Preliminary Results From a Longitudinal Study of a Tablet-Based Speech Therapy Game

Figure 1: Apraxia World is a 2D platformer speech therapy game

Adam Hair
Texas A&M University
College Station, TX, USA
adamhair@tamu.edu

Beena Ahmed
University of New South Wales
Sydney, Australia
Texas A&M University at Qatar
Doha, Qatar
beena.ahmed@unsw.edu.au

Constantina Markoulli
Penelope Monroe
Jacqueline McKechnie
Kirrie J. Ballard
University of Sydney
Sydney, Australia
kirrie.ballard@sydney.edu.au

Ricardo Gutierrez-Osuna
Texas A&M University
College Station, TX, USA
rgutier@cse.tamu.edu

CHI '20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6819-3/20/04.
http://dx.doi.org/10.1145/3334480.3382886

Abstract
We previously developed a tablet-based speech therapy game called Apraxia World to address barriers to treatment and increase child motivation during therapy. In this study, we examined pronunciation improvements, child engagement over time, and caregiver evaluation performance while using our game. We recruited ten children to play Apraxia World at home during two four-week treatment blocks, separated by a two-week break; nine of ten have completed the protocol at time of writing. In the treatment blocks, children's utterances were evaluated either by caregivers or an automated pronunciation framework. Preliminary analysis suggests that children made significant therapy gains with Apraxia World, even though caregivers evaluated pronunciation leniently. We also collected a corpus of child speech for offline examination. We will conduct additional analysis once all participants complete the protocol.

Author Keywords
Assistive technology; Computer-aided pronunciation training (CAPT); Speech Sound Disorders

CCS Concepts
•Social and professional topics → **Children; People with disabilities; •Applied computing** → *Consumer health;*

Introduction

Speech sound disorders (SSDs) are a collection of difficulties producing specific speech sounds correctly, and are typically diagnosed in childhood [2]. Although SSDs can impair communication skills development [4], children often improve speech quality and reduce symptoms by working closely with speech language pathologists (SLPs) [17]. Given that speech therapy practice must be frequent and high-intensity [12], clinic sessions should be supplemented with considerable home practice, which can be tedious. Primary caregivers typically administer home practice, but busy schedules decrease practice frequency [14]. As such, there is a need for speech therapy that allows children to self-direct their sessions with minimal caregiver involvement and makes the practice more engaging.

To address boring and infrequent home practice, our team of computer scientists and SLPs collaboratively developed Apraxia World, a mobile speech therapy game that provides pronunciation feedback [5]. In 2018, we evaluated our design with the help of 14 children with SSDs and seven children reported to be typically developing. We found that the game helped to make the speech therapy more enjoyable for the children and they were eager to keep playing. However, the pilot study consisted of a single session per participant and did not examine therapeutic benefits, which makes further longitudinal evaluations necessary.

Children often enjoy using digital therapy interventions in short-term tests, and sometimes even play beyond the required time [5, 6]; however, it remains unclear how these interventions hold children's attention over a longer period. As the perceived utility of a tool may not always match actual usage [16], longitudinal evaluations combined with iterative development are needed to better align tool functionality with stakeholder expectations.

As a follow-up to our pilot study, we are conducting a multi-month longitudinal evaluation with 10 children with SSDs. This study has three goals: examine home use and therapeutic benefit of Apraxia World, and build a corpus of in situ disordered speech from children for future speech recognition investigations. We approached this study with three primary research questions:

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

These assessments are necessary to ensure that this digital game-based therapy method constitutes a meaningful approach for home practice.

In this extended abstract, we present initial results on pronunciation improvements and pronunciation evaluator performance (RQ1 and RQ2). Preliminary analysis indicates that children are making therapy gains comparable to traditional therapy and that caregivers are evaluating child pronunciations leniently. Additionally, we have curated a collection of over 25,000 individual utterances that will be used to further investigate automatic speech recognition and pronunciation evaluation methods for child speech. Below, we presenting Apraxia World, the experimental protocol, and initial results.

