# Development of a Remote Therapy Tool for Childhood Apraxia of Speech

AVINASH PARNANDI, VIRENDRA KARAPPA, and TIAN LAN, Texas A&M University
MOSTAFA SHAHIN, Texas A&M University at Qatar
JACQUELINE MCKECHNIE and KIRRIE BALLARD, The University of Sydney
BEENA AHMED, Texas A&M University at Qatar
RICARDO GUTIERREZ-OSUNA, Texas A&M University



We present a multitier system for the remote administration of speech therapy to children with apraxia of speech. The system uses a client-server architecture model and facilitates task-oriented remote therapeutic training in both in-home and clinical settings. The system allows a speech language pathologist (SLP) to remotely assign speech production exercises to each child through a web interface and the child to practice these exercises in the form of a game on a mobile device. The mobile app records the child's utterances and streams them to a back-end server for automated scoring by a speech-analysis engine. The SLP can then review the individual recordings and the automated scores through a web interface, provide feedback to the child, and adapt the training program as needed. We have validated the system through a pilot study with children diagnosed with apraxia of speech, their parents, and SLPs. Here, we describe the overall client-server architecture, middleware tools used to build the system, speech-analysis tools for automatic scoring of utterances, and present results from a clinical study. Our results support the feasibility of the system as a complement to traditional face-to-face therapy through the use of mobile tools and automated speech analysis algorithms.

CCS Concepts: • **Computing methodologies** → **Speech recognition**; • **Applied computing** → *Computer-assisted instruction*; • **Applied computing** → **Computer-managed instruction**; • **Social and professional topics** → *Assistive technologies*; • **Social and professional topics** → **People with disabilities**

Additional Key Words and Phrases: Childhood apraxia of speech, speech therapy, automated speech analysis

**ACM Reference Format:**
Avinash Parnandi, Virendra Karappa, Tian Lan, Mostafa Shahin, Jacqueline McKechnie, Kirrie Ballard, Beena Ahmed, and Ricardo Gutierrez-Osuna. 2015. Development of a remote therapy tool for childhood apraxia of speech. ACM Trans. Access. Comput. 7, 3, Article 10 (November 2015), 23 pages.
DOI: http://dx.doi.org/10.1145/2776895

This work was made possible by NPRP Grant # [4-638-2-236] from the Qatar National Research Fund (a member of Qatar Foundation). Kirrie Ballard would also like to acknowledge financial support from the Australian Research Council Future Fellowship Scheme. The statements made herein are solely the responsibility of the authors.
A preliminary version of this paper appeared in the *Proceedings of the 15th ACM SIGACCESS International Conference on Computers and Accessibility*, October 21-23, 2013.
Authors' addresses: A. Parnandi, V. Karappa, T. Lan, and R. Gutierrez-Osuna, 506 H. R. Bright Building - Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77840-3112, United States; emails: {parnandi, vkarappa1, welkinlan, rgutier}@tamu.edu; M. Shahin and B. Ahmed, 319D - Department of Electrical and Computer Engineering, 125 Texas A&M Engineering Building, Education City, Doha 23874, Qatar; emails: {mostafa.shahin, beena.ahmed}@qatar.tamu.edu; J. McKechnie and K. Ballard, S157, C42 - Cumberland Campus, Faculty of Health Sciences, The University of Sydney, Sydney, NSW 2141, Australia; emails: {jacqueline.mckechnie, kirrie.ballard}@sydney.edu.au.

DOI: http://dx.doi.org/10.1145/2776895

## 1. INTRODUCTION

Childhood apraxia of speech (CAS) is a neurological pediatric speech sound disorder (SSD) that impairs speech motor planning/programming. CAS can delay acquisition of skills including the control of vocal pitch, intensity, and duration of speech sounds [ASHA Ad Hoc Committee on Apraxia of Speech in Children 2007]. It also impairs the child's ability to correctly pronounce sounds, syllables, and words. CAS can thus render the child unable to start articulating the first sounds, and can lead to a serious communicative disability. CAS can be difficult to diagnose and monitor due to a high comorbidity with other speech and language disorders and a lack of specific tools [Newbury and Monaco 2010].

By working intensely with a trained speech language pathologist (SLP), those with CAS can overcome their motor planning and motor programming difficulties (articulation capabilities) [ASHA Ad Hoc Committee on Apraxia of Speech in Children 2007]. However, the ratio of children with CAS to the number of qualified SLPs is growing at a high rate. According to the literature, current estimates of children with CAS fall between 3.4% and 4.3% [ASHA Ad Hoc Committee on Apraxia of Speech in Children 2007; Delaney and Kent 2004] and the estimates for developmental coordination disorder fall between 5% and 6% [Gaines et al. 2008]. Due to the increasing number of children needing intervention and the shortage of trained SLPs, there is an ever-increasing gap between the quality and duration of needed therapeutic interventions and what is available because of time constraints and expenses [ASHA Ad Hoc Committee on Apraxia of Speech in Children 2007]. Thus, there is a need for practical and cost-effective technological interventions to complement traditional face-to-face therapy sessions. CAS therapy usually consists of verbal, auditory, and visual interaction between an SLP and the child using game-like activities [Williams and Stephens 2010], making it a good candidate for technology-based alternative solutions, as these can provide not only remote and automatic monitoring but also interactive training.

As a step toward this goal, we present a multitier system that enhances the administration of traditional CAS therapy for in-home settings. We adopt the Nuffield Dyspraxia Program (NDP3), an assessment and intervention package for children with severe speech sound disorders including CAS [Williams and Stephens 2010; *ArtikPix*]. NDP3 follows a bottom-up approach, providing speech therapy that can be adapted to the child's individual needs and progress. This makes NDP3 ideal for our purposes. As illustrated in Figure 1, our system consists of three major components: (1) a mobile app running on a tablet, which allows the children to practice speech exercises in their homes; (2) a therapy management interface running on a server, which allows SLPs to assign exercises and monitor progress; and (3) a speech processing engine on the server, which performs automated diagnostics on the children's recordings. The tablet-based therapy enables remote therapy sessions for which the SLP and the child need not be in the same geographical location, and helps overcome barriers of access to speech therapy due to distances, lack of specialists, and lack of equipment. This architecture differentiates our system from existing mobile speech therapy tools [Maier et al. 2010; Bunnell et al. 2000; Wren et al. 2006; Vilozni et al. 2001], which are stand-alone applications with no automated speech assessment or remote monitoring capabilities. Automated therapy via the tablet is designed to be more engaging and motivating for the child, which facilitates compliance with the practice. As an additional benefit, the tool allows children to practice on their own, potentially allowing for a higher intensity of practice than is typically possible with parent-directed home practice.

The work presented in this article builds on our existing CAS therapy system and addresses feedback from a pilot study reported at ASSETS 2013 [Parnandi et al. 2013]. We have expanded the previous system with a new native mobile client application

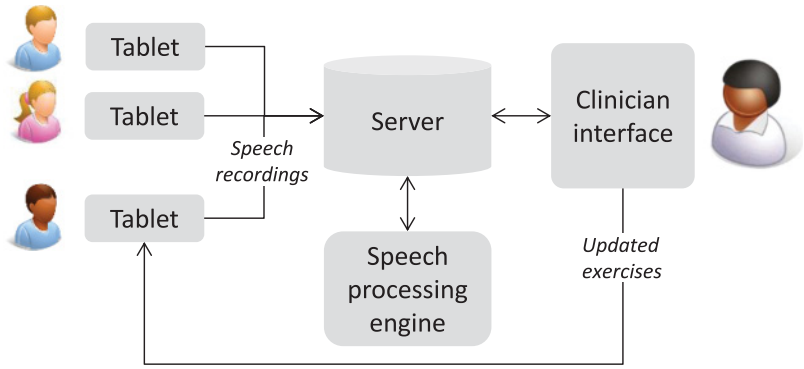


Fig. 1. General overview of the CAS therapy system showing the server, mobile clients, and the remote therapy management system.

as well as new server and speech processing tools. The new mobile app incorporates a memory game with audiovisual feedback to make the speech therapy practice more engaging for the child. It also supports offline functionality to address bandwidth and coverage limitations at the point of care. On the server side, we have added new reporting and visualization tools to assist the SLP in monitoring progress of each child. Finally, we have conducted (additional) user studies with children with CAS/SSD to test the new system during speech therapy. During these studies, children with CAS performed NDP3 exercises on the tablet under the supervision of an SLP or in the comfort of their home with their parents/guardian. We present results from a semistructured interview conducted with children, parents, and SLPs to determine if our proposed system can complement traditional therapy, explore issues related to the usability of the system (tablet and clinician interface), level of engagement of the children during therapy, preferences between tablet and paper-based therapy, and areas of improvement for the system. The study also served as a preliminary trial to compare the efficacy of speech therapy delivered via the tablet in two modes: therapy directed by an SLP 4 days/week for 3 weeks (SLP4) versus therapy directed by the SLP 1 day/week for 3 weeks and by the child the other 3 days/week (SLP1). The primary difference between the two was the additional opportunity in the SLP4 mode for detailed feedback from SLP to child on error correction strategies. The SLP1 mode is equivalent to standard care, in which a child has therapy with an SLP once a week but then completes homework 3 to 4 days a week.

The rest of the article is organized as follows. Section 2 provides a brief overview of CAS, the NDP3, and summarizes past work on computerized speech therapy. Section 3 presents our system architecture, including the mobile client, the server, clinician web interface, and speech processing engine. Section 4 describes the experimental setup for the validation studies. Survey comments and recommendations from participants, as well as results from the clinical trials, are summarized in Section 5. The article concludes with a discussion and directions for future work in Section 6.

