Exploring the use of Technology for Assessment and Intensive Treatment of Childhood Apraxia of Speech

Jacqueline Gai McKechnie

Bachelor of Arts (Languages & Linguistics, Hons 2 Div I)

The University of Western Sydney

Bachelor of Applied Science (Speech Pathology, Hons 1)

The University of Sydney

A thesis submitted in fulfilment of the

requirements for the degree of

Doctor of Philosophy

Discipline of Speech Pathology, Faculty of Health Sciences

The University of Sydney

2019

Declaration of Authorship

I, **JACQUELINE GAI MCKECHNIE**, declare that the work contained in the thesis is, to the best of my knowledge, original. No other person's work has been used without due acknowledgement within text. This work has not been submitted for any other degree at the University of Sydney or elsewhere. Approval for these studies was given by the University of Sydney Human Ethics Committee (2012/2021, 2013/703)**.** Participants and/or their caregiver were required to read a participant information statement and informed consent was given prior to data collection.

Jacqueline Gai McKechnie

Date: 28th February, 2019

Declaration of Completion

This is to certify that the thesis entitled *Exploring the use of Technology for Assessment and Intensive Treatment of Childhood Apraxia of Speech* submitted by Jacqueline Gai McKechnie in fulfilment of the requirements of the degree of Doctor of Philosophy is in a form ready for examination.

Kirrie J. Ballard

Date: 28.2.19

Acknowledgements

The studies included in this thesis were made possible by the Qatar National Research Fund which funded the development of the technology applied in the experimental studies. It was indeed a serendipitous day when I wandered into the office of Dr Kirrie Ballard, who would become my primary supervisor, and mentioned that I was interested in starting a PhD. I explained that my interests at the time were broad, I was fascinated by Childhood Apraxia of Speech, eager to explore alternative service delivery methods for all children needing the support of Speech-Language Pathologists, and having fun exploring the different mobile applications available to support speech and language intervention in my clinical practice. To which came the reply, “Well, there’s this grant with some engineers in Qatar…”. I am also grateful for funding support from an Australian Postgraduate Award and two Postgraduate Research Awards from Speech Pathology Australia.

First and foremost, I extend my eternal gratitude to my supervisors, Dr Kirrie Ballard and Dr Tricia McCabe for their support and guidance through this research degree. It has been a privilege to learn from their combined knowledge, experience and expertise. I especially wish to thank Kirrie for her continued belief in my potential as a researcher from as far back as my undergraduate days. And to Tricia for her ever practical advice and grounded perspective, especially during those periods when I most needed to be talked ‘out of my tree’.

Many thanks also to Drs Beena Ahmed and Ricardo Gutierrez-Osuna and their teams for the development of the tools and technology used in these studies and their technical support along the way.

To my friend and colleague, Elizabeth Murray, thank you for paving the way and generously sharing your expertise with the Nuffield Dyspraxia Programme and your treatment protocol, on which I modelled the intervention approach in Study 3.

To my colleague Alison Purcell, thank you for the quiet reminders that this was a marathon, not a sprint, and an investment in my career for when my baby girl was bigger. And to all the other colleagues in the department who were living proof that it is possible to complete a PhD while raising a baby.

To my HDR friends – Donna, Ellie, Angelina, Lauren, Laura, Clare E, Claire L, Maryane, Elise, Manal and Dominique. To those of you who have completed, thanks for the inspiration, to those of you still working towards completion, you can do it.

To my family – thank you for your love, support and understanding throughout this journey. It wasn't always easy but you were always there cheering me on and keeping up the supply of wine, chocolate and hugs. I love you all. Finally, to my daughter Mia – we got through this baby girl. I know you didn't always like it or understand when Mumma had lots of work to do but, for the most part, you stayed sweet, funny and honest. Thank you for being my biggest cheerleader who believed that Mumma was WonderWoman and could do anything and everything. I love you most.

Publications and Presentations Arising from this Thesis

Peer-reviewed publications

McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Murray, E., McCabe, P. & Ballard, K.J (2019, submitted) Tablet-based delivery of intensive treatment in childhood apraxia of speech: Influence of type of feedback.

McKechnie, J., Shahin, M., Ahmed, B., Murray, E., McCabe, P., Arciuli, J., Ballard, K.B. (2019, under review) An automated lexical stress tool for assessing dysprosody in childhood apraxia of speech.

Murray, E., Thomas, D.C. & **McKechnie, J.** (2018) Comorbid morphological disorder apparent in some children aged 4-5 years with Childhood Apraxia of Speech: findings from standardised testing. *Clinical Linguistics and Phonetics.* Online.

McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Monroe, P., McCabe, P., Ballard, K.J. (2018) Automated speech analysis tools for children's speech production: A systematic literature review. *International Journal of Speech-Language Pathology:* 1-17
DOI: 10.1080/17549507.2018.1477991

Peer-reviewed presentations

McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Monroe, P., McCabe, P. & Ballard, K.J. (2018) *Can speech controlled games be effective speech therapy tools?* Presentation at the Speech Pathology Australia National Conference, Adelaide, SA, Australia, May 27th to 30th, 2018.

McKechnie, J., Ballard, K.J., McCabe, P., Murray, E., Lan, T., Gutierrez-Osuna, R., Ahmed, B. (2016) *Tablet-based delivery of intensive speech therapy in children with childhood apraxia of speech: Influence of type of feedback*. Presentation at the Speech Pathology Australia National Conference, Perth, WA, Australia, May 15th to 18th, 2016.

McKechnie, J., Ballard, K.J., McCabe, P., Murray, E., Lan, T., Gutierrez-Osuna, R., Ahmed, B. (2016) *Influence of type of feedback on the effect of tablet-based delivery of intensive speech therapy in children with childhood apraxia of speech*. Poster presented at the 2016 Motor Speech Conference, Newport Beach, CA, USA, March 3rd to 6th, 2016.

McKechnie, J., Ballard, K.J., McCabe, P., Gutierrez-Osuna, R., Karappa, V., Parnandi, A., Shahin, M., Murray, E. and Ahmed, B. (2015) *Tablet-based delivery of intensive speech therapy in children with childhood apraxia of speech – Clinician, client and parent user perspectives*. Poster presented at the Speech Pathology Australia National Conference, Canberra, ACT, Australia, May 17th to 20th, 2015.

McKechnie, J., Ballard, K.J., McCabe, P., Gutierrez-Osuna, R., Karappa, V., Parnandi, A., Shahin, M., Murray, E. and Ahmed, B. (2014) *Tablet-based delivery of intensive speech therapy in children with childhood apraxia of speech : pilot phase*. Presentation at the Speech Pathology Australia National Conference, Melbourne, VIC, Australia, May 18th to 21st

McKechnie, J., Ballard, KJ., Shahin, M., Murray, E. & Ahmed, B. (2013) *An Automated Lexical Stress Classification Tool for Assessing and Treating Dysprosody In Childhood Apraxia of Speech*. Paper presented at the American Speech and Hearing Association National Conference, Chicago, IL, USA, November 13th to 16th, 2013.

Notes on Style

This is a thesis including publications. As such, three papers in the style of journal articles are embedded. These papers represent chapters 2, 4 and 6 of the thesis. Due to the need for each paper to stand alone as a complete work, there is some inherent repetition of themes across the chapters of the thesis.

Style

- The journal articles included in this thesis are presented in the styles requested by the relevant journal guidelines with the following exceptions: Figures and Tables are single spaced; embedded within papers 2 and 3 (Chapters 4 and 6); and labelled according to the thesis chapter in which they appear (e.g. Table 6.1 represents Table 1 in Chapter 6).
- The remainder of this thesis is written in accordance with APA 6th style.
- Reference lists are provided at the end of the thesis, with the exception of Chapter 2 which is an author's manuscript version of a published paper, with references appearing at the end of the paper.
- Phonetic symbols and diacritics used are from the International Phonetic Alphabet. (revised, 2015). Vowels have been transcribed in accordance with the system proposed by Mitchell & Delbridge (1956). Slanted brackets // are used to indicate canonical representations.
- Semicolons are used to refer to a child's age in years;months.

Spelling

- Paper 1 (in chapter 2) titled "Automated speech analysis tools for children's speech production: A systematic literature review" is written in Australian/British English in accordance with the instructions for authors for the *International Journal of Speech-Language Pathology*.

- Paper 2 (in chapter 4) titled “An automated lexical stress tool for assessing dysprosody in childhood apraxia of speech” is written in American English in accordance with the instructions for authors for the *Journal of Speech, Language and Hearing Research.*
- Paper 3 (in chapter 6) titled “Tablet-based delivery of intensive speech therapy in Childhood Apraxia of Speech: Influence of type of feedback” is written in British English in accordance with the instructions for authors for the *Journal of Communication Disorders.*
- The spelling for the remaining chapters in this thesis (chapters 1, 3, 5 and 7) is in Australian/British English

Reference.

Mitchell, A. G., & Delbridge, A. (1956). *The pronunciation of English in Australia* (Revised ed.). Sydney, Australia: Angus and Robertson.

Glossary of Acronyms

Acronym	Explanation
ANDOSL	Australian National Database of Spoken Language
ANOVA	Analysis of Variance
ANN	Artificial Neural Network
AOS	Acquired Apraxia of Speech
ASA	Automated Speech Analysis
ASHA	American Speech Language and Hearing Association
ASR	Automatic Speech Recognition
CAS	Childhood Apraxia of Speech
CELF-4	Clinical Evaluation of Language Fundamentals – Fourth Edition
CELF-P2	Clinical Evaluation of Language Fundamentals – Preschool – Second Edition
CNN	Convolutional Neural Network
DAS	Developmental Apraxia of Speech
DEAP	Diagnostic Evaluation of Articulation and Phonology
DIS	Disorder
DNN	Deep Neural Network
DT	Decision Tree
DTTC	Dynamic Temporal and Tactile Cueing
DTW	Dynamic Time Warping
f0	Fundamental frequency
GFTA2	Goldman-Fristoe Test of Articulation
GMM	Gaussian Mixture Models

HMM	Hidden Markov Model
KNN	k-nearest neighbor algorithm
KP	Knowledge of Performance
KR	Knowledge of Results
LALR	was not defined in the study
LDA	Linear Discriminant Analysis
LL	Language learning
LS	Lexical stress
LSR	Lexical stress ratio
LSVT	Lee Silverman Voice Treatment
MaxEnt	Maximum Entropy
MFCC	Mel-frequency cepstral coefficients
MLP	MultiLayer Perceptron
MSTP	Motor Speech Treatment Protocol
NDP3	Nuffield Dyspraxia Programme – Third Edition
NN	Neural Network
OGI	Oregon Graduate Institute Multilanguage Corpus
OS	Operating system
PCC	Percent Consonants Correct
PML	Principles of Motor Learning
PPC	Percent Phonemes Correct
PPVT-4	Peabody Picture Vocabulary Test – Fourth Edition
PVC	Percent Vowels Correct
PVI	Pairwise Variability Indices
RCT	Randomised Controlled Trial

ReST	Rapid Syllable Transition Treatment
SCED	Single Case Experimental Design
SD	Standard Deviation
SDCS	Speech Disorders Classification System
SLP	Speech-Language Pathologist
SS	Strong-Strong (equivocal stress)
SSD	Speech Sound Disorder
SVM	Support Vector Machine
SW	Strong-Weak lexical stress pattern
TD	Typically Developing
TEO	Teager Energy Operators
TORGO	Database of English speech from speakers with dysarthria. Developed by the University of Toronto and Holland-Bloorview Kids Rehab Hospital
TRAD	Traditional, paper-based, face-to-face NDP3 treatment delivery
WS	Weak-Strong lexical stress pattern

Table of Contents

List of Tables

List of Figures

Thesis Abstract

Childhood apraxia of speech (CAS) is a paediatric motor speech disorder of neurological origin. It affects the intelligibility of a child's speech, resulting in consonant and vowel omissions, substitutions and/or distortions; interrupted transitions between sounds and syllables in words and phrases; as well as prosodic difficulties. If left untreated, these difficulties with speech production can have a long-term negative impact on academic achievement and social/emotional wellbeing.

Assessment of speech sound disorders (SSD), including CAS, is traditionally perceptually-based and, anecdotally, has been reported to take up a large proportion of clinicians' time. Prosodic deficits have been established as a key predictive factor in diagnosis of CAS, yet little is known about optimal methods of assessing and evaluating prosody. Perceptually-based assessments can be subject to various sources of error and bias, however, objective methods are infrequently used.

Research indicates that best practice for CAS includes intervention frequency of 2-4 sessions per week with dose frequency of at least 100 production trials per session. However, these treatment intensities do not reflect typical services in Australia or other countries where typical session frequency is once per week or 1-2 times per month. Families face numerous barriers including service availability; service cost; and distance to services, as well as barriers of time when they are called upon to supplement their clinic visits with home practice. When home practice is implemented, research indicates that speech practice is perceived as work, some children dislike having parents as therapist, some parents do not feel confident running sessions themselves and studies of speech perception abilities in untrained adult listeners suggest that parents' ability to detect speech errors and provide accurate feedback may not be optimal.

Given the rapid advances in technology over the past decade, this thesis examines the potential for automatic speech recognition (ASR) technology to expedite the process of objective analysis of speech, particularly for lexical stress patterns. This dissertation also investigates the potential for mobile technology to bridge the gap between current service delivery models in Australia and best practice treatment intensity for CAS. To address these two broad aims, this thesis describes three main projects.

The first project is a systematic literature review of ASR technology as applied to the evaluation and modification of speech production skills in children, either in cases of speech sound disorder or foreign language learning. A systematic search and review of the literature published between January 2007 and December 2016 was conducted to explore: (i) the types of automatic speech analysis (ASA) tools being applied as well as the populations of children and aspects of speech production to which they are applied; (ii) the performance accuracy of these tools compared with human perceptual evaluation; and (iii) whether there is evidence for treatment efficacy/behaviour change when using these automated tools. Across the 32 studies included in the review, 18 different tools were identified. These tools were applied to speech sound disorders from arrange of aetiologies as well as to children learning foreign languages. The majority of tools had been developed for analysis of phonemic accuracy, with only one quarter including analysis of prosodic accuracy. Most tools were applied to word level speech, with around one third applied to phrase level speech production. ASA tools were being implemented for four main purposes. These included: (i) word recognition (i.e. whether the tool can recognise the word being spoken by the user) – these tools can be used as measures of intelligibility or overall severity of disorder; (ii) judgement of the incoming spoken word or phrase as correct or incorrect based on reference to a stored representation; (iii) classification or categorisation of the incoming speech into a category such as lexical stress pattern or phoneme error type (i.e. omission, substitution) and (iv) behaviour change – these tools were incorporated into a treatment package designed to facilitate speech modification. There was a wide range of performance accuracy values when comparing the tool's output to human perceptual judgement. The findings of the review indicated that ASA tools have clinically acceptable reliability (> 80%) with human perceptual judgement for predicting intelligibility or severity of disorder, correct/incorrect judgements of phoneme and lexical stress patterns for typical developing speech, classification of typically developing lexical stress patterns and classifying/categorising phoneme error patterns in speech sound disorder only when the tool had been specifically trained on disordered speech. Automated tools were not able to meet clinically acceptable reliability thresholds when judging phonemic pronunciation or lexical stress patterns for mispronounced words from children with speech sound disorders or children learning an additional language.

The second project is a validation study exploring the accuracy of an automated lexical stress classification tool compared with human perceptual judgment. The tool was designed by one of the co-authors and team members from electrical engineering and intended for use as one part of a multi-component speech processing engine that would analyse children's speech production attempts on a clinician server. This server and a custom designed mobile application called Tabby Talks, were designed to facilitate tablet-based home practice of speech production targets and remote monitoring by the clinician using the server. This project extended on earlier investigations of the tool's accuracy by including a larger number of participants with CAS and a wider range of three-, four- and five-syllable words; and comparing both CAS and TD speech with human perceptual judgement (rather than dictionary defined lexical stress patterns). Guided by the findings of the systematic review project, this study also explored the effects of pre-training the tool with information about specific pronunciation errors made by the children as well as the influence of within word phonetic contexts, age of the speaker and percent phoneme accuracy. The results were consistent with the findings from the systematic review that automated tools can reliability classify lexical stress patterns for TD speech when compared to human perceptual judgement. The automated tool in this study was also able to classify strong-weak (SW) words produced by children with CAS, however, classification accuracy for weak-strong words (WS) and overall classification accuracy did not reach clinically acceptable reliability thresholds. The tool classified TD speech with significantly greater accuracy than CAS speech and classified SW words with significantly greater accuracy than WS words for both experimental groups. Within-word phonetic features and phoneme/pronunciation accuracy were only weakly correlated with lexical stress classification accuracy. Unlike results from earlier research, use of a pre-trained, knowledge-driven classification algorithm offered no advantage to classification accuracy for any word type in either experimental group. The overall conclusions indicate that ASA tools require continued development and training using larger datasets of disordered speech.

