A Longitudinal Evaluation of Tablet-Based Child Speech Therapy with Apraxia World

ADAM HAIR, Texas A&M University, USA
KIRRIE J. BALLARD, CONSTANTINA MARKOULLI, PENELOPE MONROE, and JACQUELINE MCKECHNIE, The University of Sydney, Australia
BEENA AHMED, University of New South Wales, Australia
RICARDO GUTIERREZ-OSUNA, Texas A&M University, United States

3

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 4-week treatment blocks separated by a 2-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 goodness of pronunciation scoring. Our results suggest that this type of speech therapy game is a valid complement to traditional home practice.

CCS Concepts: • **Social and professional topics** → **Children**; **People with disabilities**; • **Applied computing** → *Consumer health*; • **Human-centered computing** → Tablet computers; **Accessibility systems and tools**; • **Computing methodologies** → *Speech recognition*;

Additional Key Words and Phrases: Games for health, serious games, computer-aided pronunciation training (CAPT), speech sound disorders (SSDs), childhood apraxia of speech (CAS)

This work was made possible by NPRP Grant no. [8-293-2-124] from the Qatar National Research Fund (a member of Qatar Foundation).

Authors' addresses: A. Hair, Department of Computer Science and Engineering, Texas A&M University, 506 H.R. Bright Building, College Station, Texas 77840; email: adamhair@tamu.edu; K. J. Ballard, C. Markoulli, P. Monroe, and J. McKechnie, Sydney School of Health Sciences, The University of Sydney, S157, C42 – Cumberland Campus, Sydney, NSW 2141, Australia; email: kirrie.ballard@sydney.edu.au; B. Ahmed, School of Electrical Engineering and Telecommunications, University of New South Wales, 444 Electrical Engineering Building, Sydney, NSW 2052, Australia; email: beena.ahmed@unsw.edu.au; R. Gutierrez-Osuna, Department of Computer Science and Engineering, Texas A&M University, 506 H.R. Bright Building, College Station, Texas 77840; email: rgutier@cse.tamu.edu.

1936-7228/2021/03-ART3 $15.00
https://doi.org/10.1145/3433607

ACM Reference format:
Adam Hair, Kirrie J. Ballard, Constantina Markoulli, Penelope Monroe, Jacqueline McKechnie, Beena Ahmed, and Ricardo Gutierrez-Osuna. 2021. A Longitudinal Evaluation of Tablet-Based Child Speech Therapy with Apraxia World. *ACM Trans. Access. Comput.* 14, 1, Article 3 (March 2021), 26 pages.
https://doi.org/10.1145/3433607

1 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 [American Speech-Language-Hearing Association 2007]. Children with SSDs struggle with phonological representation, phonological awareness, and print awareness, which can lead to difficulties learning to read or reading disabilities [Anthony et al. 2011], and negatively impact communication skills development [Forrest 2003]. Fortunately, children with SSDs often reduce symptoms and improve speech skills by working closely with speech-language pathologists (SLPs) to undergo speech therapy [ASHA Adhoc Commitee on CAS 2007]. For speech therapy to be effective, treatments must be "frequent, high-intensity, individualized, and naturalistic" [Maas et al. 2014] so that children can practice new habits and skills [Thomas et al. 2014]. However, scheduling appointments with SLPs can be logistically difficult [Ruggero et al. 2012; Theodoros 2008; Theodoros et al. 2008], and up to 70% of SLPs have waiting lists [McLeod and Baker 2014], 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 [McAllister et al. 2011]. 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 [Zajc et al. 2018], and have been shown to increase participant engagement and persistence [Gacňik et al. 2018; Parnandi et al. 2013]. Most importantly, research has shown that computerized and tablet-based speech therapy interventions can be as effective as traditional interventions [Ballard et al. 2019; Constantinescu et al. 2010; Jesus et al. 2019; Origlia et al. 2018; Palmer et al. 2007; Shriberg et al. 1990; Wren and Roulstone 2008], although not all digital applications outperform traditional methods [Werfel et al. 2019] or produce clinically significant results [McLeod et al. 2017]. A number of game-like applications for speech therapy have been commercially developed and are available for purchase [Furlong et al. 2017] (e.g., Apraxia Farm [Smarty Ears Apps 2017], Articulation Station [Little Bee Speech 2018], ArtikPix [Expressive Solutions 2018], Tiga Talk [Tactica Interactive 2011]). Children often enjoy using digital therapy interventions in short-term tests, and sometimes even play beyond the required time [Hair et al. 2018; Hoque et al. 2009]. 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 [Ahmed et al. 2018; Rubin et al. 2016]. 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 [Ahmed et al. 2018; Nanavati et al. 2018; Navarro-Newball et al. 2014], 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 [ASHA Adhoc Commitee on CAS 2007]. 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 [Korte 2020] 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 [Hair et al. 2018], 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 4-week treatment phases with a 2-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 presented as late-breaking work at the 2020 ACM CHI Conference on Human Factors in Computing Systems [Hair et al. 2020].