## 2. BACKGROUND AND RELATED WORK

### 2.1. Childhood Apraxia of Speech

The American Speech-Language-Hearing Association (ASHA) defines CAS as a "neurological childhood (pediatric) speech sound disorder in which the precision and consistency of movements for underlying speech are impaired in the absence of neuromuscular deficits (e.g., abnormal reflexes, abnormal tone)" [ASHA Ad Hoc Committee on Apraxia of Speech in Children 2007]. Although children with CAS usually have

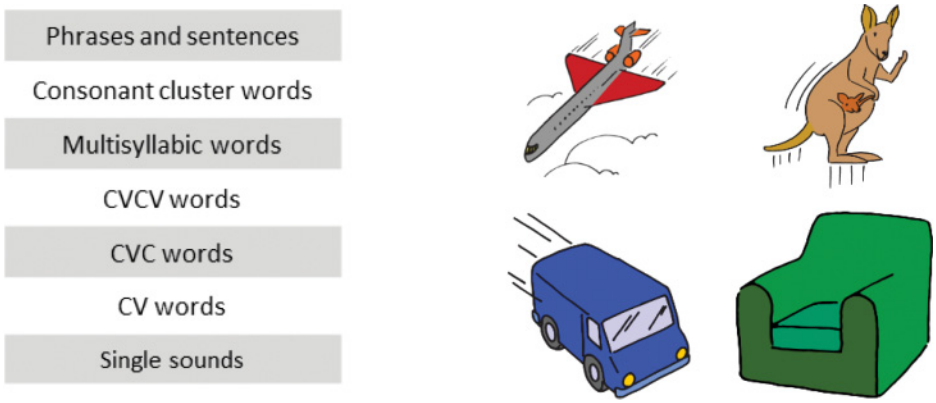


Fig. 2. (a) NDP3 "brick wall" showing the bottom-up therapy approach. C = consonant, V = vowel. (b) Sample NDP3 exercise with four stimulus images for elicitation of the utterances: airplane, kangaroo, van, chair [Jamieson et al. 2004].

no damage to muscles or nerves, the area of the brain sending signals to the muscles is damaged or not fully developed. Thus, children with CAS have marked difficulty in motor planning/programming, as well as correctly producing sounds, syllables, and words [Shriberg et al. 2003]. Specifically, these children have difficulty coordinating precise and consistent movements of the articulators (tongue, lips, jaw, and palate) required to produce speech and achieve an acceptable pronunciation of a given word [Ballard et al. 2010]. The speech of children with CAS is usually unintelligible to unfamiliar listeners due to phonemic speech errors and articulatory abnormalities. In the absence of treatment, this neuromuscular disorder can delay the acquisition of speech skills, phonological abilities, and, subsequently, reading abilities, thus causing severe communicative disability [Forrest 2003]. Hence, accurate and timely intervention is critical for children with CAS. Intervention involves repeated speech therapy sessions between the SLP and child, which can continue for several years.

### 2.2. The Nuffield Dyspraxia Programme

Our proposed system is based on the Nuffield Dyspraxia Programme (NDP3), an intervention program for children with severe speech sound disorders, including CAS [Williams and Stephens 2010, *ArtikPix*]. NDP3 has been designed to address the effects of CAS, such as articulation of individual consonants and vowels, sequencing sounds together, and maintaining prosodic accuracy [Williams and Stephens 2010; *ArtikPix*]. Therapy for CAS consists of two components: assessment and intervention, performed using a bottom-up approach (Figure 2(a)). An initial assessment provides a measure of the child's current speech skills for designing therapy goals, starting from isolated speech sounds and progressing to complex syllable structures, then to sentences [Williams and Stephens 2010, *ArtikPix*]. Readministering the tests after intervention also provides a measure of therapy effectiveness in helping a child regain a dimension of speech functionality.

The NDP3 protocol requires regular therapy sessions under the supervision of an SLP. The intervention approach consists of a therapy manual, 1800 picture cards involving 750 different images, and 550 line-drawn worksheets. This includes a set of picture cues to represent single consonants (C), vowels (V), diphthongs, and words at each of the phonotactic levels (Figure 2(b)). Word creation by joining of sounds or

syllables is facilitated by transition worksheets (e.g., bay + bee = baby), while sequencing worksheets provide repetitive practice. Guided by the instruction manual, the picture cards are presented to the child as stimuli through tabletop games to elicit the target utterance. The child is asked to produce specific sounds, syllables, or words compliant with the child's therapy level in these activities. On correctly completing an event, the SLP usually presents the child with simple rewards. On an incorrect response, the SLP follows the instructions to assist the child in eliciting the correct response. This approach works from the child's strengths and builds skills in incremental steps in a cumulative way [Williams and Stephens 2010; *ArtikPix*]. The NDP3 assessment is multilayered, with the bottom layer consisting of single speech sounds; the next layer contains CV words, followed by CVC, CVCV, and multisyllabic words, consonant clusters, and, finally, connected speech in the form of phrases and sentences (Figure 2(a)). It relies on the production of (1) all the single consonants, vowels, and diphthongs; (2) a set of 20 single words at each phonotactic structure (CV/VC, CVCV, CVC, CCV, and multisyllabics) through picture naming; and (3) phrases and sentences through imitation with pictures.

### 2.3. Previous Work on Computerized Therapy

Automated therapy is a subcategory of technological approaches to health care known as telehealth, telemedicine or telepractice. Traditional CAS therapy requires a child to undergo extended therapy sessions with a trained SLP in a clinic. This can be both logistically and financially prohibitive, paving the way for remote and automated therapy tools.

Studies have shown that computer-based therapy reduces error due to ambiguous verbal or written responses and may increase the quality of data (compared to traditional tabletop therapy) [Jamieson et al. 2004]. It allows for consistent and controlled presentation of stimuli, precise delivery of auditory information, and greater accuracy in performance measurement [Veale 1999]. Waite et al. [2006] investigated the feasibility of remote assessment of childhood SSDs by SLPs and compared it with face-to-face interaction. They found a high level of agreement between the two methods (single-word articulation: 92%; speech intelligibility: 100%; and oromotor tasks: 91%). Previous studies have also looked at the effect of teletherapy in other speech-disorder populations. As an example, in a feasibility study on online treatment delivery for speech disorder of a patient with Parkinson's disease, Constantinescu et al. [Constantinescu et al. 2010] showed that Internet-based telerehabilitation sessions can be as effective as clinic-based sessions.

A number of tools have been developed to facilitate general speech therapy. Speech recognition software tools (e.g., Dragon Dictate) have been used for the assessment of pathological disorders in which the acoustic characteristics of the voice produced are affected due to laryngeal and vocal-tract disorders [Maier et al. 2010; Kolles and Feiden 1995; Moran et al. 2006]. However, in speech sound disorders such as CAS, the child's voice quality is unaffected; they instead struggle with articulation errors. General tools to facilitate speech therapy include STAR (Speech Training, Assessment, and Remediation), which assists SLPs in treating children with articulation problems [Bunnell et al. 2000], and Ortho-Logo-Paedia (OLP), which is intended for use in in-home settings [Oster et al. 2002]. However, these systems do not cater to the specific articulation problems of children with CAS and other SSDs.

In the specific context of CAS and SSD, a few software programs and telerehabilitation tools have also become available [Georgeadis et al. 2003], such as Phoneme Factory Sound Sorter (PFSS) [Wren et al. 2006], Sound Contrasts in Phonology (SCIP) [Williams 2006], and Speech Assessment and Interactive Learning Systems (SAILS) [Rvachew and Brosseau-Lapre 2006]. These tools assist the child in developing

phonological patterns and phonemic contrasts. The main drawback of these systems, however, is the absence of automatic feedback, which makes it hard to adapt the therapy regimen on-the-fly based on the specific needs of each child.

Mobile technology provides opportunities to gain richer data and improve the experience of children undergoing clinical interventions. Touch-based devices, such as tablets and smartphones, are intuitive and engaging as compared to desktop and paper-based alternatives, and are also highly cost-effective for in-home therapy sessions. This has led to the development of generic speech therapy applications for mobile platforms, such as PocketSLP [Maier et al. 2010], ArtikPix [Bunnell et al. 2000], and Speech with Milo [Wren et al. 2006], which usually focus on articulation problems. Of particular interest in our case is Apraxiaville [Vilozni et al. 2001], a mobile tool developed for CAS. It includes features such as voice recording, self-scoring, and animated stimuli. However, this app supports only three levels of therapy (single sound, CV-VC-CVC, and multisyllabic words), whereas the NDP3 protocol also supports cluster-word and phrase and sentence formation. Furthermore, most of the current apps, including Apraxiaville, are stand-alone tools with no remote and automated speech assessment or feedback capabilities. In contrast, our system includes automated speech-processing capabilities to provide timely feedback to both the SLP and the child, as well as the ability to manage multiple children remotely.

## 3. SYSTEM ARCHITECTURE

To remotely administer NDP3 therapy at the home, the system should be able to: (1) prompt the child with the appropriate stimuli on the mobile platform; (2) record the child's speech response and stream it back to the server; (3) identify the individual consonants and vowels produced, and the errors made, through speech analysis algorithms; (4) provide feedback to the child; (5) provide reports to the SLP detailing the child's progress; and (6) facilitate the creation or modification of exercises by the SLP based on performance results. The following sections describe the system components that we have designed to meet these needs, and the software technologies that supported the development.