Apraxia World

Apraxia World is a brightly-themed 2D platformer game (see Figure 1) built from an existing game for the Unity Game Engine. It includes 40 levels, multiple characters, and an in-game store. These features align with recommendations that speech therapy systems should include

(a)

(b)

(c)

Figure 2: (a) A level from the jungle world (b) A level from the desert world (c) Speech exercise popup with both pictorial and text cue

more game-like elements [1]. Figure 2 (a) and (b) show levels from two different worlds (jungle and desert). Apraxia World was developed collaboratively with SLPs and updated based on stakeholder feedback after the pilot study.

There are different popular strategies for controlling speech therapy games: producing sustained sounds [10, 9, 11], speaking targets words corresponding with actions [16, 3], or controlling specific aspects of speech [6]. These strategies have the benefit of providing implicit feedback; however, they can be problematic if the player struggles to produce the target sounds. Additionally, it can be difficult to navigate a character through a two-dimensional world using only speech to control their movements. As such, players control Apraxia World with a traditional joystick and button combination, and speech exercises are used as a secondary input to collect in-game assets, specifically, yellow stars spread throughout the levels; see Figure 2 (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 2 (c)). Within the exercises, pronunciation attempts are evaluated and the game displays appropriate feedback (this process described below). Once the player completes the required number of attempts (either correct or incorrect), the popup disappears and the star is added to their inventory. Collecting these stars is mandatory, as the game requires the player to collect a certain number of stars before they can complete a level.

To limit therapy exposure and avoid game fatigue, children are only allowed to play one level per day. Apraxia World includes a timer that specifies how much time the child has left before their character slows down. This slight penalty encourages children to collect additional time. The player earns more time by doing speech exercises – saying a word correctly rewards the child with 10 seconds and saying a word incorrectly rewards the child with 5 seconds. In this way, the child is rewarded for all pronunciation attempts, but correct attempts are more strongly rewarded to motivate children to maintain practice effort. Once the child completes the required number of speech exercises, the game does not allow them to do any more. At this point, the child can play until they finish the level or lose, whichever comes first. The game then locks until the next day.

The current version of Apraxia World provides pronunciation feedback based on automatic pronunciation evaluation or human evaluator input via a Bluetooth keyboard. Automatic pronunciation evaluation is carried out using a template-matching (TM) framework, a well-established speech recognition technique based on dynamic time warping [19]. This method compares a test recording against target "template" recordings to determine which template 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), which is critical for child speech applications. TM has previously been used in child speech therapy applications [8, 21] and to evaluate pronunciation quality [20]. 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 calibration templates are collected by an SLP a priori using a dedicated companion application.

Experiment

We recruited 10 participants (9 males, 5–12 years old), nine of whom have completed the protocol at the time of writing. The study protocol consisted of five phases: setup, two treatment blocks playing Apraxia World, a between-treatments break, and a post-treatments break. Setup in-

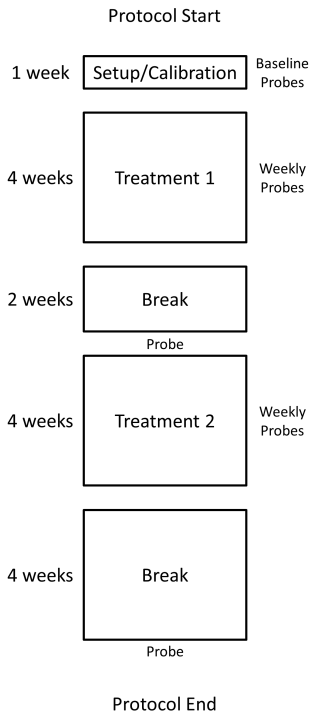

Figure 3: Experimental protocol with two treatment blocks. Pronunciation is probed before treatment, weekly during treatment, and after a one-month break following treatment.

volved selecting target words based on each child's therapy needs, calibrating the speech recognition, and familiarizing the child and caregiver with the game. Children practiced over two counterbalanced treatment phases so we can examine the effects of utterance evaluation source (caregiver versus computer). The four-week break at the end of the study is necessary to test for the longevity of any pronunciation improvements after practice ends. At the end of each treatment block, a random subset of utterances was selected for pronunciation evaluation by an SLP. The protocol timeline is illustrated in Figure 3.