The third project presented in this thesis is an intervention study exploring the effect of different types of feedback on response to intervention for children with CAS. This is a randomised control trial using an established treatment program for CAS, The Nuffield Dyspraxia Programme – Third Edition (NDP3). Treatment was delivered in the speech pathology clinic via a custom-designed mobile application, Tabby Talks, to two groups of children with CAS, both receiving treatment sessions following evidence-based treatment intensity guidelines. The intervention was designed to specifically explore the feasibility and effectiveness of using an app that, in the future, could be equipped with ASR technology to provide feedback on speech production accuracy during home practice sessions, simulating the common service delivery model in Australia. One group received app-delivered face-to-face treatment and augmented feedback from a speech pathologist four days per week for

three weeks (KP group). The home practice simulation group (KR group) received face-to-face app-delivered treatment with augmented feedback from a speech pathologist one day per week for three weeks and received only right/wrong feedback on speech production accuracy from the clinician for the remaining three days per week, simulating the type of feedback that ASR technology would provide during independent app-based home practice. Fourteen children with mild to severe CAS, aged 4;0 to 10;10 participated in the intervention. Participants were matched for age and severity and randomised to a treatment condition using stratified randomisation. Both experimental groups responded to the feedback condition they received and made positive gains in treated and untreated real word accuracy over time. Although there was no significant difference between the groups at any time point, the KP group had made significant gains in treated word accuracy immediately post-treatment, similar to traditional paper-based NDP3 treatment, while the KR group had not. Notably, both groups continued to improve over time and both groups were performing significantly above baseline levels of accuracy for treated and untreated words at long-term follow up. Clinicians, parents and children were surveyed about their experiences using mobile technology to engage with intensive speech therapy. All participants reported a general preference for app-delivered therapy compared with traditional paper-based table-top interventions. This study was the first of its kind to directly compare the effects of different types of feedback whilst maintaining the same feedback schedule between groups. The findings support the feasibility for mobile applications, that could be equipped with future ASR technology that can provide reliable and accurate feedback on speech productions, to facilitate intensive practice of speech production targets and bridge the gap between optimal treatment intensity for CAS and the realities of access to services in Australia.

Collectively, the findings from all three projects highlight the potential for ASR technology, once well-trained on disordered speech and rigorously evaluated, to support

clinicians with efficient and objective analysis of disordered speech. Mobile applications with in-built ASR have the potential to increase children’s motivation and engagement with intensive practice schedules and can be an effective supplement to face-to-face therapy with a clinician. The final chapter of this thesis discusses future directions for technology-based speech assessment and intensive speech production practice, guidelines for future development of therapy tools that include more game-based practice activities and the contexts in which children can be transferred from predominantly clinician-delivered augmented feedback to ASR-delivered right/wrong feedback and continue to make optimal gains in acquisition and retention of speech production targets.

Chapter 1:

Childhood Apraxia of Speech (CAS): Nature and Treatment Needs

Childhood Apraxia of Speech

Childhood apraxia of speech (CAS) is a subtype of speech sound disorder (SSD). Using the Speech Disorders Classification System (SDCS; Shriberg et al., 2010), CAS belongs to the typology 'motor speech disorder' and the specific subtype 'motor speech disorder – childhood apraxia of speech'. It has only been in the last decade, that consensus has been reached regarding the terminology, nature, and core features of CAS. Historically, suspected developmental apraxia of speech (DAS) was a term applied to children whose speech production patterns (a) differed from other children with speech delay; (b) took longer to normalise even with intervention; and (c) resembled the difficulties exhibited by adults with acquired apraxia of speech (AOS) (see Shriberg, Aram & Kwiatkowski, 1997a for a review). Diagnosis was made perceptually, based on the presence or absence of features from diagnostic checklists which included a wide range of speech behaviours (e.g. Davis, Jakielski, & Marquardt, 1998; Hall, Jordan, & Robin, 1993; McCabe, Rosenthal, & McLeod, 1998) that did not adequately differentiate between CAS and other types of paediatric phonological or motor speech disorders (e.g. Davis et al., 1998; McCabe et al., 1998).

In the mid 2000s, the American Speech-Language-Hearing Association (ASHA) conducted a large-scale literature review and consulted with an expert committee of researchers and consumer representatives. The resultant publication of a position statement (ASHA, 2007a) and technical report (ASHA, 2007b) declared a consensus position on the nature and features of CAS. The report defined CAS as "a neurological (pediatric) speech sound disorder in which the precision and consistency of movements underlying speech are impaired in the absence of neuromuscular deficits (e.g. abnormal reflexes, abnormal tone) …..The core impairment in planning and/or programming spatiotemporal parameters of movement sequences results in errors in speech sound production and prosody" (ASHA, 2007a; ASHA, 2007b). The consensus process provided three core features of CAS:

inconsistent errors on consonants and vowels; difficulty with co-articulatory transitions between sounds and syllables; and prosodic deficits, particularly with marking lexical or phrasal stress (ASHA, 2007b). Although these three features were not intended to be necessary or sufficient for diagnosis of CAS, they have subsequently been regularly used by researchers as minimum diagnostic criteria (e.g. Namasivayam et al., 2015; Murray, McCabe & Ballard, 2015).

Over the years, there have been some efforts to operationalise measures and/or methods for repeated and reliable measurement of the core features of CAS. For example, Shriberg and colleagues developed a range of qualitative (i.e. the Speech Disorders Classification System; Shriberg, 1993; Shriberg et al., 2010) and quantitative (e.g. the Articulation Competence Index for classifying severity of speech impairment in intervals based on percent consonants correct (PCC) or the Prosody-Voice Profile; Shriberg, 1993) methods aimed at identifying specific behavioural markers that were linked to genetic mutations (Lawrence D. Shriberg, 1993) and improving differential diagnosis of CAS (Shriberg, Aram, & Kwiatkowski, 1997b, 1997c). Murray and colleagues further explored the suite of measures that could achieve the highest predictive power with the goal of improving accuracy of clinical diagnosis of CAS (Murray, McCabe, Heard, & Ballard, 2015). In 2015, Iuzzini-Seigel and colleagues (2015) operationalised eleven commonly applied diagnostic features to encourage repeatable and reliable measurement of these features. These included: vowel error, consonant distortion, stress errors, syllable segregation, groping, intrusive schwa, voicing errors, slow rate, increased difficulty with multisyllabic words, resonance disturbance, difficulty achieving initial articulatory postures (Iuzzini-Seigel et al., 2015). However, the authors have not yet evaluated these metrics for sensitivity and specificity for CAS. Most recently, Shriberg and colleagues proposed a new behavioural marker, the Pause Marker Index, as a valid and highly sensitive and specific diagnostic marker of CAS

(Shriberg et al., 2017b). This measures the percentage of inappropriate between-word pauses from a sample of 24 utterances in a continuous speech samples that meets eligibility for coding using the Prosody-Voice Screening Profile (Shriberg et al., 2017a). The authors operationalized 'inappropriate pauses' as being either (i) linguistically inappropriate in length or location; or (ii) having articulatory, voicing or prosodic features with the pause or an adjacent sound segment (Shriberg et al., 2017a).

Assessment & Diagnosis of CAS

Assessment of CAS has traditionally been conducted via auditory-perceptual judgments of the presence or absence of features. However, reliability and validity of perceptual judgments are vulnerable to numerous sources of error and bias (see Kent, 1996) (Kent, 1996). In addition, traditional methods may not adequately differentiate disorders (Ballard, Granier & Robin, 2000; McNeil, Robin & Schmidt, 1997).

Post-assessment data analysis and paperwork is reported to be equally (McLeod & Baker, 2014) or more time consuming than the direct assessment process (Skahan, Watson, & Lof, 2007). However, computerised methods are infrequently used (McLeod & Baker, 2014; Skahan et al., 2007). It is clear that there is scope for the development of automated tools. These tools could facilitate large scale studies which would allow for the development of normative databases on specific acoustic speech measures and enable exploration of sensitivity and specificity of measures used to differentially diagnose speech disorders such as CAS (McKechnie et al., 2008; Kent & Kim, 2003). Such tools have the potential to both increase objectivity and accuracy as well as expedite the processes involved in speech analysis both for diagnosis and monitoring of post-treatment retention of skills. Reliable diagnosis of CAS is critical for ensuring that children receive timely and appropriate intervention in order to mitigate some of the recognised long-term difficulties associated with persistent CAS.

Associated difficulties and long term impact.

The speech difficulties associated with CAS have been reported to take longer to resolve than other SSDs (Forrest, 2003; Shriberg, Aram, & Kwiatkowski, 1997) and can persist throughout childhood and into adulthood (Carrigg, Parry, Baker, Shriberg, & Ballard, 2016; Lewis, Freebairn, Hansen, Iyengar, & Taylor, 2004; McCabe, Preston, Murray, Bricker, & Morgan, 2017). Some children with CAS may also demonstrate one or more additional deficits such as difficulty with auditory encoding and auditory memory skills (Shriberg, Lohmeier, Strand, & Jakielski, 2012); delays in the development of sensorimotor, sequential memory and attention skills (Nijland, Terband, & Maassen, 2015); lower verbal intelligence scores (Carrigg et al., 2016); greater reliance on auditory feedback than other children (Iuzzini-Seigel et al., 2015; Terband, van Brenk, & van Doornik-van der Zee, 2014); poorer expressive morphology (Murray, Thomas, & McKechnie, 2018); poorer expressive language skills (Lewis et al., 2004); and difficulties with phonological awareness (McNeil, Gillon, & Dodd, 2009).

Persistent CAS has been demonstrated to have a long term negative impact on the development of academic and literacy skills (Gillon & Moriarty, 2007; Lewis et al., 2004; Snowling & Stackhouse, 1983); social-emotional well-being (Carrigg, Baker, Parry, & Ballard, 2015; Carrigg et al., 2016; McCabe et al., 2017; McCormack, McAllister, McLeod, & Harrison, 2012); and vocational prospects (Carrigg et al., 2015; McCabe et al., 2017). In light of the long lasting and pervasive impact of CAS, effective treatment is necessary in order to mitigate these identified risks.

Speech Motor Control and Motor Learning

Given the consensus that CAS is a disorder of motor planning and/or programming (ASHA, 2007b), intervention protocols for CAS should be guided by the Principles of Motor Learning (PML) approach (Schmidt & Lee, 2011). These principles were developed following investigations into how healthy and typically developing individuals learn skilled limb movements and provide guidance around a number of specific practice and feedback conditions that facilitate the acquisition and/or retention of motor skills (see Schmidt & Lee, 2011).

Practice conditions

Amount of practice. The nonspeech motor literature suggests that a larger number of trials (i.e. amount of practice) leads to greater retention, however, this probably also has an interaction effect with other practice variables such as constant versus variable practice and blocked versus random practice (see Schmidt & Lee, 2011). Evidence from speech motor control studies also supports the principles of large amounts of practice (operationalised as number of trials per session) is more beneficial than fewer, specifically for CAS (e.g. Edeal & Gildersleeve-Neumann, 2011; Kim, LaPointe, & Stierwalt, 2012).

Distribution of practice. Distributing practice over a longer time period has generally been found to have greater benefit for performance and retention of nonspeech motor skills as compared with massed practice (see Schmidt & Lee, 2011). This seems to not necessarily be the case for speech production skills, with studies of distributed practice during Lee Silverman Voice Treatment (LSVT) for dysarthria associated with Parkinson’s disease (Spielman, Ramig, Mahler, Halpern & Gavin, 2007) and also during Rapid Syllable Transition Treatment (ReST) for CAS (Thomas, McCabe, & Ballard, 2014) finding comparable outcomes compared with studies using more massed practice approaches.

Practice variability. Constant practice of the same movement in the same way has been found to benefit acquisition of new skills, while variable practice, where some aspect of a movement such as timing or intensity is changed, has been found to benefit longer term learning and retention of a skill (e.g. Lai, Shea, Wulf & Wright, 2000; see also Schmidt & Lee, 2011 for a review). Some evidence from speech motor literature comparing constant versus variable practice reported equivocal results, with no difference between groups at the end of the acquisition phase (Adams & Page, 2000). Other studies of speech motor control have provided support for the benefit of variable practice (e.g. by training production of sounds in various phonetic contexts) on acquisition and transfer of skills (Ballard, Maas & Robin, 2007; Wambaugh et al., 1998, 1999; Austerman Hula et al., 2008), however, these studies did not directly compare constant with variable practice.

Practice Schedule. In nonspeech motor literature, blocked practice schedules have been demonstrated to enhance acquisition of skills while random practice enhanced longer term retention and transfer to novel skills (see Maas, 2008 and Schmidt & Lee, 2011 for reviews). Evidence supporting the use of randomised blocks of trials, where targets are presented in random order but each target is practiced in a short block before the next target is presented, found equivalent or greater benefit on performance and retention as compared with purely random practice (see Schmidt & Lee, 2011). These results suggest that randomized blocks of trials are a good middle ground between maximising the positive effects of blocked practice on acquisition of skills and of random practice on retention/learning of skills. Practice schedules have been directly compared in speech motor control of healthy young adults, with results indicating that random practice was more beneficial than blocked practice for retention of skills however there were no discernible differences between the two practice schedules when examining the effect on acquisition of skills (Adams & Page, 2000) and participants with AOS (Knock, Ballard, Robin, & Schmidt, 2000). This principle has also been directly studied with a small sample of participants with CAS (Maas & Farinella, 2012). Findings were mixed, with two children demonstrating an advantage of blocked practice, one child demonstrating an advantage from random practice and another child demonstrating no response to either practice schedule (Maas & Farinella, 2012). This principle requires further investigation using larger sample sizes and exploring the effect on speech disorders of varying aetiologies.

Movement complexity

Simple (part) versus complex (whole). Evidence from nonspeech motor literature suggests that practising part of a movement task does not generalise to improved performance of the whole task (see Schmidt & Lee, 2011). In motor speech disorders, evidence suggest that targeting complex novel behaviours facilitates generalisation to real words (e.g. Murray, McCabe, & Ballard, 2015; Schneider & Frens, 2005; Thomas et al., 2014) These findings are consistent with the main overall principle of the challenge point framework in that learners need to be challenged in order for learning to occur (Guadagnoli & Lee, 2004).

Feedback conditions

Discussion of feedback conditions here will be focused on those conditions which have been investigated in both speech and nonspeech motor literature. For a full overview of different feedback conditions see Schmidt and Lee (2011).

High versus low frequency feedback. Motor learning literature generally supports an advantage for low frequency feedback (see Schmidt & Lee, 2011; Wulf, Shea, & Lewthwaite, 2010). This is interpreted in relation to the guidance hypothesis, in the sense that frequent feedback guides the individual towards the correct response and may create a dependency such that performance degrades when feedback is removed (Salmoni, Schmidt, & Walter, 1984). Conversely, low frequency feedback provides the learner with the opportunity to evaluate their own errors (Guadagnoli & Kohl, 2001). However, feedback frequency may interact with task complexity, in that more complex skills may need more frequent feedback (Swinnen, Lee, Verschueren, Serrien & Bogaerds, 1997).

In speech, evidence from healthy speakers also supports an advantage for reduced frequency feedback when measuring retention of novel speech behaviours (Adams & Page, 2000; Kim et al., 2012). In disordered speech, studies directly comparing high frequency with low frequency feedback in AOS (Austermann Hula, Robin, Maas, Ballard, & Schmidt, 2008) and CAS (Maas, Butalla, & Farinella, 2012) have reported mixed results with some participants benefiting from high frequency and others from low frequency feedback. The principle of low frequency feedback offering an advantage to motor learning has been largely accepted and systematically applied during investigations of the ReST treatment protocol, with numerous studies supporting ReST treatment as efficacious (Ballard, Robin, McCabe, & McDonald, 2010; McCabe, Macdonald-D'Silva, van Rees, Ballard, & Arciuli, 2014; Murray, McCabe, & Ballard, 2015; Thomas et al., 2014; Thomas, McCabe, Ballard, & Lincoln, 2016).

Immediate versus delayed feedback. Nonspeech motor literature supports an advantage for delayed feedback, interpreted again in relation to the guidance hypothesis (Salmoni et al., 1984) as it can be assumed that immediate feedback interrupts any intrinsic feedback the learner may generate for themselves (see Schmidt & Lee, 2011; Maas et al., 2008 for reviews). This principle has been successfully applied to treatment for CAS using ReST (e.g. McCabe et al., 2014, Murray et al., 2015, Thomas et al., 2016) where feedback is provided at reduced frequency and following a three second delay. However, when feedback timing was directly compared in AOS, the results suggested that delayed feedback was more effective for some but not all participants (Austermann Hula et al., 2008), suggesting that further investigation of the differential effects of feedback timing is warranted with larger populations of speakers with disorders of speech motor control.

