selected one or two stimuable speech behaviors to address during treatment. Selecting stimuable behaviors increases the likelihood that the children have some internal reference of correctness, enabling them to benefit from simple KR feedback (i.e., word-level correct/incorrect feedback). Additionally, caregivers were asked to conduct five minutes of pre-practice before each home therapy session to remind the child how to produce a correct response

and interpret the feedback provided in the game. The principles of motor learning employed during practice with the game were random presentation order of stimulus, variable practice (i.e., varied phonetic contexts for each target sound), moderate complexity for the child's current production level, and high intensity (100 production attempts per session). To give clinicians flexibility when selecting target words, we curated a word pool that includes approximately 1,000 words, with both single- and compound-word targets. Each of these targets has a corresponding cartoon-style image to use as a pictorial prompt; see Figure 8 for examples of prompt images.

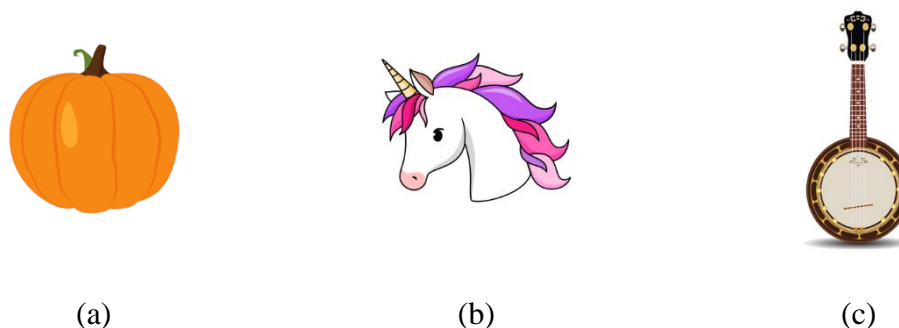


Figure 8 Pictorial prompts for (a) pumpkin, (b) unicorn, and (c) banjo.

4.4.3. Pronunciation evaluation

Apraxia World provides pronunciation feedback based on either automatic pronunciation evaluation or human evaluator input via a Bluetooth keyboard. Automatic pronunciation evaluation is carried out using template matching (TM) [152]. This method

compares a test recording against sets of “template” recordings to identify which set it most closely matches. We selected TM because it has very low data requirements (i.e., a small set of speech recordings per player), an important consideration for child speech therapy applications due to limited available data. This allows us to collect minimal speech data from each child, making the system easier for clinicians to configure, while still delivering child-specific pronunciation feedback. Additionally, TM does not require phonetic labels, making setup even simpler for clinicians. Our algorithm runs directly on the tablet, which avoids data transmission delays and allows the game to be played with limited or unstable internet connectivity.

In our approach, correct and incorrect pronunciations of a word collected from the child are used as templates when determining if a new recording of the same word is pronounced correctly. The speech processing pipeline is illustrated in Figure 9 (a). Given a recorded utterance (16 kHz), the audio signal is pre-emphasized before 13 Mel-frequency cepstral coefficients are extracted from 32 ms frames with 8 ms overlap, which are then normalized with mean cepstral normalization (MCN) [160]. Leading and trailing silence segments are removed using an energy threshold to form the final feature vector.

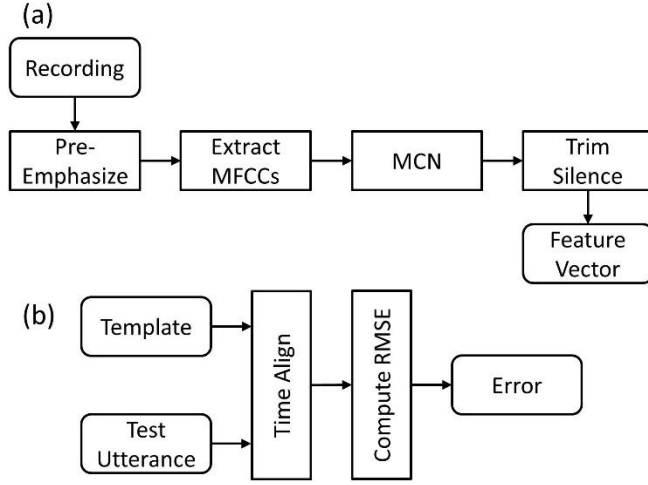


Figure 9 (a) Spectral information is extracted from an utterance, mean cepstral normalized (MCN), and trimmed (b) Template and test utterances are aligned and scored based on RMSE.

The TM process is shown in Figure 9 (b). Template t and test utterance u are aligned end-to-end using dynamic time warping (DTW). From this alignment, we compute a pronunciation distance between the two as:

$$d(t, u) = \begin{cases} \frac{\|dtw(u,t)-t\|_2}{len(t)}, & \text{if } len(t) > len(u) \\ \frac{\|dtw(t,u)-u\|_2}{len(u)}, & \text{otherwise} \end{cases}, \quad (1)$$

where $dtw(x, y)$ time-aligns the frames in x to y . To classify the test utterance, we compare its distance against those for pairs of correct and incorrect pronunciation templates for that target word. Let T_C be the set of correct pronunciation templates and T_I be the set of incorrect pronunciation templates. The correct pronunciation score s_C is the median TM distance for all unique pairs of correct pronunciation templates:

$$s_C = \text{median}(\{d(j, k) | \forall j, k \in T_C, j \neq k\}), \quad (2)$$

whereas the incorrect pronunciation score s_I is the median TM distance for all pairs of correct and incorrect pronunciation templates:

$$s_I = \text{median}(\{d(j, i) | \forall j \in T_C, \forall i \in T_I\}). \quad (3)$$

The score for a test utterance u is the median TM distance to all correct pronunciation templates:

$$s_u = \text{median}(\{d(j, u) | \forall j \in T_C\}). \quad (4)$$

In a final step, we label the test utterance pronunciation as incorrect (0) or correct (1) as:

$$\text{label}(u) = \begin{cases} 1, & |s_u - s_C| \leq |s_u - s_I| \\ 0, & \text{otherwise} \end{cases}. \quad (5)$$

To enable real-time evaluation, correct and incorrect pronunciation scores s_C and s_I are pre-computed; only the test utterance needs to be scored at runtime. Test utterances are scored against correct pronunciation templates, as we expect the child to form correct pronunciations similarly, but there are likely multiple incorrect pronunciations due to the child struggling to produce sounds consistently.

As part of the experimental setup, an SLP collects the necessary template recordings from the child. This is done using a separate companion app called Apraxia World Recorder (AWR) to make it easy for clinicians to select speech targets, which is critical when including ASR in speech therapy [161]. AWR allows the SLP to select a tailored set of target words for the child, collect calibration recordings and labels, and export the pre-processed templates for Apraxia World to use during real-time pronunciation evaluations. AWR also enables the SLP to swap target words as the child

makes progress in their therapy, which is important for customization. Figure 10 shows the recording interface for a target word in AWR.

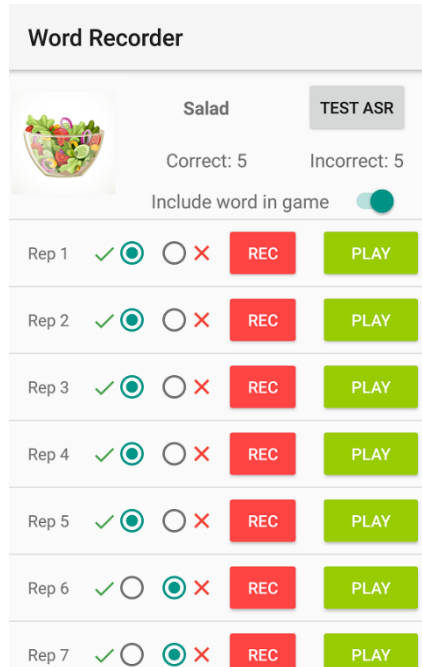


Figure 10 Word recording interface in AWR. Recordings are labeled as correctly (green check) or incorrectly (red x) pronounced.

4.5. Experimental design

4.5.1. Participants

We recruited eleven children (10 male, 5-12 years old) with SSDs in the Sydney (Australia) area via print ads in local magazines, word-of-mouth, and clinician recommendations. Although this sample size may appear small, recruiting a large number

of participants was infeasible given that the target population is limited and the protocol requires considerable time investment on the part of caregivers. All children were native Australian-English speakers with a diagnosis of SSD from their referring clinician. For the purposes of this article, SSDs were determined by difficulty producing multiple speech sounds by the expected age. All had previously received community-based therapy, but were previously discharged or on break during our study. Participants had normal receptive language, hearing and vision, and no developmental diagnosis or oral-facial structural anomalies. One participant (male) unenrolled from the study due to schedule conflicts, so his data were not included in this analysis. The remaining ten participants completed the treatment protocol. Nine participants had an idiopathic SSD (i.e., unknown cause) and the tenth had a genetic condition causing mixed CAS and dysarthria. All procedures were approved by the University of Sydney's Human Research Ethics Committee and all children and guardians provided written informed assent/consent, respectively, before participating in the study.

4.5.2. Protocol

In this study, we examined child engagement over time, pronunciation improvements, and caregiver and automated pronunciation evaluation (TM) accuracy. The study consisted of five phases: setup, two treatment blocks, a between-treatments break, and a post-treatments break. We do not report on the post-treatment break in this article, as observations from the break are addressed in a forthcoming clinician-focused manuscript. Setup involved selecting appropriate target words based on the child's therapy needs, recording the calibration utterances in AWR (see Figure 10), and familiarizing the

child and caregiver with Apraxia World. Children practiced over two counterbalanced phases (five participants received automated feedback first and five participants received caregiver feedback first) so that we could examine the effects of utterance evaluation source (caregiver versus automated system). In one treatment block, children received pronunciation assessments from their caregivers in a Wizard-of-Oz fashion (the system appears automated, but actually has a human operator). In the other treatment block, they received automatic pronunciation assessment from the TM framework. At the end of each treatment block, a representative random subset of utterances was selected for pronunciation evaluation by an SLP. The experimental protocol is illustrated in Figure 11. During the treatment blocks, children played Apraxia World as long as needed to complete their speech exercises, four days per week. The children played Apraxia World on Samsung Tab A 10.1 tablets and wore a headset with a microphone to record their speech during exercises.

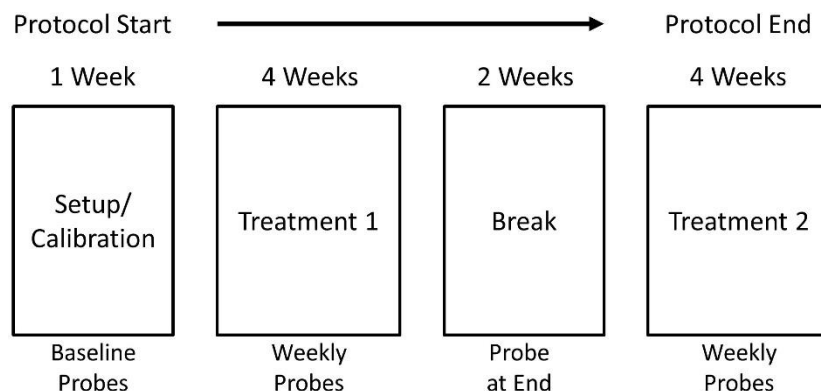


Figure 11 Experimental protocol with two treatment blocks. Pronunciation is probed before treatment and weekly during treatment.

Each treatment block repeatedly presented a different set of 10 words selected by an SLP to correspond with the child's specific speech difficulties. During gameplay, Apraxia World prompted the child to say one of their target words selected at random. Target words were not repeated until all had been presented the same number of times. In total, each child practiced 20 different words across the two treatment blocks; see Table 1. Pronunciation abilities were probed before each treatment block and weekly during the treatment blocks. Pronunciation probes contained both practiced (included in Apraxia World) and non-practiced (not included in Apraxia World) words to measure carryover effects (not reported here). A child's pronunciation ability was scored as the percentage of utterances containing the correctly produced target sound within a given probe. During the probe, children were not penalized for production errors on any sound other than the stimuable sounds selected by the SLP. Subjective questionnaires were administered twice during each treatment block and again following treatment to track and compare engagement during both treatment conditions (children were asked how hard they were trying in the game and if they wanted to continue playing; caregivers were asked if the children were engaged). Gameplay logs were captured for analysis of how children spent time in the game. Furthermore, all speech exercise attempts were recorded and stored for offline examination.

Speaker	Phase 1 words	Phase 2 words
m1	chair, chasing, cheese, chimpanzee, chopping, ginger beer, giraffe, jaguar, jam, jumping	eagle, eating, egg, elephant, kennel, key, pebble, seven, telescope, tennis
m2	bus, horse, house, kiss, mice, sail, saw, sea, seat, sun	lady, lake, lamb, lava, leaf, licking, light, lion, lip, loud
m3	binoculars, boa constrictor, kingfisher, ladder, leopard, letter, lizard, lobster, possum, stomach	biscuit, bulldozer, button, calculator, cauliflower, lettuce, pattern, pocket, salmon, scissors
m4	lair, lake, laughing, lawn mower, leak, letter, licking, lip, lobster, look	back, bat, cactus, dagger, magic, packet, pattern, shack, tap, taxi
m5	bed, bird, dirty, earth, egg, fur, girl, men, stem, ted	barber, bathroom, beehive, dinner, hammer, ladder, paper, peanut, tiger, toilet
m7	claw, climber, clip, flamingo, flash, slower, fly, glass, globe, glove	garage, garbage, jam, jumping, jungle, kitchen, teach, teacher, torch, watch
m8	shark, sharp, shed, sheep, shelf, shirt, shoe, shop, shovel, shower	chair, cheese, chicken, chocolate, chopping, jail, jam, jelly, juggle, jumping
m9	shampoo, shave, shed, sheep, shirt, shoe, shop, shore, shovel, shower	beach, giraffe, jam, jaw, jelly, jellyfish, jumping, kitchen, teacher, torch
m10	earth, earthquake, feather, mammoth, python, stethoscope, tablecloth, teeth, there, toothpaste	barber, climber, cucumber, dancer, deliver, diver, goalkeeper, kingfisher, pencil sharpener, toilet paper
f1	binoculars, burglar, caterpillar, curl, earth, hamburger, purr, purse, turkey, unicorn	chair, garbage, kitchen, peach, pencil sharpener, sponge, teacher, torch, watch, witch

Table 1 Words selected to address speaker-specific speech difficulties.

4.6. Results

We conducted four types of analysis: gameplay, therapeutic progress, audio quality, and pronunciation evaluation. To analyze gameplay, we investigated how long participants spent playing the levels, how far they progressed in the game, what slowed them down, and what they purchased in the in-game store. We also collated surveys to identify response trends; child and caregiver surveys from participant m9 were not returned, so only his game logs and audio could be explored. To examine therapeutic progress, we compared speech performance at baseline against performance at the final probe (after each treatment phase). We measured audio quality by inspecting the collected child audio and then gathered ground-truth correct/incorrect labels from an SLP for a subset of recordings. Finally, we analyzed caregiver and automated evaluations using the SLP labels as ground-truth, and compared their performance against goodness-of-pronunciation scoring.

4.6.1. *Gameplay analysis*

In a first step, we examined how long children spent within a level throughout the study. On average, participants spent just under 20 minutes per day playing a level ($\mu = 19.5$, $\sigma = 14.3$). Results are shown in Figure 12. When comparing the two treatment phases, for all participants but one⁴, there was no significant difference in the amount of

⁴ The significant difference in playtime for participant m4 arose due to a clinician reducing the number of stars required to finish the level, but increasing the number of exercises needed to earn each star. This resulted in less gameplay, while maintaining the same therapy dosage.

time spent in a level between the TM feedback phase and caregiver (CG) feedback phase. Large play time values where a child left the game unattended for long periods with a level open were excluded from the graph.

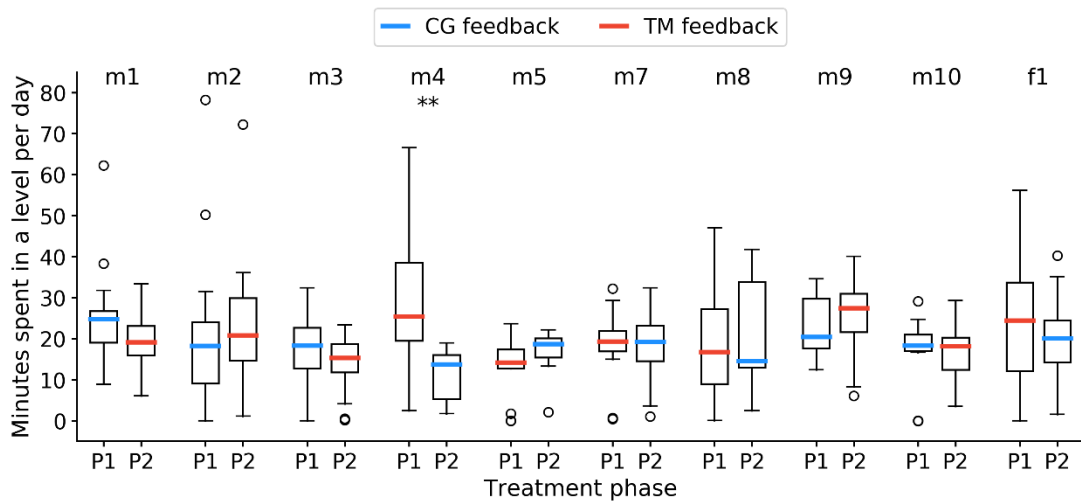


Figure 12 Minutes spent within a level per day for treatment phases one (P1) and two (P2) (indicates $p < 0.05$, two-sample t-test).**

Next, we analyzed game difficulty by examining the highest level each player was able to reach; see Table 2. Game progress was varied; four participants made it to level 25 and one progressed all the way to the penultimate level, while only two struggled to leave the first world (m4 and m8). This indicates that level 25 may be a reasonable upper limit on how far most children can progress over the two phases, which suggests that the game may support even longer treatments. Given the age range of our participants, we calculated

the correlation between progress in the game and age, and found that these factors were weakly correlated (Pearson’s $r = 0.29$, $p = 0.41$, $n = 10$). This indicates that age did not significantly influence progress, so progress was more likely affected by interest or skill with tablet-based games (e.g., the participant who made it farthest in the game was in the middle of our age range). To identify which aspect of the game prevented children from progressing through levels, we examined the causes of the in-game characters to “die.” For all participants, character deaths were significantly more likely to be caused by obstacles than by enemies ($p \ll 0.01$, paired t-test).

Participant	m1	m2	m3	m4	m5	m7	m8	m9	m10	f1
Max level	25	19	21	7	25	25	5	19	39	25

Table 2 Maximum progress in the game for each player.

We found that shopping was popular across participants, according to the number of purchases made from the in-game store and child survey responses. Caregivers also confirmed in their surveys that children enjoyed shopping in the Apraxia World store. All participants bought at least one power-up from the store. By far, the most popular power-up was flight; see Figure 13 (a). This was often used by children to navigate around challenging portions of levels, which makes sense given that the obstacles were significantly more likely to cause character “deaths.” Progress in the game and the

purchase of the flying power were weakly correlated (Pearson’s $r = 0.18$, $p = 0.62$, $n = 10$), indicating that powerups did not unduly aid players in their progress. All players purchased clothes, and most purchased additional weapons for their characters, but not all players purchased new characters. See Table 3 for the number of items purchased by each player.

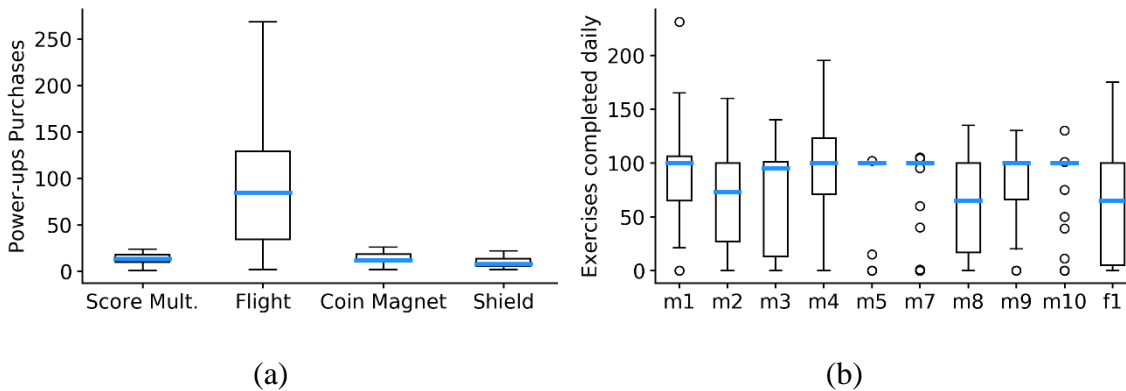


Figure 13 (a) Powerup purchases across all participants (b) Exercises completed per day.

In their survey responses, children reported enjoying the game ($n = 9$ of 9) and many indicated that they would like to continue playing ($n = 8$ of 9). Nine children actually played the game at least once after the study concluded according to the game logs, which confirms that they enjoyed AW enough to want to play without external pressure. Children also said that they were trying “very hard” while playing the game ($n = 8$ of 9), corroborating that they put effort into playing the game and stayed engaged. We

found a few repeated themes in what the children enjoyed about the game. Specifically, they reported enjoying fighting the enemies (“*defeating the big gorillas,*” “*fighting the bad guys*”), making purchases in the store (“*buying the gear,*” “*I bought a lot of characters,*” “*buying things for my character,*” “*buying clothes and accessories*”), riding animals with their character (“*I liked the fox,*” “*Level 4 had a fox – I liked that*”), and making progress through the game (“*Unlocking new levels,*” “*Moving up a level [every day],*” “*that every level has new things*”). One of the younger players (7 years old) was very proud of his progress in the game, stating “*I am up to the next map... I am up to level 10 now*” during a check-in with the SLPs. Caregivers reinforced via survey response that children enjoyed the game ($n = 9$ of 9) and some emphasized how much the children found the game motivating ($n = 8$ of 9 said motivating or highly motivating) or enjoyable. One caregiver said that their “*son wanted/asked to do practice, which [had] never happened before.*” All caregivers said that the children were engaged in the game ($n = 9$ of 9).

Although the children generally liked the game, they did dislike a few aspects. The children reported that they found the word repetitions boring (“*Getting bored because I just need to get coins and stars,*” “*Saying the same words got boring after a while*”) and that the game became too difficult (“*I didn’t like defeating some of the bad guys because it was sometimes hard,*” “*Sometimes tricky bouncing high enough,*” “*Not being able to get past a spot*”). They also disliked the software bugs (“*Game freezing,*” “*Freezing*”), which will be eliminated with further code testing.

Participant	m1	m2	m3	m4	m5	m7	m8	m9	m10	f1
Clothing	29	13	7	11	5	6	35	23	23	27
Weapons	8	5	9	7	3	6	0	6	7	2
Characters	5	1	2	1	0	3	0	2	3	6

Table 3 In-game purchases made by players during the study.