2 BACKGROUND AND RELATED WORK

2.1 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 [Koch 2018]. 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 [Koch 2018]. 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 [Maas et al. 2008]. 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 [Kewley-Port et al. 1991; Watson et al. 1989]. 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 [Dalby and Kewley-Port 1999]. 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 [Engwall et al. 2006]. 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 [Saz et al. 2009b] with three distinct components: PreLingua (basic phonation skills), Vocaliza [Vaquero et al. 2006] (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 [Parnandi et al. 2015; Shahin et al. 2015] 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 [Parnandi et al. 2013] 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. [2014] 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. [2019] 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. [2014] 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. [2017] 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. [2017, 2018] 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. [2018] 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.

2.2 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 [McKechnie et al. 2018]. The standard method for detecting mispronunciations is the goodness of pronunciation (GOP) proposed by Witt and Young [2000]. 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 [Dudy et al. 2015, 2018]. 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. [2009a] 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. [2014] 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 [Shahin et al. 2018], the authors developed a mispronunciation detection approach using one-class Support Vector Machines (SVMs). 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. [2015] 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 [McKechnie et al. 2018], 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 et al. [1997] 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 [Shobaki et al. 2000] and PF-STAR [Batliner et al. 2005] only contain typically developing speech, the PhonBank [Rose and Macwhinney 2014] collection contains corpora of disordered speech from children [Cummings and Barlow 2011; Preston et al. 2013; Torrington Eaton and Ratner 2016], but without ready-to-use recording annotations, and the recently released BioVisualSpeech corpus only contains European Portuguese speech [Grilo et al. 2020]. 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 [Reynolds 2002] 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 [Dalby and Kewley-Port 1999]. Template matching has also been successfully incorporated into child speech therapy systems as a pronunciation evaluator [Kewley-Port et al. 1987; Watson et al. 1989; Yeung et al. 2017]. Template matching evaluations have been shown to correlate with human evaluations when using high-quality productions from the speaker as pronunciation templates [Kewley-Port et al. 1987]. 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.

3 APRAXIA WORLD

3.1 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 while working 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 [Ahmed et al. 2018]. Figure 1(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 [Hair et al. 2018]. Since these initial tests, we modified the game as follows: we count all utterance attempts toward 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;

(a) (b) (c)

Fig. 1. (a) A level from the jungle world. (b) A level from the desert world. (c) Speech exercise popup with both pictorial and text cues.

and we added automatic speech processing (technical details in Section 3.3). The game mechanics are described below.

There are a handful of popular strategies for controlling speech therapy games: producing sustained sounds [Lan et al. 2014; Lopes et al. 2016, 2019], speaking target words corresponding to actions [Duval et al. 2018; Nanavati et al. 2018], or controlling specific aspects of speech [Hoque et al. 2009]. 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 1(a).

When the player attempts to collect a star by touching it with their character, the game pauses and a themed speech exercise popup appears; see Figure 1(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 2. The power-ups last only briefly and provide the player a

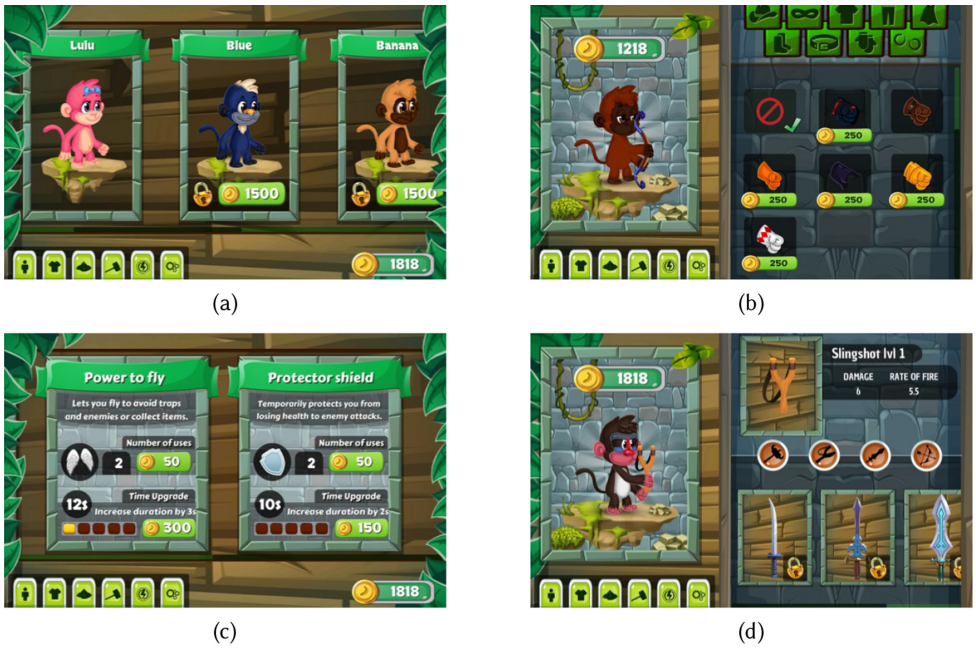

Fig. 2. (a) Various characters available for purchase. (b) Costume items to dress up the character. (c) Power-ups to give the character "superpowers." (d) Weapons with different attack behaviors.

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 [Tükel et al. 2015]. Other groups going through speech therapy may also experience difficulties with specific movements (e.g., children with Autism Spectrum Disorder [Staples and Reid 2010]). Although not implemented in this study, Apraxia World 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.