Knowledge of results versus knowledge of performance. Knowledge of results (KR) refers to the provision of summative information in regards to the accuracy of a completed movement sequence, whereas, knowledge of performance (KP) includes specific information about the nature of the movement in regards to which parts of a movement sequence were in/correct, how or why they were in/correct and how to change these parameters in order to achieve a correct movement sequence on the next attempt (Maas et al., 2008; Schmidt & Lee, 2011).

In the nonspeech motor literature, KP has been found to be beneficial when the goal or task is novel to the learner, that is, when the learner does not have any internal reference of correctness (Newell, Carlton & Antoniou, 1990). KP was found to not be more effective than KR when the goal is known (Swinnen, Walter, Lee & Serrien, 1993). In another study from around the same time, Young & Schmidt (1992) demonstrated that KP was more effective in the acquisition phase of motor learning but did not lead to improved performance on retention testing, whereas KR demonstrated an advantage for motor learning and retention.

Feedback type has never been directly compared in studies of speech motor control. The nonspeech motor literature findings of a retention advantage for KR feedback seems to have led to widespread acceptance and application of KR feedback in studies of speech motor control and learning, however, when investigating the influence of other types of feedback conditions in studies of CAS, findings have been mixed. One reason for this may be due to the use of KR feedback with children who may not yet have a stable internal reference of correctness and therefore may benefit from a period of KP to establish acquisition of novel speech motor movements before moving to KR style feedback to support long term learning.

PML and Intervention Protocols for CAS

Several motor-based treatments for CAS incorporate PML into their protocols. Those with preponderant evidence for treatment efficacy include Dynamic Temporal and Tactile Cueing [DTTC] (Strand & Debertine, 2000; Strand, Stoeckel, & Baas, 2006), ReST (Ballard et al., 2010), and the Nuffield Dyspraxia Programme – Third Edition [NDP3] (Williams & Stephens, 2004). DTTC incorporates high practice amounts, massed practice, variable practice of targets and feedback designed on a hierarchy and faded based on production accuracy (see Strand, Stoeckel & Baas, 2006). ReST incorporates high practice amounts, massed practice, randomised presentation of stimuli, and delayed, reduced frequency KR feedback (see Ballard et al., 2010; McCabe, Macdonald-D'Silva, van Rees, Ballard & Arciuli, 2014; Murray, McCabe, & Ballard, 2012). NDP3 incorporates principles aimed at facilitating acquisition of new speech behaviours and incorporates frequent KP feedback and blocked practice (see Murray et al., 2012; Williams & Stephens, 2004).

Few principles have been directly compared in CAS. Exceptions include Edeal & Gildersleeve-Neumann (2011) who examined the role of practice amount in treatment using DTTC; Maas & Farinella (2012), examining the effects of blocked versus random practice during DTTC intervention; Maas, Butalla & Farinella (2012) examining the effect of feedback frequency in DTTC intervention; Namasivayam and colleagues (2015) comparing the effects of weekly versus twice weekly intervention using the Motor Speech Treatment Protocol (MSTP); and Thomas, McCabe & Ballard (2014), exploring practice distribution using ReST intervention. To date, there has been no direct comparison of feedback type in studies of speech motor control.

The paper presented in Chapter 6 is a submitted manuscript directly examining the effects of type of feedback on treatment outcomes in CAS. The study arose from the need to explore alternative service delivery methods such as mobile applications as a way to achieve optimal practice conditions for children with CAS. Faithful application of PML, particularly the principles around practice amount, practice distribution and practice schedule is difficult to achieve within the Australian clinical context where organisation or institution policies, workload, or other barriers, as discussed in the section below, typically do not allow for the sort of intensive practice schedule needed.

Service Delivery

Intervention intensity is an influential contributing factor to treatment outcomes for SSDs in general and CAS in particular. Intervention intensity is often reported in the literature in terms of number of intervention sessions received per week with more intense treatment leading to greater outcomes for children with SSDs including CAS (Allen, 2013; Baker, 2012; Kaipa & Peterson, 2016; Namasivayam et al., 2015; Williams, 2012). However, session frequency is not the only means of conceptualising intervention intensity. Warren, Fey and Yoder (2007) identified several factors which must be considered when investigating intervention intensity. These include: dose frequency, the number of times intervention is provided per day or per week within the intervention period; dose, the number of teaching moments during an intervention session; dose form, the task or activity within which the teaching moment is delivered; total intervention duration, the time period over which the intervention is administered; and cumulative intervention intensity, an index of overall intensity which is the product of dose by dose frequency by total intervention duration (Warren et al., 2007). There is a tendency for these parameters to be under-reported in treatment research (Justice, 2018; Zeng, Law, & Lindsay, 2012). Variations may influence the treatment outcome such that the relationship may not be non-linear, and more may not always equal better (Baker, 2012).

From the best available evidence to inform clinical practice, the recommended dose frequency for SSDs in general (Sugden, Baker, Munro, Williams, & Trivette, 2018) and CAS specifically (Murray, McCabe, & Ballard, 2014) is between two and four individual sessions per week of 30-60 minutes in duration. These intensities do not reflect typical practice, either in Australia or internationally, where sessions are most often reported to be once per week or 1-2 sessions per month (Brumbaugh & Smit, 2013; Hegarty, Titterington, McLeod, & Taggart, 2018; Keilmann, Braun, & Napiontek, 2004; Oliveira, Lousada, & Jesus, 2015; Ruggero, McCabe, Ballard, & Munro, 2012; Sugden et al., 2018; To, Law, & Cheung, 2012). For treatment of CAS specifically a recent survey of Australian speech-language pathologists (SLPs) identified the most common dose frequency as once per week, with a duration of 30-45 minutes per session (Gomez, McCabe, & Purcell, 2018). Interestingly, dose frequency did not influence respondents' perception of treatment efficacy (Gomez et al., 2018).

Two recent systematic reviews of the evidence reported that recommended dose is 100 production trials per session for both SSD in general (Sugden et al., 2018) and CAS specifically (Murray, McCabe, & Ballard, 2014). In practice, a slim majority (51.9%) of Australian clinicians reported adhering to the 50-100 production trials per session as recommended in the SSD research evidence; however, a large number (44%) were not meeting this standard (Sugden et al., 2018).

Service delivery: barriers

There are several commonly reported barriers to the implementation of the high amounts of therapy that are recommended for CAS recommended amount of therapy. In Australia, one of the most frequently reported is that the number of SLPs in the workforce is unable to meet community demand. SLPs also frequently report workplace issues including productivity/workload demands, high caseloads, workplace policy/mandates and insufficient funding as influential factors to caseload management and dose frequency (Edgar & Rosa-Lugo, 2007; Gomez et al., 2018; Kenny & Lincoln, 2012; Lim, McCabe, & Purcell, 2017;

Sugden et al., 2018). High workloads and large caseloads have been reported to interact with SLPs' longevity and retention in the workforce, further compounding the difficulty with availability of services (Edgar & Rosa-Lugo, 2007). Long waiting lists (McAllister, McCormack, McLeod, & Harrison, 2011; O'Callaghan, McCallister, & Wilson, 2005; Ruggero et al., 2012) and typical operating hours (Lim et al., 2017; McAllister et al., 2011) of speech pathology services also pose challenges.

Provision of services in rural and remote areas create unique issues. Clinicians can be called upon to travel to their clients, and long travel distances cut into the clinician's available time for delivery services (Verdon, Wilson, Smith-Tamaray, & McAllister, 2011). Conversely, families often carry the burden of travel which places demands on families' time as well as the added burden of costs involved in fuel for motor vehicles or use of public transport (McAllister et al., 2011; O'Callaghan, McAllister, & Wilson, 2005; Wilson, Lincoln, & Onslow, 2002). In addition, families face barriers of access related to the cost of speech pathology services, with limitations on the number of publicly funded services and high costs involved in accessing private speech pathology services (Kenny & Lincoln, 2012; Ruggero et al., 2012; Verdon et al., 2011). Families who are accessing speech pathology services may still encounter barriers to their engagement with these services. A 2011 survey of families' experiences participating in speech pathology services (McAllister et al., 2011) found that families often report difficulty scheduling clinic-based therapy into their daily lives. However, the issues are complex, with a 2012 survey of Australian parents (Ruggero et al., 2012) reporting that families receiving services fewer than one time per week desired more.

Service delivery: potential solutions

Tele-practice.

One potential solution for overcoming barriers of access that has been increasingly researched in recent years is tele-practice. Tele-practice is defined by ASHA (n.d.) as "…the application of telecommunications technology to the delivery of speech language pathology and audiology professional services at a distance…". It includes (a) synchronous services conducted in real-time using interactive audio and/or video connections such as telephone, videophone and, most commonly, internet-based videoconferencing as well as (b) asynchronous services where images or data are collected and transferred to a clinical professional for later viewing and interpretation and (c) hybrid methods incorporating some combination of synchronous and asynchronous services (Stewart Keck & Doarn, 2014; Theodoros, 2012).

Tele-practice has been used successfully to assess and treat a variety of speech and language in both children and adults, including developmental language disorders and aphasia, phonological and motor speech disorders, stuttering, voice disorders and craniofacial anomalies (see Stewart Keck & Doarn, 2014 for a review). In CAS specifically, a phase 1 multiple baseline single case experimental design (SCED) exploring the efficacy of ReST delivered via internet-based video conferencing demonstrated that children were able to make significant gains in speech production skills, which generalised to untreated behaviours (Thomas et al., 2016). These gains were similar in magnitude to face-to-face delivery of ReST treatment with maintenance of skills at 4-month follow up (Thomas et al., 2016). Both caregivers and clinicians reported being satisfied with tele-practice for ReST treatment (Thomas et al., 2016). While tele-practice may help overcome some of the barriers of access related to distance and time, it still requires contact with the clinician and may not be an adequate solution to barriers related to availability of speech pathology services in general and, more specifically, availability of services that can provide intervention at the recommended intensity.

Parent involvement.

Parent involvement is perhaps the most commonly employed strategy for overcoming barriers to recommended intervention intensity (Lim, McCabe, & Purcell, 2017; O'Callaghan et al., 2005; Sugden et al., 2018). More than 95% of Australian SLPs report involving parents in the provision of intervention for SSDs, typically via provision of home practice activities (Pappas, McLeod, McAllister, & McKinnon, 2008; Sugden, Baker, Munro, Williams, & Trivette, 2017). Parents have been asked to undertake a wide range of home practice activities (Sugden, Baker, Munro, & Williams, 2016) and there is some evidence that parent-implemented intervention activities demonstrate equal effectiveness to clinician-delivered intervention (Lancaster, Keusch, Levin, Pring, & Martin, 2010; Lawler, Taylor, & Shields, 2013; Ruscello, Cartwright, Haines, & Shuster, 1993). Despite the regular use of parents as intervention partners, only 68.4% of clinicians reported often providing training and 30.3% reported only sometimes providing training, with 88% of clinicians acknowledging that no structured training program is used (Sugden et al., 2017). More than half of all SLPs interviewed also reported ongoing barriers related to family engagement in the therapy process and lack of completion of home practice activities (Lim et al., 2017; Sugden et al., 2017).

Parents’ perceptions and experiences of their involvement in intervention are mixed. On the one hand, some parents are generally satisfied as long as the SLP remains involved with the family and maintains primary responsibility for the outcomes (Glogowska & Campbell, 2000; Watts Pappas, McAllister, & McLeod, 2015). On the other hand, many parents have also reported barriers of time in the sense that home practice can be difficult to fit into the routine of daily life (McAllister et al., 2011; Thomas, McCabe, Ballard, & Bricker-Katz, 2018). Parents perceive the SLP as the expert (Watts Pappas et al., 2015) and demonstrate a preference for individual intervention sessions with the clinician (Ruggero et al., 2012). Only 4% of parents in Ruggero et al.'s survey reported a preference for parent training and a home program (Ruggero et al., 2012). In contrast, 93% of parents of children over the age of three years reported that they would be willing to help their child with computer-based home practice activities.

In one of the first investigations of the efficacy of parent involvement in treatment for CAS, Thomas and colleagues (2017) found that a combination of parent and clinician delivered ReST treatment was efficacious for fewer children and that fewer children generalised to untrained behaviours when compared to the participant outcomes from clinician-only intervention. The average parent fidelity of implementation of the treatment program, compared with a clinician's judgment, was 77% and the average accuracy of parent feedback on their child's speech production attempts was 78%. These figures fall short of the suggested threshold of 85% which has been historically applied when investigating the reliability between two independent evaluations of the same behaviour (Cucchiarini, 1996; Pye, Wilcox, & Siren, 1988; Shriberg & Lof, 1991). The authors concluded that overall treatment efficacy was likely influenced by a number of factors including the amount of training given to parents given that parents do not have the background in phonetic training that a clinician has (Thomas et al., 2017). The parents involved in the study expressed a number of concerns around parent-implemented intervention which fell into three main themes: that the children disliked having their parents as their therapist; that the parents were concerned about their own skill in implementing the therapy, particularly providing a model of the target words and determining the accuracy of their child's productions; and finding the time each day to conduct the therapy sessions (Thomas et al., 2018).

In another study of parent-implemented treatment, Lim and colleagues (2017) found that parent-implemented DTTC generated a wide range of treatment outcomes across the four participating parent-child dyads with only one of the four children demonstrating improvement rate that was greater than chance-level. Parent fidelity of adherence to the treatment protocol ranged from 49 to 87% (Lim et al., 2017). Parents had difficulty both with adhering to the intensive nature of the treatment protocol as well as with judging the accuracy of their child's responses and providing the appropriate level of cueing and instruction in accordance with the DTTC treatment protocol (Lim et al., 2017). While the parents reported benefit associated with spending more time with their child and learning strategies to support their child with speech production practice, all parents also reported barriers associated with finding time in their daily routine to complete the therapy activities and motivating their child to engage with the therapy activities (Lim et al., 2017). The parent perspectives in these two studies echoed those reflected in McAllister et al. (2011) of speech home practice being 'work' and something that is difficult to fit into daily life.

It is reasonable for parents to express concern over their ability to accurately judge the correctness of their children's speech production attempts. Research into the factors influencing speech perception accuracy has demonstrated that children's speech is more difficult to decipher than adults (Hearnshaw, Baker, & Munro, 2014; Markham & Hazan, 2004; Munnoch, Baker, Munro, & Hearnshaw, 2018); that individual phonemes differ in the degree of accuracy with which they are perceived (Munnoch et al., 2018; Nittrouer & Miller, 1997; Schellinger, Munson, & Edwards, 2017; Wolfe, Martin, Borton, & Youngblood, 2003); and that listener experience increases speech perception accuracy (Brunnegård, Lohmander, & van Doorn, 2009; Munson, Johnson, & Edwards, 2012; Wolfe et al., 2003). On the other hand, listeners can habituate over time, resulting in 'perceptual drift' that results in a degradation of the ability to detect subtle errors over time (see Kent, 1996). The majority of errors in speech perception accuracy involve the listener being under-sensitive to speech sound errors and judging a speech production attempt as correct even when it contained an error (Munnoch et al., 2018). In contrast, to the findings on perceptual accuracy for speech sound errors, accurate perception of syllable segregation did not differ between trained and untrained listeners (Brown, Murray, & McCabe, 2018). In that study both listener groups' perceptual accuracy was positively correlated with the degree of segregation within words (Brown et al., 2018). Thus, for both speech sound errors and syllable segregation errors, research findings suggest that perceptual accuracy may be most easily achieved for productions that differ markedly from the 'typical' correct production, but that more subtle differences may be less likely to be accurately perceived.

Advances in technology: handheld devices.

Instrumental methods have long been advocated for their potential to overcome the various sources of error and bias inherent in auditory-perceptual judgments of speech (see Kent, 1996). Acoustic and kinematic analyses have the potential to increase the objectivity of speech analysis, however these methods sometimes require specialised equipment and typically involve manual measurements which can be time and/or cost prohibitive for clinicians. Given the rapid advancement of technology over the last ten to fifteen years, and the proliferation of handheld devices and mobile applications, it is timely to re-consider the role that technology can play in overcoming some of the barriers to evidence-based service provision.

Computer-based or mobile-based approaches to assessment and treatment of SSDs, although infrequently used in clinical practice (McLeod & Baker, 2014) or home practice (Ruggero et al., 2012; Sugden et al., 2018), should be considered. Such tools, when equipped with automatic speech analysis (ASA) or recognition (ASR) software offer the potential for accessible, cost effective, and objective methods of assessing speech. ASR-equipped computer- or app-delivered intervention activities can also provide an effective supplement to face-to-face clinical sessions and an alternative to parent-delivered home practice.