Within each of the two treatment blocks, children practiced words that target different speech errors, as selected by the SLPs. Pronunciation ability was probed before each treatment block, weekly during the treatment blocks, and once more following the post-treatments break, so that we can examine pronunciation improvements during and beyond treatment. Pronunciation probes contained both practiced and non-practiced words to measure carryover effects on targeted phonemes in new contexts. Subjective questionnaires were administered twice during each treatment block, and again following treatment, to track and compare enjoyment during both treatment conditions. The children played Apraxia World on Samsung Tab A 10.1 tablets and used noise-cancelling microphone headsets to record their speech during exercises.

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 TM treatment block, they received automatic pronunciation assessment. Regardless of evaluation source, pronunciation feedback was delivered through the game. These conditions allow us to investigate if the child cares about a perceived level of autonomy, i.e., do they prefer being able to do the speech exercises on their own without a caregiver? We also scored caregiver pronunciation evaluation accuracy by measuring agreement between caregiver and pathologist evaluations.

Initial Results and Observations

Pronunciation performance was calculated as the percentage of utterances where the target speech sound was correctly produced. Preliminary analysis of the first six participants indicates that children made pronunciation improvements comparable to those reported for standard face-to-face therapy. On average, participants' production accuracy increased 42 percentage points ($95\%\ CI \pm 18.3$) in the treatment phase using TM evaluations and 45 percentage points ($95\%\ CI \pm 12.1$) in the phase with caregiver evaluations. These improvements are similar to the 40-point improvement found in a study of a more traditional speech therapy program with 13 children [15].

We collected ground truth utterance labels by having an SLP phonetically transcribe a representative subset of the utterances collected so far. Using these labels, we computed the true positive rate (TPR) and true negative rate (TNR) for the caregivers and TM evaluations pooled across the nine participants who have completed the protocol. TPR is computed as

$$TPR = \frac{true\ positives}{true\ positives + false\ negatives} \quad (1)$$

and TNR is computed as

$$TNR = \frac{true\ negatives}{true\ negatives + false\ positives}. \quad (2)$$

These results are displayed in Table 1. The TPR for caregivers is much higher than the TNR, indicating that they

	TPR	TNR
Caregiver	88.4%	26.4%
TM	29.2%	63.8%

Table 1: True positive and true negative rates for caregiver and TM evaluations

may have been overly lenient in their evaluations. The TM results are the converse; the system appears to be overly punitive, with a higher TNR and a low TPR. The TM framework also may have struggled to evaluate the utterances once the children were speaking more naturally and relaxed at home, since the calibration utterances gathered under SLP supervision were relatively stoic and consistently-produced. Further analysis on the utterance evaluations will be conducted and presented following the conclusion of our study.

In total, we have gathered 25,096 utterances from the 9 finished participants, and have manually labeled them into 5 categories: clipped (part of the recording cut off), containing background noise, unusable (speaker unintelligible), containing significant microphone noise, or good (usable for offline ASR analysis – the good label says nothing about pronunciation quality). Roughly 47% of the utterances (~12,000) were labeled as good.

We observed that when some of our participants became discouraged, they spoke too quietly for the TM to meaningfully evaluate the speech. Some children also spoke loudly and quickly when excited (either positively or negatively), which also increases processing difficulty. As such, future systems would benefit from monitoring speaking volume and speaking rate to recommend a correction. These reminders would help children produce utterances of better quality for automated speech processing, which would result in them receiving more meaningful feedback on pronunciations.

Examining the collected child speech reveals another direction for improving the system. We currently use a push-to-start/push-to-stop mechanism to record utterances, but given the number of incompletely-captured recordings, we question if this is the best interaction method. Caregivers were reminded that the children should wait until they finish speaking before stopping the recording, but the problem persisted, indicating that this mechanism is poorly suited for use with children. One potential solution is to use an endpoint detection framework to automatically determine when to stop the recording. Since incomplete recordings oftentimes result in inaccurate automated feedback, it is essential to empower children to capture the entirety of their utterance.