4.6.2. Therapy analysis

As a measure of therapy adherence, we examined the number of speech exercises completed daily by the participants, according to the game logs. Results are shown in Figure 13 (b). On average, children completed 76.0 speech exercises (i.e., word production attempts) per day during treatment ($\sigma = 43.3$). The average number of exercises completed daily was lower than the target dosage because, aside from caregiver supervision, there was nothing forcing children to complete all of their exercises before putting down the tablet for the day. As such, it is notable that children came somewhat close to the target dosage with the game being their primary motivation. Although therapy dosage was set at 100 exercises per day, children sometimes completed more exercises than prescribed, as seen in Figure 13 (b). This could have occurred if a player completed exercises in a level, exited before reaching 100 exercises (meaning the game had yet to lock for the day), started a different level, and then completed exercises in the new level.

Pronunciation improvements were measured according to the absolute percent change in correct target sounds produced in the probes immediately before and after a

treatment phase. Results are shown in Figure 14. Children experienced an average absolute improvement of 56.6 percent ($\sigma = 35.7$) when receiving TM feedback and 61.5 percent ($\sigma = 22.8$) when receiving caregiver feedback, and these differences were not statistically significant ($p = 0.73$, two-sample t-test). Children who received caregiver feedback first showed a stronger improvement across both treatment phases ($\mu = 67.3$, $\sigma = 33.5$) compared to children who received TM feedback first ($\mu = 50.8$, $\sigma = 23.3$), although the order effects were not significant; one-way Analysis of Variance: $F(2,7) = 0.85$, $p = 0.47$. Neither treatment group showed significant differences in improvement between the first and second phase of treatment (caregiver first: $p = 0.76$, TM first: $p = 0.89$, two-sample t-test).

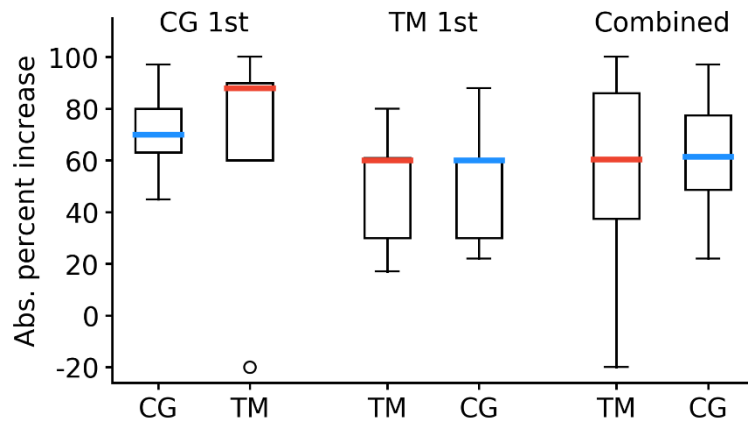


Figure 14 Absolute increase in pronunciation scores at the beginning and end of each treatment phase for caregivers (CG) and template matching (TM).

Children felt that the TM did not provide accurate feedback, which implies that they must have been doing some self-evaluation while playing the game (“*sometimes it is wrong,*” “*Game gives the wrong feedback,*” “*The computer is wrong a lot,*” “*Sometimes it is right but sometimes it is wrong*”). Regardless of how children perceived the automated feedback, they still made pronunciation improvements with both evaluation methods. Importantly, caregivers reported in their survey responses that this type of therapy generally fit easily into daily life ($n = 7$ of 9) and that they felt confident using the tablets to deliver the therapy ($n = 9$ of 9). They also responded that they were satisfied with the children’s speech therapy progress ($n = 9$ of 9 said satisfied or extremely satisfied) and that they would like to use Apraxia World either exclusively ($n = 5$ of 9) or combined with traditional paper worksheets ($n = 4$ of 9) to help with future speech practice.

4.6.3. *Quality of audio recordings*

Before we computed evaluator performance, we needed to determine the quality of the recordings to make sure that the participants were able to successfully capture entire utterances with limited background noise and distortions. Therefore, we manually listened to each recording to assign them into five categories: clipped (part of the recording cut off), containing background noise, unusable (speaker unintelligible), containing significant microphone noise, or good (usable for ASR analysis). Statistics on the gathered audio are displayed in Table 4. Overall, roughly 46% of the 27,700 recordings collected are of sufficiently good quality to use in our analysis. Clipped audio accounted for the majority of the remaining recordings (~33%). The percentage of usable recordings

compares favorably to that reported in another study where a tablet-based learning application was used to collect child audio for offline analysis [162].

Total Utterances	27,700
Good Utterances	12,742 (46%)
Clipped Utterances	9,141 (33%)
Unusable Utterances	3,878 (14%)
Background Noise	1,385 (5%)
Microphone Noise	554 (2%)

Table 4 Recorded utterances gathered during gameplay.

On average, children wore their headset during 92% of their therapy sessions (the game logged if the headphones were plugged in). Given such high level of adherence, it was surprising that many of the recordings were of low quality. This suggests that the microphone may have not been properly placed in front of the children’s mouth and was instead either too far (many of the recordings were quiet and difficult to hear) or too close (other recordings included puffs). A number of the recordings included significant distortions consistent with children accidentally holding their hand over the microphone or brushing it while speaking.

4.6.4. *Manual and automatic pronunciation evaluation*

We examined pronunciation evaluation performance using a representative subset of recordings (selected evenly from across both treatment phases) from those that had been classified as “good;” see previous subsection. Each of these recordings ($n = 2,336$) was manually labeled by an SLP, who identified if the utterance contained pronunciation errors (sound substitution or deletion). Overall, 82% of the utterances were labeled as having an error, or an average of 1.2 phoneme errors per utterance. The probability density for the number of phoneme errors per utterance is shown in Figure 15. We also identified where the phoneme errors occurred: 30% of errors occurred on the first phoneme, 27% occurred on the final phoneme, and the rest occurred in the middle of the utterance.

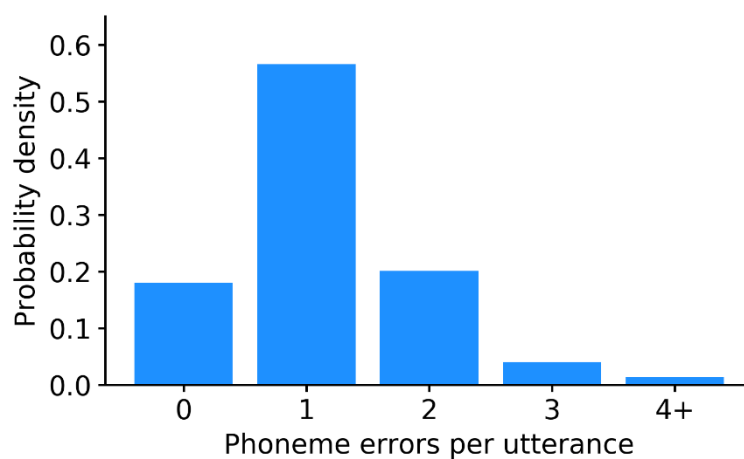


Figure 15 Probability density for the number of phoneme errors per utterance.

We used the SLP labels as ground-truth to calculate word-level performance of the TM algorithm and caregivers' pronunciation evaluation. For our calculations, we defined a true positive as a successfully identified mispronunciation and a true negative as a successfully identified correct pronunciation, which is the common notation in mispronunciation detection literature. Using these definitions, we computed the true positive rate (TPR) and true negative rate (TNR) for the caregivers and TM evaluations pooled across all participants. For caregivers, the TPR (27%) was much lower than the TNR (87%), indicating that they may have been lenient in their evaluations or that they struggled to identify mispronunciations. In contrast, the TM algorithm had higher TPR (65%) and lower TNR (28%), suggesting that the system was better at identifying mispronunciations than correct productions.

To examine if the location of the mispronounced sound affected TM performance, we took the subset of SLP-labeled utterances with only one mispronounced sound and split the recordings into three sets: error on the starting sound, error on a middle sound, or error on the final sound. We only calculated TPR because all of these utterances contain an error, so there are no true negatives. With these sets, we found that the TM yielded TPRs of 64%, 65%, and 61% for starting errors, middle errors, and ending errors, respectively. This suggests that the TM framework is somewhat robust to error location, although the detection of final sound errors was slightly less than for other error locations. We similarly split the SLP-labeled subset with only one mispronounced sound by whether the error occurred on a vowel or consonant. TM was better at identifying vowel errors (67% TPR)

than consonant errors (62% TPR). This is expected behavior, as vowels are defined by specific frequencies (formants) that show up well in the MFCC features used by our TM.

At the conclusion of the study, we compared the TM evaluation performance against a baseline algorithm based on the goodness of pronunciation (GOP) measure. We considered this to be a hard baseline since it was computed off-line on a desktop computer, whereas the TM evaluations had executed in real time on the tablet. The GOP algorithm used Kaldi acoustic models trained on the Librispeech corpus (960 hours of adult speech) [163], according to the implementation described by Witt and Young [90]. As GOP is a phoneme-level score, an utterance was labeled correctly pronounced if all phonemes scored above a specified threshold, otherwise it was labeled incorrectly pronounced. The GOP achieved similar performance detecting both incorrect and correct pronunciations, according to TPR and TNR (57% and 59%, respectively). This behavior is more balanced than that of TM, but at the cost of fewer detected mispronunciations. We also calculated the performance for a random binary classifier to show the minimum expected performance, given that our data is skewed with more incorrect than correct productions. Evaluation performance for all methods is displayed in Table 5. TM outperformed all other methods according to F1 score (harmonic mean of precision and recall); caregiver evaluations had the lowest F1 score, which was well below random classification performance. Although TM had a higher F1 score than GOP, both outperformed random classification in all measures.

	Random Classifier	TM	Caregiver	GOP
Precision	82%	80%	90%	87%
Recall	50%	65%	27%	57%
F1 Score	62%	72%	41%	69%

Table 5 Evaluator performance (True positive is an identified mispronunciation).

4.7. Discussion

In this article, we set out to investigate three research questions relating to our speech therapy game and pronunciation evaluation accuracy. Here, we discuss the results in relation to these questions.

- RQ1: Do children remain engaged in the game-based therapy practice over a long period of play?

We found that children did stay engaged in their tablet-based therapy throughout the study. For all children but one, average play time remained the same in both treatment phases, suggesting that they maintained consistent levels of effort across the protocol, rather than dawdling as time went on. Eight participants reported trying “very hard” while playing the game, which aligns with the consistent average playtime across treatment phases. On average, children spent 19.5 minutes playing a level on the days they used the game. Eight participants also responded in the surveys that they would like to continue playing, and nine participants actually played Apraxia World at least once after the treatment concluded. Playing beyond the required time, especially after two months of mandatory play, suggests that the children genuinely enjoyed the type of play offered by

Apraxia World. Additionally, all nine caregivers for whom we have surveys also said that the children were engaged with the game.

Children indicated that they liked the store aspect of the game and made numerous purchases. All children purchased clothing/costume items, which indicates that the children enjoyed being able to customize their game experience; children each purchased an average of 26 items. We found a similar positive response to game and therapy experience personalization in pilot testing for Apraxia World [25]. These purchase behaviors suggest that children are interested in tailoring their gameplay, and it is important to provide different mechanisms for customizing the game and therapy experience.

Even though the children remained engaged in their therapy during the treatment period, some found practicing a limited set of words grew boring. However, the desire for variety must be balanced against the considerable time investment to collect calibration recordings for target words. The per-speaker pronunciation verification approach used in Apraxia World allows SLPs to create highly customized therapy plans that accommodate a child's current speech production abilities, but this comes at the cost of increased setup complexity and decreased target variation. One compromise may be to configure extra target words during the initial calibration session with the clinician so that caregivers can swap out target words when they become tedious.

- RQ2: What level of pronunciation improvement do children achieve while playing Apraxia World?

In our study, participants improved their pronunciation accuracy in both feedback conditions. Children improved an average of 56.6 percent absolute with automated

feedback and 61.5 percent absolute with caregiver feedback. These improvements are similar to those reported for traditional clinician [8, 133] and clinician plus caregiver [164] speech therapy of similar intensity. They also align with results from previous studies demonstrating the efficacy of digital speech therapy applications [19, 165]. Given that Apraxia World delivers therapy through pictorial and text prompts, the game is customizable to deliver stimuli and exercises for a range of conditions (e.g., motor and phonological speech sound disorders, literacy) and across a range of skills levels (e.g., sound, word, phrase level).

While we did not detect significant order effects, the five children receiving caregiver feedback first appeared to have a greater magnitude of change across both phases (67.3 versus 50.8 percent average absolute improvement). If this trend held up in a larger study, it would suggest that children may need some initial support as they start this type of therapy, before they become more independent with TM-guided practice. This transition from high to low support is also more pedagogically valid than increasing support towards the end of treatment. As some children may need less support in the beginning, the duration of caregiver support could be adjusted to fit each child, while still ensuring that game and therapy requirements are established.

- RQ3: How accurately do caregivers and our automated system evaluate pronunciation?

We found that our TM framework was moderately successful at identifying mispronunciations (72% F1), but caregivers let many mispronunciations go unidentified (41% F1). TM outperformed caregivers and GOP (69% F1), aligning with previous results that report TM working well for child speech therapy [63, 153]. TM may also be a better

option than GOP in this application because it does not require forced alignment to score utterances. This is valuable because forced alignment segmentation can be affected by the presence of mispronunciations and inaccurate phoneme times lower pronunciation scoring accuracy. The caregivers evaluated pronunciation with high precision, but low recall, suggesting that they were more lenient than a clinician may have been. It is possible that some of the productions were on the verge of being correct and the caregivers only indicated major mispronunciations. Caregivers may have also used visual cues instead of only auditory cues when determining utterance correctness. In spite of any caregiver lenience or perceived TM severity in the utterance evaluations, children still made meaningful therapy progress.

Although the TM framework outperformed GOP on the labeled recordings set, roughly 54% of in-home recordings had quality issues. Because TM directly compares feature vectors to classify utterances, recording quality can have a large impact on its performance. Audio containing extra words or prematurely stopped recordings may be processed incorrectly by the system. These issues were also reported by Strommen and Frome [166]. They found that children's unpredictable speaking behavior and tendency to pause or repeat words lowered system performance compared to adults. Given that this method is somewhat brittle, extra care must be taken to capture high-quality recordings. If the system fails to provide accurate feedback for a child, the automatic pronunciation evaluations can always be overridden with the external keyboard.

4.7.1. Implications for future work

A potential criticism of this work is the gender imbalance (only having one female participant). In elementary-school-aged populations, males are 2.85 times more likely to have an SSD than females [6], which makes recruiting a balanced population difficult. However, this does not eliminate the need for diverse populations, especially when collecting subjective data such as enjoyment and engagement with new applications. Given that general participant solicitation (this article and references [25, 65]) has failed to provide balanced sex ratios, or even ones that approach the 2.85 to 1 ratio found in the clinical population, perhaps targeted recruitment for female participants is warranted in future work. As caregivers are the ones who need to be convinced to respond to solicitations, we should emphasize the opportunity to provide a voice to girls with SSDs in regards to what type of therapy tools they want to use. Recruiting participants for these types of studies can be challenging, but making efforts to find more female participants will yield more meaningful and generalizable results.

Even though the children wore headsets for the majority of the study, we encountered issues with microphone placement and children adjusting or touching the microphone. Additionally, we observed that when some of our participants became discouraged or excited, they spoke in ways that made it difficult for the TM to meaningfully evaluate their speech (mumbling, yelling, etc.). As such, future systems would benefit from monitoring microphone distortions, speaking volume, and speaking rate to recommend a correction. These reminders should help children produce utterances of better quality for automated speech processing, which would result in them receiving

more meaningful feedback on pronunciations. This may also have the added benefit of helping children increase self-evaluation of loudness and intelligibility.

Future speech therapy games would also benefit from adopting a different recording method than the one implemented in this version of Apraxia World. The touch-to-start/touch-to-stop mechanism proved difficult for the children to accurately control, as evidenced by the high percentage of clipped audio. Many of the clipped utterances were missing just a small portion of the utterance, so a more child-friendly mechanism could yield better recordings, which would again improve ASR performance and provide more audio for offline processing. Ahmed et al. [26] also reported that children had trouble controlling the recording mechanism in their games, but their ASRs performed better when the games used discreet start and stop actions, instead of stopping the recording automatically. As such, a better mechanism may be to start recording once the prompt is displayed and trim the audio around a window defined by the button presses extended with padding to start earlier and stop later than when the child actually pressed the buttons. Since incomplete recordings oftentimes result in inaccurate automated feedback, it is essential to empower children to capture the entirety of their utterance. This replacement recording control mechanism should be the subject of future study.

Although the TM outperformed caregivers for successfully-captured recordings, children sometimes felt the system provided inaccurate feedback. Given that around 54% of recordings had some type of quality issue, it is likely that these incorrectly-processed utterances are part of why the system behaved unexpectedly for some players. In order to build trust in intelligent systems, algorithms such as the TM framework need to offer

appropriate transparency [167, 168]; one way to move towards this goal would be to inform the player if a recording has issues that impede correct processing, rather than providing the same feedback as if a mispronunciation had been detected. Transparency could also be improved by informing the child which specific speech sound was incorrect, which would also provide actionable information for practice. This was not implemented in Apraxia World due to technological constraints and limited child speech corpora, but is the subject of ongoing work.

One benefit of Apraxia World we have yet to examine is the effect of normalizing speech therapy practice by including it in a game format not specific to children receiving therapy. In this way, children could talk about or share their experiences playing the game with their peers, without standing out as different. Children were enthusiastic about playing the game and some seemed very proud of their in-game accomplishments, which we hope they felt free to share with their friends. It could be interesting to explore how reframing speech therapy exercises as a “regular” game changes how they are perceived both by children undergoing therapy and their peers with less exposure to speech therapy.

As evidenced by the large quantity of speech samples collected in our study, digital speech-based applications may be a valuable tool when building child corpora. Although we only presented the audio collected from participants discussed in this manuscript while they completed the protocol, we actually gathered more than 5,000 additional utterances from the game for future mispronunciation detection improvements. Using digital applications to build a custom corpora extends beyond the speech therapy domain; researchers have also deployed engaging applications to gather child speech for offline

analysis of reading fluency [162] and English acquisition in foreign-language speakers [169].

One key takeaway for the human-computer interaction community is that less may be more when dealing with therapy games. We found that children enjoyed the game throughout their treatment and some even played after the study ended so that they could make additional in-game progress. By limiting the daily gameplay, we built anticipation for the next session and extended gameplay to last the entire two-month study duration; if there were no limit, children could have easily completed the game in a couple of days, depending on their skill level. We recommend other designers consider implementing this mechanic to extend therapy game engagement over lengthy treatment periods.

4.8. Conclusion

Children with speech sound disorders struggle to produce and perceive certain sounds, and typically undergo clinical speech therapy to address these difficulties. However, speech therapy is often less frequent than it needs to be for children to learn new skills. Home practice commonly complements clinic sessions to increase practice frequency, but it depends on caregiver availability and can be tedious for children. In this article, we presented Apraxia World, a speech therapy game designed to give children more independence and make therapy practice more enjoyable. Apraxia World is unique from other speech therapy games in that players control the game using traditional joystick and button inputs, while speech input is used to collect in-game assets necessary to complete the level. The game also supports pronunciation feedback provided by caregivers or an automatic evaluation framework.

To validate our game design and speech therapy delivery approach, we evaluated the long-term home use and clinical benefit of Apraxia World over a multi-month period. Children reported enjoying the game, even over the long play period. Game personalization through in-game purchases of costumes, weapons, and avatars proved to be a widely popular aspect of the game. We found that children made clinically-significant therapy gains while playing Apraxia World; this result aligns with previous studies that show computerized and tablet-based speech therapy is as effective as traditional speech therapy [19, 20]. We also found that TM outperformed GOP in detecting mispronunciations and that caregivers were lenient evaluators. The results of this examination support the use of Apraxia World to supplement home-based speech therapy by increasing practice frequency and reducing caregiver burden.

5. EVALUATING AUTOMATIC SPEECH RECOGNITION FOR CHILD SPEECH THERAPY APPLICATIONS*

5.1. Overview

Digital speech therapy games are an increasingly popular method to make speech therapy practice more engaging for children. An especially promising aspect of these applications is the potential to provide automatic pronunciation feedback, which would empower children to complete their practice with limited caregiver supervision. However, due to technological constraints, it is currently more feasible to deploy word recognition in place of phoneme-level mispronunciation detection. This would allow the therapy application to check if a child's utterance was close to the intended target, thereby verifying that they actually tried to say the word. As such, we investigated performance of two automatic speech recognition techniques on disordered speech from children. Specifically, we compared the word recognition accuracy of the open-source PocketSphinx (PS) recognizer using adapted acoustic models and a custom template-matching (TM) recognizer. In our tests, TM and the adapted models significantly outperformed the default PS model. On average, maximum likelihood linear regression and maximum a posteriori model adaptation increased PS accuracy to 63.8% and 80.0%, respectively, suggesting that the adapted models successfully captured speaker-specific

* A portion of this chapter was published at ASSETS 2019. Reprinted with permission. Hair, A., Ballard, K. J., Ahmed, B., & Gutierrez-Osuna, R. (2019, October). Evaluating Automatic Speech Recognition for Child Speech Therapy Applications. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 578-580). <https://doi.org/10.1145/3308561.3354606>

word production variations. TM reached a mean accuracy of 75.8%. These results indicate that limited training data can be used to improve ASR performance to clinically-acceptable levels, as specified by speech-language pathologists.

5.2. Introduction

Speech sound disorders (SSDs) are a group of disorders that affect development of accurate speech sound and prosody production [1]. Although SSDs can impair communication skills development [3], children often improve speech quality and reduce symptoms by working with speech-language pathologists (SLPs) [43]. Given that speech therapy practice must be frequent and high-intensity [7], clinic sessions need to be supplemented with considerable home practice, which can become tedious for children. Primary caregivers typically administer home practice, but busy schedules decrease practice frequency [14]. As such, there is a need for speech therapy tools that decrease the amount of direct caregiver involvement required and make the practice itself more engaging for children.

To address issues stemming from boring and infrequent home practice, we previously developed a mobile speech therapy game called Apraxia World [25] built upon lessons learned from designing the Tabby Talks therapy system [65]. Apraxia World includes speech exercises with pronunciation feedback, which was handled via a human operator through a Wizard of Oz protocol during pilot tests. However, in order for children to practice independently and take ownership of their therapy, speech therapy games such as Apraxia World need to include automated pronunciation verification technology. Although the eventual goal for this technology is to identify phoneme-level errors, research

into improving the accuracy of child pronunciation verification is still ongoing. As such, speech therapy systems may benefit from an interim solution: using automatic speech recognition (ASR) technology to recognize therapy target words. ASR would decrease the time that caregivers spend closely supervising home practice and this increased independence may also improve self-motivation in the children, according to self-determination theory [170]. Accordingly, the therapy application could verify if an utterance is close to the intended target, as done previously by Ahmed et al. [26]. This process ensures that the child is making an appropriate effort to say the word (they cannot say something completely different than the target word), while reserving deeper analysis (i.e., phonological) for trained SLPs. ASR word recognition accuracy has been explored with disordered speech from adults [171, 172] and typically-developing child speech [121], however, it remains unclear what levels of accuracy can be expected from disordered speech from children.