Purpose and Structure of Thesis

This thesis investigates the potential for technology to (i) overcome some of the barriers inherent in the current Australian service delivery context and (ii) offer alternative methods of access to intensive treatment for children with CAS. It is comprised of seven chapters, including publications.

Chapter 2 (paper 1) presents a systematic literature review exploring the current state of the evidence around the implementation and effectiveness of automated speech analysis and recognition software in evaluating and treating paediatric SSDs, including CAS. The literature review explored SSDs more broadly given the limited available data on CAS alone.

Chapter 3 summarises the findings from the systematic review which were specific to CAS and discusses the particular importance of designing ASA tools which can evaluate lexical stress. Lexical stress is selected as a starting point as this feature has been found to have high predictive power/validity for detecting CAS (Murray, McCabe, Heard, et al., 2015).

Chapter 4 (paper 2) presents an experimental study that aims to test and validate one ASA method - automated lexical stress classification of polysyllabic words - that could facilitate more objective assessment and diagnosis of CAS. Such testing and validation is a necessary step before such software can be integrated into apps or other clinical tools for use in standard practice.

Chapter 5 discusses intervention approaches for CAS, including the extant literature on treatment efficacy, and the necessary considerations for utilising mobile technology with or without ASA as alternative service delivery methods in CAS.

Chapter 6 (paper 3) considers how the use of mobile technology can influence the type of feedback that a child receives on their speech production attempts during practice.

This chapter presents an intervention study comparing children's response to two types of augmented feedback – KP, as provided by a clinician, and KR, as would be provided by an ASA algorithm. In both conditions, the clinician uses a tablet-based app to deliver practice exercises in a controlled clinic setting. User satisfaction with the tablet-based exercises is also explored through surveys administered to the children, their parents/caregivers and the treating clinicians.

Chapter 7 provides an overall discussion, summary and conclusions of the findings of these three studies in the context of extant literature on treatment efficacy, service delivery, and the scope within which technology can be an effective tool for assessment and treatment of CAS.

Chapter 2: Automated speech analysis tools for children's speech production: A systematic literature review

Paper 1: Automated speech analysis tools for children's speech production: A systematic literature review.

The paper presented in this chapter has been published as follows:

McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Monroe, P., McCabe, P., Ballard, K.J. (2018) Automated speech analysis tools for children's speech production: A systematic literature review. *International Journal of Speech-Language Pathology:* 1-17 DOI: 10.1080/17549507.2018.1477991

Author Contribution Statement

As co-author and primary supervisor, I confirm that Jacqueline McKechnie made the following contributions:

- Conception of the research questions in collaboration with co-authors
- Database searches
- Literature reviews
- Data entry and data analysis/interpretation in collaboration with co-authors
- Writing of the first draft of the paper, with subsequent drafts developed in collaboration with co-authors
- Journal submission
- Journal revisions and resubmissions

Kirrie J. Ballard

Date: 27.2.19

Automated speech analysis tools for children's speech production: A systematic literature review

J. MCKECHNIE[1], B. AHMED[2], R. GUTIERREZ-OSUNA[3], P. MONROE[1],
P. MCCABE[1] & K. J. BALLARD[1]

[1]Faculty of Health Sciences, University of Sydney, Lidcombe, NSW, Australia, [2]Department of Electrical and Computer Engineering, Texas A&M University, Doha, Qatar, and [3]Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA

Abstract

Purpose: A systematic search and review of published studies was conducted on the use of automated speech analysis (ASA) tools for analysing and modifying speech of typically-developing children learning a foreign language and children with speech sound disorders to determine (i) types, attributes, and purposes of ASA tools being used; (ii) accuracy against human judgment; and (iii) performance as therapeutic tools.
Method: Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were applied. Across nine databases, 32 articles published between January 2007 and December 2016 met inclusion criteria: (i) focussed on children's speech; (ii) tools used for speech analysis or modification; and (iii) reporting quantitative data on accuracy.
Result: Eighteen ASA tools were identified. These met the clinical threshold of 80% agreement with human judgment when used as predictors of intelligibility, impairment severity, or error category. Tool accuracy was typically580% accuracy for words containing mispronunciations. ASA tools have been used effectively to improve to children's foreign language pronunciation.
Conclusion: ASA tools show promise for automated analysis and modification of children's speech production within assessment and therapeutic applications. Further work is needed to train automated systems with larger samples of speech to increase accuracy for assessment and therapeutic feedback.

Keywords: automatic speech recognition; speech sound disorder; prosody

Introduction

Recent advances in automatic speech analysis technology are making the prospect of computer-driven speech assessment and intervention more viable for children with speech sound disorders (SSD). Significant barriers of access, cost and long-term engagement for children who require intensive and prolonged speech therapy have been identified (McAllister, McCormack, McLeod, & Harrison, 2011), and clients/parents have reported a desire for alternative approaches to accessing services (Ruggero, McCabe, Ballard, & Munro, 2012). In light of this, computer-driven approaches, particularly when embedded in serious games, have potential to overcome these barriers. Here, we performed a systematic search and review (Grant & Booth, 2009) to determine the types of automatic speech analysis and recognition (ASA) tools that have been developed over the past 10 years, what they are being used for in the context of speech assessment and treatment, and how they are performing. We did not aim to perform an analysis of study design and quality. Rather, our objective was to provide an overview of the current state of the field and an evaluation of the quality and accuracy of the current ASA tools; discussing feasibility for their use in clinical practice and needs for future development.

Automatic speech analysis tools

In the 1960s and 70s, the earliest ASA systems were able to process isolated words from small to medium pre-defined vocabularies using acoustic phonetics to perform: time alignment; template-based pattern recognition; or matching of the incoming speech signal with the stored reference production (Kurian, 2014). The inherent variability of the speech signal introduced by vocal tract variations across speakers and temporal variability across repeated productions

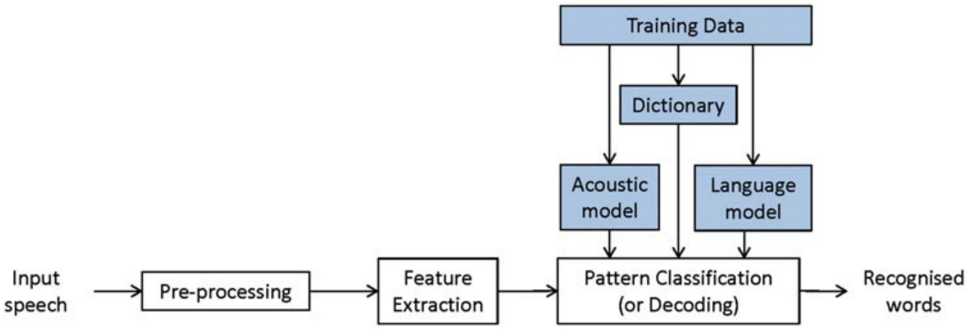


Figure 1. Basic components of a speech recognition system.

of the same word affected recognition accuracy. In the 1970s, linear predictive coding (LPC) was introduced, which could account for some of the individual variation caused by vocal tract differences (Kurian, 2014). In the 1980s, ASA tools became better able to process larger vocabularies and continuous speech, driven by the development of technology based on statistical modelling of probability that a particular set of language symbols (i.e. either phoneme sequences or word sequences) was a match to the incoming speech signal (Kurian, 2014). These systems are more robust to variations across speaker (e.g. pronunciation or accent) and environmental noise as well as temporal variations in the speech signal (Kurian, 2014). Hidden Markov models (HMMs), which perform temporal pattern recognition, are now the predominant technology behind speech recognition systems. Also in the 1990s, new innovations in pattern recognition led to discriminative training and kernel-based techniques such as Support Vector Machines (SVMs) which functioned as classifiers. Figure 1 presents a model of the component processes involved in modern ASA systems (also see Keshet, in press, in this issue; and Shaikh and Deshmukh, 2016).

Performance accuracy of ASA tools is influenced by two main components of the system (Mustafa, Rosdi, Salim & Mughal, 2015). One component is the feature extraction process, which is in turn also influenced by the type of speech (i.e. isolated words, connected speech or continuous speech); and the size of the vocabulary, with larger vocabularies associated with improved performance (Mustafa et al., 2015). Continuous speech is the most difficult to analyse because the utterances all run together and segmentation needs to be performed by the ASA in order for accurate recognition to occur (Strik & Cucchiarini, 1999). Also affecting system development and performance accuracy is the fact that availability of databases with large vocabularies is limited (Mustafa et al., 2015). The second component influencing performance accuracy is the type of speech acoustic model, which is based on speaker mode (i.e. speaker dependent, where the system is trained by the user's own speech samples; speaker independent where the system requires no additional training before use by a speaker; or speaker adaptive where the system is capable of adapting to the user over time, thus improving performance) (Mustafa et al., 2015).

Despite the remarkable improvements in ASA, particularly for adult speech, computational modelling systems continue to have difficulty adapting to the temporal and spectral variability that is introduced to the speech signal via individual differences such as vocal tract length, words in context (i.e. co-articulation effects) or environmental noise (O'Shaughnessy, 2015). These factors are particularly challenging for ASA in children, who are going through periods of growth and making developmental speech errors. In both adult and child studies, these models have also struggled with the increased within- and between-speaker variability introduced with disordered speech (Su, Wu, & Tsai, 2008). Given the rapid changes in this field, it is timely to consider the state of the field in terms of child- focussed ASA tools being developed for assessment and modification of disordered or non-native speech.

Technology

Smartphone and tablet technology are now a part of children's everyday lives. In Australian households with children under 15, 88% in major cities and 79% in remote areas have access to the Internet (Australian Bureau of Statistics, 2016). Of these, 94% access the Internet via laptop or desktop computer, 85% via mobile or smartphone and 62% via tablet (Australian Bureau of Statistics, 2016). Despite reports of infrequent use of com- puter-based or mobile-based analysis procedures or intervention activities in children with SSD (McLeod & Baker, 2014); these tools have potential to facilitate easily accessible, cost effective and objective measures of speech. This may increase clinician efficiency and assist in caseload manage- ment, and such tools may also supplement face-to- face speech-language pathology to reduce barriers to access and facilitate higher practice intensity (Baker, 2012). Technology-based approaches may also increase child engagement and motivation with learning tasks as they are colourful, can include animation and audio prompts or reinforcers, involve active manipulation of stimuli and gameplay by the child, and can incorporate speech recording, pre-recorded models, and playback of responses (Morton, Gunson, & Jack, 2012; Simmons, Paul, & Shic, 2016; Tommy & Minoi, 2016). However, to be viable, any ASA tools incorporated into diagnostic or therapeutic software need to meet the same reliability standards that we apply to human raters. Commonly accepted criteria for percent agreement on perceptual judgments of speech between two human raters or reliability of outcome across two separate evaluations of the same behav- iour is between 75 and 85% (Charter, 2003; Cucchiarini, 1996). We therefore apply an 80% threshold in evaluations of the tools identified for this review.

Assessment and treatment of SSD

Recent surveys of Australian and American paediatric speech-language pathologists (SLPs) reported that phonological process analysis, estimating intelligibility, determining phonetic inventory (independent analysis) and use of phonological processes (relational analysis) constitute essential elements of a speech assessment battery (McLeod & Baker, 2014; Skahan, Watson, & Lof, 2007). The resultant post-assessment data analysis and paperwork were reported to be equally (McLeod & Baker, 2014) or more time-consuming (Skahan et al., 2007) than the assessment process itself. Few SLPs in either study reported use of computerised analysis procedures. Scope clearly exists for automated analysis processes to be developed that could increase clinical efficiency. Such tools would ideally include: (i) high agreement with human decisions regarding word recognition, which could automate the process of intelligibility assessment; (ii) judgments of correct/ incorrect for a given speech attempt, with reference to a stored template or canonical representation, thus automating the process of relational analysis;
(iii) classification or categorisation of speech error or prosodic error patterns, useful for detecting presence of impairment; and (iv) potentially use clusters of features to differentially diagnose disorders.

If well designed, such tools could also be used to monitor and shape response to intervention over time as well as augmenting and increasing home practice. Recommended intervention frequency for SSD is 2–4 sessions per week with at least 100 trials per session (Allen, 2013; Baker & McLeod, 2011; Ballard, Robin, McCabe, & McDonald, 2010; Edeal & Gildersleeve-Neumann, 2011; Murray, McCabe, & Ballard, 2014, 2015; Thomas, McCabe, & Ballard, 2014; Williams, 2012). These treatment intensities do not, however, reflect typical practice (Keilmann, Braun, & Napiontek, 2004; McLeod & Baker, 2014; Oliveira et al., 2015; Ruggero, McCabe, Ballard, & Munro, 2012; To, Law, & Cheung, 2012). Families face barriers of service availability where community demand cannot be met by available speech-language pathol- ogy resources (Kenny & Lincoln, 2012; Lim, McCabe, & Purcell, 2017; McAllister et al., 2011; O'Callaghan, McAllister, & Wilson, 2005; Ruggero et al., 2012; Verdon, Wilson, Smith-Tamaray, & McAllister, 2011) and barriers of distance in rural and remote areas (McAllister et al., 2011; O'Callaghan et al., 2005; Ruggero et al., 2012; Verdon, Wilson, Smith-Tamaray, & McAllister, 2011). This discrepancy is further confounded by parental reports of difficulty finding time for home practice and their perception that speech homework is ''work'' (McAllister et al., 2011).

McAllister et al. (2011) found computer-based homework is provided to only 17% of families contrasting the high level of interest expressed by participants in Ruggero et al. (2012). Capitalising on this interest, as well as on the automated corrective instruction already used in second language learning contexts (e.g. Neri, Mich, Gerosa, & Giulian, 2008), ASA tools could be developed and integrated into training programmes to help facilitate independent practice (Eskenazi, 2009).

Purpose

In this review, we aim to address the following research questions:

(1) ASA tools and purposes:

   (a) What ASA tools are being used?;
   (b) For what populations of children (i.e. language learners/disordered speech; and the range of languages/disorders investigated)?;
   (c) For which aspects of production/pronunciation evaluation and what types of stimuli (i.e. sound/ word/phrase level; restricted or unrestricted stimulus sets)?

(2) Accuracy of analysis: How do these tools perform compared with human perceptual evaluation?
(3) Behaviour change: Is there evidence that improvements to children's speech sound production abilities as a response to intervention are comparable between ASA-based training tools and face-to-face training?

Method

We used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) search guidelines (Moher, Liberarti, Tetzlaff, Altman, & The PRISMA Group, 2009) when formulating our search strategy. The flow diagram of study selection is presented in Figure 2.