3.2 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

(a) (b) (c)

Fig. 3. Pictorial prompts for (a) pumpkin, (b) unicorn, and (c) banjo.

the learner of the correct response, so that they can judge the correctness of their response themselves [Shute 2008]. KR has been shown to help people using digital speech therapy systems make improvements comparable to those from traditional speech therapy [Bakker et al. 2019], 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 [Google 2018].

The speech exercises in Apraxia World are based on a Principles of Motor Learning approach [Maas et al. 2008; Schmidt and Lee 2005], 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 stimulable if the child could accurately imitate it multiple times and produce it without a model on at least five attempts within a 30-minute session. The SLP then selected one or two stimulable speech behaviors to address during treatment. Selecting stimulable 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 5 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 3 for examples of prompt images.

3.3 Pronunciation Feedback

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) [Reynolds 2002]. 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

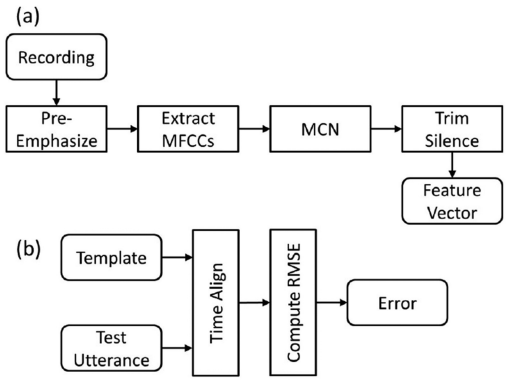

Fig. 4. (a) Spectral information is extracted from an utterance, mean cepstral normalized (MCN), and trimmed. (b) Template and test utterances are aligned and scored based on RMSE.

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 4(a). Given a recorded utterance (16 kHz), the audio signal is pre-emphasized before 13 Mel-frequency cepstral coefficients (MFCC) are extracted from 32 ms frames with 8 ms overlap, which are then normalized with mean cepstral normalization (MCN) [Furui 1981]. Leading and trailing silence segments are removed using an energy threshold to form the final feature vector.

The TM process is shown in Figure 4(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

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

where $||\cdot||_2$ is the L2 norm (Euclidean distance) and $dtw(t,u)$ time-aligns the frames in t to u. 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:

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

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

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

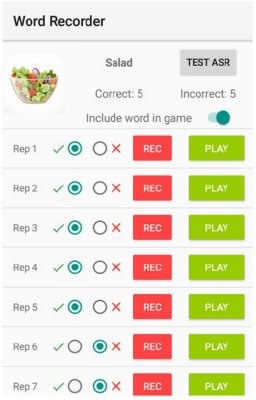

Fig. 5. Word recording interface in AWR. Recordings are labeled as correctly (green check) or incorrectly (red x) pronounced.

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

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

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

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

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 [Fager 2017]. 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 5 shows the recording interface for a target word in AWR.

4 EXPERIMENTAL DESIGN

4.1 Participants

We recruited 11 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

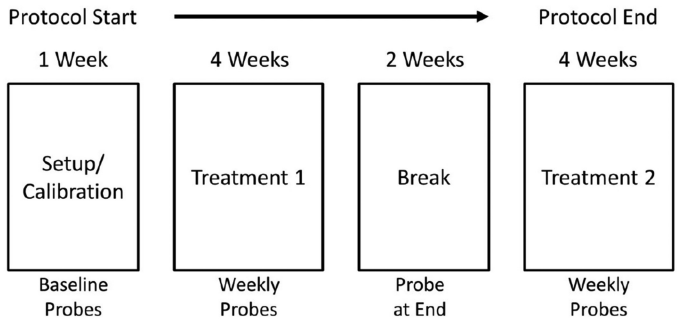

Fig. 6. Experimental protocol with two treatment blocks. Pronunciation is probed before treatment and weekly during treatment.

anomalies. One participant (male) unenrolled from the study due to schedule conflicts, so his data were not included in this analysis. The remaining 10 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.

4.2 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 5), 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 6. During the treatment blocks, children played Apraxia World as long as needed to complete their speech exercises, 4 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.

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

Table 1. Words Selected to Address Speaker-Specific Speech Difficulties

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

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 stimulable 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.

5 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.

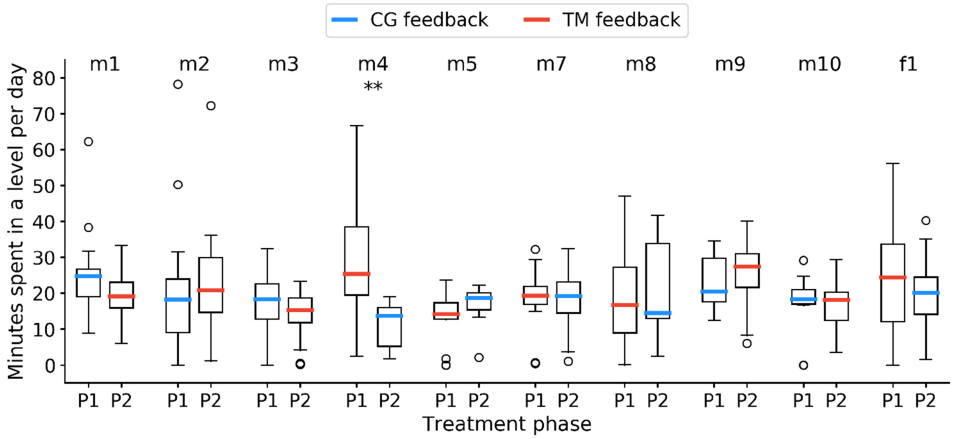

Fig. 7. Minutes spent within a level per day for treatment phases one (P1) and two (P2) (** indicates $p < 0.05$, two-sample t-test).