In this paper, we investigate ASR word recognition performance on disordered speech from children using limited child training data and mobile-device-friendly techniques. For this task, we employed a custom child speech corpus to examine two low-resource ASR methods: adapting an existing acoustic model and template matching. Both of these approaches capture speaker-specific pronunciation variants, which is important when recognizing utterances from speakers who struggle to form the canonical pronunciation. Acoustic model adaptation uses sample recordings from a speaker to update the statistical representation of sounds within the model, which creates a speaker-dependent model based on how that person speaks. To test adapted acoustic models, we

consider the PocketSphinx speech recognizer [79], which has previously been used within mobile child speech therapy applications [26, 173]. In contrast to model adaptation, template matching uses the utterances directly to represent how specific words should sound. In our approach, samples of words produced by the speaker are used as templates to determine if a new recording matches the target word. Both template matching and speaker-dependent acoustic models expand gracefully to accommodate new target words; this makes them ideal for use in speech therapy applications, since words must be replaced as a child makes therapy progress so they can practice new sounds.

We found that both template matching and PocketSphinx with an adapted model performed at the desired accuracy level; however, the adapted model significantly outperformed template matching. This suggests that both methods successfully make use of the limited training data to capture speaker-specific word production variants within disordered speech from children. The main contributions of this paper are (1) an empirical test of PocketSphinx performance on disordered speech from children and (2) recommendations for recognizing this speech when training data are limited.

5.3. Related work

To make practice more fun and increase motivation, researchers have examined incorporating speech therapy into game-like digital systems. For example, Ahmed et al. [12] evaluated five speech-controlled arcade-style therapy games with therapists, caregivers, and children. They found that children preferred games with various rewards and challenges. Lan et al. [15] created Flappy Voice, a Flappy Bird clone for prosody (stress and intonation) practice; children enjoyed the game, but reported that voice control

could be difficult. Hoque et al. [16] investigated using a turtle race game to help children with autism speak more slowly. The game successfully helped children control their speaking rate and engage with their practice. Children often enjoy using digital therapy interventions in short-term tests and sometimes even play beyond the required time [6, 16]; however, it remains unclear how these game-like applications hold children’s attention over a longer period.

McKechnie et al. [17] suggested that ASR tools show potential for improving child pronunciation within therapy applications like those described above. However, off-the-shelf ASR tools struggle to recognize speech from children, even typically-developing speech [11]. Researchers have investigated ASR performance on imperfect adult speech, such as speech from dysarthric speakers [9, 18] or deaf and hard-of-hearing speakers [10], and their findings show that ASR methods often generate inaccurate speech transcripts. One way to improve performance is through acoustic model adaptation, which uses data from a specific population or speaker to improve recognition by updating how sounds are represented in the model [19]. Speaker-dependent adaptation methods have been used to improve recognition rates on typically-developing child speech [20]; however, it remains unclear if similar improvements will arise when adapting to disordered speech from children.

Template matching is a well-established speech recognition technique [21]. This method, which is based on dynamic time warping (DTW), compares a test utterance to previously collected examples of target words (“templates”) to see which it most closely matches. This method can yield performance that comes close to that of more standard

Hidden Markov Model speech recognizers when processing adult speech [22]. Furthermore, template matching has been successfully incorporated into child speech therapy while using limited speech data [23, 24]. As such, template matching may be a viable candidate for automatic recognition of disordered speech from children.

5.4. Automatic speech recognition

Detecting mispronunciations in disordered speech from children is a developing research area (e.g., [37, 39, 143, 144]). Although similar research is ongoing in the language-learning domain (e.g., [93, 174, 175]), investigations into techniques for disordered speech from children have been slowed by limited data and difficulty processing child speech due to age-related production variance [35, 104]. While these systems are developed and accuracy is improved, speech therapy applications would benefit from an interim pronunciation verification stand-in that can be deployed using current technology. Therefore, we focus on whole word recognition instead of locating specific speech errors; this is more feasible because analyzing an utterance holistically allows systems to match against correctly-produced portions of the word during recognition. As this approach only verifies that a pronunciation attempt is close to correct, it should be paired with regular visits with SLPs who can provide sound-specific pronunciation feedback as needed. Below we describe the two ASR methods we evaluated.

5.4.1. *PocketSphinx*

PocketSphinx is a mobile-ready version of the Sphinx ASR engine developed at CMU [79], which recognizes both conversational speech or a limited set of words specified in a lexicon. We selected this ASR because it can easily be configured on-the-fly to

recognize different words, thereby allowing SLPs to swap out target words in the field; additionally, at the time of our experiments, PocketSphinx was one of the best mobile-ready ASR libraries available. Performing speech recognition directly on the mobile device is important because Internet access is not guaranteed and we want to preserve child privacy during therapy by avoiding unnecessary data transmission.

The acoustic model provides information about how specific speech elements “sound” to the ASR using a combination of hidden Markov models (HMM) and Gaussian mixture models (GMM) to represent speech. Specifically, the HMM models transitions between the sounds and GMMs are used to model the HMM states for each sound. PocketSphinx comes with an existing speaker-independent acoustic model, which can be converted into a speaker-dependent acoustic model through model adaptation. This process improves speech recognition by providing the acoustic model with samples that demonstrate how a specific speaker produces certain sounds. PocketSphinx acoustic models support two types of adaptation, maximum likelihood linear regression (MLLR) [176] and maximum a posteriori (MAP) [177]. Both methods update the acoustic model parameters based on speech from the target speaker. MLLR estimates linear transformations for the Gaussian means and variances, whereas MAP uses prior information about the parameter distribution combined with the adaptation data to re-estimate all model parameters [178]. Recognition performance can be further improved by using a constrained lexicon that only contains words the child should be practicing in their speech therapy session; this keeps the ASR from searching for irrelevant words when decoding the speech.

5.4.2. Template matching

In contrast to the statistical models used by PocketSphinx, template matching recognizes words by directly comparing a test utterance against previously-collected utterances. Prior to performing template matching, utterances must be transformed into sequences of acoustic feature vectors. Our implementation of this feature-extraction process is illustrated in Figure 16a. Given a recorded utterance (16 kHz), we trim leading and trailing silence using an energy threshold. The signal is then pre-emphasized using the filter $1 - \alpha z^{-1}$, where $\alpha = 0.97$. Spectral information is extracted as 13 Mel-frequency cepstral coefficients (MFCCs) over 32 ms (512-sample) frames with an 8 ms (128-sample) shift. We discard the first coefficient (spectral energy) to focus only on $MFCC_{1-12}$ for word recognition. Lastly, the MFCCs are normalized by applying cepstral mean normalization (CMN) [160].

The template-matching process itself is illustrated in Figure 16b. Following feature extraction, template and test utterances (t and u , respectively) are aligned end-to-end using DTW. The distance (root-mean-squared-error) between the template and test utterance is computed as:

$$d(t, u) = \begin{cases} \frac{\|dtw(u, t) - t\|_2}{len(t)} & \text{if } len(t) > len(u) \\ \frac{\|dtw(t, u) - u\|_2}{len(u)} & \text{otherwise} \end{cases} \quad (1)$$

where $dtw(t, u)$ time-aligns the frames in t to u .

The framework classifies a test utterance by comparing it against templates for all possible word classes. Let T_w be the set of templates for word w . The test utterance score

is the mean template matching score from comparing a new utterance u against all templates in T_w (Eq. 2). The test utterance is assigned to the class w with the lowest score (Eq. 3).

$$score(u, T) = mean(d(j, u) | \forall j \in T) \quad (2)$$

$$label(u) = argmin_w score(u, T_w) \quad (3)$$

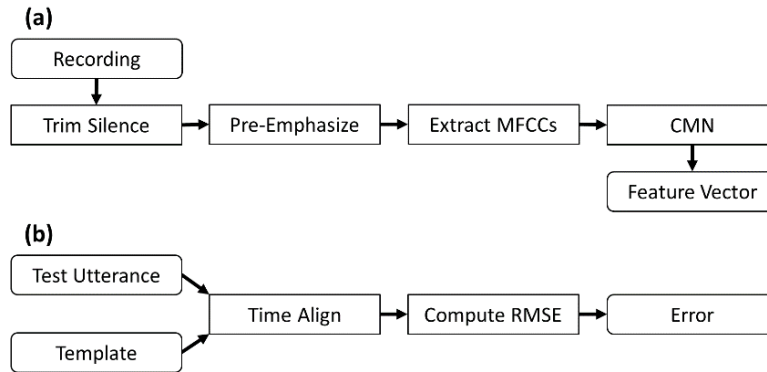


Figure 16 (a) Spectral information is extracted from a trimmed utterance and then normalized (b) Template and test utterances are aligned and scored based on RMSE

5.5. Experiments

In this section, we describe how we evaluated the two ASR systems. For this task, we used the Apraxia World speech therapy game [25] to gather disordered speech from children in their homes. Although some child speech datasets exist (e.g., the OGI kids' speech corpus [145]), they usually contain speech from typically-developing children, with

few mispronunciations. Therefore, we gathered a real-world corpus to better evaluate speech recognition performance on disordered speech. All speech recognition tests were conducted offline after collecting the data; children did not receive utterance feedback from the ASRs.

5.5.1. Data collection

Disordered speech data were collected from seven Australian children (1f, 6m, 7-9 y.o.) with speech sound disorders while they played Apraxia World for eight weeks under caregiver supervision. This provided a large set of scripted single-word recordings. Audio was recorded at 16 kHz using a headset attached to a Samsung Tab A 10.1 tablet. The children started and stopped the recordings on their own, so some recordings were stopped prematurely or “clipped.” At the time of conducting these experiments, 21,198 utterances had been recorded. Recordings that were clipped or that contained substantial background noise were discarded, leaving 10,415 recordings.

Children went through two phases of playing Apraxia World for about 30 minutes, four days per week. Phases lasted for four weeks with a two-week break between them. Each phase repeatedly drilled a different set of 10 words (for 100 words per session) selected by SLPs working with the children to target specific speech difficulties (e.g., words to practice ‘cl,’ ‘fl,’ and ‘gl,’ ‘a’ like in ‘bat’, or leading ‘l’ sounds). During a phase, Apraxia World prompted the child to say one of the 10 words selected at random and no words were repeated until all 10 words had been presented. In total, each child practiced 20 different words across the two treatment phases.

5.5.2. *Experiment setup*

To examine speech recognition performance using limited training data, we compare PocketSphinx (with a limited lexicon) against template matching (with an increasing number of templates per word). We tested PocketSphinx with both the default and speaker-dependent (adapted) acoustic models. Performance for both speech recognizers was measured as word-level accuracy.

To evaluate the template-matching framework, we developed a prototype system using the librosa audio processing library for Python [179]. For convenience, we ran tests on a desktop computer, but this framework can also be used on mobile devices and lends itself to parallelization. We randomly selected n child-specific templates per target word where $n \in [1,15]$ and used the remaining recordings of the child saying the target as test data. This process was repeated 5 times, each with new templates selected at random. Since the children only practiced 10 words at a time, each recording can only be labeled as one of the words practiced in that phase (using 10 sets of n templates).

For tests with PocketSphinx (PS), we started with the default American English acoustic model trained on adult speech⁵. To account for dialect and age differences, we created two speaker-dependent acoustic models by adapting the default model with MLLR and MAP separately. For both adaptation methods, the acoustic model was adapted using 15 samples for each of the 20 practiced words (300 utterances total per speaker), which is the maximum amount of data used in the template-matching approach. The PS decoder

⁵ <https://github.com/cmuspinyin/pocketsphinx/tree/master/model/>

was configured with a 10-word lexicon to only recognize words practiced in the respective treatment phase. The remaining 8,315 utterances were used as test data, which were passed to the PS decoder without any additional preprocessing.

5.6. Results

Figure 17 shows the per-speaker, per-word classification accuracy for template matching over the five repetitions. Even with only one template per word, template matching performs well above chance level (10%). Unsurprisingly, increasing the number of templates per word improves word recognition; however, there is no significant increase in accuracy when using nine or more templates (t-test, $p > 0.05$), where mean accuracy is between 71.3% and 75.8%. This illustrates the diminishing returns for using additional templates beyond a certain quantity; even though overall accuracy generally improves as more templates are used, the increases may not be significant. The performance plateau suggests that there is a fixed amount of information about pronunciation variation that the templates can represent with the method as implemented.

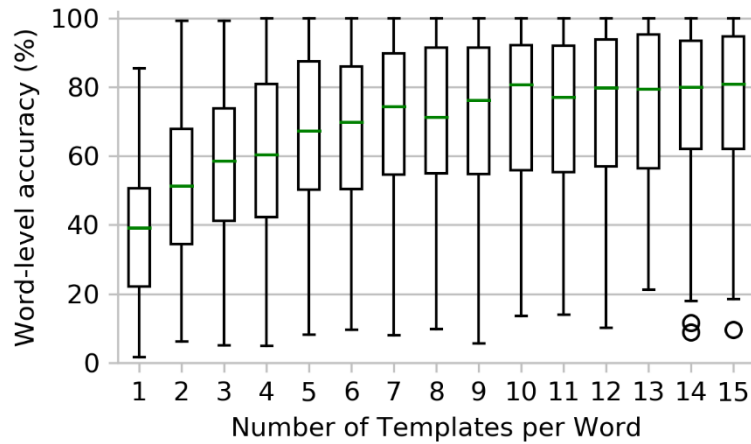


Figure 17 Per-speaker, per-word word recognition accuracy for template matching with an increasing number of templates

To compare template matching and PS performance, we tested word recognition using the same amount of speaker-dependent training/adaptation data. Figure 18 shows the per-speaker, per-word accuracy for all speakers when using 15 utterances per word, both for adapting the PS acoustic model and for template matching. The MAP-adapted models yield the best recognition performance. Both the MAP-adapted models and template matching correctly recognize all words at least some of the time; this is in contrast to PS with the non-adapted and MLLR-adapted models, which fail to recognize some of the words. Regardless, MLLR- and MAP-adapted models both performed significantly better than the non-adapted model (paired t-test, $p \ll 0.01$). Template matching performed significantly better than the non-adapted model (paired t-test, $p \ll 0.01$) and the MLLR-adapted model (paired t-test, $p \ll 0.01$). The MAP-adapted model performed significantly better than the MLLR-adapted model ((paired t-test, $p \ll 0.01$) and

template matching (paired t-test, $p = 0.03$). We used an alpha level of 0.05 for tests of significance.

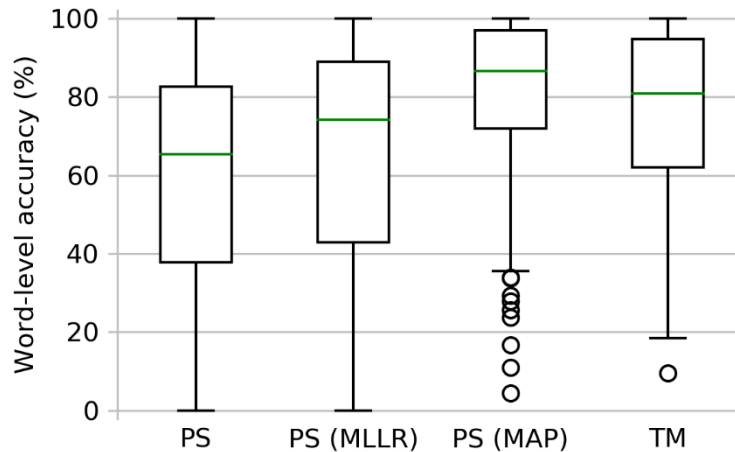


Figure 18 Per-speaker, per-word accuracy (Using 15 utterances per word for adaptation and template matching)

Table 6 displays the per-speaker word recognition accuracy for PS and template matching when using 15 utterances per word for adaptation and template matching. For all speakers, both MLLR and MAP adaptation significantly outperform the default PS baseline, which suggests that the adaptation process successfully captured how the children produce their utterances. The MAP-adapted acoustic models outperform the MLLR-adapted models for all speakers, indicating that MAP adaptation made better use of the limited training data. Due to speech quality differences resulting from age and SSD severity, word recognition accuracy varied across speakers in our corpus. As such,

recognition rate improvements should be considered within-speakers. On average, MLLR and MAP adaptation increased PS accuracy by 4.2% and 20.4%, respectively.

Speaker	PS	PS (MLLR)	PS (MAP)	TM
f1	53.8	59.1	78.2	79.3
m1	52.3	56.7	79.1	79.2
m2	29.1	31.5	47.9	35.4
m3	84.4	88.2	95.7	90.4
m4	59.1	65.7	83.0	67.3
m5	74.0	76.2	92.8	86.4
m6	54.0	57.6	72.6	84.3
avg (std)	59.6 (29.0)	63.8 (28.7)	80.0 (21.6)	75.8 (22.2)

Table 6 Average word recognition accuracy (%) using 15 utterances per word for adaptation and template matching

5.7. Discussion and conclusion

Speech therapy games, such as Apraxia World, would benefit from automatic pronunciation verification so that they can afford independent practice. While research into phoneme-level mispronunciation detection for disordered speech from children continues, therapy applications may benefit from using automatic word recognition as a temporary stand-in. As such, in this paper we examined speech recognition performance

on disordered speech from children when using limited training data. Specifically, we compared template matching word recognition and the open-source PocketSphinx recognizer with two types of speaker adaptation.

In our tests, we found that both template matching and PocketSphinx were able to recognize the child speech in our data set well above chance level and, most importantly, at the accuracy levels recommended by speech-language pathologists. Template matching performance increased as more word templates are used, but accuracy improvements level off at nine or more templates per word. Template matching was able to out-perform PocketSphinx using the default acoustic model and MLLR-adapted models. However, the PocketSphinx MAP-adapted models achieve the overall best accuracy. As PocketSphinx is easier to configure than a template matching pipeline due to documentation and online support, we recommend that accessibility developers use the off-the-shelf PocketSphinx ASR and adapt the acoustic model for increased performance. Regardless of the ASR method used, whole-word recognition is not a substitute for SLP-led speech therapy and should only be used to augment regular clinic visits where the child receives specific feedback.

Using the MAP-adapted model allows PocketSphinx to reach the suggested 80% accuracy threshold given that it updated all relevant parameters in the acoustic model. This contrasts with the MLLR-adapted models, which improve performance, but fall short of the suggested threshold because PocketSphinx treats this method as a transform applied to the original model, instead of re-estimating the parameters. It is likely that MAP-adaptation outperformed template matching because template matching is limited to only

recognize word productions for which it has a similar template. In contrast, PocketSphinx benefits from having an acoustic model that captures general production information, which helps it recognize words correctly, even if an exact match was not included in the training data. Typical variations in child speech production (e.g., high/low energy, drawing certain sounds out) are likely to lower template matching accuracy; model-based speech recognition is more robust to these variants because utterances are evaluated based on statistical features, not direct measurement against another utterance.

Based on these results, adapting acoustic models is a viable method for automatic speech recognition with limited disordered speech data. Accent and age-related speaking differences were reduced by the MAP adaptation, which significantly improved performance over the default American English acoustic model. Additional improvements may be gained by training a speaker-independent child model and adapting that to each child, but we leave that analysis for future study. Although this paper focuses only on word recognition, further work should investigate the relationship between child pronunciation quality and speech recognition accuracy with a pathologist-annotated corpus.

6. EXPLORING CLASSIFIER-BASED MISPRONUNCIATION DETECTION FOR CHILD SPEECH THERAPY

6.1. Overview

A critical component of child speech therapy is home practice, where caregivers typically lead sessions and provide feedback. However, caregivers and untrained adults have been found to struggle with accurately rating speech and generally perceiving pronunciation errors. One potential solution to inconsistent and inaccurate feedback is to use automatic mispronunciation-detection algorithms within digital speech therapy applications. To address the need for automated pronunciation evaluation within child speech therapy, we investigated classifier-based mispronunciation detection using a custom corpus of disordered speech from children with expert clinician annotations. We trained a series of phoneme-specific logistic regression classifiers (LRC) and support vector machines (SVM) on log posterior probability and log posterior ratio features. Our results show that these classifiers outperformed baseline Goodness of Pronunciation scoring by 11.1% and 10.4%, respectively. Even more importantly, in an offline test, the LRCs and SVMs outperformed student clinicians at identifying mispronunciations by 18.1% and 16.1%, respectively. These results suggest that classifier-based mispronunciation detection may be suitable to provide at-home therapy feedback for children.

6.2. Introduction

Children with speech disorders benefit from frequent and high-intensity speech therapy [7] to provide opportunities to practice new skills [8]. Clinician-led therapy

sessions are often scheduled infrequently. As such, caregiver-guided home practice is commonly employed to increase treatment dosage [56]. Home practice relies on the caregiver to lead activities and provide pronunciation feedback. However, clinicians have encountered issues with home practice delivered by caregivers, primarily, low completion rates and incorrect implementation [56]. These problems can be attributed to difficulties making time to complete the therapy practice [180, 181] and an absence of caregiver training; many caregivers feel they lack knowledge or experience to support their child themselves [182] and others report that they sometimes feel unsure how to provide proper feedback [180]. Caregivers have also been found to rate pronunciations leniently during home therapy practice [40, 41], and untrained adults may generally have difficulty perceiving errors in child speech [183]. While caregivers can be trained to deliver effective phonological interventions [184], the training takes time (on the order of a couple of months [184]) and ignores scheduling-related barriers to home practice.

A potential solution to limited caregiver availability and inconsistent pronunciation feedback is to incorporate automatic mispronunciation-detection algorithms into digital speech therapy applications, thus empowering children to practice more independently. This would allow caregivers to lightly supervise therapy practice, instead of directly administering the activities. Automatic pronunciation evaluation systems will invariably be less accurate than trained clinicians, but they may rate productions more accurately and consistently than caregivers. For example, in previous work [40] we found that automatic mispronunciation detection overwhelmingly outperformed caregivers at word-level mispronunciation detection in-the-field. Although some digital child therapy projects have

provided word-level feedback [26, 61, 68], systems like these eventually need phoneme-level feedback so that speech therapy practice can target specific problematic sounds [185]. This is a substantially more challenging task because the system needs to model individual errors, rather than matching whole utterances to a certain word label. Furthermore, even though phoneme-level mispronunciation detection (MPD) is an active research area for second-language (L2) learners (e.g., [93, 99, 186, 187]), less attention has been paid to detecting mispronunciations in disordered speech from children.

In this article, we investigate whether existing techniques from the L2 literature could be used for child speech therapy mispronunciation detection with a limited corpus of disordered speech from children collected during speech therapy practice. Specifically, we train phoneme-specific classifiers to identify mispronunciations using posterior-probability-based features. These features are a concatenation of log posterior probabilities and log posterior ratios, as proposed by Hu et al. [174]. These features are derived from an off-the-shelf acoustic model in a manner similar to the traditional Goodness of Pronunciation (GOP) score. However, these features have been shown to outperform standalone GOP when applied to mispronunciation detection for adult L2 learners [91, 174]. The ability to extract features with a generic speaker-independent acoustic model is especially important in the context of child speech therapy, as there is a general lack of corpora containing disordered speech from children for system building. Following feature generation, we trained phoneme-specific classifiers for mispronunciation detection.

Results from this study show that phoneme-specific classifiers predicted mispronunciations significantly better than a baseline GOP system, even though both

systems use features based on the same acoustic model outputs. More importantly, the classifier-based mispronunciation detection significantly outperformed student clinicians in an offline pronunciation labeling test, suggesting that our automated approach may better mimic expert clinician evaluations. As such, this type of mispronunciation detection may be useful within child speech therapy applications to improve the quality of pronunciation feedback received and alleviate caregiver scheduling burden.