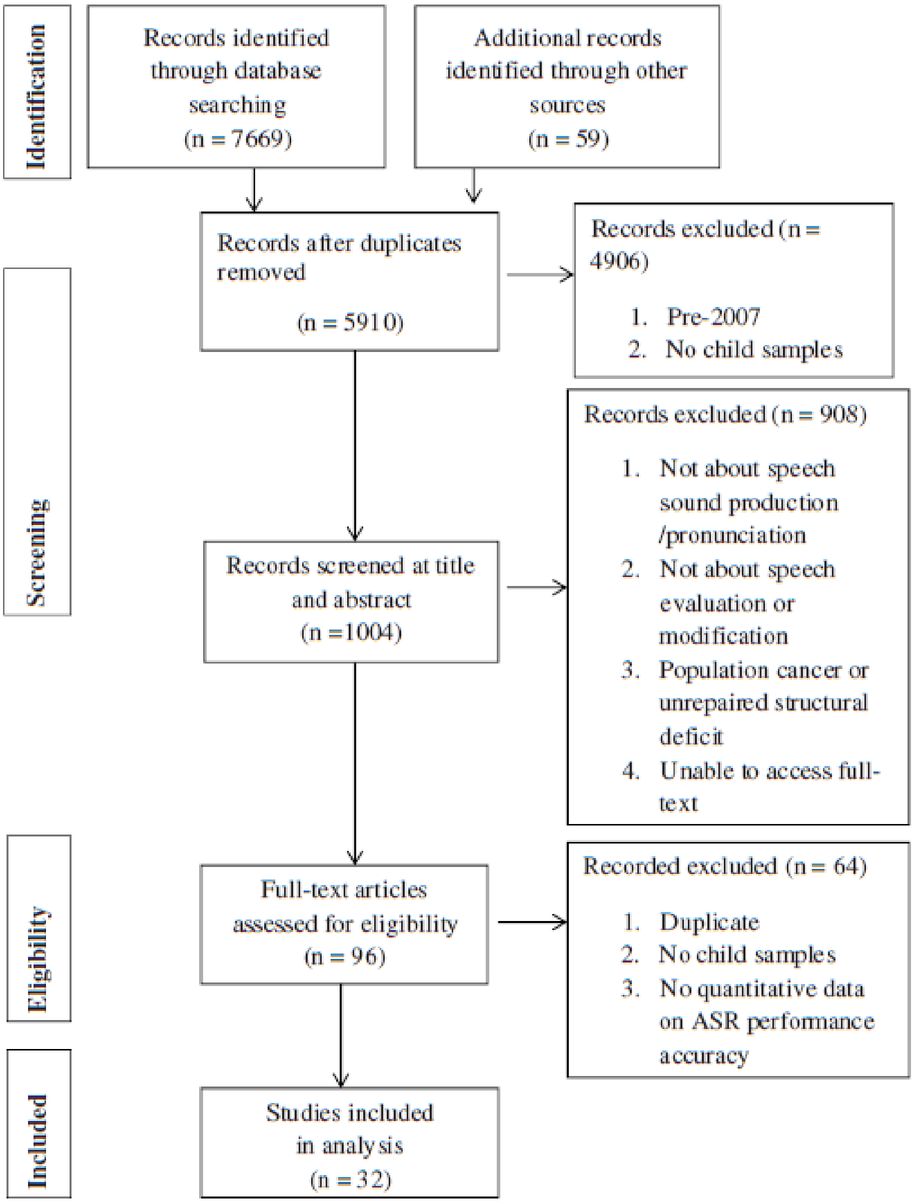

Identification
Records identified through database searching (n = 7669)
Additional records identified through other sources (n = 59)
Screening
Records after duplicates removed (n = 5910)
Records excluded (n = 4906)
1. Pre-2007
2. No child samples
Records screened at title and abstract (n =1004)
Records excluded (n = 908)
1. Not about speech sound production /pronunciation
2. Not about speech evaluation or modification
3. Population cancer or unrepaired structural deficit
4. Unable to access full-text
Eligibility
Full-text articles assessed for eligibility (n = 96)
Recorded excluded (n = 64)
1. Duplicate
2. No child samples
3. No quantitative data on ASR performance accuracy
Included
Studies included in analysis (n = 32)

Figure 2. Systematic search and review flowchart.

Evidence identification

We searched the following key databases in the fields of allied health, engineering and computer sciences to identify relevant articles: Medline, Cinahl, ERIC, Embase, Scopus, Web of Science, IEEEXplore, ACM Digital Library and Applied Science and Technology. The following search terms were used with Boolean operators, wildcards and proximity syntax: artic*; impair*; phonol*; disorder; apraxia; dyspraxia; dysarthria; speech error; patholog* speech; multilingual*; bilingual*; foreign language; language learn*; pronunciation; diagnosis; ‘‘decision making’’; instruction; therapy; intervention; training; response feedback; computer based/assisted/aided; signal processing; mobile application; app; software; speech recognition software; android; iOs; handheld; intelligent tutoring system; computer managed instruction; education* technology; electronic learning; virtual speech therapist; virtual classroom; web based instruction; computer programme; automat* speech recognition/analysis/evaluation/assessment/intelligibility assessment; speech/pronunciation verification; automat* speech error detect*/feedback/ speech processing; spoken dialogue systems; artificial intelligence; neural networks (NNs); automated pattern recognition; machine learning; acoustic-phonetic classification; corrective feedback. See Supplementary Appendix 1 for sample search strategies. Note that studies of ASA technology in foreign language learning were sought because these tools have similar goals to those designed for children with SSD (e.g. detection of phoneme mispronunciations or provision of corrective feedback for modifying productions; Saz, Lleida, & Rodríguez, 2009) that could inform development of tools for SSD diagnosis and treatment.

Studies published between January 2007 and December 2016 were considered for inclusion. Date restrictions were imposed in order to focus the review on current tools and technologies and to exclude out-dated technology that has been replaced with more advanced versions. The year 2007 was selected as it marks the release of the first iPhone, with Apple's processing speed, graphics, touch screens, and integration of app technology making them the industry front runner (Martin, 2014), and accelerating development in the field.

Screening

A total of 7669 articles were retrieved from database searching, and a further 59 from later hand searching of reference lists in articles that survived initial screening. Of these, 1759 duplicates were removed. After applying limits of (i) published between 2007 and 2016 and (ii) focussed on children's speech production, 4906 additional records were excluded. Therefore, 1004 were retained for title and abstract screening. Of these, 908 were excluded for the following reasons: (1) not dealing with paediatric speech sound production/pronunciation; (2) not explicitly focussed on evaluation or modification of speech production skills; (3) not the target population (e.g. oral or pharyngeal cancers, laryngectomy); or (4) full text record not accessible. A total of 96 papers were shortlisted for full text review.

Eligibility criteria

The review focussed on studies of ASA technology applied to the speech of typically developing (TD) children, using either their native or a non-native language (i.e. language learning; LL), or children reported to have SSD. Studies were included if they reported on the use of automated tools for speech analysis and/or speech modification delivering summative or formative feedback to the clinician or the speaker. While we acknowledge that there are numerous computer programmes and mobile applications that provide interactive and game-based presentation of stimuli such as ArtikPix (Expressive Solutions LLC, 2011), only software integrating ASA for the purpose of determining speech accuracy was included in this review, as we were interested in software with potential to act as a virtual clinician.

Studies were required to provide quantitative data on the accuracy of the tool's ASA algorithms against human judgment and, for automated speech modification tools, on treatment effects or changes to speech intelligibility, word accuracy or pronunciation accuracy. All study formats were considered, including journal articles, serials, conference papers and proceedings provided that new data were reported. The search was limited to studies written in English.

Of the 96 studies accepted for full text analysis, 32 were judged eligible for this review. Reasons for exclusion included: (1) duplicates overlooked in the initial screening process; (2) only adult participants (where this had been unclear at the screening phase); and (3) no quantitative data on ASA performance accuracy.

Analysis of evidence

To address research question (i) we extracted information on characteristics of the participants (i.e. age, sex and type of speech disorder, where appropriate); the purpose of speech analysis (i.e. phoneme or prosodic accuracy); types and attributes of ASA tools being used (i.e. technology for different ASA purposes, operating system, format of the interface and the user-feedback generated); characteristics of the speech samples used (i.e. type of speech sample and whether speech stimuli were from open or constrained sets); the speech features extracted by the tool; and the language of operation of the tool. To address question (ii) we tabulated the outcome measures used and their reported accuracy against human perceptual judgment. To answer question (iii) we tabulated details of behaviour change outcomes.

Result

All data extracted from each of the 32 publications were collated in a spreadsheet (see Supplementary Table SI). Summary tables are presented here.

ASA tools and purposes

Table I presents a summary of the tools reviewed, the speech analysis foci and participant characteristics cross the 32 studies.

Participant characteristics

Participants ranged in age from 3 to 21 years. Four studies included participants of58 years; 17 studies included participants up to 16 years and one up to 21 years. Twenty-two of the 32 articles (71%) did not report on the sex distribution of the participants in the study, therefore, these data are not discussed further. When extracting sample size data, we considered only the samples used to evaluate the tool's accuracy, not samples used for training and development of the tool. Sample sizes ranged from 1 to 1133 (n¼29 publications) with a median sample size of 37. Half of all studies had sample sizes within the range 19–119. In three publications, sample size was not stated. Tools were applied to language learning populations in 28.1% (n¼9) of articles and to disordered speech in 71.9% (n¼23).

Technology and purpose

Within the 32 articles, 18 types of ASA tools were discussed (see Figure 3(A)). Twenty-four studies

Table I. ASA technology and its purpose for each study (alphabetical) with children who have a speech disorder (DIS) or are learning a foreign language (LL).

First Author (Year)	Technology of tool[a]	Age	Sample size[b]	Sex	Population[b]	Disorder type
Phoneme-level analysis						
Azizi (2012)	HMM (4 models trained: M & F adults, F adults, F adults + kids; kids only)	5–8 yrs	13 (DIS)	not stated	DIS	articulation
Bártů (2008)	ANN (KSOM)	4–10 yrs	3 (DIS) 7 (TD)	2M, 1F 2M, 5F	DIS	developmental dysphasia
Chen (2011)	HMM Dependence network	mean 6yrs	132 (DIS)	not stated	DIS	articulation
Dudy (2015)	HMM	4–7 yrs	19 (DIS-1); 24 (DIS-2); 47 TD	not stated	DIS	DIS-1: articulation; DIS-2: speech;
Duenser (2016)	HMM incorporating Phoneme Classification; Knowledge Driven Recognition; Decision Support)	3–14 yrs	13 (DIS-mixed); 9 TD	not stated	DIS	Cerebral palsy; pre-term birth; TD
Kadi (2016)	GMM SVM GMM/SVM hybrid	not stated	19 (DIS)	16M, 3F	DIS	dysarthria
Lee (2011)	HMM	Yr 3–5	24 (TD)	12M, 12F	LL	
Maier (2008)	Unclear (OneR, DecisionStump, LDA-classifier, NativeBayes, J48, PART, RandomForest, SVM, AdaBoost)	not stated	26 (DIS)	21M, 5F	DIS	cleft lip and palate
Maier (2009a)	HMM: semi-automated using transcription data HMM: automated using trigram language model independent of transcription	mean 10.1 yrs; mean 62 yrs	31 children (DIS); 41 adults (DIS)	not stated	DIS	dysarthria; laryngectomy
Maier (2009b)	HMM	mean 9.4 yrs (DIS-1); mean 8.7 yrs (DIS-2)	26 (DIS-1); 32 (DIS-2)	not stated	DIS	cleft lip and palate
Mazenan (2015)	HMM	primary school age	20 (DIS)	not stated	DIS	not specified
Navarro-Newball (2014)	HMM	not stated	20 (DIS)	not stated	DIS	hearing impaired
Nicolao (2015)	DNN	13+ yrs	222	not stated	LL	
Obach (2012)	HMM SVM MLP	not stated	25 (TD)	18M, 7F	LL	
Pantoja (2014)	KNN	not stated	not stated	not stated	LL	
Parnandi (2015)	HMM (phoneme decoder)	7–10 yrs	7 (DIS)	6M, 1F	DIS	CAS
Saz (2009)	HMM (ASR) Confusion network (pronunciation verification)	11–21 yrs (DIS); 10–18 yrs (TD)	14 (DIS); 168 (TD)	7M, 7F (DIS); 73M 95F (TD)	DIS	dysarthria
Schipor (2012)	HMM	preschool & young school age	not stated	not stated	DIS	dyslalia
Shahin (2014)	GMM-HMM DNN-HMM	4–10 yrs (DIS); K – Yr 10 (TD);	5 (DIS); 110 (TD);	not stated	DIS	CAS
Shahin (2015)	HMM (posterior probability) HMM (lattice based phoneme verification)	4–16 (DIS); not stated (TD)	2 (DIS); 4 (TD)	not stated	DIS	CAS
Singh (2015)	SVM	8–16 yrs	20 (DIS)	not stated	DIS	not specified
Suanpirintr (2007)	HMM Phoneme based speech recognition (PSR) HMM (word-based speech recognition) HMM (pause reduced word-based recognition)	7–13 yrs (DIS); 8–11 yrs (TD)	4 (DIS); 2 (TD)	2M, 2F (DIS); 1M, 1F (TD)	DIS	dysarthria
Ting (2008)	MLP	8 yrs	1 (DIS)	1M	DIS	articulation
Wielgat (2008)	DTW (phoneme based); DTW (word based); HMM (whole word); HMM (phoneme level)	not stated	not stated	not stated	DIS	speech disorder
Prosodic analysis						
Delmonte (2009)	LALR parser	not stated	not stated	not stated	LL	
Ferrer (2015)	HMM (GMM) Decision Trees Neural Networks	10–14 yrs	168 (TD); 329 (TD approximating errors)	not stated	LL	

(continued)

Table I. Continued

First Author (Year)	Technology of tool[a]	Age	Sample size[b]	Sex	Population[b]	Disorder type
Parnandi (2015)	MLP (lexical stress classifier)	7–10 yrs	7 (DIS)	6M, 1F	DIS	CAS
Shahin (2012)	ANN SVM MaxEnt	K – Yr 10	196 (TD)	not stated	LL	
Shahin (2015)	MLP (lexical stress) SVM (lexical stress) MaxEnt (lexical stress)	4–16 (DIS); not stated (TD)	2 (DIS); 4 (TD)	not stated	DIS	CAS
Shahin (2016)	DNN CNN	K – Yr 10; adult	110 (TD); not stated (adults)	not stated	LL	
Sztaho (2010)	HMM	10–14 yrs	19 (DIS)	not stated	DIS	speech impaired
van Santen (2009)	not specified	not stated	15 (ASD); 13 (TD); 15 (DIS)	not stated	DIS	ASD; non-ASD
Both phoneme-level and prosody analysis						
de Wet (2009)	HMM	not stated	90 (TD)	not stated	LL	
Hacker (2007)	LDA (ADABOOST)	10–11 yrs	28 (TD)	15M, 13F	LL	
Voicing delay						
Parnandi (2015)	Intensity threshold (VAD)	7–10 yrs	7 (DIS)	6M, 1F	DIS	CAS
Shahin (2015)	Intensity threshold (VAD)	4–16 (DIS); not stated (TD)	2 (DIS); 4 (TD)	not stated	DIS	CAS

[a] In alphabetical order, ANN: artificial neural network; CNN: convolutional neural network; DNN: deep neural network; DTW: dynamic time warping; GMM: Gaussian mixture models; HMM: hidden Markov model; KNN: K-nearest neighbour algorithm; KSOM: Kohonen self-organising map; LALR: look-ahead left-right parser; LDA: linear discriminant analysis; MaxEn: maximum entropy; MLP: multilayer perceptron; SVM: support vector machine; VAD: voice activity detector.

[b] ASD: autism spectrum disorder; CAS: childhood apraxia of speech; TD: typically developing children.

(75%) described tools for phoneme level analysis of pronunciation, eight studies (25%) described tools for prosodic aspects of pronunciation and two studies (6.25%) described tools that simultaneously analysed phonemic and prosodic aspects of pronun- ciation (See Table I).

Twelve publications evaluated two or more ASA tools. Some studies compared the performance of two or more tools for a specific analysis purpose; for example, comparing classification accuracy for dysarthria severity using Gaussian Mixture Models (GMM), a SVM or a hybrid of the two (Kadi, Selouani, Boudraa, & Boudraa, 2016). Other studies reported an ASA system comprised of multiple automated analysis modules, each performing a different task, for example, a HMM-based phoneme segmentation/forced alignment module and a dependence network for subsequent phoneme error classification accuracy (Chen, 2011). For details, see Supplementary Table SI.

Figure 3(A) also presents data on the proportion of tools addressing the different analysis foci of the ASA tools. The majority of tools (17/18) were designed to analyse a specific feature of speech (i.e. intelligibility, correctness, classification of phoneme error or lexical stress pattern). Nine tools across 8/32 studies (25%) measured speech recognition rates. These studies reported on whether the tool recognised the input as the target word or phoneme. These tools could be applied to automated intelli- gibility assessment or evaluation of the degree of disorder or mispronunciation. Success of classifying speech into different categories was reported in twenty-five of the included studies (25/32; 78%). This included classification of speech input as correct or incorrect based on reference to a stored representation as well as classification to a specific category, such as lexical stress patterns (e.g. strong- weak or weak-strong) or phoneme error type (e.g. substitution or omission). Two studies (2/32; 6.25%) reported duration measures including total voicing/utterance duration and voicing delay. Voicing delay was defined as a measure of response latency or delayed initiation of speech following presentation of stimulus.

No studies reported on tools designed for identifying a syndrome or differentiating different speech disorders. Only three systems were designed for speech modification within a treatment or learning package (Delmonte, 2009; Lee et al., 2011; Navarro-Newball et al., 2014).