Table 2. Maximum Progress in the Game for Each Player

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

5.1 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 7. When comparing the two treatment phases, for all participants but one,[1] there was no significant difference in the amount of time spent in a level between the TM feedback phase and the 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.

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).

[1]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.

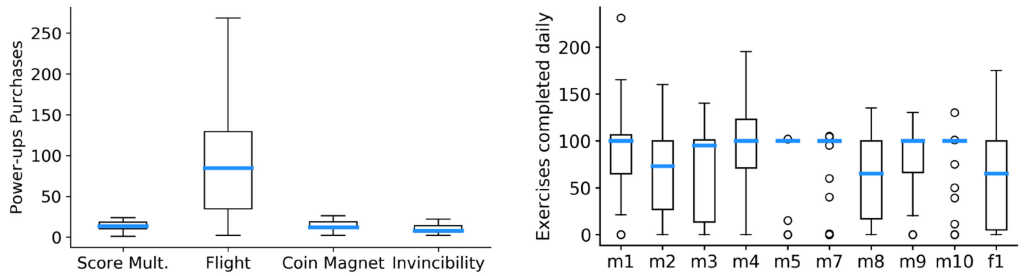

Fig. 8. (a) Power-up purchases across all participants. (b) Exercises completed per day.

Table 3. In-Game Purchases Made by Players During the Study

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

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 8(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 power-ups 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.

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*

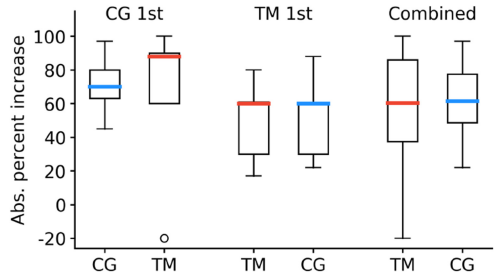

Fig. 9. Absolute increase in pronunciation scores at the beginning and end of each treatment phase for caregivers (CG) and template matching (TM).

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.

5.2 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 8(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 8(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 9. Children experienced an average absolute improvement of 56.6% ($\sigma = 35.7$) when receiving TM feedback and 61.5% ($\sigma = 22.8$) when receiving caregiver feedback; the difference between feedback methods was 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).

Some of the 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

Table 4. Recorded Utterances Gathered During Gameplay

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%)

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.

5.3 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 [Loukina et al. 2019].

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 mouths 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.

5.4 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 10. 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.

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

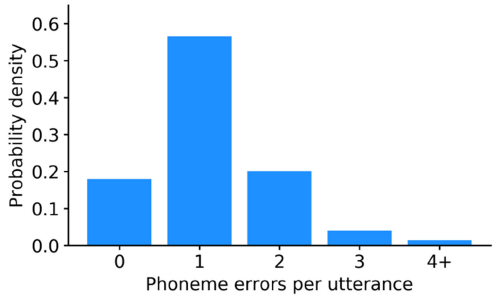

Fig. 10. Probability density for the number of phoneme errors per utterance.

Table 5. Evaluator Performance (True Positive is an Identified Mispronunciation)

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

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 GOP measure. We considered this to be a hard baseline since it was computed offline 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) [Panayotov et al. 2015], according to the implementation described by Witt and Young [2000]. 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.

6 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 2 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 [Hair et al. 2018]. 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 that 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% absolute with automated feedback and 61.5% absolute with caregiver feedback. These improvements are similar to those reported for traditional clinician [Murray et al. 2015; Thomas et al. 2014] and clinician plus caregiver [Thomas et al. 2018] speech therapy of similar intensity. They also align with results from previous studies demonstrating the efficacy of digital speech therapy applications [Jesus et al. 2019; Wren and Roulstone 2008]. 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% 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 toward 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 [Kewley-Port et al. 1987; Yeung et al. 2017]. 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 timestamps 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 [1993]. 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.

6.1 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 [McKinnon et al. 2007], 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 Hair et al. [2018] and Parnandi et al. [2015]) 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 regard 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. [2018] 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 [Springer and Whittaker 2019; Zhou and Chen 2018]; one way to move toward 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 article 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 [Loukina et al. 2019] and English acquisition in foreign-language speakers [Baur et al. 2014].

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 2-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.

7 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 [Jesus et al. 2019; Palmer et al. 2007]. 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.

ACKNOWLEDGMENTS

The authors wish to thank the children and families who gave us their time and invaluable feedback, without which this work could not have been completed. The authors are also grateful to Dr. Guanlong Zhao for providing the GOP framework for our analysis. The statements made herein are solely the responsibility of the authors.

REFERENCES

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: https://doi.org/10.1080/17549507.2018.1513562

American Speech-Language-Hearing Association. 2007. Speech Sound Disorders. Retrieved November 5, 2018 from https://www.asha.org/public/speech/disorders/SpeechSoundDisorders/.