6.3. Background

Current research efforts within mispronunciation detection can generally be grouped into three categories: posterior-based, classifier-based, or rule-based. Posterior-based mispronunciation detection methods score phoneme segment pronunciation quality according to the posterior likelihood output of the production matching the target phoneme. These continuous-valued scores are often converted into binary pronunciation classifications by comparing against a set threshold [188, 189], which yields the same output as classifier-based methods. Posterior probabilities are often derived from the output of an automatic speech recognizer acoustic model and frequently take the form of a Goodness of Pronunciation (GOP) metric [90]. These methods are commonly used as mispronunciation detection baselines [94, 190], but have also served as the foundation for novel methods [93, 191]. For example, the GOP has been used as a standalone method to process L2 speech [192] and disordered speech [92] from adults. For child speakers, Dudy et al. [38] combined the GOP with rule-based error modeling and explicit acoustic modeling of the phonetic errors. Saz et al. [143] also deployed posterior-based

mispronunciation detection for child speakers and increased likelihood score separation by using speaker normalization and acoustic model adaptation.

Classifier-based approaches treat mispronunciation detection as a binary classification problem, where a phoneme can either be correct or incorrect (mispronunciation) [193]. Individual phoneme segments are converted into feature vectors, which are passed through a classifier to obtain a pronunciation prediction [91, 94]. Features vectors may consist of Mel-Frequency Cepstral Coefficients [194], speech attribute scores [195, 196], or even posterior probabilities [91, 174]. Researchers have explored a variety of classification methods, including decision trees [185, 197], support vector machines [95, 198, 199], and more recently, various neural network architectures [187, 200, 201]. These methods have also been used with child speakers. For example, Shahin and colleagues [36, 144] explored a classifier-based approach using a one-class SVM trained on phonetic attribute features to detect anomalous phoneme pronunciations. Wang et al. [196] also tested classifier-based mispronunciation detection for child speech, wherein they trained binary pronunciation classifiers on the distance from the expected phoneme, as measured by a Siamese network.

Rule-based methods take existing knowledge of mispronunciation patterns to identify errors, usually by including these errors to the ASR decoder lattice [97, 100, 202]. Obtaining the necessary error patterns requires expert manual curation [97, 100] or using large quantities of speech to identify the patterns in a data-driven fashion [99, 203]. Shahin et al. [37] deployed rule-based mispronunciation detection for child speech by including expected errors as provided by an SLP to the decoding path. They later made the system

more generic by including an alternative garbage node at each phoneme along the decoding path [66].

6.4. Methods

The proposed mispronunciation detection process involves three components: a speaker-independent acoustic model to generate posterior probabilities, phonetic segment feature generation, and a set of speaker-independent, phoneme-specific mispronunciation detection classifiers.

6.4.1. *Acoustic modeling and posterior probabilities*

We use a deep neural network (DNN) acoustic model to generate the posterior probabilities for each speech frame [204]. The acoustic model is trained on the Librispeech corpus [163], which contains 960 hours of adult English speech, mostly American English. This corpus is not used for any other training or testing. Specifically, we use the Kaldi Librispeech recipe⁶ to train a DNN that contains five fully-connected hidden layers (5,000 neurons) using the p -norm non-linearity ($p = 2$). After the final hidden layer, there is a 14,000-node softmax layer that is group-summed to produce the final output across 5,816 senones. We extract 13-dimension Mel-Frequency Cepstral Coefficients (MFCCs) with 7-frame context, which are transformed with LDA to create a 40-dimension feature vector, and these vectors are concatenated into nine-frame inputs (40×9) for the DNN; final input features are decorrelated using a fixed linear transform. The DNN acoustic model

⁶ <https://github.com/kaldi-asr/kaldi/tree/master/egs/librispeech>

output represents the senone posterior probabilities conditioned on an input observation, i.e., $P(s|\mathbf{o})$ [75].

6.4.2. Feature generation and classification

Our feature generation process follows Hu et al. [174], wherein each phoneme segment is represented by a single feature vector containing two types of features: Log Posterior Probabilities (LPP) and Log Posterior Ratios (LPR). The LPP is a log posterior normalized over the phoneme duration:

$$LPP(p|\mathbf{o}) = \log P(p|\mathbf{o}; t_s, t_e) \approx \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} \log P(p|\mathbf{o}_t) \quad (1)$$

where the posterior for phoneme p is obtained according to:

$$P(p|\mathbf{o}) = \sum_{s \in p} P(s|\mathbf{o}) \quad (2)$$

for each senone s associated with phoneme p , i.e., a senone shared by a tied-state triphone where the center phoneme is p [205]. The posterior $P(s|\mathbf{o})$ comes directly from the DNN acoustic model. The LPR is the difference of the LPPs for phonemes p_i and p_j , given the same observation \mathbf{o} :

$$LPR(p_j|p_i, \mathbf{o}) = LPP(p_j|\mathbf{o}) - LPP(p_i|\mathbf{o}). \quad (3)$$

For each phoneme segment, we compute a series of LPPs and LPRs to form a feature vector. LPPs are calculated for all N phoneme classes and LPRs are calculated for all pairs p_i, p_j where p_i is the expected phoneme class and $j \in N$. The final feature vector $f(p_i|\mathbf{o}; t_s, t_e)$ is the concatenation of LPPs and LPRs:

$$f(p_i|\mathbf{o}; t_s, t_e) = \begin{bmatrix} LPP(p_1|\mathbf{o}), LPP(p_2|\mathbf{o}), \dots, LPP(p_N|\mathbf{o}), \\ LPR(p_1|p_i, \mathbf{o}), LPR(p_2|p_i, \mathbf{o}), \dots, LPR(p_N|p_i, \mathbf{o}) \end{bmatrix}^T. \quad (4)$$

Recordings are force-aligned against the canonical pronunciation using a pre-trained aligner [206] to automatically generate the phoneme segments. Silence segments are discarded, leaving only speech segments for our analysis. Features are extracted by passing individual segments to the acoustic model to generate the posterior probabilities, which are transformed into the final feature vector according to the above equations; this process is shown in Figure 19. These features are used to train supervised phoneme-specific classifiers with examples of correct and incorrect phoneme pronunciations. For classification, we used support vector machines (SVM) and logistic regression classifiers (LRC); SVMs are commonly deployed for mispronunciation detection (e.g., [95, 96, 194]) and neural LRCs have also been used successfully for this task [91, 174]. However, given our data constraints, we use a traditional LRC instead of a neural-network-based classifier.

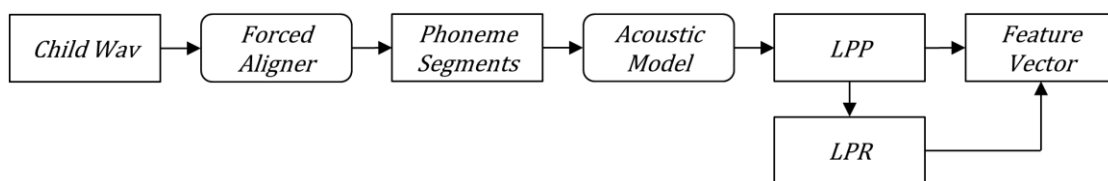


Figure 19 Phoneme-level feature vector feature extraction pipeline

The LRC and SVM were implemented using the Scikit-learn Python library [207]. The LRC used an L2 penalty and was allowed iterate until convergence during training for

each phoneme. The SVM used a fourth-degree polynomial kernel. These hyperparameters were determined empirically. Forced alignment was performed with the Montreal Forced Aligner [206]. We used a 40-phoneme set, so each feature vector contained 80 features: 40 LPPs and 40 LPRs.

6.4.3. Goodness of pronunciation baseline

In this work, the GOP serves as a baseline for automatic mispronunciation detection. Originally, the GOP was defined as the normalized log posterior of phoneme p , which was computed as the ratio between the likelihood of the expected phoneme and the most probable phoneme [90]. Given the assumption that priors $P(q_i) \approx P(q_j)$ for any phonemes q_i, q_j , and that the sum in the denominator can be approximated by its maximum, the GOP is canonically defined as:

$$GOP(p|\mathbf{o}) = \frac{1}{t_e - t_i + 1} \log \frac{P(\mathbf{o}|p)P(p)}{\sum_{q \in Q} P(\mathbf{o}|q)P(q)} \approx \log \frac{P(\mathbf{o}|p; t_s, t_e)}{\max_{q \in Q} P(\mathbf{o}|q; t_s, t_e)} \quad (5)$$

for segment observation \mathbf{o} , canonical phoneme p , start and stop frame indices t_s and t_e , and phoneme set Q . Each probability $P(\mathbf{o}|p; t_s, t_e)$ is computed as:

$$P(\mathbf{o}|p; t_s, t_e) = \frac{1}{t_e - t_i + 1} \sum_{t=t_s}^{t_e} \sum_{s \in p} P(\mathbf{o}_t|s) \quad (6)$$

where s is a senone associated with the phoneme p and the likelihood $P(\mathbf{o}_t|s)$ is traditionally obtained from a GMM-HMM acoustic model. However, given that we use a DNN acoustic model that directly outputs senone posteriors, the original GOP equation needs to be modified slightly. Therefore, we use the GOP computation proposed by Hu et

al. [91], where the score is the ratio between the LPPs for the expected phoneme and the highest posterior across all phonemes:

$$GOP(p|\mathbf{o}) = LPP(p|\mathbf{o}) - \max_{q \in Q} LPP(q|\mathbf{o}) \quad (7)$$

Following calculation, GOP scores are converted into a binary evaluation by comparing the score against a threshold (see section 6.6); if the score is greater than the threshold, the phoneme segment is labeled as correctly pronounced, otherwise, the segment is labeled as incorrectly pronounced.

6.5. Experiments

For our mispronunciation detection tests, we use a custom corpus of disordered speech from children. This corpus is an expert-annotated subset of larger collection of speech therapy audio recordings, which were gathered as part of a longitudinal evaluation of a tablet-based speech therapy game [41]. This corpus contains 2,336 recordings of prompted single or compound word utterances from nine children with speech sound disorders (native Australian-English speakers), each practicing 20 words. These recordings were captured at 16kHz in the children’s homes and contain some distortions and excited speech. Children generally spoke at a normal volume. The corpus contains 10,059 non-silence phonemes, 27.0% of which are mispronounced. The phoneme ZH is not represented in this corpus and W only has correct samples; all other phonemes have samples with mispronunciations. Table 7 shows the 15 most common phonemes in the corpus. Each utterance was annotated for phoneme-level errors by a speech-language pathologist at the University of Sydney. All annotations were collected offline and are

binary labels of correctness for each phoneme; they do not provide the actual sound produced in the case of a substitution error.

Phoneme	Frequency	Phoneme	Frequency	Phoneme	Frequency
L	8.7%	IH	4.8%	M	3.5%
AH	7.6%	P	4.8%	N	3.4%
ER	6.8%	AE	4.1%	SH	3.4%
T	6.1%	S	3.8%	EH	3.2%
K	5.6%	IY	3.7%	B	3.0%

Table 7 Top 15 phonemes in the corpus as percent of total non-silence phonemes

In this article, we define a true positive (TP) as a pronunciation error that was correctly labeled as a pronunciation error, and a true negative (TN) as a correct pronunciation labeled as correct. Within speech therapy, providing accurate feedback on both correct and incorrect pronunciations is critical for children to make progress. As such, we report wholistic system performance according to a combined F1 score (eq. (8)), which averages the F1 scores calculated for correct and incorrect pronunciation detection. Additionally, since the proposed mispronunciation detection systems cannot handle insertion errors, we focus only on substitution and deletion errors.

$$F1_{combined} = \frac{TN}{2TN + FN + FP} + \frac{TP}{2TP + FN + FP} \quad (8)$$

6.6. Results

To ensure that there were enough samples to train and test the classifiers, we only examined phonemes which had at least 60 samples of correct and incorrect pronunciations in the child corpus ($n = 8$ phonemes). For each phoneme, we trained two phoneme-specific classifiers: an LRC and an SVM. To accommodate our small corpus, we used 5-fold stratified cross-validation (each fold contains the same class distribution) when evaluating classifier performance. For each fold, both classifiers were trained, labels were predicted for the test data, and the predictions from each classifier were scored against the expert labels. As a baseline, we also computed the performance of GOP scoring at each fold. Phoneme-specific GOP thresholds were found by exhaustively checking between the minimum and maximum scores in the training samples for the threshold that maximized the combined F1 score. This threshold was then used to convert the test segment scores into labels, which were compared against the expert labels. Because each phoneme has a different correct/incorrect class distribution, we also calculated the performance of a random binary classifier as a measure of chance level. The average combined F1 scores for all phonemes are shown in Table 8. All three methods performed above chance level ($p < 0.05$, paired t-test). Both the LRC and SVM achieved significantly higher combined F1 scores than the GOP baseline ($p < 0.05$, paired t-test). The LRC and SVM demonstrated 11.1% and 10.4% relative increases, respectively, compared to GOP. Although the SVM outperformed the LRC for six phonemes, on average, there was no significant difference between the LRC and SVM ($p > 0.05$, paired t-test). The SVM and GOP each failed to

classify one phoneme correctly; the SVM had problems with CH and the GOP struggled with ER.

Given the varied number of samples for each phoneme, we also looked at correlation between classification performance and sample size. For the LRC, performance and sample size were not correlated ($r = -0.21$). However, for the SVM and GOP, these variables were moderately correlated ($r = -0.48$ and $r = -0.53$, respectively). For the SVM, this appears to be explained by the poor performance on the phoneme L, which all methods struggled to classify; when L is excluded from the correlation calculations, sample size and performance is no longer correlated for the SVM ($r = -0.30$).

	LRC	SVM	GOP	Chance
AH	57.2 (2.0)	58.0 (2.3)	54.6 (1.4)	48.5 (0.0)
CH	55.7 (2.4)	43.6 (3.1)	58.4 (1.1)	47.6 (0.0)
EH	77.7 (1.1)	78.3 (1.6)	69.1 (1.5)	49.7 (0.0)
ER	58.8 (1.9)	62.9 (2.0)	44.6 (1.6)	49.3 (0.0)
IY	61.0 (4.4)	62.7 (2.7)	50.9 (1.2)	46.7 (0.0)
L	50.7 (1.5)	52.0 (2.3)	50.4 (1.6)	42.1 (0.0)
S	75.0 (2.0)	77.4 (2.0)	52.4 (3.6)	49.9 (0.0)
SH	59.8 (3.4)	57.6 (2.0)	66.0 (1.4)	49.9 (0.0)
All	62.0 (1.6)	61.6 (1.9)	55.8 (1.4)	48.0 (0.4)

Table 8 Average combined F1 score from 5-fold cross validation (std. err.)

6.6.1. Comparison against human raters

To put our mispronunciation detection results in context for speech therapy, we compared our performance against that of an independent set of human evaluators. For this purpose, we asked 32 student clinicians to annotate a subset of 154 recordings in our corpus; due to the annotation process, each evaluator labeled a slightly different quantity of the 154 recordings. As the final step in our analysis, we compared these student clinician labels against classifier predictions, only considering the eight phonemes analyzed above. Phoneme-specific SVMs and LRCs were trained using phoneme samples from all recordings in the corpus not annotated by the student clinicians. We treated evaluator annotations as another set of predictions, which were scored against the expert annotations, which were treated as ground truth. For each evaluator, we calculated their performance and the chance level for the phoneme set they annotated. Additionally, each of the 32 sets of student-annotated phonemes were labeled by the LRCs and SVMs; these predictions were also compared against the expert annotations.

Average F1 performance on the 154-recording subset for student evaluators and classifiers is displayed in Table 9. The student clinicians labeled the phoneme segments well above chance level ($p \ll 0.05$, paired t-test), however, both automated approaches significantly outperformed the students ($p \ll 0.05$, paired t-test). The LRC and SVM obtained combined F1 scores 18.1% and 16.1% higher, respectively, relative to the student clinicians. On this subset, the LRC achieved a significantly higher combined F1 score compared to the SVM ($p < 0.05$, paired t-test).

	Student Clinician	LRC	SVM	Chance
F1 Combined	69.0 (1.6)	81.5 (0.4)	80.1 (0.4)	48.9 (0.1)

Table 9 Average combined F1 score for each set of student clinician annotations (std. err.)

6.7. Discussion and conclusion

Our results show that phoneme-specific classifiers trained using posterior-probability-based features identify mispronunciations in field-collected disordered speech from children significantly better than a baseline GOP system. This follows results presented by Hu et al. [174], even though they used a neural-network-based classifier and we used traditional classifiers. We found no significant difference between LRC and SVM mispronunciation detection on the entire corpus. Notably, both types of phoneme-specific classifiers significantly outperformed student clinicians at identifying mispronunciations in a subset of our corpus. This suggests that these automated methods may approximate expert clinician evaluations better than students with some training. These results further strengthen the argument that child speech therapy systems should include automated mispronunciation detection to improve the quality of feedback received at home.

In this investigation, although classifiers were trained with phoneme-specific data, we set global classifier hyperparameters (e.g., SVM kernel, LRC penalty). However, future work may benefit from setting hyperparameters on a per phoneme basis. Speech production is a complex process, with many variables contributing to the final sound

(place, manner, voicing, etc.). Accordingly, phoneme-specific hyperparameters may help classifiers better identify pronunciation errors.

Our goal with this type of system is not to replace clinicians or clinic visits, but to better approximate clinician evaluations at home. This is especially important given the difficulty some adults have identifying errors in child speech [183] and some caregivers have been shown to evaluate word-level pronunciation below chance level [41]. Additionally, even though caregivers are motivated to help their child, some are reluctant to take the lead and want clinicians to do the decision making during therapy practice [208]; an automated system that imitates clinician ratings helps to fill this desire. Although there is still significant work to be done in the speech therapy mispronunciation domain, the results presented in this article suggest that phoneme-specific classifiers perform well over chance level and can even outperform student clinicians when comparing against expert evaluations. As such, child speech therapy application designers could use these methods to provide automated feedback in their systems. Significantly, this can reduce caregiver scheduling burdens by allowing them to lightly supervise instead of directly managing home practice, thereby increasing the quantity of speech therapy children receive.

7. CONCLUSIONS FROM THIS DISSERTATION

This dissertation represents an effort to improve the home speech therapy practice experience for children with a novel digital speech therapy game called Apraxia World; the game increases engagement through extended gameplay and affords independent practice with automated pronunciation verification technology. The first two manuscripts (Chapters 3 and 4) present the development and evaluation of Apraxia World over a pilot and longitudinal study. The last two manuscripts (Chapters 5 and 6) investigate speech processing on disordered speech from children. This chapter summarizes the findings from the four manuscripts, discusses limitations, offers ideas for future work, and ends with a final conclusion.

7.1. Summary

Chapter 3 presented the initial prototype of Apraxia World and examined exercise integration and delivery. The game offers speech exercises delivered during the level as the player collects in-game assets, or at the end of the level when the player crosses the finish line. Pronunciation evaluation was handled in a Wizard-of-Oz manner, where the administering clinician provided binary utterance ratings via a Bluetooth keyboard paired to the tablet running the game. This pilot study examined if children enjoyed the game, if the speech exercises detracted from gameplay, and when children wanted the exercises delivered. Children were enthusiastic about playing Apraxia World, enjoying both the gameplay (e.g., exploring, fighting) and personalization (character costumes). Questionnaire responses suggest that neither exercise delivery method (during or after)

dramatically altered the game difficulty and that the game made speech exercises more fun than normal (paper-based). Exercise delivery timing preferences were mixed; 13 out of 21 children preferred exercises after the level, and the remaining 8 preferred exercises during the level. However, neither exercise delivery method encouraged children to complete more than a few extra speech exercises beyond the required minimum, suggesting that children are unlikely to do more speech therapy than required.

Chapter 4 described the full version of Apraxia World and a corresponding longitudinal study. This version of the game improved upon the prototype used in pilot testing by rewarding all speech exercises and including automatic pronunciation evaluation based on template matching. Although the study in Chapter 2 revealed that children preferred to complete their exercises at the end of the level, after consultation with clinicians, it was decided that the exercises should be delivered during the level. This allowed for a tighter integration of gameplay and the rewards from completing exercises (especially once the “energy” timer was added to the game). Delivering exercises during the level also avoided a game-first-exercises-later paradigm, which decouples the speech exercises from rewards and negates the benefits of a having custom game. The study explored the long-term use of Apraxia World, speech therapy benefits arising from gameplay, and both caregiver and automated framework pronunciation evaluation accuracy. Even over the long period, children remained engaged in the game-based therapy. Children reported that they would like to continue playing (eight out of nine returned questionnaires) and nine children actually played Apraxia World at least once after the study concluded, which suggests that they genuinely enjoy the game. Caregivers

also confirmed in their questionnaires that children were engaged with the game. Over the game-based treatment period, children achieved pronunciation accuracy improvements on-par with those reported for traditional clinician and clinician-plus-caregiver speech therapy of similar intensity. Finally, results suggested that caregivers were lenient evaluators, while the template-matching framework was moderately successful at identifying mispronunciations. The template-matching framework also out-performed Goodness of Pronunciation scoring for word-level mispronunciation detection in an offline test.

Chapter 5 explored using limited population-specific data to improve word-level ASR accuracy on disordered speech from children. In this way, speech therapy applications can verify that the child produced an utterance close to the target, making sure that they maintain appropriate effort during practice, while leaving deeper analysis for trained clinicians. This chapter compared two approaches: acoustic model adaptation for the PocketSphinx ASR engine, and a custom word recognizer based on template matching. Both template matching and maximum-a-posteriori-adapted acoustic models demonstrated accuracy close to or above the target threshold of 80% for 6 out of 7 test speakers. On average, the maximum-a-posteriori-adapted acoustic model yielded a higher accuracy than template matching. However, both outperformed the maximum-likelihood-linear-regression-adapted and non-adapted acoustic models.

Chapter 6 investigated using an existing phoneme-level mispronunciation detection technique from the L2 literature on disordered speech from children. These methods are a way to lighten the supervision responsibilities for caregivers and allows children to complete their practice more independently. This chapter compared classifier-

based mispronunciation detection against the standard Goodness of Pronunciation (GOP) baseline and student clinician evaluations. Phoneme-specific classifiers were trained on posterior-based features extracted from child speech samples gathered during the longitudinal evaluation of Apraxia World in Chapter 4. Results showed that these classifiers significantly outperformed the GOP approach at identifying mispronunciations in field-collected disordered speech from children. More importantly, the phoneme-specific classifiers detected mispronunciations significantly better than student clinicians.

7.2. Contributions

This dissertation contains the following main contributions:

- Developing Apraxia World, a novel speech therapy game that children play with traditional controls and incorporates speech input as a secondary control mechanic
- Conducting a longitudinal evaluation of Apraxia World that indicates children make therapy improvements with the game comparable to traditional home practice
- Showing that the game held children's attention over a two-month treatment period, with some even continuing to play of their own accord post-study
- Finding that children prefer Apraxia World to traditional therapy practice and that caregivers would like to include the game in future home practice
- Showing that limited child speech data can be used to increase word recognition rates to clinically-desirable levels for pronunciation verification
- Displaying that classifier-based mispronunciation detection outperforms both Goodness of Pronunciation scoring and student clinician evaluations on the collected disordered speech from children

7.3. Limitations

While this dissertation presented encouraging results that suggest Apraxia World is effective at increasing engagement and improving pronunciation accuracy, there are a few limitations to the evaluations, which are discussed below.