Operating system

The operating system (OS) for the ASA tool was not defined in 20 publications (62.5%). Three papers described Web-based tools and servers (Lee et al., 2011; Maier et al., 2009b; Parnandi et al., 2015), four described tools that run on a desktop or laptop computer (Duenser, 2016; Pantoja, 2014; Shahin, Ahmed, & Ballard, 2012; Shahin, Ahmed,

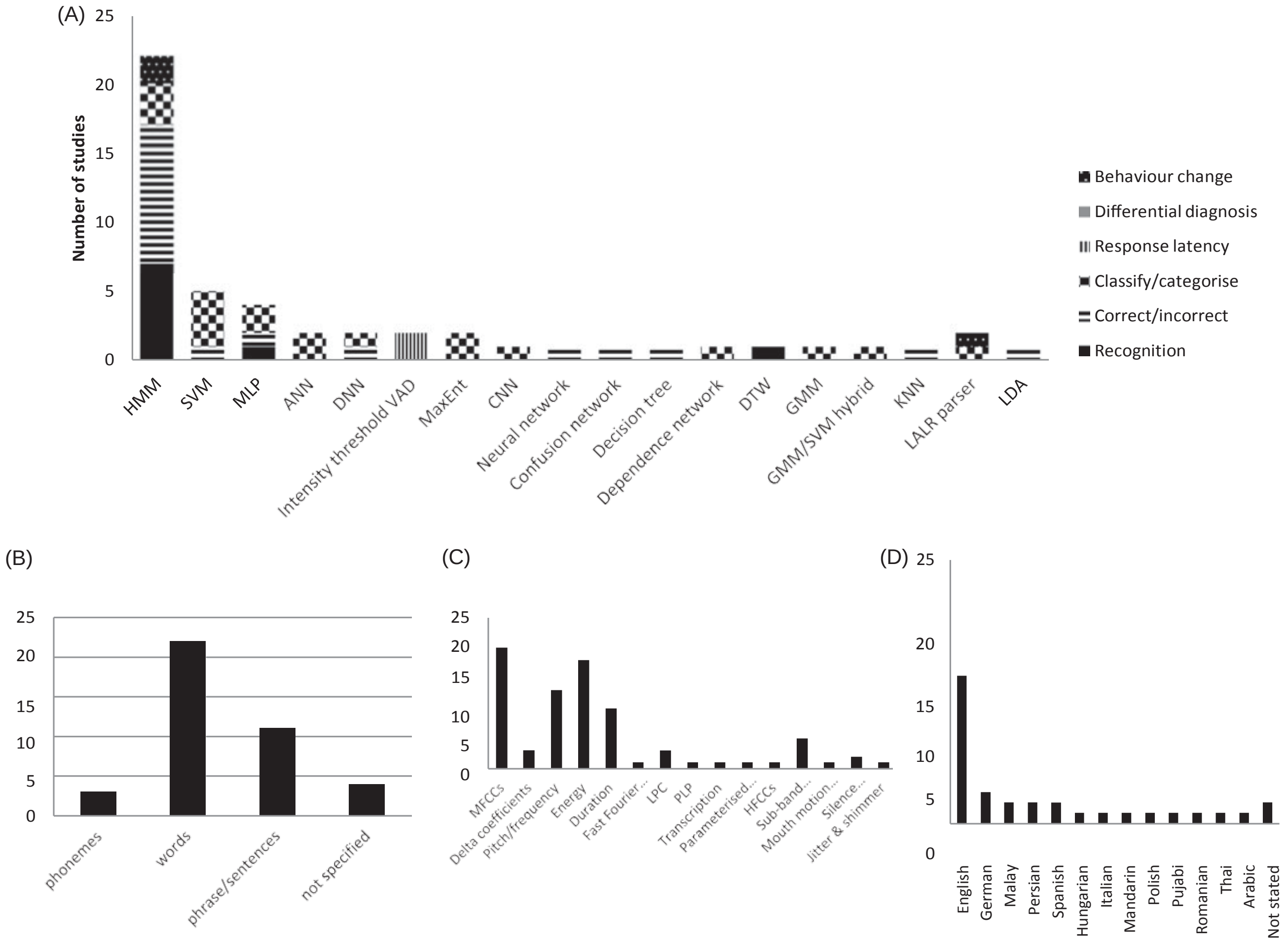


Figure 3. Frequency across the 32 studies of A. each automated technology used and proportion of tools addressing each analysis focus (HMM ¼ Hidden Markov Models; SVM ¼ Support Vector Machine; MLP ¼ MultiLayer Perceptron; ANN ¼ Artificial Neural Network; DNN ¼ Deep Neural Network VAD ¼ Voice Activity Detector; MaxEnt ¼ Maximum Entropy; CNN ¼ Convolutional Neural Network; DTW ¼ Dynamic Time Warping; GMM ¼ Gaussian Mixture Models; KNN ¼ k-nearest neighbour algorithm; LALR parser was not defined in the study; LDA ¼ Linear Discriminant Analysis); B. each type of speech sample elicited; C. use for each feature extraction method (MFCCs ¼ mel-frequency cepstral coefficients; LPC ¼ linear predictive coding coefficients; PLP ¼ perceptual linear prediction coding coefficients; HFCCs ¼ human frequency cepstral coefficients); D. each language represented.

McKechnie, Ballard, & Gutierrez-Osuna, 2014), two specified Windows OS (Navarro-Newball et al., 2014; Sztaho, Nagy, & Vicsi, 2010), one ran on the Mac OS (Delmonte, 2009), one on the Android OS (Parnandi et al., 2015), and one was a cross-platform tool that could operate in Windows, Mac, Linux and Android (Ferrer et al., 2015).

Interface: user input and output

In four studies, ASA was embedded in an application incorporating both a clinician/teacher interface and a child interface (Maier et al., 2009a; Navarro-Newball et al., 2014; Parnandi et al 2015; Saz et al., 2009). That is, the ASA potentially could be used to deliver feedback on speech productions to the child or to provide analysis of performance to a remote clinician/teacher. Of these, two studies addressed dysarthria (Maier et al., 2009a; Saz, Yin, et al., 2009); one addressed childhood apraxia of speech (CAS) (Parnandi et al., 2015); and one included children with hearing loss (Navarro-Newball et al., 2014). Two studies focussed on describing a speech processing engine, which was being developed for later integration into a programme with both clinician/teacher and child interfaces; one for language learning (Hacker, Cincarek, Maier, HeBler, & Noth, 2007) and one for CAS (Shahin et al., 2015). Two tools, both designed for foreign language learning, had only a child interface (Delmonte, 2009; Lee et al., 2011). The ASA system in the remaining 16 studies had been evaluated in its development phase, without reference to the user interface.

Regarding the child interface, three studies described game-based programmes through which the children recorded their speech samples (Lee et al., 2011; Navarro-Newball et al., 2014; Parnandi et al., 2015). All other studies used non-game speech sampling methods such as picture naming or word reading, or provided insufficient informa- tion to determine the method used.

Of the six studies that reported an ASA system already integrated into a child interface, four used the speech analysis output to provide feedback to the child. In three of these studies, all using HMM- based ASA systems, the feedback was on accuracy (i.e. correct/incorrect) of phonemes in picture naming (Saz, Yin, et al., 2009), syllable string repetition (Navarro-Newball et al., 2014) or sentence level (Lee et al., 2011) tasks. The language-learning system in Lee et al. (2011) also provided feedback in the form of a model and recast. The fourth system with a child interface, an LALR parser system, was designed for children learning English pronunciation and provided feedback on accuracy of lexical and phrasal stress assignment, as well as performance-based feedback such as ‘speak more slowly’ (Delmonte, 2009). Two other language–learning studies, with systems not yet integrated into a child–friendly interface, provided feedback on pronunciation accuracy. The system in Pantoja (2014) focussed on phonemic accuracy and the system in Hacker et al. (2007) analysed both phonemic and prosodic input features to provide the child with feedback on pronunciation accuracy.

Speech sample characteristics

Figure 3(B) presents data on the elicited speech samples used to develop and evaluate the tools in the included studies. Most commonly, ASA tools were developed and evaluated using single word stimuli (n¼22 studies). When multi-word utterances were used, they ranged from three word phrases to sentences. Ten tools, across seven publications, were tested using both single and multi-word utterances (see Supplementary Table SI). The majority (75%) of ASA tools were tested with a constrained stimulus set (n¼24 studies), meaning participants were produced a specific set of words or sentences rather than spontaneous speech. In seven studies, it was unclear whether the stimulus set was open or constrained.

There was large variability across the selected studies in number of speech tokens used to evaluate a tool. The median was 1750 (range 78–54,080), with 50% of studies reporting between 340 and 8400. Six publications did not report number of tokens per participant or total number.

Features extracted

Figure 3(C) summarises the feature extraction data from the studies. The majority of tools, in 20/32 publications, used Mel-frequency cepstral coefficients (MFCCs), often in combination with other features. MFCCs map spectral information from the speech signal onto the Mel scale, which approximates the way the human auditory system perceives frequencies. For three tools feature extraction was not reported (de Wet, Van der Walt, & Niesler, 2009; Duenser et al., 2016; Lee et al., 2011).

Language

ASA systems were developed for thirteen different languages, most commonly English (14/32 or 43.75%) (see Figure 3(D)). Of the studies targeting English, 9/14 were designed for children learning English as a non-native language and 5 were for English-speaking children with a speech disorder. For the other 12 languages addressed, 2 studies were tools for second language learning and 15 for helping children with disorders in their native language. One study did not specify the language used to train and test the tool.

Accuracy of analysis

The accuracy of speech recognition or classification against human judgment was reported in a number of ways including word recognition rate, percent agreement, correlation and measures used in signal detection (e.g. true/false positive rates, sensitivity, specificity). A summary of the ASA technology, outcome measure, and accuracy of analysis and population studies is in Supplementary Table SII.

Word recognition rate

Word recognition rates for TD children ranged from 69.4% to 98% (Azizi, Towhidkhah, & Almasganj, 2012; Suanpirintr & Thubthong, 2007, respectively). For SSD/LL speech, word recognition rates ranged from 48.5% for speakers with dysarthria (Suanpirintr & Thubthong, 2007) to 91.67% for children learning another language (Wielgat, Zieliński, Woźniak, Grabias, & Król, 2008). Ting and Mark (2008) achieved high recognition rates of 97–100% for isolated vowel phonemes in a SSD/LL speaker. Mazenan et al. (2015) reported high recognition rates on a range of isolated phonemes (88.19–96.92%) and at the whole word level (95–100%); however, the population was not specified.

Percent agreement with human judgment

Accuracy in classifying phoneme-level pronunciation as (in)correct against human judgment ranged from 45.7% for mispronounced words for a combined group of TD and SSD speakers (Dudy, Asgari, & Kain, 2015) to 95.67% for LL speakers (Obach & Cordel, 2012). Tools categorising phoneme error type in SSD speech showed from 91.13% agreement with human judgment (Singh, Thakur, & Vir, 2015) to 99.6% (Maier, Honig, Hacker, Schuster, & Noth, 2008).

One study reported on a dual-component tool in which an HMM-based component decoded the sequence of incoming phonemes and compared this input to a stored representation of the target word; and a Dependence Network component classified the input sequence to a particular phon- eme error category (e.g. substitution or omission) (Chen, 2011). Accuracy for automated vs. manual phoneme labelling accuracy of the HMM tool ranged from 46.32% for mispronounced words, where the sequence of phonemes produced violated the phonotactic rules/permissible sequences of the target language, to 88.7% for correctly pronounced words (Chen, 2011).

.

Regarding percent agreement for lexical stress classification, four studies of TD children reported values ranging from 53–70% (Sztaho et al., 2010) to 93.4% (Shahin et al., 2016). Shahin et al. (2012) reported higher agreement for words with strong-weak stress (93.8%) than words with weak-strong stress (75%). For two studies of TD and SSD/LL children combined, overall accuracy ranged from 77.6% (Shahin et al., 2015) to 88.4% (Duenser et al., 2016). For nine studies examining only SSD/LL speech, percent agreement ranged between 10 and 71% (Sztaho et al., 2010) up to 93.5% (Ferrer et al., 2015).

Considering phonemic and prosodic features simultaneously for determining word accuracy, Hacker et al. (2007) reported 74.2% agreement with human judgment for SSD/LL speakers and 89% for the pooled TD and SSD/LL.

For intensity threshold-based voice activity detection tools, percent agreement for automated vs. manual calculation ranged from 96% in SSD/LL speech (Parnandi et al., 2015) to 96.6%. These studies considered TD and SSD/LL speech combined (Shahin et al., 2015). For calculations of total utterance duration, accuracy of the tool ranged from 94% for SSD/LL speech (Parnandi et al., 2015) to 94.8% (Shahin et al., 2015) for TD and SSD/LL speech combined. These measures were explored in only two studies from the same research team, which may account for the narrow range of percent agreement values.

Correlation

Eight of the 32 studies reported human–machine correlations for the evaluation of pronunciation at the phoneme-level in SSD/LL speech. Correlations ranged from a non-significant or weak correlation (range 0.02–0.40; de Wet, Van der Walt, & Niesler, 2009) to a strong correlation of 0.89 (Maier et al., 2008, 2009a,b). One study exploring prosodic accuracy in a sample of pooled TD and SSD/LL speakers reported moderate to strong correlations (0.66–0.86) between automatic and human assessments (van Santen, Prud'hommeaux, & Black, 2009).

Signal detection theory measures

Thirteen of the 32 studies reported more detailed information on classification accuracy of the tool versus the ‘‘gold standard’’ of human judgment. Six reported on true positive rate (i.e. sensitivity – all items included in a category truly do belong in that category); two reported on precision (i.e. the probability that an item truly belongs in the assigned category); one reported on true negative rate (i.e. specificity – all items excluded from a category truly do not belong in that category); four reported true and false positive/negative rates; and one reported equal errors rates (i.e. the threshold where likelihood of false acceptance and false rejection is equal).

For SSD/LL phoneme-level classification accuracy, true positive rates ranged from 52.6% (SSD) in Maier et al. (2009b) to 100% (LL) in Obach & Cordel (2012). For TD speakers, true positive rate was reported at 96% (Shahin et al., 2014). Classification true negative rate for phoneme-level analysis in SSD/LL speakers ranged from 53.8% (Shahin et al., 2014) to 82–95% (Chen, 2011). For TD speakers, Shahin et al. (2014) reported a true negative rate of 74.6%. Classification precision rates for phoneme-level pronunciation accuracy ranged from 87 to 100% for LL speech in Obach and Cordel (2012). For TD and SSD speakers combined, classification precision was reported at 91.1% by Shahin et al. (2015). The ASA tool from three studies reported multiple measures including sensitivity, specificity, false positive and/or false negative rates for SSD/LL speakers. False positive rates ranged from 19.5% (Duenser et al., 2016) to 70.5% (Saz, Yin, et al., 2009). The lowest false negative rates were reported by Saz et al. (2009) at 1.5% for speaker-dependent conditions (i.e. where the ASA tool had been trained for each impaired speaker). For speaker-independent conditions (i.e. where the tool had been trained on unimpaired speakers), false negative rates ranged from 6.1% (Shahin et al., 2014) to 12.3% (Saz, Yin, et al., 2009). Shahin et al. (2014) reported 16.3% false positives; and 4% false negatives for their tool's analysis of phoneme-level accuracy in TD speakers. Equal error rates ranged from 14 to 25.3% across a range of speaker-dependent and speaker-independent conditions analysed by Saz et al. (2009).

Behaviour change

Only three publications reported on changes in speech production following practice with an ASA-based tool providing feedback on accuracy: one tool was an LALR parser (Delmonte, 2009) and the other two studies both developed and evaluated an HMM-based ASA system (Lee et al., 2011; Navarro-Newball et al., 2014). Delmonte (2009) reported that 20 LL children improved their production of lexical and phrasal stress after 10 hours of training but no statistics were reported to substantiate this claim. Lee et al. (2011) reported significant improvement in mean pronunciation scores in 21 beginner and intermediate LL students, with a large effect size of 0.90. Navarro-Newball et al. (2014) studied a single child with hearing loss who acquired all trained two to three syllable consonant-vowel combinations within eight sessions. No studies compared performance of the children using ASA-based tools against traditional clinician-delivered intervention. Given the variability in outcome measurement across these three studies and the absence of raw data/statistical analyses in two studies, we were unable to report on pooled results.

Discussion

The over-arching aim of this review was to examine the use and effectiveness of ASA tools in analysing and/or modifying children's speech production. To that end, we addressed the following sub-goals: 1. (a) to examine the types of automatic speech analysis (ASA) and recognition (ASR) tools used for speech analysis/modification; (b) the populations and (c) goals/purposes to which they have been applied; 2. to determine the accuracy of ASA tools' analyses of speech in typically developing (TD) children, children with speech sound disorders (SSD), or TD children learning a foreign language (LL); and 3. to determine whether currently there is evidence that changes in children's speech production accuracy is comparable between of ASA-based training tools and face-to-face training.

ASA tools and purpose

Based on the data extracted from the studies included in this review, HMMs are the most studied automated analysis tools to date. SVMs, NNs and GMMs were also frequently described with outcomes meeting or exceeding clinical thresholds. These tools apply probability or likelihood measures that are better able to adapt to temporal variability in the speech signal and nonlinear interactions between speech input and other environmental acoustic variables (Deng & Li, 2013). ASA-based tools have been most often applied to phonemic accuracy at single word level and infrequently at utterance level. Less commonly, tools evaluated lexical or phrasal stress at both word and utterance level. These tools have been applied to populations of children with SSDs in their native language and typically developing children learning to speak additional languages.