Jason L. Anthony, Rachel Greenblatt Aghara, Martha J. Dunkelberger, Teresa I. Anthony, Jeffrey M. Williams, and Zhou Zhang. 2011. What factors place children with speech sound disorders at risk for reading problems?*American Journal of Speech-Language Pathology* 20, 2 (2011), 146–160. DOI: https://doi.org/10.1044/1058-0360(2011/10-0053)

ASHA Adhoc Commitee on CAS. 2007. Childhood apraxia of speech. Retrieved November 5, 2018 from https://www.asha.org/policy/TR2007-00278/.

Marjoke Bakker, Lilian Beijer, and Toni Rietveld. 2019. Considerations on effective feedback in computerized speech training for dysarthric speakers. *Telemedicine and e-Health* 25, 5 (2019), 351–358. DOI: https://doi.org/10.1089/tmj.2018.0050

Kirrie J. Ballard, Nicole M. Etter, Songjia Shen, Penelope Monroe, and Chek Tien Tan. 2019. Feasibility of automatic speech recognition for providing feedback during tablet-based treatment for apraxia of speech plus aphasia. *American Journal of Speech-Language Pathology* 28, 2S (2019), 818–834. DOI: https://doi.org/10.1044/2018_AJSLP-MSC18-18-0109

Anton Batliner, Mats Blomberg, Shona D'Arcy, Daniel Elenius, Diego Giuliani, Matteo Gerosa, Christian Hacker, Martin Russell, Stefan Steidl, and Michael Wong. 2005. The PF_STAR children's speech corpus. In *Ninth European Conference on Speech Communication and Technology (EUROSPEECH'05)*. ISCA, 2761–2764.

Claudia Baur, Manny Rayner, and Nikos Tsourakis. 2014. Using a serious game to collect a child learner speech corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), 2726–2732.

Gabriel J. Cler, Talia Mittelman, Maia N. Braden, Geralyn Harvey Woodnorth, and Cara E. Stepp. 2017. Video game rehabilitation of velopharyngeal dysfunction: A case series. *Journal of Speech, Language, and Hearing Research : JSLHR* 60, 6 Suppl (2017), 1800–1809. DOI: https://doi.org/10.1044/2017_JSLHR-S-16-0231

Gabriella A. Constantinescu, Deborah G. Theodoros, Trevor G. Russell, Elizabeth C. Ward, Stephen J. Wilson, and Richard Wootton. 2010. Home-based speech treatment for Parkinson's disease delivered remotely: A case report. *Journal of Telemedicine and Telecare* 16, 2 (2010), 100–104. DOI: https://doi.org/10.1258/jtt.2009.090306

Alycia E. Cummings and Jessica A. Barlow. 2011. A comparison of word lexicality in the treatment of speech sound disorders. *Clinical Linguistics & Phonetics* 25, 4 (2011), 265–286. DOI: https://doi.org/10.3109/02699206.2010.528822

Jonathan Dalby and Diane Kewley-Port. 1999. Explicit pronunciation training using automatic speech recognition technology. *CALICO Journal* 16, 3 (1999), 425–445.

Shiran Dudy, Meysam Asgari, and Alexander Kain. 2015. Pronunciation analysis for children with speech sound disorders. In *Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'15)*. IEEE, 5573–5576. DOI: https://doi.org/10.1109/EMBC.2015.7319655

Shiran Dudy, Steven Bedrick, Meysam Asgari, and Alexander Kain. 2018. Automatic analysis of pronunciations for children with speech sound disorders. *Computer Speech & Language* 50 (2018), 62–84. DOI: https://doi.org/10.1016/j.csl.2017.12.006

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: https://doi.org/10.1145/3170427.3186494

Jared Scott Duval, Zachary Rubin, Elizabeth Goldman, Nick Antrilli, Yu Zhang, Su-Hua Wang, and Sri Kurniawan. 2017. Designing towards maximum motivation and engagement in an interactive speech therapy game. In *Proceedings of the 2017 Conference on Interaction Design and Children*. ACM, 589–594. DOI: https://doi.org/10.1145/3078072.3084329

Olov Engwall, Olle Bölter, Anne-Marie Öster, and Hedvig Kjellström. 2006. Designing the user interface of the computer-based speech training system ARTUR based on early user tests. *Behaviour & Information Technology* 25, 4 (2006), 353–365. DOI: https://doi.org/10.1080/01449290600636702

Expressive Solutions. 2018. ArtixPix. Retrieved February 28, 2020 from http://expressive-solutions.com/artikpix/.