Study sizes: The pilot study in Chapter 3 presented results for 21 participants (14 with SSDs) and the longitudinal study in Chapter 4 presented results for 10 participants (all with SSDs). The low numbers reflect difficulties recruiting children for these types of studies, especially considering that they take time away from caregivers, as well. The small study sizes mean that results must be interpreted cautiously and taken as a precursor to larger-scale studies.

Gender imbalance: In elementary-school-aged populations, males are 2.85 times more likely to have an SSD than females [6], which makes recruiting balanced populations difficult. However, this does not eliminate the need for diverse populations, especially when collecting subjective data such as enjoyment and engagement with new applications. Given that general participant solicitation (Chapters 3 and 4) failed to provide balanced sex ratios, or even ones that approach the 2.85 to 1 ratio found in the clinical population, targeted recruitment for female participants is warranted in future work. As caregivers are the ones who need to be convinced to respond to solicitations, researchers should emphasize the opportunity to provide a voice to girls with SSDs in regards to what type of therapy tools they want to use. Recruiting participants for these types of studies can be challenging, but making efforts to find more female participants will yield more meaningful and generalizable results.

Novel technology: It is possible to argue that children's enthusiasm about Apraxia World in Chapter 2 was partially due to it being a novel technology and game they had never played before, instead of actual excitement about the game itself. However, it appears that whatever novelty factor impacted their opinion of the game was relatively limited, as in Chapter 3, children still reported enjoying Apraxia World over a two-month period and some even played beyond the formal study conclusion. Couse and Chen [209] found similar behavior in a study examining tablet use for early childhood education, where children remained excited to use tablets for educational activity both over two short study sessions and then informally for the remaining two months of the school year. However, the effect of novelty on child speech therapy applications remains an important consideration.

No control group (traditional speech therapy): Although the pronunciation improvement results reported in Chapter 4 are similar to those reported in studies of more traditional speech therapy practice, the study itself did not include a control phase with traditional speech therapy exercises. However, Apraxia World is presented as a supplement to other forms of practice, not a replacement for traditional practice. As such, comparing Apraxia World directly to other forms of speech therapy practice remains open for further investigation.

Lack of comparable corpora: An issue brought up in Chapters 5 and 6 is the lack of available child corpora, especially those containing disordered speech from children. Because the speech tests presented in this dissertation use a custom-curated corpus that cannot be distributed due to human research ethics committee restrictions, it may be

difficult for future researchers to replicate the results. Providing the child speech processing community with a standard corpus of disordered speech from children is still an open challenge.

7.4. Future work

The primary focus of this dissertation was to design an engaging speech therapy game incorporating mispronunciation detection for disordered speech from children. The findings from this work introduce potential directions and implications for further research.

7.4.1. *Game work*

Additional therapy game genres: This dissertation demonstrated the success of employing a side-scrolling adventure game for speech therapy. Given the variety of gaming preferences, future work should go into developing a wide range of speech therapy games across genres to give children the option to select the one that they enjoy most, with special emphasis placed on providing non-gendered options. These could be additional adventure games, building games, puzzle games, social games, racing games, etc. Some ideas for how to include speech into these game genres are shown in Table 10. However, these suggestions focus only on a few voice interaction techniques. Allison et al. [210] described 25 different voice interaction paradigms for games, demonstrating that there are many more ways to include speech beyond keyword repetition. Regardless of the speech integration method, it is important to design games that offer replay value and allow the player to make continual progress over a long period. These new genres and additional speech integration techniques offer exciting directions for future investigation.

Game genre	Genre example	Speech Integration
Action-Adventure	The Legend of Zelda	<ul style="list-style-type: none"> • Say words correctly to give the character's attack extra strength • Use speech to unlock special items • Say "magic words" to heal the character
Building	Minecraft	<ul style="list-style-type: none"> • Say words to place blocks • Say words to purchase building materials
Social	The Sims	<ul style="list-style-type: none"> • Complete speech exercises to earn money to buy clothing or decorative items for a virtual home • Say commands to make the character complete tasks
Racing	Mario Kart	<ul style="list-style-type: none"> • Say words at consistent volume and prosody to get a speed boost

Table 10 Possible speech therapy game genres and speech integration methods

Points, badges, and leaderboards: Three key elements of gamification (applying game strategies to non-game scenarios to increase engagement) are points, badges, and leaderboards [211]. These strategies have been demonstrated to increase feelings of confidence and task meaningfulness when completing task in a simulated environment [212], suggesting that they may do the same for digital speech therapy tasks. Although creating speech therapy games is more along the lines of traditional game development, instead of gamification, the fact that these three aspects are singled out for their ability to increase engagement means that researchers should pay special attention to how these are implemented within their therapy games.

One could argue that Apraxia World already includes a points system for completing speech exercises by offering high and low rewards (in-game currency and time on the "energy" timer) for correct and incorrect pronunciations, respectively. However,

the effect of badges and leaderboards has yet to be examined. Badges could be awarded for completing a certain number of speech exercises, improving accuracy beyond a threshold, or playing the game X days per week (similar to how some mobile games encourage frequent play with daily rewards or a play streak counter). Leaderboards have also been shown to increase engagement across a variety of otherwise mundane tasks [213]. However, introducing a leaderboard into speech therapy applications must be done very carefully, as not all information can be shared. Speech therapy outcomes are protected health information and could not be displayed on a leaderboard. Game points or progress could be shared between all players, but that rewards gameplay ability, rather than speech therapy effort. This dissertation demonstrated that gameplay ability varies greatly, so some children may find it discouraging to fall behind their peers. As badges and leaderboards are established methods for improving interest in dull tasks, their use in speech therapy applications remains an interesting, albeit challenging, research area.

Collaborative play: Providing other children or caregivers with a complementary role in the game, such as a helper character, may increase motivation and turn the therapy practice into a social experience. For example, Ganzeboom et al. [113] used separate player roles to encourage elderly people with dysarthria to give instructions through a speech therapy game, while their speech is analyzed for therapy feedback. In-game collaboration may not even need to be with a real human; Sailer et al. [212] demonstrated that including virtual teammates and a story motivating simulated tasks increased feelings of social relatedness. Virtual social motivation and motivating stories are demonstrated in the SpokeIt [68, 141] and Talking to Teo [32] child speech therapy games, where players must

speak to help other characters in the game, thus giving meaning to the actions. Further studies should explore the extent to which collaborative play increases a child's willingness to complete therapy exercises.

Accessible controls: Even though the controls employed in Apraxia World are standard for tablet games, they may not be completely accessible for populations undergoing speech therapy. A subset of children with movement-based speech disorders, such as childhood apraxia of speech, have limb coordination difficulties; some children during the pilot study were observed to have difficulty with game controls, extraneous limb movements, and rapidly timed double clicks. Other groups going through speech therapy may also experience difficulties with specific movements (e.g., children with Autism Spectrum Disorder [155]). As such, the touch-screen-based controls may not be an accessible control strategy and compensatory strategies for these factors, such as an external joystick, should be addressed in future speech therapy games.

Therapy normalization: One benefit of Apraxia World yet to be examined is the effect of normalizing speech therapy practice by including it in a game format not specific to children receiving therapy. In this way, children could talk about or share their experiences playing the game with their peers, without standing out as different. Children were enthusiastic about playing the game and some seemed very proud of their in-game accomplishments, which hopefully they felt free to share with their friends. It could be interesting to explore how reframing speech therapy exercises as a “regular” game changes how they are perceived both by children undergoing therapy and their peers with less exposure to speech therapy.

7.4.2. *Speech work*

Audio quality checks: Even though the children wore headsets for the majority of the longitudinal study in Chapter 4, there were issues with microphone placement and children adjusting or touching the microphone. Additionally, when some of the participants became discouraged or excited, they spoke in ways that made it difficult for the template matching to meaningfully evaluate their speech (mumbling, yelling, etc.). As such, future systems would benefit from monitoring microphone distortions, speaking volume, and speaking rate to recommend a correction. These reminders should help children produce utterances of better quality for automated speech processing, which would result in them receiving more meaningful feedback on pronunciations. This may also have the added benefit of helping children increase self-evaluation of loudness and intelligibility.

Combine word-level and phoneme-level verification: In order for phoneme-level mispronunciation detection to work correctly, systems must be able to accurately segment phonemes for analysis. However, incomplete utterances or productions that vary too much from the expected pronunciation are likely to be incorrectly segmented using automated approaches such as forced alignment. As such, it may be beneficial to the learner to offer feedback that the utterance had an overall issue, rather than trying to process an utterance that the system cannot accurately provide feedback for. One way to accomplish this would be to adopt the word-level verification approach presented in Chapter 5 as a precursor to deeper examination; this method uses ASR system with a speaker-dependent acoustic model to make sure that the child attempted to say something close to the target. If the

utterance passes this check, it can be passed along for segmentation and phoneme-level analysis. Otherwise, the child would receive appropriate word-level feedback asking them to try again. Combining the above proposed audio quality checks with this word-level verification could allow the system to better communicate why the child received specific utterance feedback; this would provide additional transparency and help build trust in the intelligent system [168].

Recording control mechanism: The touch-to-start/touch-to-stop mechanism implemented in Apraxia World proved difficult for the children to accurately control, as evidenced by the high percentage of clipped audio collected in the longitudinal study (Chapter 4). Many of the clipped utterances were missing just a small portion of the utterance, so a more child-friendly mechanism could yield better recordings, which would again improve ASR performance and provide more audio for offline processing. Ahmed et al. [26] also reported that children had trouble controlling the recording mechanism in their games, but their ASRs performed better when the games used discreet start and stop actions, instead of stopping the recording automatically. As such, a better mechanism may be to start recording once the prompt is displayed and trim the audio around a window defined by the button presses extended with padding to start earlier and stop later than when the child actually pressed the buttons. Since incomplete recordings oftentimes result in inaccurate automated feedback, it is essential to empower children to capture the entirety of their utterance. This replacement recording control mechanism should be the subject of future study.

One-Class Neural Network: Recent work by Shahin et al. [36, 144] suggests that an anomaly detection approach to mispronunciation detection may be able to identify child pronunciation errors, without the need for large amounts of error annotations. Their one-class SVM successfully discriminates between correct and incorrect pronunciations in disordered speech from children. Although their results are not replicated in this dissertation, this method poses interesting further steps, specifically the investigation of one-class neural networks, which have been shown to outperform one-class SVMs for anomaly detection tasks [214]. These performance improvements may also apply to the mispronunciation detection domain. Additionally, the one-class neural network can process higher-dimensional inputs better than one-class SVMs [214]; this means that features such as phonetic posteriorgrams, which describe the phonetic content in fine detail, could be used for anomaly-detection-based pronunciation evaluation. Phonetic posteriorgrams have been shown to better represent phonetic content than MFCCs during frame matching [215], suggesting that they may be appropriate for anomaly detection. Shahin et al. [36, 144] implemented phonetic attribute features for their mispronunciation detection pipeline, which requires a dedicated feature extraction network. However, phonetic posteriorgrams can be obtained from pre-existing, high quality acoustic models, which may simplify the mispronunciation detection framework.

REFERENCES

- [1] American Speech-Language-Hearing Association. (2007, November 5, 2018). *Speech Sound Disorders*. Available: <https://www.asha.org/public/speech/disorders/SpeechSoundDisorders/>
- [2] J. L. Anthony, R. G. Aghara, M. J. Dunkelberger, T. I. Anthony, J. M. Williams, and Z. Zhang, "What factors place children with speech sound disorders at risk for reading problems?," *American Journal of Speech-Language Pathology*, vol. 20, no. 2, pp. 146-160, 2011.
- [3] K. Forrest, "Diagnostic criteria of developmental apraxia of speech used by clinical speech-language pathologists," *American Journal of Speech-Language Pathology*, vol. 12, no. 3, pp. 376-380, 2003.
- [4] A. S.-L.-H. Association. (2020, May 9, 2020). *Speech Sound Disorders-Articulation and Phonology*. Available: <https://www.asha.org/Practice-Portal/Clinical-Topics/Articulation-and-Phonology/>
- [5] J. Law, J. Boyle, F. Harris, A. Harkness, and C. Nye, "Prevalence and natural history of primary speech and language delay: findings from a systematic review of the literature," *International Journal of Language and Communication Disorders*, vol. 35, pp. 165-188, 2000.
- [6] D. H. McKinnon, S. McLeod, and S. Reilly, "The prevalence of stuttering, voice, and speech-sound disorders in primary school students in Australia," *Language, Speech, and Hearing Services in Schools*, vol. 38, no. 1, pp. 5-15, 2007.
- [7] E. Maas, C. Gildersleeve-Neumann, K. J. Jakielski, and R. Stoeckel, "Motor-based intervention protocols in treatment of childhood apraxia of speech (CAS)," *Current developmental disorders reports*, vol. 1, no. 3, pp. 197-206, 2014.
- [8] D. C. Thomas, P. McCabe, and K. J. Ballard, "Rapid syllable transitions (ReST) treatment for childhood apraxia of speech: The effect of lower dose-frequency," *Journal of communication disorders*, vol. 51, pp. 29-42, 2014.
- [9] L. Ruggero, P. McCabe, K. J. Ballard, and N. Munro, "Paediatric speech-language pathology service delivery: An exploratory survey of Australian parents," *International Journal of Speech-Language Pathology*, vol. 14, no. 4, pp. 338-350, 2012/08/01 2012.
- [10] D. Theodoros, T. Russell, and R. Latifi, "Telerehabilitation: current perspectives," *Studies in health technology and informatics*, vol. 131, pp. 191-210, 2008.
- [11] D. G. Theodoros, "Telerehabilitation for service delivery in speech-language pathology," *Journal of Telemedicine and Telecare*, vol. 14, no. 5, pp. 221-224, 2008.
- [12] V. Joffe and T. Pring, "Children with phonological problems: A survey of clinical practice," *International Journal of Language & Communication Disorders*, vol. 43, no. 2, pp. 154-164, 2008.
- [13] S. Mcleod and E. Baker, "Speech-language pathologists' practices regarding assessment, analysis, target selection, intervention, and service delivery for

- children with speech sound disorders," *Clinical linguistics & phonetics*, vol. 28, no. 7-8, pp. 508-531, 2014.
- [14] L. McAllister, J. McCormack, S. McLeod, and L. J. Harrison, "Expectations and experiences of accessing and participating in services for childhood speech impairment," *International Journal of Speech-Language Pathology*, vol. 13, no. 3, pp. 251-267, 2011.
- [15] M. Zajc, A. Istenič Starčič, M. Lebeničnik, and M. Gačnik, "Tablet game-supported speech therapy embedded in children's popular practices," *Behaviour & Information Technology*, vol. 37, no. 7, pp. 693-702, 2018.
- [16] A. Parnandi *et al.*, "Architecture of an automated therapy tool for childhood apraxia of speech," in *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, 2013, p. 5: ACM.
- [17] M. Gačnik, A. I. Starčič, J. Zaletelj, and M. Zajc, "User-centred app design for speech sound disorders interventions with tablet computers," *Universal Access in the Information Society*, vol. 17, no. 4, pp. 821-832, 2018.
- [18] G. A. Constantinescu, D. G. Theodoros, T. G. Russell, E. C. Ward, S. J. Wilson, and R. Wootton, "Home-based speech treatment for Parkinson's disease delivered remotely: a case report," *Journal of Telemedicine and Telecare*, vol. 16, no. 2, pp. 100-104, 2010.
- [19] L. M. Jesus, J. Martinez, J. Santos, A. Hall, and V. Joffe, "Comparing Traditional and Tablet-Based Intervention for Children With Speech Sound Disorders: A Randomized Controlled Trial," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 11, pp. 4045-4061, 2019.
- [20] R. Palmer, P. Enderby, and M. Hawley, "Addressing the needs of speakers with longstanding dysarthria: computerized and traditional therapy compared," *International journal of language & communication disorders*, vol. 42, no. S1, pp. 61-79, 2007.
- [21] L. D. Shriberg, J. Kwiatkowski, and T. Snyder, "Tabletop versus microcomputer-assisted speech management: Response evocation phase," *Journal of Speech and Hearing Disorders*, vol. 55, no. 4, pp. 635-655, 1990.
- [22] A. Origlia *et al.*, "Evaluating a multi-avatar game for speech therapy applications," in *Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good*, 2018, pp. 190-195: ACM.
- [23] K. J. Ballard, N. M. Etter, S. Shen, P. Monroe, and C. T. Tan, "Feasibility of Automatic Speech Recognition for Providing Feedback During Tablet-Based Treatment for Apraxia of Speech Plus Aphasia," *American Journal of Speech-Language Pathology*, vol. 28, no. 2S, pp. 818-834, 2019.
- [24] M. E. Hoque, J. K. Lane, R. e. Kaliouby, M. Goodwin, and R. W. Picard, "Exploring Speech Therapy Games with Children on the Autism Spectrum," in *Tenth Annual Conference of the International Speech Communication Association*, 2009, pp. 1455-1458.
- [25] A. Hair, P. Monroe, B. Ahmed, K. J. Ballard, and R. Gutierrez-Osuna, "Apraxia World: A Speech Therapy Game for Children with Speech Sound Disorders," in

- Proc. of the 2018 Conference on Interaction Design and Children*, Trondheim, Norway, 2018, pp. 119-131: ACM.
- [26] B. Ahmed, P. Monroe, A. Hair, C. T. Tan, R. Gutierrez-Osuna, and K. J. Ballard, "Speech-driven mobile games for speech therapy: User experiences and feasibility," *International Journal of Speech-Language Pathology*, vol. 20, no. 6, pp. 644-658, 2018.
- [27] Z. Rubin, S. Kurniawan, T. Gotfrid, and A. Pugliese, "Motivating Individuals with Spastic Cerebral Palsy to Speak Using Mobile Speech Recognition," in *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*, 2016, pp. 325-326: ACM.
- [28] L. Furlong, S. Erickson, and M. E. Morris, "Computer-based speech therapy for childhood speech sound disorders," *Journal of communication disorders*, vol. 68, pp. 50-69, 2017.
- [29] Smarty Ears Apps. (2017). *Apraxiaville*. Available: <http://smartyearsapps.com/apraxia-ville/>
- [30] Little Bee Speech. (2018). *Articulation Station*. Available: http://littlebeespeech.com/articulation_station.php
- [31] Expressive Solutions. (2018). *ArtikPix*. Available: <http://expressive-solutions.com/artikpix/>
- [32] A. A. Navarro-Newball *et al.*, "Talking to Teo: Video game supported speech therapy," *Entertainment Computing*, vol. 5, no. 4, pp. 401-412, 2014/12/01/ 2014.
- [33] A. Nanavati, M. B. Dias, and A. Steinfeld, "Speak Up: A Multi-Year Deployment of Games to Motivate Speech Therapy in India," in *Proc. of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC, Canada, 2018, pp. 1-12: ACM.
- [34] T. Lan, S. Aryal, B. Ahmed, K. Ballard, and R. Gutierrez-Osuna, "Flappy voice: an interactive game for childhood apraxia of speech therapy," in *Proc. of the first ACM SIGCHI annual symposium on Computer-human interaction in play*, 2014, pp. 429-430: ACM.
- [35] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of children's speech: Duration, pitch and formants," in *Fifth European Conference on Speech Communication and Technology*, 1997.
- [36] M. Shahin and B. Ahmed, "Anomaly detection based pronunciation verification approach using speech attribute features," *Speech Communication*, vol. 111, pp. 29-43, 2019.
- [37] M. Shahin, B. Ahmed, J. McKechnie, K. Ballard, and R. Gutierrez-Osuna, "A Comparison of GMM-HMM and DNN-HMM Based Pronunciation Verification Techniques for Use in the Assessment of Childhood Apraxia of Speech," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [38] S. Dudy, M. Asgari, and A. Kain, "Pronunciation analysis for children with speech sound disorders," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 5573-5576: IEEE.

- [39] S. Dudy, S. Bedrick, M. Asgari, and A. Kain, "Automatic analysis of pronunciations for children with speech sound disorders," *Computer speech & language*, vol. 50, pp. 62-84, 2018.
- [40] A. Hair *et al.*, "A Longitudinal Evaluation of Tablet-Based Child Speech Therapy with Apraxia World," *ACM Transactions on Accessible Computing*, no. Under Review, pp. 1-25, 2020.
- [41] A. Hair *et al.*, "Preliminary Results From a Longitudinal Study of a Tablet-Based Speech Therapy Game," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing*, 2020: ACM.
- [42] A. Hair, K. J. Ballard, B. Ahmed, and R. Gutierrez-Osuna, "Evaluating Automatic Speech Recognition for Child Speech Therapy Applications," in *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 2019, pp. 578-580.
- [43] ASHA Adhoc Committee on CAS, "Childhood Apraxia of Speech," American Speech-Language-Hearing Association 2007.
- [44] L. D. Shriberg, D. Austin, B. A. Lewis, J. L. McSweeney, and D. L. Wilson, "The percentage of consonants correct (PCC) metric: Extensions and reliability data," *Journal of Speech, Language, and Hearing Research*, vol. 40, no. 4, pp. 708-722, 1997.
- [45] B. L. Davis, K. J. Jakielski, and T. P. Marquardt, "Developmental apraxia of speech: Determiners of differential diagnosis," *Clinical Linguistics & Phonetics*, vol. 12, no. 1, pp. 25-45, 1998.
- [46] K. J. Ballard, D. A. Robin, P. McCabe, and J. McDonald, "A treatment for dysprosody in childhood apraxia of speech," *Journal of Speech, Language, and Hearing Research*, 2010.
- [47] J. Iuzzini-Seigel, T. P. Hogan, and J. R. Green, "Speech inconsistency in children with childhood apraxia of speech, language impairment, and speech delay: Depends on the stimuli," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 5, pp. 1194-1210, 2017.
- [48] E. Murray, P. McCabe, R. Heard, and K. J. Ballard, "Differential diagnosis of children with suspected childhood apraxia of speech," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 1, pp. 43-60, 2015.
- [49] H. Terband and B. Maassen, "Speech motor development in childhood apraxia of speech: Generating testable hypotheses by neurocomputational modeling," *Folia Phoniatrica et Logopaedica*, vol. 62, no. 3, pp. 134-142, 2010.
- [50] J. Iuzzini and K. Forrest, "Evaluation of a combined treatment approach for childhood apraxia of speech," *Clinical linguistics & phonetics*, vol. 24, no. 4-5, pp. 335-345, 2010.
- [51] R. Prezas and B. Hodson, "The cycles phonological remediation approach," *Interventions for speech sound disorders in children*, pp. 137-158, 2010.
- [52] L. D. Shriberg and J. Kwiatkowski, "Phonological disorders II: A conceptual framework for management," *Journal of speech and Hearing Disorders*, vol. 47, no. 3, pp. 242-256, 1982.