### 3.1. Mobile Client

The mobile client provides a supplement to (and is modeled after) the face-to-face session between the SLP and the child. The application provides visual stimuli to the child, records the child's speech response, and provides feedback. Its user interface is specially tailored for children following recommendations in the literature [Mich 2009; Rick et al. 2009; Anthony et al. 2012]. Children have lower manual dexterity and generally less experience with tablets than adults. This results in unique behaviors when compared to those observed with adults, such as *holdovers* (hitting a button a few times for a single action), unintentional swipes, and problems in target acquisition, for example, difficulty in performing the precise motion for selecting a target and lifting without slipping [Froehlich et al. 2007]. Thus, we designed the user interface (UI) to minimize these types of errors: as an example, to avoid holdovers, the interface has distinct buttons to start, stop, and replay the recording of the speech response. Once the *record* button is pressed, the *stop* button shows up on the screen and the *record* button becomes inactive. Thus, even if the child presses the *record* button multiple times, it will not lead to multiple recordings. After stopping, the *play* and *record* buttons are reactivated to play the recorded speech or to record a new utterance.

To assist the user during the course of a therapy session, the client provides additional visual cues. For example, when the child is recording, the *stop* button flashes to remind the child how to stop; after the child completes an exercise, a banner pops up guiding the child to go back to the front page. To guide the child in producing the correct

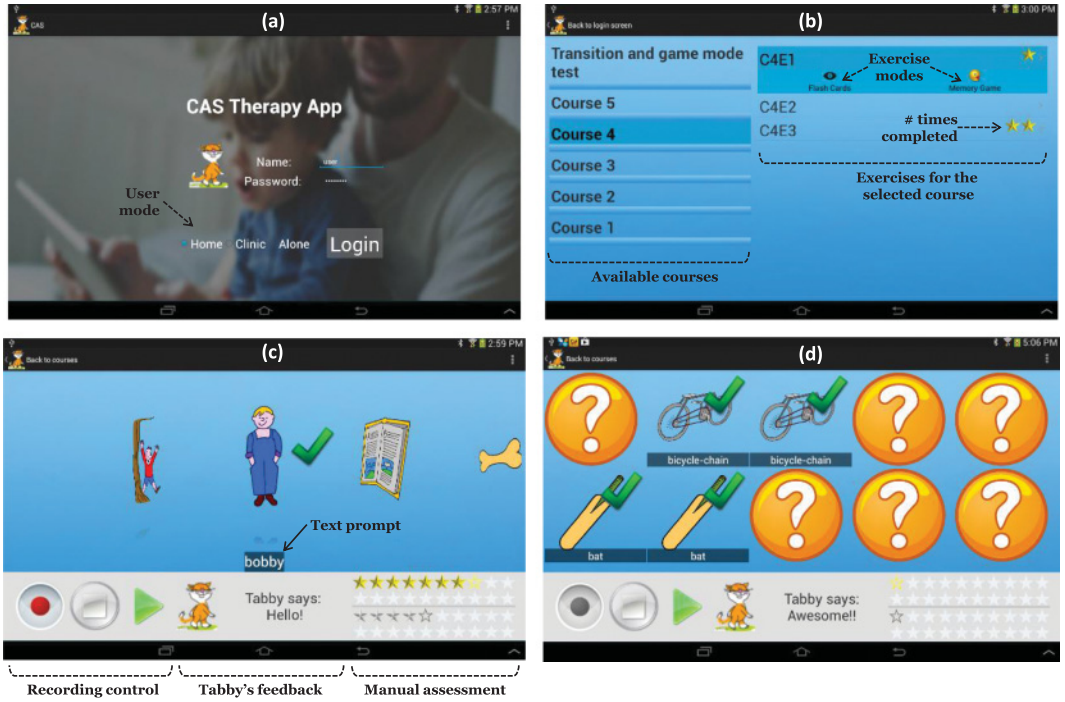


Fig. 3. (a) Log-in screen with input areas for username/password and radio buttons for the three user modes. (b) Home screen shows the different courses that a child is enrolled in, and the exercises for one of those courses (Course 4). (c) Exercise in flashcard mode. (d) Exercise in memory-game mode.

utterance, the exercises have audio and text prompts associated with each of the NDP3 stimulus images; as noted by the SLPs in our pilot study [Parnandi et al. 2013], this also helps the child to gradually learn letter-to-sound relationships.

Figure 3(a) shows the login screen for the client app. The screen allows the user to select one of three modes: "Home" (intended to be used when the child is with a caregiver), "Clinic" (use by the SLP) and "Alone" (when the child is practicing by oneself). The first two modes allow the caregiver/SLP to annotate the child's productions, as we discuss later. Figure 3(b) show the home screen, which allows the user to browse through the courses that have been remotely assigned by the SLP, each course comprised of multiple NDP3 exercises. Once the child selects a particular exercise in a course, the child can then select either "flash card" or "memory game" mode Figures (3(c) and 3(d), respectively). Both modes have the same control panel, located at the bottom of the screen, which contains the recording control (record, stop, and replay buttons), a manual assessment panel, and a Tabby feedback panel.

—*Recording control*. While in an exercise, tapping on the red circle will start the recording. Once finished, the child can play back the recorded sound, practice the stimulus again, or move to next image by tapping on the corresponding image. To assist the child, tapping on an image plays the desired utterance.
—*Manual assessment panel*. If the user logs in under the "Home" or "Clinic" modes, the app provides parents/SLPs the option to manually assess the goodness of each utterance by means of a gold star (good) or a silver star (fair). This serves both as an annotation tool and as a reward system for the child, since the gold stars accumulate with the completion of each correct utterance. These annotations are saved

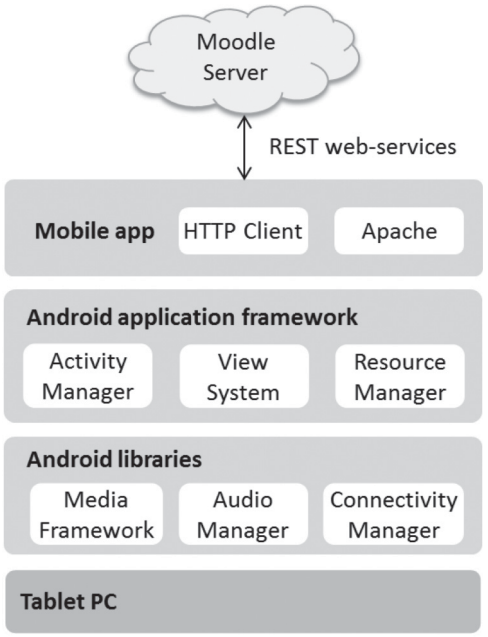


Fig. 4. Mobile client architecture showing the software tools used for developing the mobile app user interface and protocol used for communicating with the server.

for later use as ground truth and uploaded to the server along with the recording. All recordings are time-stamped and uploaded to the server asynchronously in the background.

—*Tabby feedback panel*. An animated character (a tabby cat) provides written and audio encouragement to the child in the form of expressions such as *Well done!* and *Excellent!*.

An important addition with respect to the mobile client used in our pilot study [Parnandi et al. 2013] is an interactive memory game that complements the original flashcard mode. In the flashcard mode (Figure 3(c)), the child is presented with a set of stimulus images. The child records an utterance corresponding to the current stimulus and moves to the next image by tapping or swiping. In the memory game mode (Figure 3(d)), the child is presented with five pairs of NDP3 images hidden behind bubbles; the goal is to find/match all pairs of stimuli. After uncovering each bubble, the child has to record an utterance before being able to uncover the next bubble, in this way encouraging speech production.

The mobile client, which was originally implemented as an HTML5 web application [Parnandi et al. 2013], has also been refactored as a native Android application. This provides a stable and more efficient access to the resources on the tablet, thus greatly enhances the responsiveness and robustness of the mobile client. In contrast to the previous HTML5 web application, the native client app is able to access the device's hardware features (e.g., microphone, network, graphic display, and storage) using the Android APIs (e.g., MediaRecorder, ConnectivityManager, and ImageView, and File) without importing any external plug-ins (Figure 4). To further reduce latency, the client app stores the entire collection of images in the NDP3 protocol on the tablet. Therefore, upon assigning a new exercise to a child, only the metadata and not the complete images need to be downloaded from the server. The client app also supports

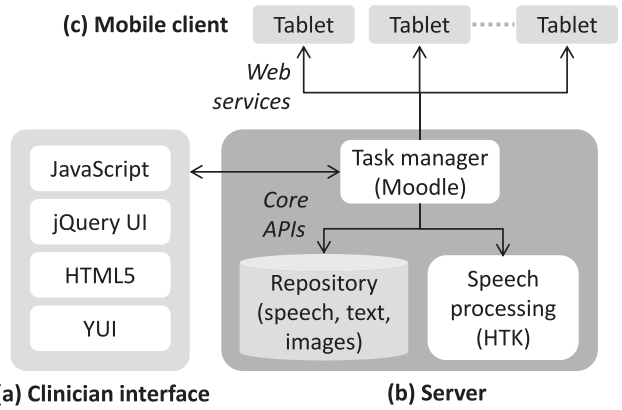


Fig. 5. Software components used to develop (a) the clinician interface and (b) server architecture; (c) mobile clients.

offline functionality to improve access in remote settings and eliminate bandwidth issues. This has been facilitated using asynchronous handshaking and data transfer mechanisms between client and server. During a therapy session, all recordings and metadata are stored in a local repository on the tablet. A background Wi-Fi watchdog service checks the network status every 10min and uploads the data when connected to the Internet. The local repository also serves as a backup for audio recordings in case of intermittent network coverage.

### 3.2. Server

The server provides logic control to manage therapy for multiple children, create and store therapy exercises, and store incoming speech recordings from children. It also hosts the speech analysis engine; upon receipt of each recording, the task manager invokes the appropriate speech-analysis routines and stores the results on a centralized relational database. The database also provides storage for the profile for each child, including each child's history of speech recordings, results of speech processing, and SLP's assessments and annotations.