Future Work

In this article, we presented preliminary results while reserving in-depth discussion of game engagement for future publication. Further and more in-depth analysis will be conducted once the final participant ($n = 10$) has completed the protocol and all data is collated. Specifically, we will examine engagement according to the time they spend navigating levels, how often their character dies, when they make store purchases, if they use power-ups, etc. Motivation will be evaluated based on responses to the questionnaires. This information allows us to analyze how children spend their time in the game, which is valuable given that Apraxia World should remain enjoyable as long as possible during home practice to extend the life of the intervention. Additionally, we will present therapy gains for all children and examine if the order of evaluator (caregiver and automated system) had an effect on improvement magnitude.

We plan use the collected child speech corpus to examine child ASR improvements like neural network transfer learning [13] or signal modification [22] for future therapy applications. We expect this line of work to generate separate publications for speech-specific venues.

Conclusion

Speech sound disorders are problematic for children, but can be addressed by clinical speech therapy. However, speech therapy is often less frequent than it needs to be. Home practice is a common complement to clinic sessions, but it depends on caregiver availability and can be tedious. To give children more independence and make therapy more enjoyable, we developed Apraxia World.

In this study, we evaluated the home use and clinical benefit of Apraxia World over a multi-month period, while also collecting a child speech corpus. Encouragingly, our results so far align with previous studies that show computerized and tablet-based speech therapy is as effective as traditional speech therapy [18, 7]. We also found that caregivers may be overly lenient and the TM framework as implemented may be too punitive. In spite of any bias in the utterance evaluations, children are still making meaningful therapy progress. This suggests that word repetitions with imperfect feedback may help children improve pronunciation.

Based on our findings so far, we recommend that future HCI researchers include recording quality safeguards, such as volume and speaking rate monitors. Researchers should also be careful to select child-appropriate recording control methods, as we found that the push-to-start/push-to-stop paradigm was difficult for children to grasp. We believe that including speech as a secondary input may help researchers design games that are not limited by speech quality. Finally, we suggest using a professional game engine like the Unity Game Engine and either purchasing an existing game or working with an experienced game designer to quickly build a game prototype that matches the quality of games children play daily, which can be important for engagement. We expect that results from this study will inform future work on speech therapy games and their role in clinical practice, which benefits developers, caregivers, and especially children.

Acknowledgments

This work was made possible by NPRP Grant # [8-293-2-124] from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors

REFERENCES

[1] Beena Ahmed, Penelope Monroe, Adam Hair, Chek Tien Tan, Ricardo Gutierrez-Osuna, and Kirrie J. Ballard. 2018. Speech-Driven Mobile Games for Speech Therapy: User Experiences and Feasibility. *International journal of speech-language pathology* 5, 20 (2018), 644–658. DOI: http://dx.doi.org/10.1080/17549507.2018.1513562

[2] American Speech-Language-Hearing Association. 2007. Speech Sound Disorders. (2007). https://www.asha.org/public/speech/disorders/SpeechSoundDisorders/.

[3] Jared Scott Duval, Elena Márquez Segura, and Sri Kurniawan. 2018. SpokeIt: A Co-Created Speech Therapy Experience. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 1–4. DOI: http://dx.doi.org/10.1145/3170427.3186494

[4] Karen Forrest. 2003. Diagnostic Criteria of Developmental Apraxia of Speech Used by Clinical Speech-Language Pathologists. *American Journal of Speech-Language Pathology* 12, 3 (2003), 376–380. DOI: http://dx.doi.org/10.1044/1058-0360(2003/083)

[5] Adam Hair, Penelope Monroe, Beena Ahmed, Kirrie J. Ballard, and Ricardo Gutierrez-Osuna. 2018. Apraxia World: A Speech Therapy Game for Children with Speech Sound Disorders. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*. ACM, 119–131. DOI: http://dx.doi.org/10.1145/3202185.3202733

[6] Mohammed E Hoque, Joseph K Lane, Rana El Kaliouby, Matthew Goodwin, and Rosalind W Picard. 2009. Exploring Speech Therapy Games with Children on the Autism Spectrum. In *Proc. Interspeech 2009*. ISCA, 1455–1458.