- [53] A. A. Tyler and L. C. Tolbert, "Speech-language assessment in the clinical setting," *American Journal of Speech-Language Pathology*, 2002.
- [54] K. M. Brumbaugh and A. B. Smit, "Treating children ages 3–6 who have speech sound disorder: A survey," *Language, Speech, and Hearing Services in Schools*, 2013.
- [55] E. Sugden, E. Baker, N. Munro, A. L. Williams, and C. M. Trivette, "Service delivery and intervention intensity for phonology - based speech sound disorders," *International journal of language & communication disorders*, vol. 53, no. 4, pp. 718-734, 2018.
- [56] E. Sugden, E. Baker, N. Munro, A. L. Williams, and C. M. Trivette, "An Australian survey of parent involvement in intervention for childhood speech sound disorders," *International journal of speech-language pathology*, vol. 20, no. 7, pp. 766-778, 2018.
- [57] H. K. Ezell, "What Educators Can Teach Speech-Language Pathologists About Effective Homework Practices," *Journal of Children's Communication Development*, vol. 19, no. 1, pp. 63-69, 1997.
- [58] A. M. Piper, N. Weibel, and J. D. Hollan, "Introducing multimodal paper-digital interfaces for speech-language therapy," in *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*, 2010, pp. 203-210.
- [59] T. Burns. (2019, May 14, 2020). *How Can Parents Help Their Child with Apraxia at Home*. Available: https://www.apraxia-kids.org/apraxia_kids_library/the-importance-of-parent-involvement-in-the-speech-therapy-process/
- [60] A. S.-L.-H. Association. (May 14, 2020). *Activities to Encourage Speech and Language Development*. Available: <https://www.asha.org/public/speech/development/activities-to-Encourage-speech-and-Language-Development/>
- [61] C. S. Watson, D. J. Reed, D. Kewley-Port, and D. Maki, "The Indiana Speech Training Aid (ISTRA) I: Comparisons between human and computer-based evaluation of speech quality," *Journal of Speech, Language, and Hearing Research*, vol. 32, no. 2, pp. 245-251, 1989.
- [62] D. Kewley-Port, C. Watson, M. Elbert, D. Maki, and D. Reed, "The Indiana speech training aid (ISTRA) II: Training curriculum and selected case studies," *Clinical Linguistics & Phonetics*, vol. 5, no. 1, pp. 13-38, 1991.
- [63] D. Kewley-Port, C. Watson, D. Maki, and D. Reed, "Speaker-dependent speech recognition as the basis for a speech training aid," in *Proc. of the 1987 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1987, vol. 12, pp. 372-375: IEEE.
- [64] J. Dalby and D. Kewley-Port, "Explicit pronunciation training using automatic speech recognition technology," *CALICO journal*, pp. 425-445, 1999.
- [65] A. Parnandi *et al.*, "Development of a Remote Therapy Tool for Childhood Apraxia of Speech," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 7, no. 3, p. 10, 2015.
- [66] M. Shahin *et al.*, "Tabby Talks: An automated tool for the assessment of childhood apraxia of speech," *Speech Communication*, vol. 70, pp. 49-64, 2015.

- [67] J. Duval *et al.*, "SpokeIt: building a mobile speech therapy experience," in *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2018, pp. 1-12.
- [68] J. S. Duval, E. Márquez Segura, and S. Kurniawan, "SpokeIt: A Co-Created Speech Therapy Experience," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1-4: ACM.
- [69] J. Duval, "A mobile game system for improving the speech therapy experience," in *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2017, pp. 1-3.
- [70] C. T. Tan, A. Johnston, K. Ballard, S. Ferguson, and D. Perera-Schulz, "sPeAK-MAN: towards popular gameplay for speech therapy," in *Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death*, 2013, pp. 1-4.
- [71] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357-366, 1980.
- [72] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, 1992, vol. 1, pp. 137-140.
- [73] M. Gales and S. Young, *The application of hidden Markov models in speech recognition*. Now Publishers Inc, 2008.
- [74] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2016.
- [75] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82-97, 2012.
- [76] CMU Sphinx. (2017, September 20, 2020). *Basic concepts of speech recognition*. Available: <https://cmusphinx.github.io/wiki/tutorialconcepts/>
- [77] K.-F. Lee, H.-W. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, pp. 35-45, 1990.
- [78] P. Lamere *et al.*, "The CMU SPHINX-4 speech recognition system," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003)*, Hong Kong, 2003, vol. 1, pp. 2-5.
- [79] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006, vol. 1, pp. I-I: IEEE.
- [80] CMU Sphinx. (2020, May 22, 2020). *pocketsphinx 5prealpha*. Available: <https://sourceforge.net/projects/cmusphinx/files/pocketsphinx/5prealpha/>
- [81] cmusphinx. (2020, May 22, 2020). *PocketSphinx 5prealpha*. Available: <https://github.com/cmusphinx/pocketsphinx>
- [82] CMUSphinx, "Update on CMUSphinx Project," in *CMUSphinx*, ed. github.io, 2019.

- [83] C. Gaida, P. Lange, R. Petrick, P. Proba, A. Malatawy, and D. Suendermann-Oeft, "Comparing open-source speech recognition toolkits," *Tech. Rep., DHBW Stuttgart*, 2014.
- [84] D. Povey *et al.*, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011: IEEE Signal Processing Society.
- [85] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of DNNs with natural gradient and parameter averaging," *arXiv preprint arXiv:1410.7455*, 2014.
- [86] D. Can, V. R. Martinez, P. Papadopoulos, and S. S. Narayanan, "Pykaldi: A Python Wrapper for Kaldi," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5889-5893: IEEE.
- [87] M. Ravanelli, T. Parcollet, and Y. Bengio, "The PyTorch-Kaldi Speech Recognition Toolkit," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6465-6469: IEEE.
- [88] E. Silva, "Compile Kaldi for Android," in *Clean Blog* vol. 2017, ed, 2017.
- [89] alphacephei. (2020, May 22, 2020). *vosk-android-demo*. Available: <https://github.com/alphacep/vosk-android-demo>
- [90] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95-108, 2000.
- [91] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154-166, 2015.
- [92] T. Pellegrini, L. Fontan, J. Mauclair, J. Farinas, and M. Robert, "The goodness of pronunciation algorithm applied to disordered speech," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [93] H. Ryu and M. Chung, "Mispronunciation Diagnosis of L2 English at Articulatory Level Using Articulatory Goodness-Of-Pronunciation Features," in *SLaTE*, 2017, pp. 65-70.
- [94] V. Arora, A. Lahiri, and H. Reetz, "Phonological feature-based speech recognition system for pronunciation training in non-native language learning," *The Journal of the Acoustical Society of America*, vol. 143, no. 1, pp. 98-108, 2018.
- [95] S. Wei, G. Hu, Y. Hu, and R.-H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communication*, vol. 51, no. 10, pp. 896-905, 2009.
- [96] T. Zhao, A. Hoshino, M. Suzuki, N. Minematsu, and K. Hirose, "Automatic Chinese pronunciation error detection using SVM trained with structural features," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, 2012, pp. 473-478: IEEE.
- [97] A. M. Harrison, W. Y. Lau, H. M. Meng, and L. Wang, "Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

- [98] A. M. Harrison, W.-K. Lo, X.-j. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *International Workshop on Speech and Language Technology in Education*, 2009.
- [99] S. Mao, X. Li, K. Li, Z. Wu, X. Liu, and H. Meng, "Unsupervised discovery of an extended phoneme set in 12 english speech for mispronunciation detection and diagnosis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6244-6248: IEEE.
- [100] H. Meng, Y. Y. Lo, L. Wang, and W. Y. Lau, "Deriving salient learners' mispronunciations from cross-language phonological comparisons," in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2007, pp. 437-442: IEEE.
- [101] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Workshop on Speech and Language Technology in Education*, 2007.
- [102] M. Russell, S. D'Arcy, and L. Qun, "The effects of bandwidth reduction on human and computer recognition of children's speech," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1044-1046, 2007.
- [103] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 65-78, 2002.
- [104] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455-1468, 1999.
- [105] G. Yeung and A. Alwan, "On the Difficulties of Automatic Speech Recognition for Kindergarten-Aged Children," in *Interspeech*, 2018, pp. 1661-1665.
- [106] Tiga Talk. (2011). *Tiga Talk Speech Therapy Games*. Available: <http://tigatalk.com/app/>
- [107] Pocket SLP. (2018). *Pocket SLP*. Available: <http://pocketslp.com>
- [108] M. M. Mustaquim, "Automatic speech recognition- an approach for designing inclusive games," *Multimedia Tools and Applications*, journal article vol. 66, no. 1, pp. 131-146, September 01 2013.
- [109] H. Gürkök, G. Hakvoort, M. Poel, and A. Nijholt, "User expectations and experiences of a speech and thought controlled computer game," presented at the Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology, Lisbon, Portugal, 2011.
- [110] B. Lange, S. Flynn, and A. Rizzo, "Initial usability assessment of off-the-shelf video game consoles for clinical game-based motor rehabilitation," *Physical Therapy Reviews*, vol. 14, no. 5, pp. 355-363, 2009/10/01 2009.
- [111] J. Mora-Guiard, C. Crowell, N. Pares, and P. Heaton, "Lands of Fog: Helping Children with Autism in Social Interaction through a Full-Body Interactive Experience," presented at the Proceedings of the The 15th International Conference on Interaction Design and Children, Manchester, United Kingdom, 2016.
- [112] C. J. Cai, R. C. Miller, and S. Seneff, "Enhancing speech recognition in fast-paced educational games using contextual cues," in *SLaTE*, 2013, pp. 54-59.

- [113] M. Ganzeboom, E. Yilmaz, C. Cucchiarini, and H. Strik, "On the Development of an ASR-based Multimedia Game for Speech Therapy: Preliminary Results," in *Proceedings of the 2016 ACM Workshop on Multimedia for Personal Health and Health Care*, 2016, pp. 3-8: ACM.
- [114] M. Lopes, J. Magalhães, and S. Cavaco, "A voice-controlled serious game for the sustained vowel exercise," in *Proc. of the 13th International Conference on Advances in Computer Entertainment Technology*, 2016, pp. 1-6: ACM.
- [115] A. J. Sporka, S. H. Kurniawan, M. Mahmud, and P. Slavík, "Non-speech input and speech recognition for real-time control of computer games," in *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, 2006, pp. 213-220: ACM.
- [116] A. J. Sporka and P. Slavík, "Vocal Control of a Radio-Controlled Car," *ACM SIGACCESS Accessibility and Computing*, no. 91, pp. 3-8, 2008.
- [117] S. Harada, J. A. Landay, J. Malkin, X. Li, and J. A. Bilmes, "The vocal joystick: evaluation of voice-based cursor control techniques," in *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, 2006, pp. 197-204: ACM.
- [118] S. Harada, J. O. Wobbrock, and J. A. Landay, "Voice games: investigation into the use of non-speech voice input for making computer games more accessible," in *IFIP Conference on Human-Computer Interaction*, 2011, pp. 11-29: Springer.
- [119] B. House, J. Malkin, and J. Bilmes, "The VoiceBot: a Voice Controlled Robot Arm," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, 2009.
- [120] C. T. Tan, A. Johnston, A. Bluff, S. Ferguson, and K. J. Ballard, "Retrogaming as visual feedback for speech therapy," in *SIGGRAPH Asia 2014 Mobile Graphics and Interactive Applications*, 2014, p. 4: ACM.
- [121] J. Kennedy *et al.*, "Child speech recognition in human-robot interaction: evaluations and recommendations," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 82-90: ACM.
- [122] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511-1522, 1999.
- [123] O. Saz, S.-C. Yin, E. Lleida, R. Rose, C. Vaquero, and W. R. Rodríguez, "Tools and technologies for computer-aided speech and language therapy," *Speech Communication*, vol. 51, no. 10, pp. 948-967, 2009.
- [124] A. Alessandrini, V. Loux, G. F. Serra, and C. Murray, "Designing ReduCat: Audio-Augmented Paper Drawings Tangible Interface in Educational Intervention for High-Functioning Autistic Children," in *Proceedings of the The 15th International Conference on Interaction Design and Children*, Manchester, United Kingdom, 2016, pp. 463-472, 2930675: ACM.
- [125] G. Ferri, W. Sluis-Thiescheffer, D. Booten, and B. Schouten, "Playful Cognitive Behavioral Therapy Apps: Design Concepts and Tactics for Engaging Young Patients," presented at the Proceedings of the The 15th International Conference on Interaction Design and Children, Manchester, United Kingdom, 2016.

- [126] J. H. Annema, M. Verstraete, V. V. Abeele, S. Desmet, and D. Geerts, "Video games in therapy: a therapist's perspective," *International Journal of Arts and Technology*, vol. 6, no. 1, pp. 106-122, 2012.
- [127] L. Nguyen, W. Lu, E. Y.-L. Do, A. Chia, and Y. Wang, "Using digital game as clinical screening test to detect color deficiency in young children," presented at the Proceedings of the 2014 conference on Interaction design and children, Aarhus, Denmark, 2014.
- [128] C. Vaquero, O. Saz, E. Lleida, J. Marcos, C. Canalís, and C. P. De Educación, "VOCALIZA: An application for computer-aided speech therapy in Spanish language," *IV Jornadas en Tecnología del Habla*, pp. 321-326, 2006.
- [129] D. G. Jamieson, G. Kranjc, K. Yu, and W. E. Hodgetts, "Speech intelligibility of young school-aged children in the presence of real-life classroom noise," *Journal of the American Academy of Audiology*, vol. 15, no. 7, pp. 508-517, 2004.
- [130] D. Umanski, D. Kogovšek, M. Ozbič, and N. Schiller, "Development of a voice-based rhythm game for training speech motor skills of children with speech disorders," in *Proceedings 8th International Conference Disability, Virtual Reality and Associated Technologies*, 2010.
- [131] M. Shtern, M. B. Haworth, Y. Yunusova, M. Baljko, and P. Faloutsos, "A Game System for Speech Rehabilitation," in *Motion in Games: 5th International Conference, MIG 2012, Rennes, France, November 15-17, 2012. Proceedings*, M. Kallmann and K. Bekris, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 43-54.
- [132] P. Williams and H. Stephens, *Nuffield Centre Dyspraxia Programme 2004*. Miracle Factory, for the Speech & Language Therapy Department, Nuffield Hearing and Speech Centre, Royal National Throat, Nose and Ear Hospital, 2004.
- [133] E. Murray, P. McCabe, and K. J. Ballard, "A Randomized Controlled Trial for Children With Childhood Apraxia of Speech Comparing Rapid Syllable Transition Treatment and the Nuffield Dyspraxia Programme—Third Edition," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 3, pp. 669-686, 2015.
- [134] Y. Gao, B. M. L. Srivastava, and J. Salsman, "Spoken English Intelligibility Remediation with PocketSphinx Alignment and Feature Extraction Improves Substantially over the State of the Art," *arXiv preprint arXiv:1709.01713*, 2017.
- [135] J. Korte, "Patterns and Themes in Designing with Children," *Foundations and Trends® in Human-Computer Interaction*, vol. 13, no. 2, pp. 70-164, 2020.
- [136] C. L. Koch, *Clinical Management of Speech Sound Disorders*. Jones & Bartlett Learning, 2018.
- [137] E. Maas *et al.*, "Principles of motor learning in treatment of motor speech disorders," *American Journal of Speech-Language Pathology*, vol. 17, no. 3, pp. 277-298, 2008.
- [138] O. Engwall, O. Bälter, A.-M. Öster, and H. Kjellström, "Designing the user interface of the computer-based speech training system ARTUR based on early user tests," *Behaviour & Information Technology*, vol. 25, no. 4, pp. 353-365, 2006.

- [139] V. Lopes, J. Magalhaes, and S. Cavaco, "Sustained Vowel Game: a computer therapy game for children with dysphonia," in *Proc. Interspeech 2019*, 2019, pp. 26-30: ISCA.
- [140] G. J. Cler, T. Mittelman, M. N. Braden, G. H. Woodnorth, and C. E. Stepp, "Video Game Rehabilitation of Velopharyngeal Dysfunction: A Case Series," *Journal of Speech, Language, and Hearing Research : JSLHR*, vol. 60, no. 6 Suppl, pp. 1800-1809, 2017.
- [141] J. Duval *et al.*, "Designing Towards Maximum Motivation and Engagement in an Interactive Speech Therapy Game," in *Proceedings of the 2017 Conference on Interaction Design and Children*, 2017, pp. 589-594: ACM.
- [142] J. McKechnie, B. Ahmed, R. Gutierrez-Osuna, P. Monroe, P. McCabe, and K. Ballard, "Automated speech analysis tools for children's speech production: A systematic literature review," *International journal of speech-language pathology*, pp. 1-17, 2018.
- [143] O. Saz, E. Lleida, and W.-R. Rodríguez, "Avoiding speaker variability in pronunciation verification of children's disordered speech," presented at the Proceedings of the 2nd Workshop on Child, Computer and Interaction, Cambridge, Massachusetts, 2009. Available: <https://doi.org/10.1145/1640377.1640388>
- [144] M. Shahin, B. Ahmed, J. X. Ji, and K. Ballard, "Anomaly Detection Approach for Pronunciation Verification of Disordered Speech Using Speech Attribute Features," *Proc. Interspeech 2018*, pp. 1671-1675, 2018.
- [145] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The OGI kids' speech corpus and recognizers," in *Sixth International Conference on Spoken Language Processing*, 2000.
- [146] A. Batliner *et al.*, "The PF_STAR children's speech corpus," in *Ninth European Conference on Speech Communication and Technology*, 2005.
- [147] Y. Rose and B. MacWhinney, "The PhonBank Project: Data and software-assisted methods for the study of phonology and phonological development," in *The Oxford handbook of corpus phonology* Oxford, UK: Oxford University Press, 2014, pp. 380-401.
- [148] C. Torrington Eaton and N. B. Ratner, "An exploration of the role of executive functions in preschoolers' phonological development," *Clinical linguistics & phonetics*, vol. 30, no. 9, pp. 679-695, 2016.
- [149] A. E. Cummings and J. A. Barlow, "A comparison of word lexicality in the treatment of speech sound disorders," *Clinical linguistics & phonetics*, vol. 25, no. 4, pp. 265-286, 2011.
- [150] J. L. Preston, M. Hull, and M. L. Edwards, "Preschool speech error patterns predict articulation and phonological awareness outcomes in children with histories of speech sound disorders," *American Journal of Speech-Language Pathology*, 2013.
- [151] M. Grilo *et al.*, "The BioVisualSpeech European Portuguese Sibilants Corpus," in *International Conference on Computational Processing of the Portuguese Language*, 2020, pp. 23-33: Springer.

- [152] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. of the 2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, vol. 4, no. 4, pp. 4072-4075.
- [153] G. Yeung, A. Afshan, K. E. Ozgun, C. Kaewtip, S. M. Lulich, and A. Alwan, "Predicting Clinical Evaluations of Children's Speech with Limited Data Using Exemplar Word Template References," in *Proc. 7th ISCA Workshop on Speech and Language Technology in Education*, 2017, pp. 161-166: IEEE.
- [154] Ş. Tükel, H. Björelus, G. Henningsson, A. McAllister, and A. C. Eliasson, "Motor functions and adaptive behaviour in children with childhood apraxia of speech," *International journal of speech-language pathology*, vol. 17, no. 5, pp. 470-480, 2015.
- [155] K. L. Staples and G. Reid, "Fundamental movement skills and autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 40, no. 2, pp. 209-217, 2010.
- [156] V. J. Shute, "Focus on formative feedback," *Review of educational research*, vol. 78, no. 1, pp. 153-189, 2008.
- [157] M. Bakker, L. Beijer, and T. Rietveld, "Considerations on Effective Feedback in Computerized Speech Training for Dysarthric Speakers," *Telemedicine and e-Health*, vol. 25, no. 5, pp. 351-358, 2019.
- [158] Google. (August 23, 2018). *Cloud Text-to-Speech*. Available: <https://cloud.google.com/text-to-speech>
- [159] R. A. Schmidt and T. D. Lee, *Motor Learning and Control: A Behavioral Emphasis*. Champaign, IL, United States: Human Kinetics, 2005.
- [160] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254-272, 1981.
- [161] S. K. Fager, "Speech Recognition as a Practice Tool for Dysarthria," in *Seminars in speech and language*, 2017, vol. 38, no. 03, pp. 220-228: Thieme Medical Publishers.
- [162] A. Loukina *et al.*, "Automated Estimation of Oral Reading Fluency During Summer Camp e-Book Reading with MyTurnToRead," in *Proc. Interspeech 2019*, 2019, pp. 21-25: ISCA.
- [163] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206-5210: IEEE.
- [164] D. C. Thomas, P. McCabe, and K. J. Ballard, "Combined clinician-parent delivery of rapid syllable transition (ReST) treatment for childhood apraxia of speech," *International journal of speech-language pathology*, vol. 20, no. 7, pp. 683-698, 2018.
- [165] Y. Wren and S. Roulstone, "A comparison between computer and tabletop delivery of phonology therapy," *International Journal of Speech-Language Pathology*, vol. 10, no. 5, pp. 346-363, 2008.

- [166] E. F. Strommen and F. S. Frome, "Talking back to big bird: Preschool users and a simple speech recognition system," *Educational Technology Research and Development*, vol. 41, no. 1, pp. 5-16, 1993.
- [167] A. Springer and S. Whittaker, "Progressive disclosure: empirically motivated approaches to designing effective transparency," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, pp. 107-120.
- [168] J. Zhou and F. Chen, *Human and machine learning: Visible, explainable, trustworthy and transparent*. Springer, 2018.
- [169] C. Baur, E. Rayner, and N. Tsourakis, "Using a serious game to collect a child learner speech corpus," in *Ninth international conference on language resources and evaluation (LREC)*, 2014.
- [170] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being," *American psychologist*, vol. 55, no. 1, p. 68, 2000.
- [171] F. Ballati, F. Corno, and L. De Russis, "Assessing Virtual Assistant Capabilities with Italian Dysarthric Speech," in *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, 2018, pp. 93-101: ACM.
- [172] R. Fok, H. Kaur, S. Palani, M. E. Mott, and W. S. Lasecki, "Towards More Robust Speech Interactions for Deaf and Hard of Hearing Users," in *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, 2018, pp. 57-67: ACM.
- [173] Z. Rubin, S. Kurniawan, and T. Tollefson, "Results from using automatic speech recognition in cleft speech therapy with children," in *International Conference on Computers for Handicapped Persons*, 2014, pp. 283-286: Springer.
- [174] W. Hu, Y. Qian, and F. K. Soong, "A new neural network based logistic regression classifier for improving mispronunciation detection of L2 language learners," in *The 9th International Symposium on Chinese Spoken Language Processing*, 2014, pp. 245-249: IEEE.
- [175] W. Li, K. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Detecting Mispronunciations of L2 Learners and Providing Corrective Feedback Using Knowledge-Guided and Data-Driven Decision Trees," in *Interspeech*, 2016, pp. 3127-3131.
- [176] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech & language*, vol. 9, no. 2, pp. 171-185, 1995.
- [177] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE transactions on speech and audio processing*, vol. 2, no. 2, pp. 291-298, 1994.
- [178] S. Young *et al.*, "The HTK book," *Cambridge university engineering department*, vol. 3, p. 175, 2002.
- [179] B. McFee *et al.*, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18-25.