The server runs Moodle [Mich 2009], an open-source learning management software, which acts as the task manager. Moodle's modular and object-oriented architecture facilitates information sharing and integration with other software. It also provides course management functionalities such as user profiles, course pages, secure access by the user, and scheduling of events. Moodle is an example of a LAMP stack (Linux, Apache, MySQL, and Perl) [Moore and Churchward 2010]. It comes with a web server (Apache), a database (MySQL) and a scripting interpreter (PHP). In our current version, the task manager and other applications run on a Linux machine with Ubuntu Server OS 12.04.

Our framework makes extensive use of Moodle's web services and core APIs (Figure 5). *Web services* allow us to create fully parameterized generic methods and provide seamless communication between the server, mobile client, and external applications. For example, the mobile client uses web services to download therapy exercises from the server, and to upload the recorded speech files in the audio repository on the server for speech analysis. Along with these, we have developed a web service to facilitate real-time feedback from the speech-analysis engine, which allows us to provide rewards to the child based on progress during a therapy session. In turn, Moodle's *core APIs* provide a number of tools for data manipulation, enrollment, secure access management, reporting, and so forth. Our implementation uses the data manipulation

APIs to manage therapy courses, enrollment APIs to manage enrollment of children to these courses, access APIs to provide secure access to the SLPs, and file APIs to store NDP3 images and related files into the server. We have also developed a reporting API that provides data for the reporting UI in the clinician interface.

Communication between the Moodle database and the speech-processing module is facilitated by a background process that we have created for this purpose. The background process avoids software dependency and concurrency issues between the Apache server and the speech-processing module, and allows asynchronous communication. It periodically queries the audio repository for recent uploads of audio files, invokes the speech-analysis scripts, and uploads the results to the Moodle database.

### 3.3. Clinician Interface

The clinician UI provides three basic functionalities: managing children, creating new exercises, and monitoring progress, all done remotely over the web. The *child management* functionality allows the SLP to add/remove children and view their profiles. Each profile stores the name, picture, and date of birth of the child, parent's name and contact, and time/date of last access. It also stores all the courses to which the child has been enrolled and the currently available (unenrolled) courses. The interface also provides the means for the SLP to *create (or edit) exercises* based on the NDP3 protocol. The exercise-creation functionality consists of a canvas, which allows the SLP to create exercises using a drag-and-drop selection of stimuli (i.e., NDP3 images), and a textbox to provide comments and instructions for each exercise.

The third functionality, *viewing results and reports,* allows the SLP to analyze the performance of each child, play recordings of individual utterances, and produce comprehensive reports. A profile page within the interface provides a summary of the child's progress along the various courses assigned by the SLP (Figure 6(a)). Detailed performance results can be accessed by clicking on the course links in the summary (Figure 6(b)). Should the SLP need it, the recording of every attempt made by the child is also accessible through a drop-down menu. Finally, the interfaces can generate a comprehensive summary of the child's progress as a PDF document, including the number of weeks into therapy, number of sessions in clinic and at home, tabulated and graphical representations of the child's progress, and any additional comments that the SLP wishes to provide.

The clinician's UI is implemented using JavaScript, PHP, YUI, and HTML5 (Figure 5). The user/course/exercise creation and reporting pages are built using HTML5, JavaScript, and YUI (Yahoo interface library). The drag–drop box interface on the exercise creation page uses YUI. YUI is an open-source JavaScript and CSS library that uses techniques including Ajax and DOM scripting/handling to build interactive web applications. Finally, the functionality for data handling and communication with the server is provided by PHP.

### 3.4. Speech Analysis

The speech-analysis module identifies errors made in the child's utterance based upon criteria recommended by ASHA for the analysis of CAS [ASHA Ad Hoc Committee on Apraxia of Speech in Children 2007; Forrest 2003], and quantifies them for presentation to the SLP. The three segmental and suprasegmental features of CAS validated by ASHA are (1) inconsistent errors on consonants and vowels in repeated productions of syllables or words and differential use of a certain phoneme or sound class in different word positions (i.e., inconsistency and variability) [ASHA Ad Hoc Committee on Apraxia of Speech in Children 2007]; (2) lengthened and disrupted coarticulatory transitions or struggle between sounds and syllables (i.e., articulatory struggle) [Forrest 2003], and (3) inappropriate prosody, especially in the realization of lexical

(a)

| Courses | Percentage Completion | Average Correctness | Average Delay |
|---|---|---|---|
| CV Fricative Mixed | 75% | 24.42% | 0.26s |
| CV Words | 18.09% | 21.78% | 0.29s |
| CV fronting pairs | 100% | 4.08% | 0.04s |

(b)

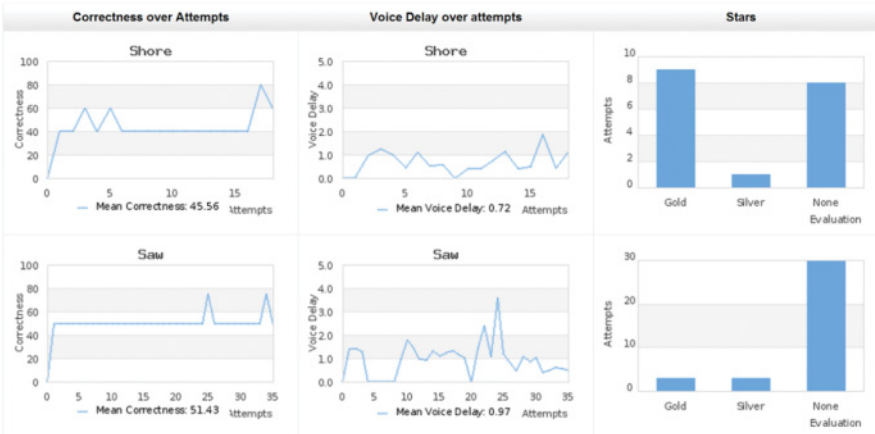


Fig. 6. Reporting screen for a child on the clinician interface.

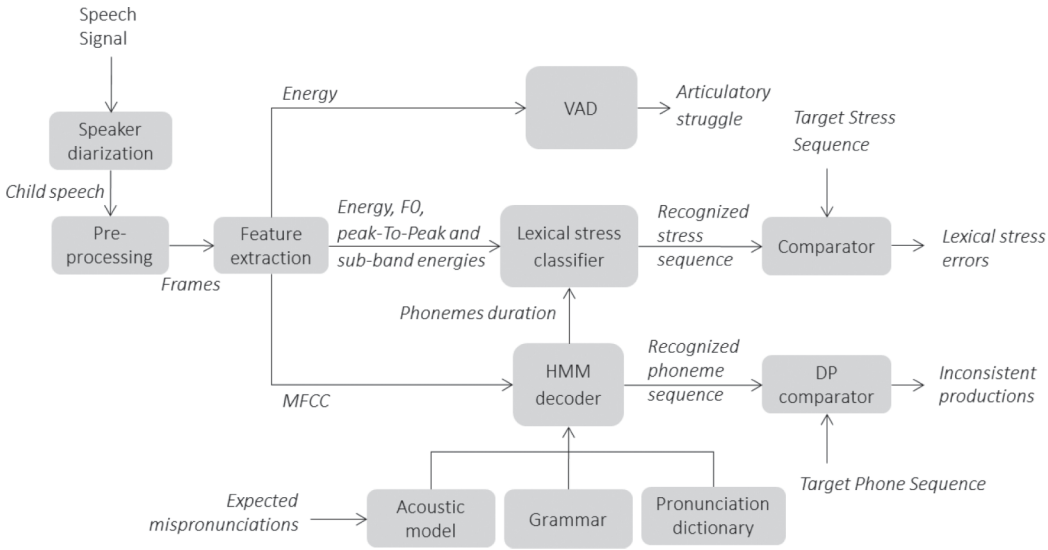


Fig. 7. Block diagram of the speech-analysis engine.

stress [Shriberg et al. 2003; Ballard et al. 2010]. Figure 7 shows a block diagram of the speech processing steps used to identify these errors. These steps include:

—**Speaker diarization (SD):** As the child may use the client app with the help of an SLP or caretaker, the recorded sample may contain speech segments from these

other speakers. The goal of the SD module is to detect the presence of multiple speakers in the recording and extract those portions that contain the child's speech. For this purpose, the recording is first divided into speech segments according to the silence positions and then Mel Frequency Cepstral Coefficients (MFCCs) features are extracted for each segment. The distance between all combinations of segments is calculated using a weighted Euclidian distance [Kwon and Narayanan 2002]. Segments with a distance exceeding a certain threshold are marked as belonging to different speakers. These segments are then classified by measuring the Euclidian distance between the average MFCC coefficients of each group of segments that are recognized to be from the same speaker and the average MFCC coefficients of the segments recorded from the child and the parent/SLP. Each group of segments is then classified as belonging to the class for which the Euclidian distance to the corresponding reference segment is the shortest.

—*Preprocessing:* This stage removes the DC offset, applies a pre-emphasis filter, and segments the speech signal into 25ms frames (15ms overlap).

—*Feature extraction:* Several kinds of features are extracted for each frame. The average energy is calculated and used for voice activity detection. The maximum and average pitch, peak-to-peak amplitude and Bark-scale subband energies [Zwicker 1961] are also calculated and fed to a lexical stress classifier along with the average energy. Finally, MFCCs are extracted and fed as input to a speech decoder.