Susan Koch Fager. 2017. Speech recognition as a practice tool for dysarthria. *Seminars in Speech and Language* 38, 03 (2017), 220–228. DOI: https://doi.org/10.1055/s-0037-1602841

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: https://doi.org/10.1044/1058-0360(2003/083)

Lisa Furlong, Shane Erickson, and Meg E. Morris. 2017. Computer-based speech therapy for childhood speech sound disorders. *Journal of Communication Disorders* 68 (2017), 50–69. DOI: https://doi.org/10.1016/j.jcomdis.2017.06.007

Sadaoki Furui. 1981. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29, 2 (1981), 254–272. DOI: https://doi.org/10.1109/TASSP.1981.1163530

Mateja Gačnik, Andreja Istenič Starčič, Janez Zaletelj, and Matej Zajc. 2018. User-centred app design for speech sound disorders interventions with tablet computers. *Universal Access in the Information Society* 17, 4 (2018), 821–832. DOI: https://doi.org/10.1007/s10209-017-0545-9

Google. 2018. Cloud Text-to-Speech. Retrieved August 23, 2018 from https://cloud.google.com/text-to-speech

Margarida Grilo, Isabel Guimarães, Mariana Ascensão, Alberto Abad, Ivo Anjos, João Magalhães, and Sofia Cavaco. 2020. The BioVisualSpeech European Portuguese sibilants corpus. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language*. Springer, 23–33. DOI: https://doi.org/10.1007/978-3-030-41505-1_3

Adam Hair, Constantina Markoulli, Penelope Monroe, Jacqueline McKechnie, Kirrie J. Ballard, Beena Ahmed, and Ricardo Gutierrez-Osuna. 2020. 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* (2020-04-25). ACM, 1–8. DOI: https://doi.org/10.1145/3334480.3382886

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: https://doi.org/10.1145/3202185.3202733

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 *Proceedings of Interspeech 2009*. ISCA, 1455–1458.

Luis M. T. 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: https://doi.org/10.1044/2019_JSLHR-S-18-0301

Diane Kewley-Port, Charles S. Watson, Mary Elbert, Daniel Maki, and Daniel Reed. 1991. The Indiana speech training aid (ISTRA) II: Training curriculum and selected case studies. *Clinical Linguistics & Phonetics* 5, 1 (1991), 13–38. DOI: https://doi.org/10.3109/02699209108985500

Diane Kewley-Port, Charles S. Watson, Daniel Maki, and Daniel Reed. 1987. Speaker-dependent speech recognition as the basis for a speech training aid. In *Proceedings of the 1987 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'87)*, Vol. 12. IEEE, 372–375. DOI: https://doi.org/10.1109/ICASSP.1987.1169661

Carol L. Koch. 2018. *Clinical Management of Speech Sound Disorders*. Jones & Bartlett Learning, Burlington, Massachusetts.

Jessica Korte. 2020. Patterns and themes in designing with children. *Foundations and Trends® in Human–Computer Interaction* 13, 2 (2020), 70–164. DOI: https://doi.org/10.1561/1100000079

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 *Proceedings of the 1st ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play*. ACM, 429–430. DOI: https://doi.org/10.1145/2658537.2661305

Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. 1997. Analysis of children's speech: Duration, pitch and formants. In *Proceedings of the 5th European Conference on Speech Communication and Technology*. ISCA, 473–476.

Little Bee Speech. 2018. Articulation Station. Retrieved February 28, 2020 from http://littlebeespeech.com/articulation_station.php.

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

Vanessa Lopes, João Magalhães, and Sofia Cavaco. 2019. Sustained vowel game: A computer therapy game for children with dysphonia. In *Proceedings of Interspeech 2019*. ISCA, 26–30. DOI: https://doi.org/10.21437/Interspeech.2019-3017

Anastassia Loukina, Beata Beigman Klebanov, Patrick Lange, Yao Qian, Binod Gyawali, Nitin Madnani, Abhinav Misra, Klaus Zechner, Zuowei Wang, and John Sabatini. 2019. Automated estimation of oral reading fluency during summer camp e-book reading with myturntoread. In *Proceedings of Interspeech 2019*. ISCA, 21–25. DOI: https://doi.org/10.21437/Interspeech.2019-2889

Edwin Maas, Christina E. 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: https://doi.org/10.1007/s40474-014-0016-4

Edwin Maas, Donald A. Robin, Shannon N. Austermann Hula, Skott E. Freedman, Gabriele Wulf, Kirrie J . Ballard, and Richard A. Schmidt. 2008. Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology* 17, 3 (2008), 277–298. DOI: https://doi.org/10.1044/1058-0360(2008/025)

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: https://doi.org/10.3109/17549507.2011.535565

Jacqui McKechnie, Beena Ahmed, Ricardo Gutierrez-Osuna, Penelope Monroe, Patricia McCabe, and Kirrie J. Ballard. 2018. Automated speech analysis tools for children's speech production: A systematic literature review. *International Journal of Speech-Language Pathology* 20, 6 (2018), 583–598. DOI: https://doi.org/10.1080/17549507.2018.1477991

David H. McKinnon, Sharynne McLeod, and Sheena Reilly. 2007. The prevalence of stuttering, voice, and speech-sound disorders in primary school students in Australia. *Language, Speech, and Hearing Services in Schools* 38, 1 (2007), 5–15. DOI: https://doi.org/10.1044/0161-1461(2007/002)

Sharynne McLeod and Elise Baker. 2014. Speech-language pathologists' practices regarding assessment, analysis, target selection, intervention, and service delivery for children with speech sound disorders. *Clinical Linguistics & Phonetics* 28, 7–8 (2014), 508–531. DOI: https://doi.org/10.3109/02699206.2014.926994

Sharynne McLeod, Elise Baker, Jane McCormack, Yvonne Wren, Sue Roulstone, Kathryn Crowe, Sarah Masso, Paul White, and Charlotte Howland. 2017. Cluster-randomized controlled trial evaluating the effectiveness of computer-assisted intervention delivered by educators for children with speech sound disorders. *Journal of Speech, Language, and Hearing Research* 60, 7 (2017), 1891–1910.