Most tools are being used to analyse single words in one language and have been tested using constrained word sets. Such tools are limited in their generalisability to other contexts without extensive training and re-testing. Accessing or collecting large samples of speech from specific user groups/populations in order to comprehensively train the ASA module to better adapt to speaker variability can be difficult (Lee et al., 2011; O'Shaughnessy, 2008). Task-dependent and/or speaker-dependent models such as the HMM + Confusion Network model in Saz et al. (2009), demonstrated clinically acceptable performance accuracy; however, their reliance on a specific set of vocabulary items significantly limits transferability to other populations, languages and word sets. Using a limited vocabulary, particularly one with few easily confused words (e.g. neighbours such as ''pat'' and ''bat'') will increase analysis/recognition accuracy at the cost of reducing breadth of application, which places limits on their wider use in assessment and treatment.

None of the studies included in this review demonstrated the use of ASA methods to differentially diagnose disorders. This is an area of particular clinical need, particularly for disorders that have historically been difficult to differentiate, for example, CAS and inconsistent phonological disorder (Dodd, 2013; Murray, McCabe, Heard, & Ballard, 2015) or some types of dysarthria (Kent & Kim, 2003).

Accuracy of analysis

ASA-based tools built on HMM architectures that extract Mel-frequency cepstral coefficients (MFCCs) from the speech signal correlate well with human judgment and can accurately predict intelligibility/severity ratings for child speech (Maier et al., 2009a; Saz, Yin, et al., 2009). For both phoneme- and prosody-level judgments of correct/incorrect, accuracy was particularly high when tools were applied to correctly pronounced words in groups of TD speakers or groups of SSD/LL speakers (Chen, 2011; Duenser et al., 2016; Ferrer et al., 2015; Shahin et al., 2012, 2016). Mixed results were obtained when evaluating the perform- ance accuracy of HMM-based tools on combined samples of TD and SSD/LL speakers (Hacker et al., 2007; Obach & Cordel, 2012; Parnandi et al., 2015; Shahin et al., 2015). It is possible that, in studies reporting high rates of classification accuracy for combined samples of TD and SSD/LL speakers, high accuracy for correctly pronounced words from TD speakers may have masked potentially poorer performance of the tool with SSD/LL speech. Classification of incorrectly pronounced words did not reach the 80% threshold for TD, LL, or SSD speakers at phoneme- or prosodic-level analysis (Chen, 2011; Ferrer et al., 2015; Shahin et al., 2014).

For tools which demonstrated high rates of classification/categorisation accuracy for phoneme error patterns (Dependence Network based tool, Chen, 2011; HMM-based tool, Maier et al., 2009b, SVM-based tool, Singh et al., 2015) or severity level (GMM-based tool, Kadi et al., 2016), results need to be interpreted with caution, as overall sensitivity can be low when datasets contain few samples with errors (Maier et al., 2008, 2009b). Wider clinical applicability of these particular tools (Singh et al., 2015; Kadi et al., 2016) will be limited as each tool is language specific, disorder specific and word-list specific.

Regarding tools which classify/categorise lexical stress patterns, tools meeting clinically acceptable standards when applied to TD speakers (ANN-based tool, Shahin et al., 2012; CNN-based tool, Shahin et al., 2016) or approaching clinically acceptable accuracy when applied to a combined group of TD and SSD/LL speakers (MLP-based tools, Parnandi et al., 2015; Shahin et al., 2015) need to be validated on SSD/LL speakers to determine their accuracy on speech samples where the likelihood of mispronunciations is high.

Taken together, these findings suggest that ASA methods are able to meet/exceed clinically acceptable thresholds for correctly-pronounced words but do not meet clinically acceptable standards when evaluating words containing mispronunciations, particularly in the case of impaired speech. Of the best performing ASA tools in the reviewed studies, two HMM-based tools (Duenser et al., 2016; Obach & Cordel, 2012), one GMM-based tool (Kadi et al., 2016), one SVM-based tool (Singh et al., 2015) and one HMM plus Dependence Network tool (Chen, 2011) were trained on populations of LL or SSD speakers, which may account for their increased performance accuracy. Of these five tools, two incorporated Knowledge Driven recognition systems that had been trained specifically for the types of errors those speakers were likely to produce (Chen, 2011; Duenser et al., 2016). For performance accuracy to increase for mispronounced words, ASA models need to be trained on a larger corpus of speech containing incorrectly pronounced words. Until this happens, clinical applicability of these tools to speech disordered populations will be limited, particularly in the case of disorders with a motor basis where errors may be less predictable and consistent than in disorders with a linguistic basis that follow largely predictable ‘‘rules’’.

Behaviour change

To date, the focus on tools for automated speech analysis (ASA) have been mainly at the development stage and for evaluation of accuracy or error type in speech production. Given the varied success of these tools, it is not surprising that very few studies have yet explored their utility or appropriateness for changing behaviour. We found only three studies documenting the ability of these tools to facilitate changes to speech production/pronunciation abilities of the child. For two of these studies (LALR parser, Delmonte, 2009; HMM-based ASA, Navarro-Newball, et al., 2014), the exact nature of the intervention and performance measurement was unclear and the effect size for the intervention was not reported. For these reasons, pooled data on effect sizes could not be reported. The HMM-based tool in Lee et al. (2011) was reported to facilitate significant improvement in mean pronunciation accuracy with large effect sizes; however, the exact measure of pronunciation accuracy was not defined. None of the studies compared ASA-based instruction and feedback to face-to-face instruction.

The absence of information about the quality and accuracy of the ASA-based feedback in many studies reporting quantitative changes to speech production (Neri et al., 2008; Wang & Young, 2015) makes it difficult to determine the true agent of change in these studies. Qualitative data suggests that, to be effective, feedback must be both ‘‘correct’’ i.e. not reject an utterance that a human listener would accept, and ‘‘adequate’’, i.e. specific to the error made by the user (Engwall & Balter, 2007). The quantitative data reviewed here leads us to question the capacity of ASA tools to meet both these criteria, especially for children and impaired speakers.

Surprisingly, only one of the studies included in this review described the development of a mobile application for speech analysis and modification (Parnandi et al., 2015), despite the proliferation of speech therapy apps over the last 10 years. In that study, the application offered a digital, interactive method of stimulus presentation and a method for assigning rewards for correct productions, but the speech processing unit was located on a separate server and automated analysis of the child’s speech attempts was conducted offline. Therefore, the user relied on traditional feedback from a trained clinician or parent (Parnandi et al., 2015). Most therapy apps for paediatric speech disorders simply provide an alternative method of stimulus presentation and rely on a SLP, therapy assistant or parent/caregiver to provide feedback and shaping of responses. One possible reason for the current scarcity of apps equipped with in-built real-time ASA-based evaluation and feedback is that mobile devices have limited computational capacity to perform those functions with high reliability (Lee, Lee, Kim, & Kang, 2017).

Limitations and future directions

While the demand for ASA continues to grow, its rate of growth depends on successfully closing the performance gap between human and machine recognition, a need that has been described for 10 years (O’Shaughnessy, 2008). Some authors have investigated the effects of applying vocal tract length normalisation to samples of children’s speech to improve the recognition accuracy of ASA models trained on adult speech (Azizi et al., 2012). Dudy et al. (2015) demonstrated that training a standard Goodness-of-Pronunciation model (GOP) on explicit samples of correct and incorrect pronunciations produced a statistically-significant increase in the rate of agreement between ASA and human experts’ classification; however, the modified GOP algorithm continued to perform below clinical ‘‘gold standard’’. Phonetically-based systems are, by necessity, language-specific as the set of phonemes and the range of allowable phoneme sequences is specific to individual languages (Delmonte, 2009). By extension, this could be applied to impaired speech. Future research should focus on optimising the performance of automated tools for phoneme labelling, classification of correct/incorrect, and sensitivity for error identification in populations with impaired speech production abilities where high instances of mispronounced words are likely.

We acknowledge the risk of publication bias and English language bias as a result of restricting our database search terms to title and abstract fields, limiting the date range, restricting the search to articles published in English, and to tools that have been evaluated in scholarly journals. Further investigation is needed to identify potentially useful ASA tools developed for languages other than English.

Although outside the date range of this review, two papers were recently published on video-game delivered (Cler, Mittelman, Braden, Woodnorth, & Stepp, 2017) and app-delivered (Byun et al., 2017) biofeedback for treatment of speech sound disorders. Notably, these studies both focussed on discrete aspects of speech production (velopharyngeal valving and production of the /r/ phoneme, respectively). This suggests tools more narrowly focussed to specific speech sounds or discrete bioacoustic features may have greater potential for success, at least in the short-term.

Conclusion

ASA shows promise for automated assessments of intelligibility or automated classification of impairment severity level. In order for ASA systems to be useful to users, false acceptance and rejection rates need to be low to avoid frustration for the user, and error detection accuracy and feedback capabilities need to be high in order to avoid potentially harmful effects of inaccurate guidance for shaping a student's behaviour. Quantitative data presented in this review suggest that clinical transferability of the described ASA tools is limited at this time. This is due to subpar performance on mispronounced words combined with highly constrained speech sample sets, as well as heterogeneous languages on which these systems have been trained. The proliferation of language learning and speech therapy apps suggests that automated feedback from computer and tablet-based gaming as speech therapy is an area of keen interest and we should expect to see the body of literature growing in the near future. With continued research interest and effort, these tools have real potential to assist children to achieve high intensity and engaging speech practice outside the clinic and can help overcome service delivery barriers. It is feasible that serious games with integrated ASA could soon be used to assist children with SSD to achieve rapid speech change by facilitating high frequency, high quality, engaging home practice with ASA-generated feedback on performance.

Acknowledgements

The statements made herein are solely the responsibility of the authors. Jacqueline McKechnie wishes to thank Yulia Ulyannikova from the University of Sydney Health Sciences Library for her time and expertise in database search syntax.

Declaration of interest

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Qatar National Research Fund (a member of the Qatar Foundation) [NPRP # 8-293-2-124].

References

Allen, M.M. (2013). Intervention efficacy and intensity for children with speech sound disorder. Journal of Speech, Language & Hearing Research, 56, 865–877. doi: 1092-4388(2012/11-0076)

Australian Bureau of Statistics. (2016). Household use of information technology, Australia, 2014-15 (Vol. 2017). Canberra, Australia: Australian Bureau of Statistics.

Azizi, S., Towhidkhah, F., & Almasganj, F. (2012). Study of VTLN method to recognize common speech disorders in speech therapy of Persian children. Paper presented at the 2012 19th Iranian Conference of Biomedical Engineering, ICBME 2012.

Baker, E. (2012). Optimal intervention intensity in speech-language pathology: Discoveries, challenges, and unchartered territories. International Journal of Speech-Language Pathology, 14, 478–485. doi:10.3109/17549507.2012.717967

Baker, E., & McLeod, S. (2011). Evidence-based practice for children with speech sound disorders: Part 1 Narrative review. Language, Speech & Hearing Services in Schools, 42, 102–139. doi:10.1044/0161-1461(2010/09-0075)

Ballard, K.J., Robin, D.A., McCabe, P., & McDonald, J. (2010). A treatment for dysprosody in childhood apraxia of speech. Journal of Speech, Language & Hearing Research, 53, 1227–1245. doi:1092-4388(2010/09-0130)

Bártů, M., & Tucková, J. (2008). A classification method of children with developmental dysphasia based on disorder speech analysis. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 5164 LNCS, pp. 822–828).

Byun, T.M., Campbell, H., Carey, H., Liang, W., Park, T.H., & Svirsky, M. (2017). Enhancing intervention for residual rhotic errors via app-delivered biofeedback: A case study. Journal of Speech, Language, and Hearing Research, 60, 1810–1817. doi:10.1044/2017_JSLHR-S-16-0248

Charter, R.A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. The Journal of General Psychology, 130, 290–304. doi:10.1080/00221300309601160

Chen, Y.J. (2011). Identification of articulation error patterns using a novel dependence network. IEEE Transactions on Biomedical Engineering, 58, 3061–3068. doi:10.1109/TBME.2011.2135352

Cler, G.J., Mittelman, T., Braden, M.N., Woodnorth, G.H., & Stepp, C.E. (2017). Video game rehabilitation of velopharyngeal dysfunction: A case series. Journal of Speech, Language, and Hearing Research, 60, 1800–1809. doi:10.1044/2017_JSLHR-S-16-0231

Cucchiarini, C. (1996). Assessing transcription agreement: Methodological aspects. Clinical Linguistics & Phonetics, 10, 131–155. doi:10.3109/02699209608985167

de Wet, F., Van der Walt, C., & Niesler, T.R. (2009). Automatic assessment of oral language proficiency and listening comprehension. Speech Communication, 51, 864–874. doi:10.1016/j.specom.2009.03.002

Delmonte, R. (2009). Prosodic tools for language learning. International Journal of Speech Technology, 12, 161–184. doi:10.1007/s10772-010-9065-1

Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. IEEE Transactions on Audio, Speech & Language Processing, 21, 130.

Dodd, B. (2013). Differential diagnosis and treatment of children with speech disorder. West Sussex, UK: Wiley.

Dudy, S., Asgari, M., & Kain, A. (2015). Pronunciation analysis for children with speech sound disorders. Paper presented at the Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS.

Duenser, A., Ward, L., Stefania, A., Smith, D., Freyne, J., Morgan, A., & Dodd, B. (2016). Feasibility of technology enabled speech disorder screening. In A. Georgiou, L. K. Schaper, & S. Whetton (Eds.), Digital health innovation for consumers, clinicians, connectivity and community (Vol. 227, pp. 21–27). Amsterdam, Netherlands: IOS Press.

Edeal, D.M., & Gildersleeve-Neumann, C.E. (2011). The importance of production frequency in therapy for childhood apraxia of speech. American Journal of Speech-Language Pathology, 20, 95–110. doi:10.1044/1058-0360(2011/09-0005)

Engwall, O., & Balter, O. (2007). Pronunciation feedback from real and virtual language teachers. Computer Assisted Language Learning, 20, 235–262. doi:10.1080/09588220701489507

Eskenazi, M. (2009). An overview of spoken language technology for education. Speech Communication, 51, 832–884. doi:10.1016/j.specom.2009.04.005

Expressive Solutions LLC. (2011). ArtikPix (Version 2.0) [Mobile Application]: Expressive Solutions LLC. Retrieved from http://itunes.apple.com

Ferrer, L., Bratt, H., Richey, C., Franco, H., Abrash, V., & Precoda, K. (2015). Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. Speech Communication, 69, 31–45. doi:10.1016/j.specom.2015.02.002

Grant, M.J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. Health Information & Libraries Journal, 26, 91–108. doi:10.1111/j.1471-1842.2009.00848.x

Hacker, C., Cincarek, T., Maier, A., HeBler, A., & Noth, E. (2007, April 15–20). Boosting of prosodic and pronunciation features to detect mispronunciations of non-native children. Paper presented at the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP '07.

Kadi, K.L., Selouani, S.A., Boudraa, B., & Boudraa, M. (2016). Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge. Biocybernetics and Biomedical Engineering, 36, 233–247. doi:10.1016/j.bbe.2015.11.004

Keilmann, A., Braun, L., & Napiontek, U. (2004). Emotional satisfaction of parents and speech-language therapists with outcome of training intervention in children with speech annd language disorders. Folia Phoniatrica et Logopaedica, 56, 51–61. doi:10.1159/000075328

Kenny, B., & Lincoln, M. (2012). Sport, scales, or war? Metaphors speech-language pathologists use to describe caseload management . International Journal of Speech-Language Pathology, 14, 247–259. doi:10.3109/17549507.2012.651747

Kent, R.D., & Kim, Y.J. (2003). Toward an acoustic typology of motor speech disorders. Clinical Linguistics & Phonetics, 17, 427–445. doi:10.1080/0269920031000086248

Keshet, J. (in press). Automatic speech recognition: A primer for speech pathology researchers. International Journal of Speech-Language Pathology.

Kurian, C. (2014). A review on technological development of automatic speech recognition. International Journal of Soft Computing and Engineering, 4, 2231–2307.