—*Voice activity detector (VAD):* An energy-based VAD is used to discriminate between speech and nonspeech (silence) segments on a frame-by-frame basis based on average energy. The energy-based VAD is well known and achieves high accuracy in recordings with high signal-to-noise ratio [Kristjansson et al. 2005]. For each recording, a silence threshold (T) is calculated as:

$$T = P(05) + 0.2 \times [P(95) - P(05)], \tag{1}$$

where P(05) and P(95) are the 5th and 95th percentile values of the average frame energy over the entire speech recording for each speaker. All frames whose average energy exceeds this threshold are marked as speech, and all other frames are marked as silence [Boersma 2002]. Frames identified by the VAD as containing nonspeech indicate the presence of articulatory struggle, specifically silent struggle while preparing for speech.

—*Lexical stress classifier:* This module is used to classify the child's productions into strong–weak (SW) and weak–strong (WS) stress patterns. Most of the existing work on stress detection focuses on detecting the most stressed syllable in the pronunciation of a multisyllabic word [Li et al. 2013], not lexical stress patterns, that is, SW or WS patterns. There is very limited work on the automatic classification of biphonemic stress patterns. Kim and Beutnagel [2011] implemented four different machine-learning algorithms to classify different lexical stress patterns of 3- and 4-syllable English words; when compared on an adult female healthy speaker, the authors report an accuracy of 83.3%. The highest accuracy was achieved with Support Vector Machine (SVM) and Maximum Entropy (MaxEnt) classifiers. Accordingly, we compared the performance of three different classifiers: SVM, MaxEnt and a multilayer perceptron, with the best accuracy obtained using the multilayer perceptron. The input feature vector consisted of the following acoustic measures: mean and maximum energy over nucleus (the syllable vowel), mean and maximum pitch over nucleus, peak-to-peak amplitude over nucleus, durations of the syllable and nucleus, and 21 Bark-scale subband energies [Shahin et al. 2012]. Our choice of features was motivated by the nature of lexical stress production, which occurs through variations in syllable duration, intensity, and fundamental frequency [Fletcher 2010]. The features related to the energy, pitch, and duration of the syllable nucleus (vowel) are

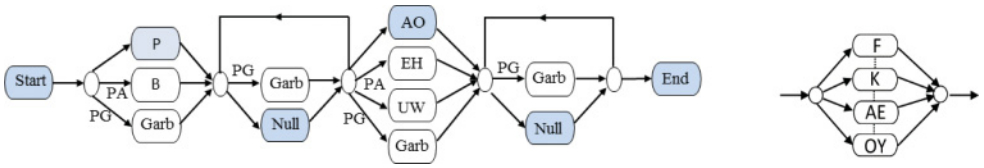


Fig. 8. (a) Lattice example for the word "paw." Garb denotes a garbage node, and filled nodes represent the correct phoneme sequence. (b) Construction of the garbage node.

commonly used in speech stress classification algorithms [Tepperman and Narayanan 2005]. Energy subband features have been used successfully in other stress detectors [Chen and He 2007; Rahurkar et al. 2002]. Duration information is obtained from the decoder, which provides the recognized phoneme sequence along with time boundaries. A pairwise variability index (PVI), a common method for measuring degree of lexical stress contrastivity across languages [Fletcher 2010], is calculated for each acoustic measure to determine the degree of asymmetry across pairs of neighboring syllables and to make the features speaker-independent [Ballard et al. 2010]. Given acoustic feature $x_i$, its PVI is given by:

$$PVI_i = \frac{x_i^{(1)} - x_i^{(2)}}{\left(x_i^{(1)} + x_i^{(2)}\right)/2}, \tag{2}$$

where $x_i^{(1)}$, $x_i^{(2)}$are the acoustic features of the first and second (consecutive) syllables, respectively. The classified stress patterns are then compared against the correct (target) stress patterns to identify any lexical stress errors made by the child [Shahin et al. 2012].

—*HMM decoding:* Mispronunciations in the child's utterance are detected by means of a hidden Markov model (HMM) decoder. Currently, there are two state-of-the-art approaches for phone-level pronunciation verification (PV). The first approach, the confidence-based PV method, computes an evaluation value for each expected phone [Yin et al. 2009]. The second approach, lattice-based, creates a mispronunciation lattice and/or dictionary, and aligns the produced speech to the best phoneme sequence in the lattice. Mispronunciation lattices are used widely for PV in computer-aided pronunciation-learning (CAPL) applications, particularly in second-language learning [Harrison et al. 2009]. The advantage of this method is that it can not only detect the occurrence of pronunciation error but also specify the type of the mispronunciation error. We implemented the lattice structure, as it can give detailed feedback to the therapist regarding the articulation errors detected in the child's speech. We use a grammar lattice consisting of the correct phoneme sequence (known from the exercise's prompt) and expected mispronunciations of each phoneme (generated by an SLP after assessment of 20 children with CAS). Figure 8 shows an example of a lattice created for the word "paw." For each phoneme, a set of expected alternative phonemes are inserted as parallel arcs to collect any expected substitution errors, while the unexpected substitution errors and the insertion errors can be collected by the garbage node. A null arc is also added to catch any occurrence of deletion errors. Penalty costs PA and PG are added to the alternative and garbage nodes, respectively, to prevent the decoder from aligning the speech to the alternative or garbage nodes unless it is confident enough. The generated lattice is then used by an HMM decoder along with acoustic models to generate a sequence of phonemes from the child's utterance. The HMM acoustic models consist of tied-state triphones

trained using 40h of child speech corpus from Oregon Graduate Institute of Science and Technology (OGI) [Shobaki et al. 2000]. The recognized phoneme sequence is then compared to the target phoneme sequence through a dynamic-programming string-alignment procedure using the HResults tool in the HTK toolkit [Young et al. 2006]. Three kinds of mispronunciations are identified—insertion, deletion, and substitution mispronunciations—and used to identify inconsistent and variable speech in the child's productions.

## 4. VALIDATION STUDIES

We designed a user study to determine if our proposed system would be a reliable alternative to traditional therapy. We wanted to observe the kinds of interaction between the child and the tablet app in the context of a speech therapy session, and explore the advantages (and shortcomings) of tablet-based therapy as compared to traditional paper-based exercises. The study also served as a preliminary trial comparing the efficacy of speech therapy delivered via the tablet in two modes:

—*SLP4:* Therapy directed by a speech-language pathologist (SLP) 4 days/week for 3 weeks, versus
—*SLP1:* Therapy directed by the SLP 1 day/week for 3 weeks and by the child the other 3 days/week.

In both cases, all activities were presented via the tablet. The primary difference between both modes was the additional opportunity in the SLP4 mode for detailed feedback from SLP to child on error correction strategies, that is, knowledge of performance (KP). As some studies have suggested [Schmidt and Lee 2005; Ballard et al. 2012], KP may be unnecessary or even interfere with long-term learning relative to simply providing feedback on correct/incorrect production, that is, knowledge of results (KR). On the contrary, other studies have found that KP is beneficial when the learner has a poor internal representation of the target behavior or has less well-developed error detection and correction skills [Newell et al. 1990; Maas et al. 2008]. The SLP1 mode is equivalent to standard care, in which a child has therapy with an SLP once a week but then completes homework alone or with the parent 3 or 4 days a week. In this mode, feedback may be limited to KR feedback. Based on this prior knowledge, we posed the following two hypotheses:

—*H1:* Children, parents, and SLPs express increased satisfaction with a tablet-based delivery over their prior experiences with paper-based activities.
—*H2:* Children improve by a similar degree from pre- to posttreatment, regardless of the treatment mode (SLP1 and SLP4), but children in the SLP4 mode may show marginally better improvement and/or retention due to the facilitative effect of KP feedback when the internal representation of the target behavior is poor.

In what follows, we describe our experimental protocol, participant demographics, and results from the study. It is beyond the scope of this article to include a full-scale efficacy study comparing these two methods. Such a study is currently underway and will be reported separately. The pilot data are summarized here and provide preliminary support for our hypotheses.

### 4.1. Participants and Design

Eight children (7 male, 1 female) aged 4 to 10 years and diagnosed with CAS were recruited for the study (Table I). Inclusion criteria included no reported impairment in cognition, language comprehension, hearing or vision, orofacial structure, or lower-level movement programming/execution (i.e., dysarthria). A comprehensive assessment battery was completed with each child prior to commencing treatment to rule out these

Table I. Age, Sex, and Mode of Therapy Received by the Seven Children in the Trial

| Participant ID | Gender, Age | Mode |
|---|---|---|
| C1 | Male, 10 yr | SLP4 |
| C2 | Male, 10 yr | SLP1 |
| C3 | Male, 7 yr | SLP4 |
| C4 | Male, 7 yr | SLP1 |
| C5 | Male, 4 yr | SLP4 |
| C6 | Male, 4 yr | SLP4 |
| C7 | Female, 10 yr | SLP1 |
| C8 | Male, 5 yr | SLP1 (withdrew) |

conditions and confirm the CAS diagnosis [ASHA Ad Hoc Committee on Apraxia of Speech in Children 2007]. Hence, the SLPs were aware of the speech sound skills and disabilities of each child. Likewise, the children had prior experience with traditional picture-card–based speech therapy but no experience with the tablet-based therapy. We received approval from the Institutional Review Board prior to the study. During the study, parents were given the option to accompany the children for the session or leave the clinic; one of the parents opted to leave the child with the SLP for the session. No compensation was provided to the participants.