[7] Luis MT Jesus, Joana Martinez, Joaquim Santos, Andreia Hall, and Victoria Joffe. 2019. Comparing Traditional and Tablet-Based Intervention for Children With Speech Sound Disorders: A Randomized Controlled Trial. *Journal of Speech, Language, and Hearing Research* 62, 11 (2019), 4045–4061. DOI: http://dx.doi.org/10.1044/2019_JSLHR-S-18-0301

[8] Diane Kewley-Port, C Watson, Daniel Maki, and Daniel Reed. 1987. 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)*, Vol. 12. IEEE, 372–375. DOI: http://dx.doi.org/10.1109/ICASSP.1987.1169661

[9] Tian Lan, Sandesh Aryal, Beena Ahmed, Kirrie J. Ballard, and Ricardo Gutierrez-Osuna. 2014. 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*. ACM, 429–430.

[10] Marta Lopes, João Magalhães, and Sofia Cavaco. 2016. A voice-controlled serious game for the sustained vowel exercise. In *Proc. of the 13th International Conference on Advances in Computer Entertainment Technology*. ACM, 32. DOI: http://dx.doi.org/10.1145/3001773.3001807

[11] Vanessa Lopes, João Magalhães, and Sofia Cavaco. 2019. Sustained Vowel Game: a computer therapy game for children with dysphonia. In *Proc. Interspeech 2019*. ISCA, 26–30. DOI:http://dx.doi.org/10.21437/Interspeech.2019-3017

[12] Edwin Maas, CE Gildersleeve-Neumann, Kathy J Jakielski, and Ruth Stoeckel. 2014. Motor-Based Intervention Protocols in Treatment of Childhood Apraxia of Speech (CAS). *Current Developmental Disorders Reports* 1, 3 (2014), 197–206. DOI: http://dx.doi.org/10.1007/s40474-014-0016-4

[13] Marco Matassoni, Roberto Gretter, Daniele Falavigna, and Diego Giuliani. 2018. Non-Native Children Speech Recognition Through Transfer Learning. In *Proc. of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6229–6233. DOI: http://dx.doi.org/10.1109/ICASSP.2018.8462059

[14] Lindy McAllister, Jane McCormack, Sharynne McLeod, and Linda J Harrison. 2011. Expectations and experiences of accessing and participating in services for childhood speech impairment. *International Journal of Speech-Language Pathology* 13, 3 (2011), 251–267. DOI: http://dx.doi.org/10.3109/17549507.2011.535565

[15] Elizabeth Murray, Patricia McCabe, and Kirrie J. Ballard. 2015. 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 58, 3 (2015), 669–686. DOI: http://dx.doi.org/10.1044/2015_JSLHR-S-13-0179

[16] Amal Nanavati, M Bernardine Dias, and Aaron Steinfeld. 2018. 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*. ACM, 318. DOI: http://dx.doi.org/10.1145/3173574.3173892

[17] ASHA Adhoc Commitee on CAS. 2007. Childhood apraxia of speech. (2007). https://www.asha.org/policy/TR2007-00278/.

[18] Rebecca Palmer, Pam Enderby, and Mark Hawley. 2007. Addressing the needs of speakers with longstanding dysarthria: computerized and traditional therapy compared. *International Journal of Language & Communication Disorders* 42, S1 (2007), 61–79. DOI: http://dx.doi.org/10.1080/13682820601173296

[19] Douglas A. Reynolds. 2002. An Overview of Automatic Speaker Recognition Technology. In *Proc. of the 2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4072–4075. DOI: http://dx.doi.org/10.1109/ICASSP.2002.5745552

[20] Charles S Watson, Daniel J Reed, Diane Kewley-Port, and Daniel Maki. 1989. The Indiana Speech Training Aid (ISTRA) I: Comparisons between human and computer-based evaluation of speech quality. *Journal of Speech, Language, and Hearing Research* 32, 2 (1989), 245–251. DOI: http://dx.doi.org/10.1044/jshr.3202.245

[21] Gary Yeung, Amber Afshan, Kaan Ege Ozgun, Canton Kaewtip, Steven M Lulich, and Abeer 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*. ISCA, 161–166. DOI: http://dx.doi.org/10.21437/SLaTE.2017-28

[22] Gary Yeung and Abeer Alwan. 2019. A Frequency Normalization Technique for Kindergarten Speech Recognition Inspired by the Role of fo in Vowel Perception. In *Proc. Interspeech 2019*. ISCA, 6–10. DOI:http: //dx.doi.org/10.21437/Interspeech.2019-1847