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: https://doi.org/10.1044/2015_JSLHR-S-13-0179

Amal Nanavati, Mary Bernardine Dias, and Aaron Steinfeld. 2018. Speak Up: A multi-year deployment of games to motivate speech therapy in India. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 318. DOI: https://doi.org/10.1145/3173574.3173892

Andrés Adolfo Navarro-Newball, Diego Loaiza, Claudia Oviedo, Andrés Castillo, Anita Portilla, Diego Linares, and Gloria Ineś Álvarez. 2014. Talking to Teo: Video game supported speech therapy. *Entertainment Computing* 5, 4 (2014), 401–412. DOI: https://doi.org/10.1016/j.entcom.2014.10.005

Antonio Origlia, Federico Altieri, Giorgia Buscato, Alice Morotti, Claudio Zmarich, Antonio Rodá, and Piero Cosi. 2018. 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*. ACM, 190–195. DOI: https://doi.org/10.1145/3284869.3284913

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: https://doi.org/10.1080/13682820601173296

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15)*. IEEE, 5206–5210. DOI: https://doi.org/10.1109/ICASSP.2015.7178964

Avinash Parnandi, Virendra Karappa, Tian Lan, Mostafa Shahin, Jacqueline McKechnie, Kirrie J. Ballard, Beena Ahmed, and Ricardo Gutierrez-Osuna. 2015. Development of a remote therapy tool for childhood apraxia of speech. *ACM Transactions on Accessible Computing (TACCESS)* 7, 3 (2015), 10. DOI: https://doi.org/10.1145/2776895

Avinash Parnandi, Virendra Karappa, Youngpyo Son, Mostafa Shahin, Jacqueline McKechnie, Kirrie J. Ballard, Beena Ahmed, and Ricardo Gutierrez-Osuna. 2013. Architecture of an automated therapy tool for childhood apraxia of speech. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 5. DOI: https://doi.org/10.1145/2513383.2513450

Jonathan L. Preston, Margaret Hull, and Mary Louise Edwards. 2013. Preschool speech error patterns predict articulation and phonological awareness outcomes in children with histories of speech sound disorders. *American Journal of Speech-Language Pathology* 22, 2 (2013), 173–184. DOI: https://doi.org/10.1044/1058-0360(2012/12-0022)

Douglas A. Reynolds. 2002. An overview of automatic speaker recognition technology. In *Procedings of the 2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'02)*. IEEE, 4072–4075. DOI: https://doi.org/10.1109/ICASSP.2002.5745552

Yvan Rose and Brian Macwhinney. 2014. The PhonBank Project: Data and software-assisted methods for the study of phonology and phonological development.

Zak Rubin, Sri Kurniawan, Taylor Gotfrid, and Annie Pugliese. 2016. 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*. ACM, 325–326. DOI: https://doi.org/10.1145/2982142.2982203

Leanne Ruggero, Patricia McCabe, Kirrie J. Ballard, and Natalie Munro. 2012. Paediatric speech-language pathology service delivery: An exploratory survey of Australian parents. *International Journal of Speech-Language Pathology* 14, 4 (2012), 338–350. DOI: https://doi.org/10.3109/17549507.2011.650213

Oscar Saz, Eduardo Lleida, and William Ricardo Rodríguez. 2009a. Avoiding speaker variability in pronunciation verification of children's disordered speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*. ACM, 1–5. DOI: https://doi.org/10.1145/1640377.1640388

Oscar Saz, Shou-Chun Yin, Eduardo Lleida, Richard Rose, Carlos Vaquero, and William Ricardo Rodríguez. 2009b. Tools and technologies for computer-aided speech and language therapy. *Speech Communication* 51, 10 (2009), 948–967. DOI: https://doi.org/10.1016/j.specom.2009.04.006

Richard A. Schmidt and Timothy D. Lee. 2005. *Motor Control and Learning: A Behavioral Emphasis*. Human Kinetics, Champaign, Illinois.

Mostafa Shahin, Beena Ahmed, Jim X. Ji, and Kirrie J. Ballard. 2018. Anomaly detection approach for pronunciation verification of disordered speech using speech attribute features. In *Proceedings of Interspeech 2018*. ISCA, 1671–1675. DOI: https://doi.org/10.21437/Interspeech.2018-1319

Mostafa Shahin, Beena Ahmed, Jacqueline McKechnie, Kirrie J. Ballard, and Ricardo Gutierrez-Osuna. 2014. A comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech. In *Proceedings of Interspeech 2014*. ISCA, 1583–1587.

Mostafa Shahin, Beena Ahmed, Avinash Parnandi, Virendra Karappa, Jacqueline McKechnie, Kirrie J. Ballard, and Ricardo Gutierrez-Osuna. 2015. Tabby talks: An automated tool for the assessment of childhood apraxia of speech. *Speech Communication* 70 (2015), 49–64.

Khaldoun Shobaki, John-Paul Hosom, and Ronald A. Cole. 2000. The OGI kids' speech corpus and recognizers. In *Proceedings of the 6th International Conference on Spoken Language Processing*. ISCA, 258–261.