- [180] E. Sugden, N. Munro, C. M. Trivette, E. Baker, and A. L. Williams, "Parents' experiences of completing home practice for speech sound disorders," *Journal of Early Intervention*, vol. 41, no. 2, pp. 159-181, 2019.
- [181] R. Goodhue, M. Onslow, S. Quine, S. O'Brian, and A. Hearne, "The Lidcombe Program of early stuttering intervention: mothers' experiences," *Journal of Fluency Disorders*, vol. 35, no. 1, pp. 70-84, 2010/03/01/ 2010.
- [182] K. E. Davies, J. Marshall, L. J. Brown, and J. Goldbart, "Co-working: Parents' conception of roles in supporting their children's speech and language development," *Child Language Teaching and Therapy*, vol. 33, no. 2, pp. 171-185, 2017.
- [183] B. Munson, J. M. Johnson, and J. Edwards, "The role of experience in the perception of phonetic detail in children's speech: A comparison between speech-language pathologists and clinically untrained listeners," *American Journal of Speech-Language Pathology*, 2012.
- [184] E. Sugden, E. Baker, A. L. Williams, N. Munro, and C. M. Trivette, "Evaluation of Parent-and Speech-Language Pathologist-Delivered Multiple Oppositions Intervention for Children With Phonological Impairment: A Multiple-Baseline Design Study," *American Journal of Speech-Language Pathology*, vol. 29, no. 1, pp. 111-126, 2020.
- [185] H. Strik, K. Truong, F. De Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech communication*, vol. 51, no. 10, pp. 845-852, 2009.
- [186] K. Li, X. Qian, and H. Meng, "Mispronunciation detection and diagnosis in 12 english speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193-207, 2017.
- [187] W.-K. Leung, X. Liu, and H. Meng, "CNN-RNN-CTC Based End-to-end Mispronunciation Detection and Diagnosis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8132-8136: IEEE.
- [188] Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," in *Fifth European Conference on Speech Communication and Technology*, 1997.
- [189] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," in *Sixth European Conference on Speech Communication and Technology*, 1999.
- [190] Z. Wang, J. Zhang, and Y. Xie, "L2 Mispronunciation Verification Based on Acoustic Phone Embedding and Siamese Networks," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2018, pp. 444-448: IEEE.
- [191] H. Huang, H. Xu, X. Wang, and W. Silamu, "Maximum F1-score discriminative training criterion for automatic mispronunciation detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 787-797, 2015.

- [192] S. Kanters, C. Cucchiarini, and H. Strik, "The Goodness of Pronunciation Algorithm: a Detailed Performance Study," in *ISCA International Workshop on Speech and Language Technology in Education*, Warwickshire, England, 2009: ISCA.
- [193] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," in *International Symposium on automatic detection on errors in pronunciation training*, 2012.
- [194] J. Van Doremalen, C. Cucchiarini, and H. Strik, "Automatic detection of vowel pronunciation errors using multiple information sources," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, 2009, pp. 580-585: IEEE.
- [195] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6135-6139: IEEE.
- [196] J. Wang, Y. Qin, Z. Peng, and T. Lee, "Child Speech Disorder Detection with Siamese Recurrent Network using Speech Attribute Features," *Proc. Interspeech 2019*, pp. 3885-3889, 2019.
- [197] A. Ito, Y.-L. Lim, M. Suzuki, and S. Makino, "Pronunciation error detection method based on error rule clustering using a decision tree," in *Ninth European Conference on Speech Communication and Technology*, 2005.
- [198] S.-Y. Yoon, M. Hasegawa-Johnson, and R. Sproat, "Landmark-based automated pronunciation error detection," in *Eleventh annual conference of the international speech communication association*, 2010.
- [199] H. Franco, L. Ferrer, and H. Bratt, "Adaptive and discriminative modeling for improved mispronunciation detection," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7709-7713: IEEE.
- [200] R. Duan, T. Kawahara, M. Dantsujii, and J. Zhang, "Pronunciation error detection using DNN articulatory model based on multi-lingual and multi-task learning," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2016, pp. 1-5: IEEE.
- [201] W. Li, N. F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Improving Mispronunciation Detection for Non-Native Learners with Multisource Information and LSTM-Based Deep Models," in *INTERSPEECH*, 2017, pp. 2759-2763.
- [202] X. Qian, H. Meng, and F. Soong, "A two-pass framework of mispronunciation detection and diagnosis for computer-aided pronunciation training," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 6, pp. 1020-1028, 2016.
- [203] Y.-B. Wang and L.-s. Lee, "Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 564-579, 2015.
- [204] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *2014 IEEE international*

- conference on acoustics, speech and signal processing (ICASSP), 2014, pp. 215-219: IEEE.
- [205] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy modelling," in *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994.
- [206] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," *Proc. Interspeech 2017*, pp. 498-502, 2017.
- [207] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825-2830, 2011.
- [208] N. Watts Pappas, L. McAllister, and S. McLeod, "Parental beliefs and experiences regarding involvement in intervention for their child with speech sound disorder," *Child Language Teaching and Therapy*, vol. 32, no. 2, pp. 223-239, 2016.
- [209] L. J. Couse and D. W. Chen, "A tablet computer for young children? Exploring its viability for early childhood education," *Journal of research on technology in education*, vol. 43, no. 1, pp. 75-96, 2010.
- [210] F. Allison, M. Carter, M. Gibbs, and W. Smith, "Design Patterns for Voice Interaction in Games," presented at the Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play, Melbourne, VIC, Australia, 2018. Available: <https://doi.org/10.1145/3242671.3242712>
- [211] K. Werbach and D. Hunter, *For the Win: How Game Thinking Can Revolutionize Your Business*. Philadelphia, Pennsylvania, United States: Wharton Digital Press, 2012.
- [212] M. Sailer, J. U. Hense, S. K. Mayr, and H. Mandl, "How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction," *Computers in Human Behavior*, vol. 69, pp. 371-380, 2017.
- [213] A. Alter, *Irresistible: The rise of addictive technology and the business of keeping us hooked*. Penguin, 2017.
- [214] L. Ruff *et al.*, "Deep one-class classification," in *International conference on machine learning*, 2018, pp. 4393-4402.
- [215] G. Zhao and R. Gutierrez-Osuna, "Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1649-1660, 2019.
- [216] M. Russell. (2006, May 18, 2020). *The PF-STAR British English Children's Speech Corpus*. Available: <http://www.thespeechark.com/pf-star-child-corpus-v1.0.pdf>
- [217] Boulder Learning Inc. (2019, May 16, 2020). *The MyST Children's Conversational Speech Corpus*. Available: <http://boulderlearning.com/products/details-of-the-myst-corpus/>
- [218] A. Eshky *et al.*, "UltraSuite: a repository of ultrasound and acoustic data from child speech therapy sessions," *arXiv preprint arXiv:1907.00835*, 2019.
- [219] B. MacWhinney, "The Talkbank Project," in *Creating and Digitizing Language Corpora*: Springer, 2007, pp. 163-180.

APPENDIX A

HUMAN RESEARCH ETHICS COMMITTEE MATERIALS

This section contains the consent forms and questionnaires used during the longitudinal study of Apraxia World in Chapters 3. The consent forms and questionnaires used for the pilot study in chapter 2 were largely similar to those shown here, so only the latest versions are presented here. They are ordered as follows:

- Caregiver consent form
- Child consent form
- Caregiver information sheet
- Child information sheet
- Caregiver questionnaire (Longitudinal study only)
- Child questionnaire

During the longitudinal study, the caregivers only answered questions 13-16 on their questionnaire at the end of the study. All other questions were answered both during and after the study



Discipline of Speech Pathology
Faculty of Health Sciences

ABN 15 211 513 464

Professor Kirrie Ballard

The University of Sydney
NSW 2006 AUSTRALIA
Telephone: +61 2 9351 9879
Facsimile: +61 2 9351 9173
Email: kirrie.ballard@sydney.edu.au
Web:
sydney.edu.au/health_sciences/staff/kirrie_ballard

Apraxia World: An interactive technology-based tool for remote speech therapy of childhood speech sound disorders

PARENT/CARER CONSENT FORM

I, [PRINT PARENT'S/CARER'S NAME], consent to my child
..... [PRINT CHILD'S NAME] participating in this research study.

In giving my consent I state that:

- ✓ I understand the purpose of the study, what my child will be asked to do, and any risks/benefits involved.
- ✓ I have read the Information Statement and have been able to discuss my child's involvement in the study with the researchers if I wished to do so.
- ✓ The researchers have answered any questions that I had about the study and I am happy with the answers.
- ✓ I understand that being in this study is completely voluntary and my child does not have to take part. My decision whether to let them take part in the study will not affect our relationship with the researchers or anyone else at the University of Sydney now or in the future.
- ✓ I understand that my child can withdraw from the study at any time.
- ✓ I understand that my child may stop the assessment at any time if they do not wish to continue, and that unless I indicate otherwise any recordings will then be erased and the information provided will not be included in the study. I also understand that my child may refuse to answer any questions they don't wish to answer.
- ✓ I understand that personal information about my child that is collected over the course of this project will be stored securely and will only be used for purposes that I have agreed to. I understand that information about my child will only be told to others with my permission, except as required by law.

- ✓ I understand that the results of this study may be published, and that publications will not contain my child's name or any identifiable information about my child.
- ✓ I understand that I will be given a tablet to use for the duration of the study only. I will return the tablet to the researchers when my involvement in the study ends.

I consent to:

- **The team contacting our speech pathologist to obtain copies of my child's previous speech pathology reports and get advice on speech exercises** YES NO
- **Audio-recording of my child** YES NO
- **Receiving feedback about my child's personal results** YES NO

Would you like to receive feedback about the overall results of this study?

YES NO

If you answered **YES** to providing the research team with copies of previous speech pathology reports, please provide the name of contact details for your child's speech pathologist(s):

Name: _____
 Address: _____
 Phone / Email: _____

Name: _____
 Address: _____
 Phone / Email: _____

If you answered **YES** to receiving feedback about your child's personal results and/or the overall results of the study, please indicate your preferred form of feedback and address:

- Postal: _____

- Email: _____

Parent's/carer's signature:

.....
Signature

.....
PRINT name **Date**



ABN 15 211 513 464

Professor Kirrie Ballard

The University of Sydney
NSW 2006 AUSTRALIA
Telephone: +61 2 9351 9879
Facsimile: +61 2 9351 9173
Email: kirrie.ballard@sydney.edu.au
Web:
sydney.edu.au/health_sciences/staff/kirrie_ballard

Apraxia World: An interactive technology-based tool for remote speech therapy of childhood speech sound disorders

CONSENT FORM

If you are happy to be in the study, please

- **write your name** in the space below
- **sign your name** at the bottom of the next page
- put the **date** at the bottom of the next page.

You should only say 'yes' to being in the study if you know what it is about and you want to be in it. If you don't want to be in the study, don't sign the form.

I,[PRINT NAME], am happy to be in this research study.

In saying yes to being in the study, I am saying that:

- ✓ I know what the study is about.
- ✓ I know what I will be asked to do.
- ✓ Someone has talked to me about the study.
- ✓ My questions have been answered.
- ✓ I know that I don't have to be in the study if I don't want to.
- ✓ I know that I can pull out of the study at any time if I don't want to do it anymore.
- ✓ I know that I don't have to answer any questions that I don't want to answer.

✓ I know that the researchers won't tell anyone what I say when we talk to each other, unless I talk about being hurt by someone or hurting myself or someone else.

Now we are going to ask you if you are happy to do a few other things in the study. Please circle 'Yes' or 'No' to tell us what you would like.

Are you happy for us to **record** your voice with the table / iPad? **Yes** **No**

Do you want us to tell you what we **learnt** in the study? **Yes** **No**

.....
Signature

.....
Date

ABN 15 211 513 464

Professor Kirrie Ballard

The University of Sydney
NSW 2006 AUSTRALIA
Telephone: +61 2 9351 9879
Facsimile: +61 2 9351 9173
Email: kirrie.ballard@sydney.edu.au
Web: sydney.edu.au/health_sciences/staff/kirrie_ballard

***Apraxia World: An interactive technology-based tool for remote speech therapy of
childhood speech sound disorders***

PARTICIPANT INFORMATION STATEMENT

(1) What is this study about?

You and your child are invited to take part in a research study looking into the development of a tablet-based tool for the treatment of children with speech sound disorders. The purpose of this study is to identify whether a tablet-based tool with interactive speech games can help children engage in speech therapy that is higher intensity and more fun.

You and your child are invited to participate in this study because your child has been diagnosed with a speech sound disorder (i.e. others sometimes/always find your child difficult to understand). This Participant Information Statement tells you about the research study. Knowing what is involved will help you decide if you and your child want to take part in the research. Please read this sheet carefully and ask questions about anything that you don't understand or want to know more about.

Participation in this research study is voluntary.

By giving your consent you are telling us that you:

- ✓ Understand what you have read.
- ✓ Agree for your child to take part in the research study as outlined below.
- ✓ Agree to the use of your child's personal information as described.

You will be given a copy of this Parental Information Statement to keep.

(2) Who is running the study?

The study is being carried out by the following researchers:

- Professor Kirrie Ballard, University of Sydney (experienced speech pathologist)
- Dr Beena Ahmed, Texas A&M University in Qatar (electrical engineer)
- Dr Ricardo Gutierrez-Osuna, Texas A&M University in USA (computer scientist)
- Penelope Monroe, University of Sydney (speech pathologist)
- Adam Hair, Texas A&M University in USA (PhD student)

This study is being funded by the Qatar National Research Foundation.

(3) What will the study involve?

If your child is currently attending speech therapy, you will be asked to suspend the regular therapy sessions for the duration of the trial only. You will instead be asked to engage in weekly skype or phone sessions with a speech pathologist from the University of Sydney and to use a therapy app on a Samsung tablet for at-home practice. If needed (e.g. technical problems), you will be able to come in to the university clinic for some sessions. You and your child will be taught how to use the tablet app. The tablet and app will be provided free of charge. You will need to use your home internet connection for the tablet to automatically send audio recordings of your child's speech attempts during home practice to us. All your child's speech attempts are recorded by the tablet and uploaded to a secure site for the research team to listen, check progress, and update the exercises if needed. ~~These recordings will also be analysed by the research team.~~

You will be asked to use the app at home for two 4-week blocks over 10 weeks and then return the tablet to the research team. After the 10 weeks, we will ask both you, and your child (with your help), to complete an online survey about your impressions and experiences with the table and app.

You will be loaned a tablet to use for the duration of the study only. You will need to return the tablet to the researchers, when your involvement in the study ends.

To help the research team set up the app for your child and interpret how different children respond to the tablet app and games, we will ask your permission to obtain copies of previous speech pathology reports describing your child's speech difficulties, what they have worked on in therapy before, and how they have progressed in previous therapy. If you provide this permission, we may also speak directly with your regular speech pathologist to explain the study to them, get copies of previous assessment and therapy reports, and get advice on what speech exercises to use in the app for your child. You will be able to participate in the study, whether or not you give us permission to obtain these reports and information.

(4) How much time will the study take?

The study will take 10 weeks. We ask that you suspend your regular visits to your speech pathologist for the duration of the study. We will Skype / phone you and your child once a week for up to 1 hour. In between these calls, we recommend that your child use the app at home for 45 minutes of practice a day for any 4 days each week. This could be done in one 45 minute block, in several shorter periods, or spread out over more than 4 days a week. The most important thing as that your child aim for the total of 3 hours practice in a week.

(5) Who can take part in the study?

Children aged between 6 and 12 years who –

- Have a diagnosis of a speech sound disorder
- Have good understanding of what is said to them and normal hearing
- Have no other developmental diagnosis
- Speak Australian English and have at least one parent who speaks English as a first language

(6) Can my child and I withdraw from the study once they have started?

Being in this study is completely voluntary and you/your child do not have to take part. Your decision whether to participate will not affect your relationship with your speech pathologist or researchers or anyone else at the University of Sydney now or in the future.



If you/your child decide to take part in the study and then change your mind later (or no longer wish to take part), you are free to withdraw from the study at any time. Just let one of the researchers know and we can arrange for the tablet to be returned.

You/your child are free to stop the trial at any time. Unless you say that you want us to keep them, any recordings will be erased and the information you/your child have provided will not be included in the study results.

(7) Are there any risks or costs associated with being in the study?

No. The tablet and server access are provided for free for the duration of the study only.

(8) Are there any benefits associated with being in the study?

We cannot guarantee that you/your child will receive any direct benefits from being in the study. You will need to return the tablet and your access to the app will be stopped after the trial; this is because the system is still in development and is not ready for use without the research team's supervision.

(9) What will happen to information that is collected during the study?

By providing your consent, you are agreeing to us collecting personal information about you/your child for the purposes of this research study. This personal information will only be used for the purposes outlined in this Participant Information Statement, unless you consent otherwise.

Your/your child's information will be stored securely and identity/information will be kept strictly confidential, except as required by law. Study findings may be published, but you/your child will not be individually identifiable in these publications.

We will keep the information we collect for this study, and we may use it in future projects. By providing your consent you are allowing us to use your information in future projects. We don't know at this stage what these other projects will involve. We will seek ethical approval before using the information in these future projects.

(10) Can I / my child or tell other people about the study?

Yes, you are welcome to tell other people about the study.

(11) What if I would like further information about the study?

When you have read this information, Professor Kirrie Ballard will be available to discuss it with you further and answer any questions you may have. If you would like to know more at any stage during the study, please feel free to contact Professor Kirrie Ballard (Kirrie.ballard@sydney.edu.au; ph: 0431-416-936).

(12) Will we be told the results of the study?

You have a right to receive feedback about the overall results of this study. You can tell us that you wish to receive feedback by ticking the feedback box on the consent form. This feedback will be in the form of a one page summary. You will receive this feedback after the study is finished.

(13) What if we have a complaint or any concerns about the study?

Research involving humans in Australia is reviewed by an independent group of people called a Human Research Ethics Committee (HREC). The ethical aspects of this study have been



approved by the HREC of the University of Sydney [INSERT protocol number once approval is obtained]. As part of this process, we have agreed to carry out the study according to the *National Statement on Ethical Conduct in Human Research (2007)*. This statement has been developed to protect people who agree to take part in research studies.

If you are concerned about the way this study is being conducted or wish to make a complaint to someone independent from the study, please contact the university using the details outlined below. Please quote the study title and protocol number.

The Manager, Ethics Administration, University of Sydney:

- **Telephone:** +61 2 8627 8176
- **Email:** ro.humanethics@sydney.edu.au
- **Fax:** +61 2 8627 8177 (Facsimile)

This information sheet is for you to keep

ABN 15 211 513 464

Professor Kirrie Ballard

The University of Sydney
NSW 2006 AUSTRALIA
Telephone: +61 2 9351 9879
Facsimile: +61 2 9351 9173
Email: kirrie.ballard@sydney.edu.au
Web: sydney.edu.au/health_sciences/staff/kirrie_ballard

STUDY INFORMATION SHEET:

Apraxia World: An interactive technology-based tool for remote speech therapy of childhood speech sound disorders

Hello. Our names are

- Kirrie Ballard, Beena Ahmed, Ricardo Gutierrez-Osuna, Penelope Monroe and Adam Hair



We are doing a research study to help make an app for children who have trouble with talking. The app has activities and games for children to practice their talking.

We are asking you to be in our study because you have trouble with talking.

You can decide if you want to take part in the study or not. You don't have to - it's up to you.

This sheet tells you what we will ask you to do if you decide to take part in the study. Please read it carefully so that you can make up your mind about whether you want to take part.

If you decide you want to be in the study and then you change your mind later, that's ok. All you need to do is tell us that you don't want to be in the study anymore.

If you have any questions, you can ask us or your family or someone else who looks after you. If you want to, you can call us any time on 0431-416-936.

What will happen if I say that I want to be in the study?

If you decide that you want to be in our study, we will ask you to do these things:

- Not see your regular speech pathologist for 10 weeks.
- Skype or call Kirrie or Penny once a week and do practice at home on a tablet.
- We will lend you a tablet (like an iPad) to do your speech exercises at home.
- Kirrie or Penny will teach you how to use our tablet and our new app that is on the tablet.
- You don't have to pay for the tablet but you need to give it back to us when the study finishes.
- If you say it's ok, we will record what you say using the tablet.
- The recordings are sent over the internet to us so we can check how you're going. If the exercises are too easy or too hard, we will change them for you.

Will anyone else know what I say in the study?

We won't tell anyone else what you say to the app, except if you talk about someone hurting you or about you hurting yourself or someone else. Then we might need to tell someone to keep you and other people safe.



All of the information that we have about you from the study will be stored in a safe place and we will look after it very carefully. We will write a report about the study and show it to other people but we won't say your name in the report and no one will know that you were in the study, unless you tell us that it's ok for us to say your name.

How long will the study take?



We will lend you the tablet for 10 weeks. You can use it to do your speech therapy homework on. We think it is best to do your homework activities and games on the tablet with your carer for about 45 mins a day, for 4 days each week (that's 3 hours all together). But you can spread out the practice each day to a few shorter practices; or you can spread it out over more than 4 days.

Are there any good things about being in the study?



Because you are doing your speech therapy homework, you may get better at saying the words that we have put into the activities and games. You will also be helping us do our research.

Are there any bad things about being in the study?



It might be hard to say some of the words correctly; that's why you are practising them. This study will take up some of your time, but we don't think it will be bad for you or cost you anything.

Will you tell me what you learnt in the study at the end?

Yes, we will if you want us to. There is a question on the next page that asks you if you want us to tell you what we learnt in the study. If you circle Yes, when we finish the study we will tell you what we learnt.

What if I am not happy with the study or the people doing the study?



If you are not happy with how we are doing the study or how we treat you, then you or the person who looks after you can:

- Call the university on +61 2 8627 8176 or
- Write an email to ro.humanethics@sydney.edu.au

This sheet is for you to keep.

The pictures we used in this sheet are from Microsoft Clip Art and from the people at Inspired Services Publishing (www.inspiredservices.org.uk). They said it's ok for us to use them.