Children were matched in pairs by age and the children in a pair allocated randomly, without replacement, to one or the other treatment mode (SLP4, SLP1). One child withdrew after the first week of treatment due to illness. Consistent with the NDP3 program [Murray et al. 2012], each child was assigned three treatment goals for practice with the tablet, based on individual speech error patterns. Each child had a goal(s) addressing consonant or vowel accuracy in words (e.g., initial consonant sounds *k* and *s*, in CVC words) and a goal(s) on stress production in multisyllabic words or phrases (e.g., stress the appropriate syllable in the word ***di****nosaur* vs. *po****ta****to*). Five words were selected from the NDP3 stimuli to work on each goal. The data reported here are part of a larger ongoing randomized clinical trial testing these two training methods, in which children retested at three time points: pretraining and immediately and 1 month posttraining to measure acquisition and retention of new skills.

### 4.2. Experiments

During the experiments, each child underwent an NDP3 session with the SLP using a tablet (Samsung Galaxy Tab 2 10.1, Android 4.1) running the CAS mobile app. Prior to the start of the session, the SLP briefed the participants about the goal of the study, and then asked the parent(s) to sign a consent form. The tablet-based therapy sessions were similar to previous sessions with the SLP, with the exception that paper-based exercise sheets were replaced by the tablet and exercises were assigned using the SLP's interface on the Moodle server based on the NDP3 protocol.

At the start of each session, the child would sit with the SLP/parent with the tablet placed on a desk in front of them. The SLP/parent would explain the current activity (single sound, transition, and so on) to the child, following which the speech practice would start.

All sessions involved practicing the words for each goal, with goal and word order randomized each session. Each goal was practiced for 15min, with short breaks, with about 100 child responses per session. In the SLP4 condition, all 12 sessions had the same format: the SLP gave KR feedback on every trial and gave KP feedback to shape correct production on error trials. In the SLP1 condition, the SLP followed the protocol for SLP4 for the first session each week but gave KR feedback only on every trial for the remaining three sessions each week. As the ASR module was not fully functional for

Table II. Exit Questionnaire for Children

| Child questionnaire |
|---|
| Did you enjoy using the tablet for the speech therapy activities? |
| Did you need any help completing the activities on the tablet? |
| Were you able to maintain focus/attention on the exercises? |
| How motivating did you find the therapy sessions? |
| What did you like about the exercises on the tablet? |
| What did you dislike about the exercises on the tablet? |
| How easy was it to fit the therapy program into your daily life? |
| In the future, would you prefer to do these exercises using tablet or paper worksheets? |
| If tablet-based exercises were available to you, how often would you want to use them? |

Table III. Exit Questionnaires for SLP and Parents

| SLP and parent questionnaire |
|---|
| Did you/your child enjoy using the tablet for their speech therapy activities? |
| Did you/your child need any help completing the activities on the tablet? |
| Was your child able to maintain focus/attention on the exercises? |
| How motivating did your child find the therapy sessions? |
| What did you/your child like about the exercises on the tablet? |
| What did you/your child dislike about the exercises on the tablet? |
| How easy was it to fit the therapy program into your daily life? |
| How satisfied are you with the child's speech progress during the therapy program? |
| In future would you prefer to do these exercises using: tablet or paper worksheets? |
| If tablet-based exercises were available to you, how often would you want to use them with your child? |

Table IV. Survey Results Indicating the Proportion of Participants Who Answered "Yes"

| Question | Child | Parent | SLP |
|---|---|---|---|
| Did you enjoy using the tablet? | 6/6 | 6/7 | 2/3 |
| Did the child need help? | 3/6 | 4/7 | 1/3 |
| Was the child able to maintain focus? | 6/6 | 4/7 | 2/3 |
| Did you prefer tablet over paper? | 6/6 | 6/7 | 3/3 |
| Were the tablet sessions motivating? | 4/6 | 6/7 | Neither |

providing automated feedback to the child on response accuracy, the SLP was always present in both modes to indicate correct/incorrect production (i.e., KR) for each trial by selecting a gold (correct) or silver (incorrect) star on the display. The children, parents, and SLPs completed usability surveys (Table II and Table III) after the treatment to indicate their satisfaction with tablet-based presentation of activities.

## 5. DISCUSSION OF RESULTS

### 5.1. Survey Results

The survey was completed by 6 children (of the 7 children who participated in the study), by a parent for all 7 children, and by 3 SLPs. Overall, and consistent with hypothesis H1, we found that all participants (children, parents, and SLPs) enjoyed working with the tablet, as summarized in Table IV. In addition, the adult participants noted that the majority of the children were able to maintain focus on the activities.

When asked "What did you/your child like about the exercises on the tablet?" the most common response from the children was "I liked listening to myself and recording." Child C1 liked that "it (the app) taught learning how to say it (words)" while child C3 liked the pictures "the dinosaur, museum" in reference to the images in the exercises. Child C4 commented, "you get to do stars and get to play memory game" in reference to the stars and badges (rewards given on completing an exercise). Child C5 also liked the memory game and commented that it was "fun because you can move around." The SLPs and parents also indicated that being able to record/playback their speech was very appealing. SLP T1 liked the memory game and audio playback. SLP T2 and T3 both liked the stars and audio playback as well as recording. The parents also liked the ability to record and play back the speech. In addition, they liked the rewards (stars), the ease of use, and graphics. Parent P5 liked "that the tablet was used to facilitate the speech exercises." Along the same lines, Parent P6 commented that his child "loves anything that involves tablets/computers, etc., the interactivity of it." P7 also liked the system and commented that is was "a new and interesting way to combine his speech work with a device he enjoys using."

When asked "How motivating did your child find the therapy sessions?" 4 out of 6 children found the sessions to be either motivating or highly motivating while 2 children found the sessions to be discouraging. Six out of 7 parents found the sessions to be motivating. The SLPs, however, had a different viewpoint toward the tablet-based therapy system. Two of the three SLPs found the sessions to be between "neither motivating" and "motivating" while one found it motivating. SLP T1 commented that "for young children, particularly, the activities are not interesting enough; the child did not know/was not familiar with tablets." SLP T2 commented that "rewards need to be bigger or more colors and more motivating. Pictures need to move" in reference to the rewards and animations.

When asked "Did you or your child require any help completing the activities on the tablet?" the majority of children reported that they did not require any assistance. Two of the children needed help with accessing the audio model, selecting an exercise, and navigating back to home screen. One of the SLPs (T3) responded that a child needed continued help in using the application, that is, "selecting an exercise; moving between images/activities; starting/stopping the recording; accessing the audio model." She suggested that this might be due to the child's age (C6 is one of the youngest children in the study). Four out of 7 parents found that their child needed help in completing the exercise on the tablet. They observed that children needed assistance in starting/stopping the recording and selecting an exercise. One parent (P3) noted that his child wanted to cheat and move through the exercises, and commented that "The app seemed within his ability, but he needed a lot of support to apply himself repetitively to the task and complete all the requirements." In contrast, one of the parents (P4) said that "only minimal instructions were told and my son easily picked them up." Generally, children needed some initial demonstration and help if they got confused with the UI, but otherwise were able to perform the tasks independently. The system in its current form seems suitable for children to use in their homes.

When asked "What did you/your child dislike about the exercise on the tablet?" Child C3 commented "no better with mic" in reference to the external microphone used for accurate/noiseless audio capture. Child C4 said that "you have to keep going back out and into a different area," indicating a scope for improvement in the mobile app UI. Parents P3, P6, and P7 found the repetitive nature of the exercise on tablet to be difficult, as evident from P7's comment: "the amount of times he had to repeat a single word before moving to the next." Parent P5 commented that "he found it difficult to concentrate at times; he wanted to play with the tablet and explore it." One parent, P3, suggested that "having some humor in the exercise will help." He

also found the rewards to be quite abstract and showed preference for more physical rewards/activities, for example, "putting a card in a box, rubbing out a picture on the blackboard."

When asked "What could we do to make the overall application/system more usable?" Child C2 said "enjoyed it but took a lot of time. It was hard to fit in. I would have rather come in every two days. I liked working with the clinicians," in reference to the frequency of visits to/sessions with the SLP. Child C5 made a comment about the external microphone: "microphone–hurts ears and tangles in hair." Parent P5 commented on the need for a visual showing speed and emphasis: "maybe a visual on the tablet which shows the speed or emphasis on the parts or chunks of the word." Parent P6 commented on the importance of being able to get therapist feedback: "I would love to be able to use this kind of program/work at home, but at times I found it difficult to know if her answer was correct or not." Along the same lines, Parent P2 commented: "It was very difficult watching—struggle + dealing with his frustration levels on the day when he had to use the tablet and receive no feedback from the therapist." One of SLPs (T1) noted an issue with the manual assessment panel: "problem with stars; there weren't enough stars" to accommodate for assessment of multiple utterances corresponding to a prompt. SLP (T2) commented that "some images were missing." In addition, she commented that "prompts in the exercise needs to (be) audio as well as text for nonreaders." Our application already presents the prompts in both audio and text format, but perhaps this was not readily apparent. SLP (T2) also pointed out the lack of a scheduling feature and commented: "I would want an alert to tell the child to stop practice if s/he is off target. Maybe cue the child to stop practice after making 3 or 4 errors on one item." SLP T1 also noted that "the child seemed to not know that she could select and enlarge different images to record." These points indicate some usability issues with the mobile application that have been addressed in the most recent version of the app.