Lawrence D. Shriberg, Joan Kwiatkowski, and Tereza Snyder. 1990. Tabletop versus microcomputer-assisted speech management: Response evocation phase. *Journal of Speech and Hearing Disorders* 55, 4 (1990), 635–655. DOI: https://doi.org/10.1044/jshd.5504.635

Valerie J. Shute. 2008. Focus on formative feedback. *Review of Educational Research* 78, 1 (2008), 153–189. DOI: https://doi.org/10.3102/0034654307313795

Smarty Ears Apps. 2017. Apraxia Farm. Retrieved February 28, 2020 from http://smartyearsapps.com/apraxia-ville/.

Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: Empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 107–120. DOI: https://doi.org/10.1145/3301275.3302322

Kerri L. Staples and Greg Reid. 2010. Fundamental movement skills and autism spectrum disorders. *Journal of Autism and Developmental Disorders* 40, 2 (2010), 209–217. DOI: https://doi.org/10.1007/s10803-009-0854-9

Erik F. Strommen and Francine S. Frome. 1993. Talking back to big bird: Preschool users and a simple speech recognition system. *Educational Technology Research and Development* 41, 1 (1993), 5–16.

Tactica Interactive. 2011. Tiga Talk Speech Therapy Games. Retrieved February 28, 2020 from https://tactica.ca/project/tiga-talk-speech-therapy-games/.

Deborah Theodoros. 2008. Telerehabilitation for service delivery in speech-language pathology. *Journal of Telemedicine and Telecare* 14, 5 (2008), 221–224. DOI: https://doi.org/10.1258/jtt.2007.007044

Deborah Theodoros, Trevor Russell, and R. Latifi. 2008. Telerehabilitation: Current perspectives. *Studies in Health Technology and Informatics* 131 (2008), 191–210.

Donna C. Thomas, Patricia McCabe, and Kirrie J. Ballard. 2014. Rapid syllable transitions (ReST) treatment for childhood apraxia of speech: The effect of lower dose-frequency. *Journal of Communication Disorders* 51 (2014), 29–42. DOI: https://doi.org/10.1016/j.jcomdis.2014.06.004

Donna C. Thomas, Patricia McCabe, and Kirrie J. Ballard. 2018. Combined clinician-parent delivery of rapid syllable transition (ReST) treatment for childhood apraxia of speech. *International Journal of Speech-language Pathology* 20, 7 (2018), 683–698. DOI: https://doi.org/10.1080/17549507.2017.1316423

Catherine Torrington Eaton and Nan Bernstein Ratner. 2016. An exploration of the role of executive functions in preschoolers' phonological development. *Clinical Linguistics & Phonetics* 30, 9 (2016), 679–695. DOI: https://doi.org/10.1080/02699206.2016.1179344

Şermin Tükel, Helena Björelius, Gunilla Henningsson, Anita McAllister, and Ann Christin Eliasson. 2015. Motor functions and adaptive behaviour in children with childhood apraxia of speech. *International Journal of Speech-Language Pathology* 17, 5 (2015), 470–480. DOI: https://doi.org/10.3109/17549507.2015.1010578

Carlos Vaquero, Oscar Saz, Eduardo Lleida, José Manuel Marcos, and César Canalís. 2006. VOCALIZA: An application for computer-aided speech therapy in spanish language. In *IV Jornadas en Tecnología del Habla*. Zaragoza, Spain, 321–326.

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: https://doi.org/10.1044/jshr.3202.245

Krystal L. Werfel, Marren C. Brooks, and Lisa Fitton. 2019. The comparative efficiency of speech sound interventions that differ by delivery modality: Flashcards versus tablet. *Communication Disorders Quarterly* 0, 0 (2019), 1–9. DOI: https://doi.org/10.1177/1525740119859520

Pamela Williams and Hilary Stephens. 2004. *Nuffield Centre Dyspraxia Programme 2004*. The Miracle Factory, for the Speech & Language Therapy Department, Nuffield Hearing and Speech Centre, Royal National Throat, Nose and Ear Hospital, Windsor, United Kingdom.

Silke M. Witt and Steve J. Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication* 30, 2–3 (2000), 95–108. DOI: https://doi.org/10.1016/S0167-6393(99)00044-8

Yvonne Wren and Sue Roulstone. 2008. A comparison between computer and tabletop delivery of phonology therapy. *International Journal of Speech-Language Pathology* 10, 5 (2008), 346–363. DOI: https://doi.org/10.1080/17549500701873920

Gary Yeung, Amber Afshan, Kaan Ege Ozgun, Canton Kaewtip, Steven M. Lulich, and Abeer Alwan. 2017. Predicting clinical evaluations of children's speech with limited data using exemplar word template references. In *Proceedings of the 7th ISCA Workshop on Speech and Language Technology in Education*. ISCA, 161–166. DOI: https://doi.org/10.21437/SLaTE.2017-28

Matej Zajc, Andreja Istenič Starčič, Maja Lebeničnik, and Mateja Gačnik. 2018. Tablet game-supported speech therapy embedded in children's popular practices. *Behaviour & Information Technology* 37, 7 (2018), 693–702. DOI: https://doi.org/10.1080/0144929X.2018.1474253

Jianlong Zhou and Fang Chen. 2018. *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*. Springer.

Received April 2020; revised September 2020; accepted November 2020