Satisfaction and Software Usability Survey: CARER

Child's ID: _____

Date: _____

Child's gender (circle) M F

1. Did your child enjoy using the tablet for the speech therapy activities?

Yes

No

2. Did your child need any help completing the activities on the tablet?

Yes

No

(2a) If yes, please tell us what you/your child needed help with:

Selecting an exercise

Moving between images/activities/games

Starting the recording

Stopping the recording

Accessing the hints (stored audio)

Navigating back to the home page

Internet access / connectivity for uploading recordings to the server

Other.... (please specify)

7. What did your child dislike about the exercises (activities and / or games) on the tablet?

8. How easy was it to fit the therapy program into your daily life?

|_____||_____||_____||_____||

Very easy Easy Neither easy nor difficult Difficult Very difficult

9. How satisfied were you with your frequency of contact with the speech pathologist?

|_____||_____||_____||_____||

Extremely Satisfied Satisfied Neither satisfied nor dissatisfied Dissatisfied Extremely dissatisfied

10. How easy was it for you to contact your speech pathologist in between sessions (or tick Not Applicable)?

|_____||_____||_____||_____||

Very easy Easy Neither easy nor difficult Difficult Very difficult

Not Applicable

11. How easy was it to complete the exercises on the tablet at home?

|_____||_____||_____||_____||

Very easy Easy Neither easy nor difficult Difficult Very difficult

12. I felt confident to use the tablet activities and games with my child at home:

|_____||_____||_____||_____||

Totally Agree Neither agree Disagree Totally
agree nor disagree disagree disagree

13. How satisfied are you with how your child's speech progressed during the therapy program?

|_____||_____||_____||_____||

Extremely Satisfied Neither satisfied Dissatisfied Extremely
Satisfied nor dissatisfied dissatisfied

14. In future, would you prefer to do your child's speech home practice:

On the tablet Using a different app

Using paper cards/worksheets A combination of the above

(14a) Please tell us why you answered this way:

15. What could we do to make the exercises more engaging for your child and the overall application more usable?

16. If tablet-based exercises were available to you in the future, how often would you want to use them with your child?

- | | |
|--|---|
| <input type="checkbox"/> Never | <input type="checkbox"/> 2 or 3 times a week |
| <input type="checkbox"/> Once a month | <input type="checkbox"/> 4 or more times a week |
| <input type="checkbox"/> 2 or 3 times a month | <input type="checkbox"/> Once a day |
| <input type="checkbox"/> 4 or more times a month | <input type="checkbox"/> 2 or 3 times a day |
| <input type="checkbox"/> Once a week | <input type="checkbox"/> 4 or more times a day |
| <input type="checkbox"/> Other (please specify): | |
-

17. Do you have any other comments or feedback for us?

Thanks for your feedback 😊



Apraxia World Questionnaire (Child Survey)

Child ID: _____

Date: _____ Are you a boy or a girl? ___BOY ___GIRL

How old are you? _____ What grade are you in? _____

Condition [ASR / Parent]: _____

Would you mind helping us make the game you played better for other children by answering some questions?

1. How enjoyable was the game?

1	2	3	4	5
_____	_____	_____	_____	_____
not very				very




2. How difficult was the game?

1	2	3	4	5
_____	_____	_____	_____	_____
very				not at all

3. With some more practice I could play the game by myself

1	2	3	4	5
_____	_____	_____	_____	_____
no		maybe		could already




4. How hard were you trying while playing the game?

1	2	3	4	5
				
not very				very

5. What did you like about playing the game? You might say something about speaking to control the game, the amount of time you had to say each word, or anything else you want to talk about.




6. What did you not like about playing the game? You might say something about speaking to control the game, the amount of time you had to say each word, or anything else you want to talk about.

7. What did you think about buying items for your character?

1	2	3	4	5
				
didn't like		it was ok		liked it a lot




Why?

8. Was it more fun doing speech exercises in the game than how you normally do them?

1	2	3	4	5
----- ----- ----- -----				
				
no		about the same		yes

Why?

9. Would you like to keep playing this game?

1	2	3	4	5
----- ----- ----- -----				
				
no		maybe		yes

Why?

10. Did the speech exercises make the game harder to play?

1	2	3	4	5
_____	_____	_____	_____	
yes		a bit		no

Why?

11. What would make this game more fun?

Game Usability Survey: CHILD

Do you have any other comments you'd like to make?

Ask the child How well do you think the game is recognising your speech / do you think the game is giving you the right scores for your speech (i.e. good job vs not quite or try again)

Thank you!



APPENDIX B

APRAXIA WORLD USER GUIDE

B.1. Overview

Apraxia World has 40 full levels, 8 bonus levels that consist only of coins and power-up collectables, and a training level in World One where the player cannot die. Players can buy additional characters (1,500 coins each), items of clothing (250 coins per item), weapons (300 – 6000 coins each), or power-ups (50 coins per use, 150 – 2,400 coins to increase duration). Full levels have between 7 and 9 regenerative stars (10-second delay between award of star and regeneration).

B.2. Installation

To properly run the game, you need an Android 6 or above device. First, copy the Apraxia World Images and Pronunciation Models folders (both will be provided) to the root folder of the device. Download the app from the Google Play Store to install.

B.3. Main screens

When you open the app for the first time, or after it has been force-closed, you will see the start screen (Figure 20). Once you press Play, select the username associated with the calibration profile you want and then press “Ok” (Figure 21). If you want to later select a different calibration profile, you need to force close the app so the start screen will appear again.



Figure 20 Apraxia World start screen only appears when the app starts anew, not after the application is paused.



Figure 21 Select the user profile associated with the calibration data you want to use.

Once you have selected a user profile, you will see the world selection screen (Figure 22). After selecting the desired world, the level selection screen will display (Figure 23). The shown levels are for World One, which includes a training level (marked by the T) where the player cannot die.



Figure 22 World selection screen. Bottom menu has buttons for the character store, costume store, world selection screen, weapon store, powerup store, and settings screen.



Figure 23 Level selection example. This is World One, which has a training level marked by the T.

B.4. Game settings

All exercise settings are located in a single settings page (indicated by the gear tab at the bottom of all menu screens). Any options changed on this screen are saved and will persist until changed again. The right half of Figure 24 shows the Exercise Parameter selection. The “Stars required” option selects how many stars are necessary to complete a level (increments/decrements one at a time). “Exercises per star” allows the SLP to determine how many utterances must be spoken before awarding a star (increments/decrements one at a time). The game has two evaluation options, ASR or keyboard. These two options are toggled by tapping the evaluation source. The “Coins per star” option determines how many bonus coins are awarded on exercise completion

(increments/decrements five at a time). The Word List settings are no longer used, since Apraxia World imports exercises from Apraxia World Recorder.

When the keyboard option is selected, during speech exercises, the game will wait to advance the prompt until it receives external evaluation. If a keyboard is not connected, the game will not advance past the speech exercise.

Some administrative options are hidden to the right in the settings screen and are accessed by dragging the screen to the left. These options are shown in Figure 25. The “Ignore playtime restrictions” toggle will keep the game from limiting game access when selected. The “Reset everything” button resets all game settings such that it acts like a new installation again. The “Reset progress” resets the progress the player has made in the game, while leaving the settings alone.



Figure 24 Settings page with the word list selected for each level (left) and exercise-specific parameters (right).



Figure 25 Administrative options hidden to the right in the settings page.

B.5. Gameplay

The game is controlled by overlaid buttons, shown in Figure 26. The joystick on the left controls motion, A activates the weapon, and B makes the character jump. The character can do a double jump, that is, press B, wait for the character to get into the air, and press B again to jump even higher. The heads-up display shows (clockwise from bottom left) collected coins, character health (yellow bar), points, stars, weapon selected, how much time is left before the character slows down, and if the exercises are done (“Say more words” or “Exercises done!”). If game timer runs out, the character will move at half speed until more time is earned by doing exercises. Figure 27 shows the level checkpoint, represented as a blue anchor. Once you touch the anchor, your character will restart here instead of the beginning if you die. Each level is won by going to the right to reach the finish line (Figure 29) once you have completed enough exercises.



Figure 26 Example level with heads-up display and overlaid controls.



Figure 27 The blue anchor represents the checkpoint. After crossing this point, the player will reset here if they die.

B.6. Exercise delivery

The game delivers exercises during gameplay. To simplify the explanations below, E is “Exercises per star,” S is “Stars per level,” and C is “Coins per star,” all as defined by the SLP in the settings page.

The game displays an exercise popup when the player attempts to collect a star and then the player must complete E prompts. An example exercise popup is displayed in Figure 28. Each correct utterance adds 10 seconds to the game timer, and each incorrect utterance adds 5 seconds to the game timer. When the child has completed E prompts (correct or incorrect), the popup window disappears, a star is awarded, and C coins are awarded. If the player attempts to complete the level before completing $E \times S$ prompts, then they see the text banner shown in Figure 29. Once the child completes their exercises and they cross the goal line, the level ends.

If a child says a word incorrectly three times in a row, the game will present a new prompt (a skip). If three words are skipped during an attempt to collect a star, the exercise will end without awarding the star or coins. Therefore, the exercise ending without reward will only happen if E is greater than nine. All prompts are randomly selected from the exercise list such that they do not repeat until all words have been prompted.

It is important to capture the entire utterance when doing the exercises (press start, speak, press stop). If the child needs help with the pronunciation, they can press the Pronunciation Example button to hear a sample pronunciation.



Figure 28 Exercise popup with an image and text prompt.



Figure 29 A prompt that the child should go collect more stars before finishing the level.

B.7. Keyboard Evaluation

When the evaluation source is set to keyboard, someone must evaluate the child's utterances with a Bluetooth keyboard. All evaluations are binary: if the utterance is correct, the evaluator should press "C"; if the utterance is incorrect, the evaluator should press "I". The game will only act on the human evaluation once both human and ASR evaluations have been captured. Both evaluations are saved to the game logs.

B.8. Logging

Apraxia World is set to upload that day's audio and game logs to the server any time the app is paused (home button pressed, screen turned off), as long as a username has been selected. This means that some days may have multiple uploads, but the most recent upload can be identified by the timestamp in its filename. This will all happen automatically while the tablet is connected to the internet, unless turned off in the settings.

APPENDIX C

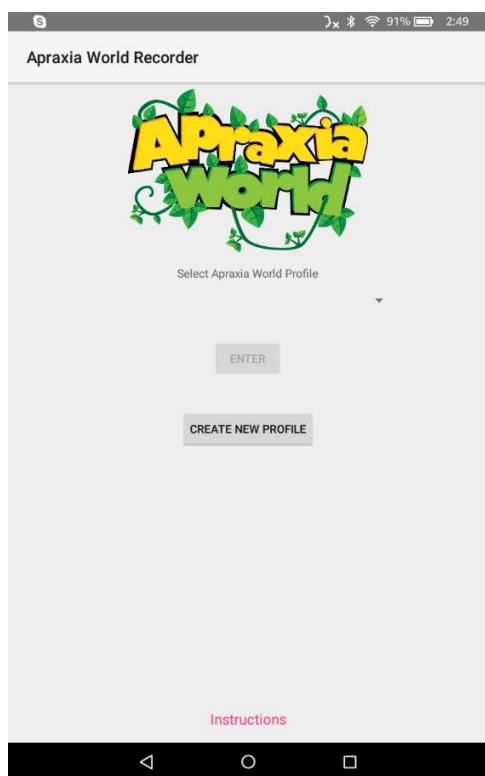
APRAXIA WORLD RECORDER USER GUIDE

C.1. Summary

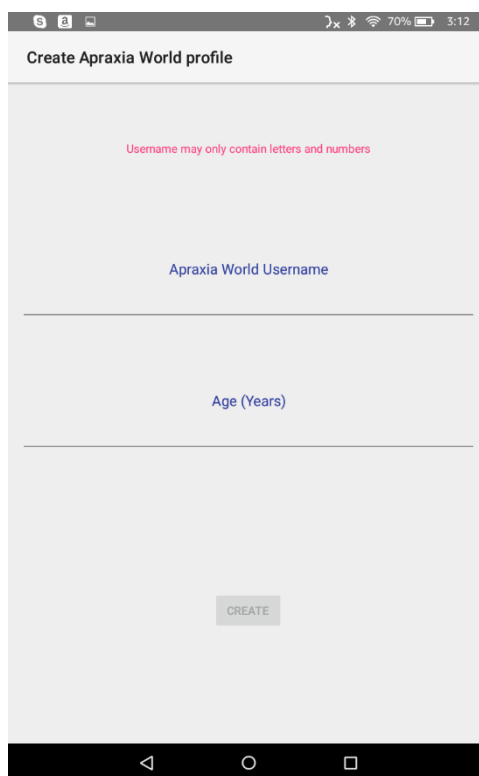
Apraxia World Recorder is how clinicians and caregivers configure which target words are included in Apraxia World and collect the appropriate calibration data. The app allows for different usernames, each of which can be configured with different therapy targets; these are the usernames that display when first opening the Apraxia World game. Apraxia World Recorder also contains a probe function, where the clinician can probe target words and collect recordings to track a child's progress.

C.2. Start screen

Before you can do anything in Apraxia World Recorder, you must select the username you want to work with. Pick the desired username from the dropdown list and press "Enter" (see Figure 30 a). If no username has been created, or you simply need a new one, select "Create new profile" and enter the desired username and child's age, and then select "Create" (Figure 30 b) Both selecting a username and creating a new one will take you to the main screen, where you can modify the therapy words or complete a probe.



(a)



(b)

Figure 30 Apraxia World Recorder start (a) and username creation (b) screens.

C.3. Calibration

Once you have created or selected a username, you'll see the main screen. The menu button (Figure 31 a) shows additional functions, such as probe and export. The word list is initially empty, but will fill in automatically as you type (Figure 31 b). Selecting a word takes you to the recording screen (Figure 31 c). For a word to be included in Apraxia World, you must record five correct and five incorrect pronunciation samples of the target. The order in which these are recorded does not matter. Once you've recorded the 10

utterances, press “Test ASR” to see the effect size between scores for correct utterances and incorrect utterances. In general, the system does well with words that show an effect size greater than one. After testing the ASR, you can select the “Include word in game” option, which will not appear until you’ve tested the ASR. Only words that are marked to be included in the game will export to Apraxia World or appear in probes; if a word should no longer be used, simply deselect “Include word in game.” For quick access of words marked for inclusion in Apraxia World, you can use the “Show Selected Words” option in the menu on the main screen (Figure 31 a).

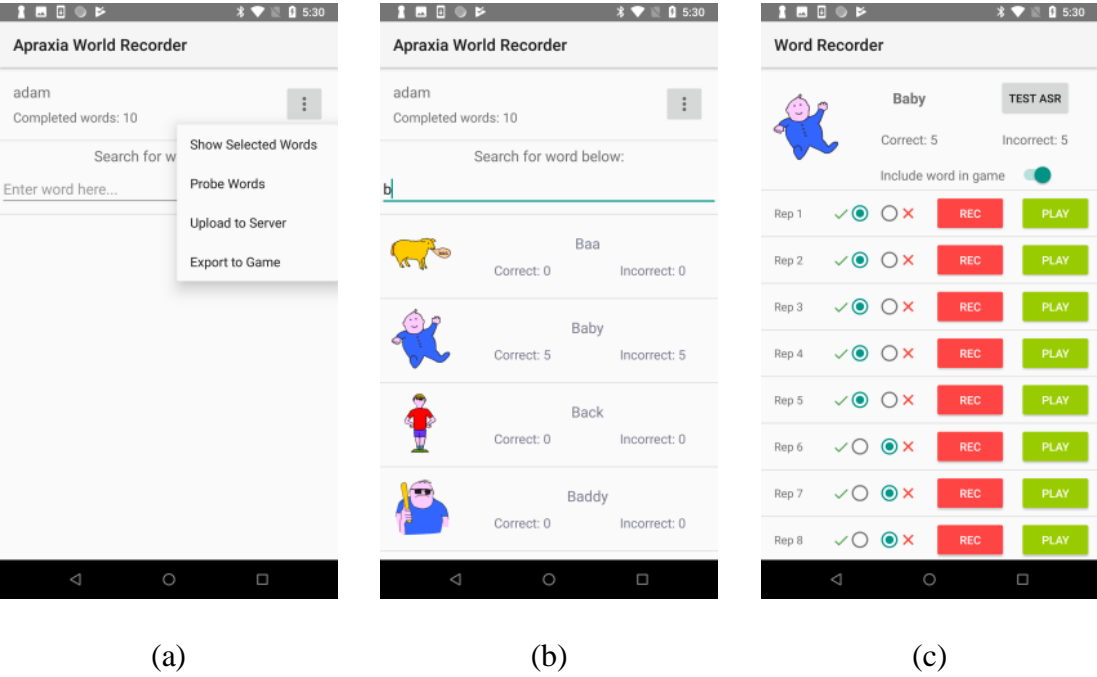


Figure 31 Apraxia World Recorder prompt calibration screens.

Once you have marked at least 10 words to be included in Apraxia World, you can select “Export to Game” from the menu on the main screen (Figure 31 a). This option pre-processes the recordings so that the ASR will run faster during gameplay and configures the necessary metadata. You will see an indicator showing that the game is processing the export and it will tell you once it has successfully completed exporting the words. At this point, you’re ready to play Apraxia World with the selected targets.

C.4. Probes

This functionality allows clinicians to record pronunciation probes for later analysis. Similarly to the audio export function, at least 10 words must be marked to be included in the game before you can access probe functionality. To start or view a probe, select “Probe Words” from the menu on the main screen (Figure 31 a). You will then see a list of past probes and the option to create a new probe (Figure 32 a); previous probes are named with the format DAY-MONTH-YEAR HOUR_MINUTE_SECOND. When you create a new probe, you’ll see the words marked to be included in the game (Figure 32 b). These words can be selected in any order, so you can match them to any external prompting (PowerPoint, booklet, etc.). Once you select a word, a popup will appear with a pictorial and text prompt, recording and playback functions, and a correct or incorrect label (Figure 32 c). When you press the record button to start the recording, the label will change to “Stop;” press the button again to stop the recording and then select the appropriate label (green check for correct, red x for incorrect). Once the word is recorded and labeled, press OK to dismiss the popup. After recording and labeling an utterance for each word, you can press “Upload Probe” to send the probe to our server. The files are

also stored locally. The ASR does not run during probes, so only clinician labels are recorded.

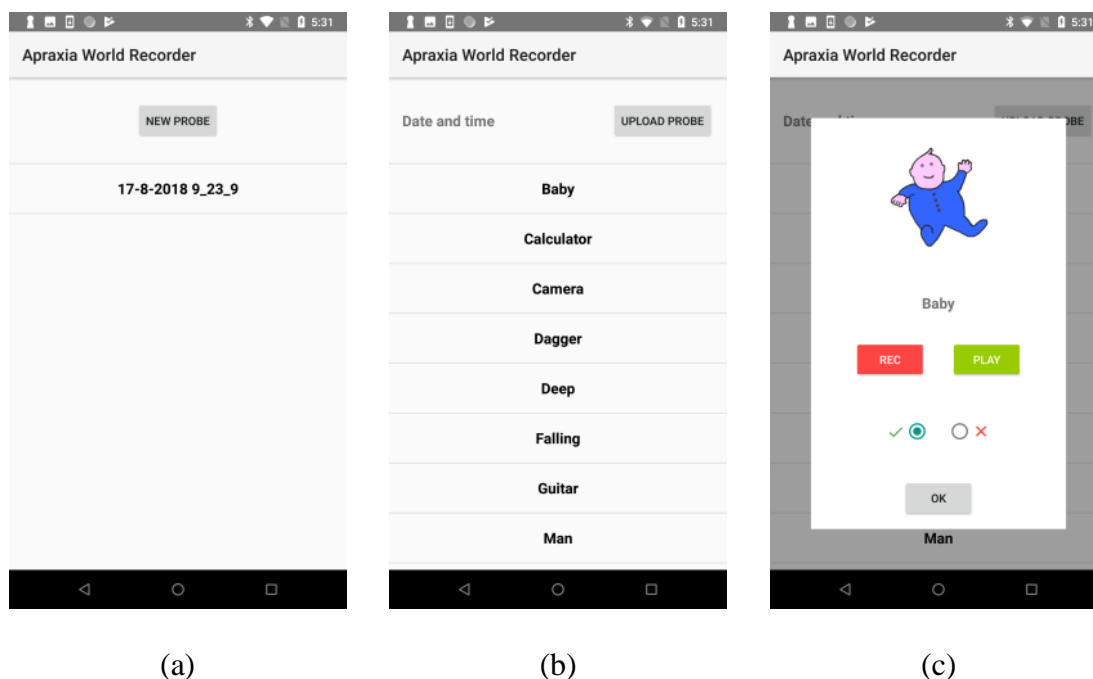


Figure 32 Apraxia World Recorder probe screens.

If you select an existing probe, you'll see the same screen as when you created a new prompt (Figure 32 b). This screen will contain the recordings and labels from that probe session so that you can review the probe. The probe labels can be changed here, so be careful not to change labels when viewing past probes.

APPENDIX D

CHILD SPEECH CORPORA

As a majority of speech recognition work focuses on adults, the amount of available adult speech data vastly outweighs the amount of child speech data. However, there are a handful of notable corpora that can be used within child speech recognition research.

D.1. Typically-developing speech

The Oregon Graduate Institute (OGI) Kids' Corpus [145]: This corpus contains both prompted and spontaneous American English speech collected from approximately 1,100 children ranging in school grade from Kindergarten all the way to 10th grade. The prompted speech contains 205 isolated words, 100 sentences, and 10 numeric sequences. The spontaneous speech contains open-ended responses to questions asked by the experimenter; each child recorded between eight to ten minutes of spontaneous speech, which was orthographically transcribed. In total, the corpus contains 101 hours of child speech (70 hours scripted).

The PF-STAR Corpus [146, 216]: This corpus contains prompted British English child speech collected from 158 children between the ages of 4 and 14. The speech prompts included 30 sentences, 40 isolated words, 20 “generic phrases,” and 20 digit triples. The corpus contains orthographic transcriptions and contains 7.5 hours of child speech.

The Boulder Learning MyST Corpus [217]: This recently-released corpus contains conversational speech collected from 1,371 third through fifth grade students. The conversations took place between the student and a virtual science tutor, and as such, the conversations focus on basic science topics including physics, chemistry, astronomy, and

biology. In total, the corpus contains over 393 hours of child speech and 197 hours have been transcribed at the word-level, with more transcriptions being added as a community effort.

D.2. Disordered speech from children

UltraSuite [218]: This corpus contains three collections of prompted Scottish English speech, one from typically-developing children and two from children with speech sound disorders. In addition to speech data, the corpus also contains ultrasound files showing the midsagittal view of child's tongue. Prompts consist of words, non-words used to elicit certain phonemes, sentences, phonemes produced at differing speeds, and non-speech (swallowing, coughs). The corpus contains a limited quantity of conversational speech from children with speech sound disorders, but not from typically-developing children. A small portion of the disordered speech has been annotated by a clinician to note the boundaries of words and phonemes of interest, but error tags are unavailable. Some of the typically-developing speech is transcribed, and the disordered speech has been aligned to the expected pronunciation. In total, the corpus contains 13 hours child speech (11 hours of disordered speech from children).

PhonBank Clinical Corpora [147]: TalkBank [219] is a project focused on sharing and studying spoken communication, and as such, contains databases of speech representing a variety of populations (e.g., people with dementia, second-language learners, students). Within the TalkBank system, PhonBank a collection of databases to facilitate research on child phonology. The clinical database contains speech from children with speech sound disorders, although not all corpora are from English-speakers and some

contain speech from a relatively limited number of speakers. Of specific interest are the corpora from Torrington Eaton and Bernstein Ratner [148], Cummings and Barlowe [149], and Preston et al. [150]. The Torrington Eaton corpus contains typically-developing and disordered speech from children completing picture naming tasks and non-word repetitions, in addition to spontaneous speech from a play session. Recordings were collected from 51 children between the ages of four and five years old. The Cummings corpus contains disordered speech from children during clinical probes with single-word utterances. Recordings were collected from 30 children between the ages of three and six years old. The Preston corpus contains disordered speech from children completing a picture naming task. Recordings were collected from 44 children between the ages of four and five. All corpora contain phonetic transcriptions of actual and expected productions, however, these transcripts are not time-aligned and recording quality varies between recording sessions; as such, these corpora were not in a usable state at the time of completing this dissertation.