When asked "In the future, would you prefer to do these exercises on the tablet or paper-based worksheets?" the three SLPs showed preference for the tablet-based therapy. These sentiments were echoed by the parents and children, with the majority of parents (6/7) and children (6/6) showing preference for the tablet. Child C3 gave the reason behind this, saying that it was easier to work on tablet, while C2 said: "because I can listen to myself and play it back; more interesting." Child C4 said: "because you can't do the memory board on the cardboard." Parent P3 said: "motivation and focus are a real issue. Paper is more adaptable and provides more variety," justifying his preference for paper-based exercises. In contrast, Parent P4 was in favor of tablet-based therapy and said: "more on the tablet although paper cards/worksheets are quite helpful but because of the convenience it could be a helpful tool in augmenting speech therapy for CAS." On a similar note, P5 said: "using the tablet was definitely more motivating especially when he was able to record himself and give himself a tick." P6 said: "being able to hear the word is far more helpful than just an image." SLP T1 said that tablets are "easy to use; saves time making flashcards; audio playback is easy to use." SLP T2 noted that "it is easier to users from home.". Overall, children, parents, and SLPs in our pilot study had a positive response towards the tablet when compared to paper-based CAS therapy.

### 5.2. Testing and Dependent Measures

Speech sound and stress production accuracy were measured six times: once prior to treatment, once a week during treatment, and 1 week and 1 month posttreatment. Data from the pretreatment and the two posttreatment tests are presented here as percent change from baseline (Figure 9). Due to the small sample size, individual data for each child is discussed. We hypothesized (H2) that the children receiving SLP4

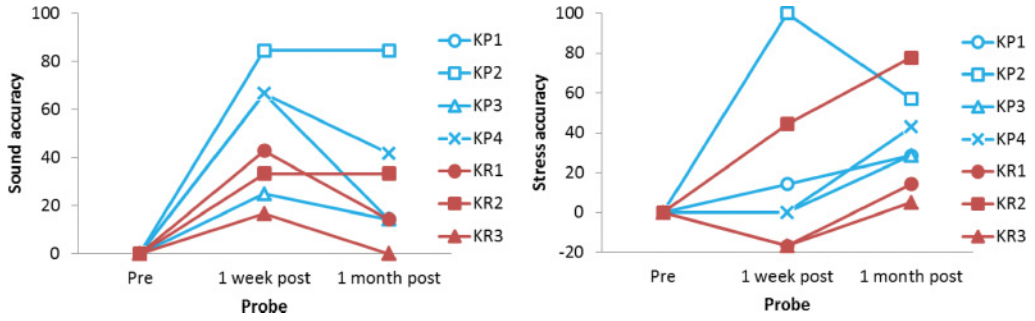


Fig. 9. Change in percent accuracy on (a) speech sounds in words and (b) stress production in multisyllabic words for children receiving either the SLP4 or the SLP1 mode of treatment with the tablet application. Probes represent pretreatment (1), 1 week posttreatment (2) and 1 month posttreatment (3).

would show similar improvement and retention of gains as children receiving SLP1 due to the facilitative effects of KP feedback when the internal representation of target behaviors is undeveloped. Our results show a small benefit for the SLP4 (KP) over the SLP1 (KR) condition at 1 week posttreatment. While there was considerable variability in the magnitude of children's responses to treatment, the two conditions performed similarly at 4 weeks posttreatment (Figure 9), validating hypothesis H2. While a larger-scale study is needed, these initial findings support the tablet-based therapy with just once-a-week contact with the SLP for detailed feedback to guide learning, a mode that allows high-intensity practice but reduced strain on SLP resources and family budget.

In summary, when therapy focused on improving speech sound accuracy, the four children in the SLP4 group showed substantial improvement from pretreatment to 1 week posttreatment (median 70% change from baseline; Figure 9(a)). The SLP1 group showed a smaller percent change (median 40%), consistent with the lesser guidance and support from the SLP in the KR feedback mode. At 4 weeks posttreatment, accuracy declined a similar degree for both groups but remained above baseline. When therapy focused on production of stress patterns in multisyllabic words, the SLP4 children showed small improvement from pretreatment to 1 week posttreatment (median 20%) while the SLP1 children showed no change or even slight deterioration (−10%; Figure 9(b)). However, both groups improved substantially from 1 week to 1 month posttreatment.

The children responded better, in the short-term, to the sound production activities compared with the stress production activities. Sound goals emphasize spatial accuracy of articulator movement while stress goals emphasize temporal accuracy. Ballard et al. [2010] have shown that temporal control of speech is particularly challenging for children with CAS, and stress errors often persist after sound accuracy has normalized. Similar to McCabe et al. [2014], however, stress gains continued improving to 1 month posttreatment, while sound accuracy tended to deteriorate. The reason for this is not yet clear but is unrelated to the use of the tablet. Although these results are encouraging, we are recruiting a larger sample to confirm these findings.

### 5.3. Speech Processing

The speech module was tested on disordered speech and results compared against manual markings made by the therapist. Table V summarizes the performance for the four main modules. The diarization module was assessed by its ability to classify segments containing either SLP or child speech. Voice activity detection was considered to be correct if the difference between the estimated timing (delay in voicing, total production time) and ground truth was less than 100ms. Results for the lexical stress

Table V. Performance Evaluation of the Speech Processing Modules

| Module | Evaluation | |
|---|---|---|
| Speaker diarization | Accuracy | 74% |
| VAD | Delay in voice | 96% |
| | Total production time | 94% |
| Lexical stress classifier | SW | 78% |
| | WS | 77% |
| | Overall | 78% |
| HMM decoder | Phone level accuracy | 89% |

detector were computed individually for each pattern (SW, WS) and overall; the two SLPs had a strong level of agreement (>90%).

Finally, results from the HMM decoder were based on recognition rates at the phoneme level. Overall, these results indicate that the speech modules are able to accurately identify voicing delays in children's productions as well as pronunciation and prosodic errors, the main error types associated with CAS.

## 6. CONCLUSION

Technology-based interventions have tremendous potential in improving the delivery of speech therapy. They can also help overcome geographical barriers and address the issues of shortage of SLPs and equipment. In this article, we have presented an automated speech therapy system for childhood apraxia of speech. The system provides mechanisms to not only deliver therapy exercises but also to remotely monitor the child's performance and modify the child's therapy regimen.

We reviewed CAS and existing work on therapy tools for CAS, including the NDP3 protocol. We then described our system architecture, including the software modules used for developing the mobile therapy app, speech analysis engine, and SLP's interface. Finally, we conducted a user study to validate the system and discussed results from a semistructured interview of the participants. Overall, we found that the tablet app was widely liked by the participants. Feedback from study participants indicates that additional features such as animations, rewards, and audio and visual aid on the tablet would make the therapy more engaging and speed up the learning process. Implementation of these features is currently underway for the next release of the software. We also have plans to incorporate games and puzzles as part of the speech therapy; we believe that these improvements will further increase the appeal of (and compliance with) regular practice.

In contrast to existing mobile tools [Maier et al. 2010; Bunnell et al. 2000; Wren et al. 2006; Vilozni et al. 2001], which are stand-alone apps, our system includes an automated speech analysis engine that provides quantitative speech assessment results to the SLPs. This enables the SLP to remotely monitor progress and adapt the therapy regimen as needed. This is particularly advantageous because each child and his/her speech disability is unique, and therefore requires individualized care. Results from our pilot study support the feasibility of our system as a supplement to traditional face-to-face speech therapy.

We hypothesized that children, parents, and SLPs would express increased satisfaction with a tablet-based delivery over their prior experiences with paper-based activities (hypothesis H1). In general, feedback received from the users was positive, which validates our hypothesis. Most children found it engaging and fun, and particularly enjoyed listening to their own productions. The memory game was enjoyed; the intention is to continue adding therapeutic games to increase variety and continued engagement in therapy multiple times a week. Children may need training to increase

independence in accessing the audio model and starting/stopping self-recordings. While some parents found the tablet activities repetitive, this was not identified as an issue for children or SLPs. Further refinement of the tablet is warranted given that 88% of respondents indicated they would use the tablet for therapy multiple times a week. This supports a primary goal of increasing the amount of weekly/daily practice in this population, which requires intensive practice over months or years to normalize speech production [ASHA Ad Hoc Committee on Apraxia of Speech in Children 2007]. Trials of child/parent-directed therapy at home using the tablet are underway. In addition, we argued that the two modes of therapy (SLP1 and SLP4) would have similar effects on the children's performance (hypothesis H2). From the small samples tested here, both modes of treatment had similar effects on the performance from pretreatment to 1 month posttreatment. This is encouraging given the many barriers to providing intensive practice in the clinic and the current standard mode of care being once a week with in-home practice.

Although most participants in the study enjoyed working with the system, they also provided information about areas for improvement. One of the SLPs noted that the activity with a tablet may get boring with time. More game-like features and animations would increase the patient's interest, especially for younger children. Further, constant evolution of the app is also necessary to maintain the novelty factor, which is important to keep children engaged in the long term [de Sá et al. 2012]. Another point worth noting is that the proposed therapy tool requires parents and clinicians to have knowledge of using tablets and computers. This may be a concern for some of the participants for whom computers/tablets are still an unfamiliar technology, though we believe that this can be addressed by providing prior training. During our experiments and consequent discussions, we saw a need for standards and guidelines to ensure that a remote mobile application, such as ours, does not compromise the standard of therapy compared to clinical settings. Finally, the training experiment involved a small sample of children, thus statistical comparison of performance across the two groups was not possible and the potential influence of variables such as child's age or impairment severity could not be explored. It is possible that the results of the ongoing larger study will identify statistical differences in these two training methods.

## REFERENCES

L. Anthony, Q. Brown, J. Nias, B. Tate, and S. Mohan. 2012. Interaction and recognition challenges in interpreting children's touch and gesture input on mobile devices. In *ACM Conference on Interactive Tabletops and Surfaces*. 225–234.

*ArtikPix*. Retrieved October 1, 2015 from http://rinnapps.com/artikpix/.

ASHA Ad Hoc Committee on Apraxia of Speech in Children. American Speech-Language-Hearing Association. 2007. Childhood apraxia of speech [Technical Report]. Available from www.asha.org/policy.

K. J. Ballard, D. A. Robin, P. McCabe, and J. McDonald. 2010. A treatment for dysprosody in childhood apraxia of speech. *Journal of Speech, Language, and Hearing Research* 53, 1227–1245.

K. J. Ballard, H. D. Smith, D. Paramatmuni, P. McCabe, D. G. Theodoros, and B. E. Murdoch. 2012. Amount of kinematic feedback affects learning of speech motor skills. *Motor Control* 16, 106–119.

P. Boersma. 2002. Praat, a system for doing phonetics by computer. *Glot International* 5, 341–345.

H. T. Bunnell, D. M. Yarrington, and J. B. Polikoff. 2000. STAR: Articulation training for young children. In *International Conference on Spoken Language Processing*. 85–88.

N. Chen and Q. He. 2007. Using nonlinear features in automatic English lexical stress detection. In *International Conference on Computational Intelligence and Security Workshops, 2007 (CISW'07)*. 328–332.

G. A. Constantinescu, D. G. Theodoros, T. G. Russell, E. C. Ward, S. J. Wilson, and R. Wootton. 2010. Home-based speech treatment for Parkinson's disease delivered remotely: A case report. *Journal of Telemedicine and Telecare* 16, 100–4.

A. L. Delaney and R. D. Kent. 2004. Developmental profiles of children diagnosed with apraxia of speech. Presented at the Annual Convention of the American-Speech-Language-Hearing Association.

M. de Sá, L. Carriço, J. Faria, and I. Sá. 2012. Children psychotherapy with mobile devices. In *Human-Computer Interaction: The Agency Perspective*, Studies in Computational Intelligence, M. Zacarias and J. V. de Oliveira, eds. Springer, 85–109.

J. Fletcher. 2010. The prosody of speech: Timing and rhythm. In *The Handbook of Phonetic Sciences* (2nd ed.),W. J. Hardcastle, J. Laver, F. E. Gibbon, eds. Wiley, Hoboken, NJ, 521–602.

K. Forrest. 2003. Diagnostic criteria of developmental apraxia of speech used by clinical speech-language pathologists. *American Journal of Speech-Language Pathology* 12, 376–380.

J. Froehlich, J. Wobbrock, and S. Kane. 2007. Barrier pointing: Using physical edges to assist target acquisition on mobile device touch screens. In *ACM SIGACCESS Conference on Computers and Accessibility*. 19–26.

R. Gaines, C. Missiuna, M. Egan, and J. McLean. 2008. Educational outreach and collaborative care enhances physician's perceived knowledge about Developmental Coordination Disorder. *BMC Health Services Research* 8, 1–9.

A. Georgeadis, D. M. Brennan, L. N. Barker, and C. R. Baron. 2003. Telerehabilitation and its effect on story retelling by adults with neurogenic communication disorders. In *Clinical Aphasiology Conference*. 639–652.

A. M. Harrison, W.-K. Lo, X. Qian, and H. Meng. 2009. Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In *SLaTE*. 45–48.

D. G. Jamieson, G. Kranjc, K. Yu, and W. E. Hodgetts. 2004. Speech intelligibility of young school-aged children in the presence of real-life classroom noise. *Journal of the American Academy of Audiology* 15, 7, 508–517.

Y.-J. Kim and M. C. Beutnagel. 2011. Automatic assessment of American English lexical stress using machine learning algorithms. In *SLaTE*. 93–96.

H. Kolles and W. Feiden. 1995. Computer-assisted speech recognition in diagnostic pathology. Development of the DragonDictate. *Pathologe* 16, 6, 439–442.

T. Kristjansson, S. Deligne, and P. Olsen. 2005. Voicing features for robust speech detection. *Entropy* 2, 3.

S. Kwon and S. S. Narayanan. 2002. Speaker change detection using a new weighted distance measure. In *International Conference on Spoken Language Processing (ICSLP'02).* 2537–2540.

K. Li, X. Qian, S. Kang, and H. Meng. 2013. Lexical stress detection for L2 English speech using deep belief networks. In *INTERSPEECH*. 1811–1815.

E. Maas, D. A. Robin, S. N. A. Hula, S. E. Freedman, G. Wulf, K. J. Ballard, et al. 2008. Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology* 17, 277–298.

A. Maier, T. Haderlein, F. Stelzle, E. Nöth, E. Nkenke, and F. Rosanowski, et al. 2010. Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. *EURASIP Journal on Audio, Speech, and Music Processing* 1, Article ID: 926951.

P. McCabe, A. G. Macdonald-D'Silva, L. J. van Rees, K. J. Ballard, and J. Arciuli. 2014. Orthographically sensitive treatment for dysprosody in children with Childhood Apraxia of Speech using ReST intervention. *Developmental Neurorehabilitation* 17, 137–145.

O. Mich. 2009. Evaluation of software tools with deaf children. In *International ACM SIGACCESS Conference on Computers and Accessibility.* 235–236.

J. Moore and M. Churchward. 2010. *Moodle 1.9 Extension Development*. Packt Publishing, Birmingham, UK.

R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy. 2006. Telephony-based voice pathology assessment using automated speech analysis. *IEEE Transactions on Biomedical Engineering* 53, 468–477.

E. Murray, P. McCabe, and K. J. Ballard. 2012. A comparison of two treatments for childhood apraxia of speech: Methods and treatment protocol for a parallel group randomised control trial. *BMC Pediatrics* 12, 112.

D. Newbury and A. Monaco. 2010. Genetic advances in the study of speech and language disorders. *Neuron* 68, 309–320.

K. Newell, M. Carlton, and A. Antoniou. 1990. The interaction of criterion and feedback information in learning a drawing task. *Journal of Motor Behavior* 22, 536–552.

A. M. Oster, D. House, A. Protopapas, and A. Hatzis. 2002. Presentation of a new EU project for speech therapy: Ortho-Logo-Paedia. Presented at the *Proceedings of TMH-QPSR, Fonetik*.

A. Parnandi, V. Karappa, Y. Son, M. Shahin, J. McKechnie, K. Ballard, et al. 2013. Architecture of an automated therapy tool for childhood apraxia of speech. In *15th International ACM SIGACCESS Conference on Computers and Accessibility.* 5.

M. A. Rahurkar, J. H. Hansen, J. Meyerhoff, G. Saviolakis, and M. Koenig. 2002. Frequency band analysis for stress detection using a teager energy operator based feature. In *INTERSPEECH*.

J. Rick, A. Harris, P. Marshall, R. Fleck, N. Yuill, and Y. Rogers. 2009. Children designing together on a multi-touch tabletop: an analysis of spatial orientation and user interactions. In *Conference on Interaction Design and Children*. 106–114.

S. Rvachew and F. Brosseau-Lapre. 2006. Speech perception intervention. In *Interventions for Speech Sound Disorders in Children*, S. McLeod, (ed.). Brookes Publishing, Baltimore, MD.

R. A. Schmidt and T. Lee. 2005. *Motor Control and Learning*, 4th ed. Human Kinetics, Champaign, IL.

M. A. Shahin, B. Ahmed, and K. J. Ballard. 2012. Automatic classification of unequal lexical stress patterns using machine learning algorithms. In *2012 IEEE Spoken Language Technology Workshop (SLT)*. 388–391.

K. Shobaki, J. P. Hosom, and R. A. Cole. 2000. The OGI kids' speech corpus and recognizers. In *International Conference on Spoken Language Processing*.

L. D. Shriberg, T. F. Campbell, H. B. Karlsson, R. L. Brown, J. L. Mcsweeny, and C. J. Nadler. 2003. A diagnostic marker for childhood apraxia of speech: The lexical stress ratio. *Clinical Linguistics & Phonetics* 17, 549–574.

J. Tepperman and S. Narayanan. 2005. Automatic Syllable Stress Detection Using Prosodic Features for Pronunciation Evaluation of Language Learners. In *ICASSP (1)*, 937–940.

T. K. Veale. 1999. Targeting temporal processing deficits through fast ForWord® language therapy with a new twist. *Language, Speech, and Hearing Services in Schools* 30, 353–362.

D. Vilozni, M. Barker, H. Jellouschek, G. Heimann, and H. Blau. 2001. An interactive computer-animated system (SpiroGame) facilitates spirometry in preschool children. *American Journal of Respiratory and Critical Care Medicine* 164, 2200–2205.

M. Waite, L. Cahill, D. Theodoros, S. Busuttin, and T. Russell. 2006. A pilot study of online assessment of childhood speech disorders. *Journal of Telemedicine and Telecare* 92–94.

A. Williams. 2006. Multiple oppositions intervention. In *Interventions for Speech Sound Disorders in Children*, A. L. Williams, S. McLeod, R. J. McCauley, et al. (eds.). Brookes Publishing, Baltimore, MD.

P. Williams and H. Stephens. 2010. Nuffield Centre Dyspraxia Programme. In *Interventions for Speech Sound Disorders in Children*, A. L. Williams, S. McLeod, R. J. McCauley, et al. (eds.). Brookes Publishing, Baltimore, MD.

Y. Wren, S. Roulstone, and A. L. Williams. 2006. Computer-Based Interventions. In *Interventions for Speech Sound Disorders in Children*, S. McLeod (ed.), Brookes Publishing, Baltimore, MD.

S.-C. Yin, R. Rose, O. Saz, and E. Lleida. 2009. A study of pronunciation verification in a speech therapy application. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4609–4612.

S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, and G. Moore, et al. 2006. *The HTK Book, version 3.4*. Cambridge University, Cambridge, UK.

E. Zwicker. 1961. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *The Journal of the Acoustical Society of America* 33, 248–248.

Received June 2014; revised March 2015; accepted May 2015