Lee, J., Lee, C.H., Kim, D.-W., & Kang, B.-Y. (2017). Smartphone-assisted pronunciation learning technique for ambient intelligence. IEEE Access, 5, 312–325. doi:10.1109/ACCESS.2016.2641474

Lee, S., Noh, H., Lee, J., Lee, K., Lee, G.G., Sagong, S., & Kim, M. (2011). On the effectiveness of Robot-Assisted Language Learning. ReCALL, 23, 25–58. doi:10.1017/S0958344010000273

Lim, J.M., McCabe, P., & Purcell, A. (2017). Challenges and solutions in speech-language pathology service delivery across Australia and Canada. European Journal for Person Centred Healthcare, 5, 120–128. doi:10.5750/ejpch.v5i1.1244

Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., & Nöth, E. (2009a). PEAKS – A system for the automatic evaluation of voice and speech disorders. Speech Communication, 51, 425–437. doi:10.1016/j.specom.2009.01.004

Maier, A., Honig, F., Bocklet, T., Noth, E., Stelzle, F., Nkenke, E., & Schuster, M. (2009b). Automatic detection of articulation disorders in children with cleft lip and palate. Journal of the Acoustical Society of America, 126, 2589–2602. doi:10.1121/1.3216913

Maier, A., Honig, F., Hacker, C., Schuster, M., & Noth, E. (2008). Automatic evaluation of characteristics of speech disorders in children with cleft lip and palate. Paper presented at the Interspeech 2008 – International Conference on Spoken Language Processing, Brisbane, Australia.

Martin, T. (2014). The evolution of the smartphone. Pocketnow, 2017 (25th June). Retrieved from Pocketnow.com website: http://pocketnow.com/2014/07/28/the-evolution-of-the-smartphone

Mazenan, M. N., Swee, T. T., & Soh, S. S. (2015). Recognition test on highly newly robust Malay corpus based on statistical analysis for Malay articulation disorder. Paper presented at the BMEiCON 2014 – 7th Biomedical Engineering International Conference.

McAllister, L., McCormack, J., McLeod, S., & Harrison, L.J. (2011). Expectations and experiences of accessing and participating in services for childhood speech impairment. International Journal of Speech-Language Pathology, 13, 251–267. doi:10.3109/17549507.2011.535565

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

Moher, D., Liberarti, A., Tetzlaff, J., & Altman, D.G. & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Medicine, 6, e1000097. doi:10.1371/journal.pmed.1000097

Morton, H., Gunson, N., & Jack, M. (2012). Interactive language learning through speech-enabled virtual scenarios. Advances in Human-Computer Interaction, 2012, 389523.

Murray, E., McCabe, P., & Ballard, K.J. (2014). A systematic review of treatment outcomes for children with childhood apraxia of speech. American Journal of Speech-Language Pathology, 23, 486–504. doi:10.1044/2014_AJSLP-13-0035

Murray, E., McCabe, P., & Ballard, K.J. (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 & Hearing Research, 58, 669–686. doi:10.1044/2015_JSLHR-S-13-0179

Murray, E., McCabe, P., Heard, R., & Ballard, K.J. (2015). Differential diagnosis of children with suspected childhood apraxia of speech. Journal of Speech, Language & Hearing Research, 58, 43–60. doi:10.1044/2014_JSLHR-S-12-0358

Mustafa, B.M., Rosdi, F., Salim, S.S., & Mughal, M.U. (2015). Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker. Expert Systems with Applications, 42, 3924–3932. doi:10.1016/j.eswa.2015.01.033

Navarro-Newball, A.A., Loaiza, D., Oviedo, C., Castillo, A., Portilla, A., Linares, D., & Álvarez, G. (2014). Talking to Teo: Video game supported speech therapy. Entertainment Computing, 5, 401–412. doi:10.1016/j.entcom.2014.10.005

Neri, A., Mich, O., Gerosa, M., & Giuliani, D. (2008). The effectiveness of computer assisted pronunciation training for foreign language learning by children. Computer Assisted Language Learning, 21, 393–408. doi:10.1080/09588220802447651

Nicolao, M., Beeston, A.V., & Hain, T. (2015 April 19–24). Automatic assessment of English learner pronunciation using discriminative classifiers. Paper presented at the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

O'Callaghan, C., McAllister, L., & Wilson, L. (2005). Barriers to accessing rural paediatric speech pathology services: Health care consumers' perspectives. Australian Journal of Rural Health, 13, 162–171. doi:10.1111/j.1440-1854.2005.00686.x

O'Shaughnessy, D. (2008). Invited paper: Automatic speech recognition: History, methods and challenges. Pattern Recognition, 41, 2965–2979. doi:10.1016/j.patcog.2008.05.008

O'Shaughnessy, D. (2015, 28-30 October, 2015). Automatic speech recognition. Paper presented at the 2015 Chilean Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), Santiago, Chile.

Obach, D.D., & Cordel, M.O. (2012, 19-22 Nov. 2012). Performance comparison of ASR classifiers for the development of an English CAPT system for Filipino students. Paper presented at the TENCON 2012 IEEE Region 10 Conference.

Oliveira, C., Lousada, M., & Jesus, L.M.T. (2015). The clinical practice of speech and language therapists with children with phonologically based speech sound disorders. Child Language Teaching & Therapy, 31, 173–194. doi:10.1177/0265659014550420

Pantoja, M. (2014). Automatic pronunciation assistance on video. Paper presented at the PIVP 2014 - Proceedings of the 1st International Workshop on Perception Inspired Video Processing, Workshop of MM 2014.

Parnandi, A., Karappa, V., Lan, T., Shahin, M., McKechnie, J., Ballard, K., . . . Gutierrez-Osuna, R. (2015). Development of a remote therapy tool for childhood apraxia of speech. ACM Transactions on Accessible Computing, 7, 10. doi:10.1145/2776895

Ruggero, L., McCabe, P., Ballard, K.J., & Munro, N. (2012). Paediatric speech language pathology service delivery: An exploratory survey of Australian parents. International Journal of Speech-Language Pathology, 14, 338–350. doi:10.3109/17549507.2011.650213

Saz, O., Lleida, E., & Rodríguez, W. R. (2009). Avoiding speaker variability in pronunciation verification of children's disordered speech. Paper presented at the Proceedings of the 2nd Workshop on Child, Computer and Interaction, WOCCI '09.

Saz, O., Yin, S.C., Lleida, E., Rose, R., Vaquero, C., & Rodríguez, W.R. (2009). Tools and technologies for computer-aided speech and language therapy. Speech Communication, 51, 948–967. doi:10.1016/j.specom.2009.04.006

Schipor, O.A., Pentiuc, S.G., & Schipor, M.D. (2012). Automatic assessment of pronunciation quality of children within assisted speech therapy. Automatinis vaikų tarsenos kokybJs vertinimas pagalbinio kalbJjimo terapijoje, (122), 15–18.

Shahin, M., Ahmed, B., & Ballard, K. J. (2012). Automatic classification of unequal lexical stress patterns using machine learning algorithms. Paper presented at the 2012 IEEE Workshop on Spoken Language Technology, Miami, FL, USA.

Shahin, M., Ahmed, B., McKechnie, J., Ballard, K., & Gutierrez-Osuna, R. (2014). Comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech. Paper presented at the Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.

Shahin, M., Ahmed, B., Parnandi, A., Karappa, V., McKechnie, J., Ballard, K.J., & Gutierrez-Osuna, R. (2015). Tabby Talks: An automated tool for the assessment of childhood apraxia of speech. Speech Communication, 70, 49–64. doi:10.1016/j.specom.2015.04.002

Shahin, M., Epps, J., & Ahmed, B. (2016). Automatic classification of lexical stress in English and Arabic languages using deep learning. Paper presented at the Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.

Shaikh, N., & Deshmukh, R.R. (2016). Speech recognition system – a review. IOSR Journal of Computer Engineering, 18, 1–9.

Simmons, E.S., Paul, R., & Shic, F. (2016). Brief Report: A mobile application to treat prosodic deficits in autism spectrum disorder and other communication impairments: A pilot study. Journal of Autism & Developmental Disorders, 46, 320–327. doi:10.1007/s10803-015-2573-8

Singh, S., Thakur, A., & Vir, D. (2015). Automatic articulation error detection tool for Punjabi language with aid for hearing impaired people. International Journal of Speech Technology, 18, 143–156. doi:10.1007/s10772-014-9256-2

Skahan, S.M., Watson, M., & Lof, G.L. (2007). Speech-language pathologists' assessment practices for children with suspected speech sound disorders: results of a national survey. American Journal of Speech-Language Pathology, 16, 246–259. doi:10.1044/1058-0360(2007/029)

Strik, H., & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. Speech Communication, 29, 225–246. doi:10.1016/S0167-6393(99)00038-2

Su, H. Y., Wu, C. H., & Tsai, P. J. (2008). Automatic assessment of articulation disorders using confident unit-based model adaptation. Paper presented at the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings.

Suanpirintr, S., & Thubthong, N. (2007). The effect of pauses in dysarthric speech recognition study on Thai cerebral palsy children. Paper presented at the Proceedings of the 1st international convention on Rehabilitation engineering &#38; assistive technology: in conjunction with 1st Tan Tock Seng Hospital Neurorehabilitation Meeting, Singapore.

Sztaho, D., Nagy, K., & Vicsi, K. (2010). Subjective tests and automatic sentence modality recognition with recordings of speech impaired children. Paper presented at the Proceedings of the Second international conference on Development of Multimodal Interfaces: active Listening and Synchrony, Dublin, Ireland.

Thomas, D.C., McCabe, P., & Ballard, K.J. (2014). Rapid Syllable Transitions (ReST) treatment for childhood apraxia of speech: The effect of lower dose-frequency. Journal of Communication Disorders, 51, 29–42. doi:10.1016/j.jcomdis.2014.06.004

Ting, H. N., & Mark, K. M. (2008). Speaker-dependent Malay vowel recognition for a child with articulation disorder using multi-layer perceptron. Paper presented at the IFMBE Proceedings.

To, C.K., Law, T., & Cheung, P.S.P. (2012). Treatment intensity in everyday clinical management of speech sound disorders in Hong Kong. International Journal of Speech-Language Pathology, 14, 462–466. doi:10.3109/17549507.2012.688867

Tommy, C. A., & Minoi, J. L. (2016, 4-8 Dec. 2016). Speech therapy mobile application for speech and language impairment

.

children. Paper presented at the 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES).

van Santen, J.P.H., Prud'hommeaux, E.T., & Black, L.M. (2009). Automated assessment of prosody production. Speech Communication, 51, 1082–1097. doi:10.1016/j.specom.2009.04.007

Verdon, S., Wilson, L., Smith-Tamaray, M., & McAllister, L. (2011). An investigation of equity of rural speech-language pathology services for children: A geographical perspective. International Journal of Speech-Language Pathology, 13, 239–250. doi:10.3109/17549507.2011.573865

Wang, Y.H., & Young, S.S.C. (2015). Effectiveness of feedback for enhancing English pronunciation in an ASR-based CALL System. Journal of Computer Assisted Learning, 31, 493–504. doi:10.1111/jcal.12079

Wielgat, R., Zieliński, T.P., Woźniak, T., Grabias, S., & Król, D. (2008). Automatic recognition of pathological phoneme production. Folia Phoniatrica et Logopaedica, 60, 323–331. doi:10.1159/000170083

Williams, L.A. (2012). Intensity in phonological intervention: Is there a prescribed amount?. International Journal of Speech-Language Pathology, 14, 456–461. doi:10.3109/17549507.2012.688866

Chapter 3:

Automatic speech recognition tools for childhood apraxia of speech: The importance of lexical stress

Chapter 2 reported on findings from a systematic review of the literature on the use of automated speech analysis (ASA) tools for children's speech production. Overall, these findings indicated that ASA tools are unable to meet the clinical acceptable thresholds of accuracy of judgement when applied to disordered speech, either for phoneme level judgments or judgments of prosody, including lexical stress (McKechnie et al., 2018). To date, there have been proportionately more studies exploring ASA for phoneme level analysis than prosodic level analysis (approximately 75% compared to 25% respectively) (McKechnie et al., 2018). Of the 32 papers reported in the literature review, only three reported on tools for CAS and these were all from our research team (Parnandi et al., 2015; Shahin, Ahmed, McKechnie, Ballard, & Gutierrez-Osuna, 2014; Shahin et al., 2015). Of these, two included analysis of lexical stress (Parnandi et al., 2015; Shahin et al., 2015) which has emerged as a critical characteristic of CAS.

Lexical stress

Lexical stress has been established as an important aspect of speech and language development. It has emerged as an influential component in models of how typically developing children learn to read aloud (Arciuli, Monaghan, & Seva, 2010) and models of how humans make lexical decisions and segment the speech stream (Mattys, 1997). It has also demonstrated importance for the diagnosis and treatment of several childhood disorders such as CAS (Ballard, Robin, McCabe, & McDonald, 2010; Shriberg et al., 2003); language disorder (Aguilar-Mediavilla, Sanz-Torrent & Serra-Raventos, 2002); literacy difficulties (Leitão, Hogben, & Fletcher, 1997); and autism spectrum disorder (e.g. McCann & Peppe, 2003; Paul, Augustyn, Klin & Volkmar, 2005). Difficulty with lexical stress production has been found to negatively impact intelligibility (Field, 2005; Klopfenstein, 2009) and perceptions about social and communicative competence (Paul et al., 2005) and has even been linked with reduced likelihood living independently (Shriberg & Widder, 1990)

In English stressed (i.e., strong) and unstressed (i.e., weak) syllables tend to alternate both within and across words within a phrase or sentence (Fletcher, 2010; Greenberg, 1999) There are three lexical stress patterns in English. Most words of more than one syllable are classified as having either a strong–weak (SW) pattern, for example the word CONduct, or a weak–strong (WS) pattern, for example the word conDUCT. The strong–strong (SS) pattern, such as in the word FOOTBALL, is less common. There is a subset of English homographs in which lexical stress serves to distinguish between grammatical word classes such as noun and verb (e.g., CONduct vs. conDUCT). Lexical stress can therefore also provide critical information during online spoken word recognition (Arciuli & Slowiaczek, 2007; Cooper, Cutler, & Wales, 2002; Slowiaczek, 1990). English speakers typically develop an awareness to these differing stress patterns by adulthood (e.g. Arciuli & Cupples, 2004, 2006, 2007). Vowels in stressed syllables tend to be longer (Fletcher, 2010; Greenberg, 1999), louder, and higher in pitch than unstressed syllables.

The acoustic features used to measure lexical stress include vowel duration (msec), vocal intensity (dB SPL), and fundamental frequency (*f*0 in Hz). Vowel quality also contributes to lexical stress categorisation, with vowels in weak syllables typically reduced to schwa. English speakers use all three of these acoustic correlates to mark lexical stress. The prominence of any one of the three acoustic correlates can vary according to factors such as grammatical structure and word position within the sentence (Turk & Sawusch, 1997; Van Kuijk & Boves, 1999).

Lexical stress in CAS

Prosodic difficulties have been included in descriptions of children with CAS from as early as 1972 (see Skinder, Strand & Mignerey, 1999). Shriberg and colleagues first suggested stress production difficulties as a potential diagnostic marker for CAS during development of the Prosody-Voice Screening Profile (Shriberg, Kwiatkowski & Rasmussen, 1990). These authors conducted a series of studies in 1997, which found a higher proportion of lexical stress errors in children with suspected CAS when compared with typically developing children (Shriberg, Aram, & Kwiatkowski, 1997c). They concluded that lexical and phrasal stress errors, particularly ‘excessive/equal/misplaced’ stress, was a valid diagnostic marker for a subtype of CAS (Shriberg et al., 1997c).

Initial investigations by Skinder, Strand and Mignerey (1999) found support for perceptual differences in lexical and phrasal stress accuracy when comparing children with CAS to typically developing children but did not find any differences between the two groups in the use of acoustic features to mark stress. In later studies, Skinder and colleagues also reported acoustic differences between children with CAS and TD children in their marking of lexical and phrasal stress (Skinder, Connaghan, Strand, & Betz, 2000; Skinder, Strand, Stoel-Gammon, Mignerey & Betz, 1999b). In one study, they found that children with CAS were not able to use acoustic cues to mark lexical stress as effectively as TD children, with particular difficulty observed with reducing duration to mark an unstressed syllable (Skinder, Strand, Stoel-Gammon et al., 1999). In a later study, the same researchers also reported that correctly stressed words could be differentiated from incorrectly stressed words using the acoustic correlates of peak fundamental frequency and peak amplitude (Skinder et al., 2000).

To further their work around identifying potential diagnostic markers for CAS, Shriberg and colleagues (2003) developed and validated a metric for acoustic analysis of lexical stress called the Lexical Stress Ratio (LSR). The LSR is a composite score statistically derived from the ratios of three acoustic variables (frequency area, amplitude area and duration) to quantify relative prominence across adjacent syllables in bisyllabic words (Shriberg et al., 2003). A high LSR indicates excess stress on the stressed syllable and a low LSR indicates reduced stress on the stressed syllable (Shriberg et al., 2003). Loss of data due to children purposefully lengthening the second syllable of words with weak-strong (WS) and strong-strong (SS) stress patterns led to the authors analysing only the strong-weak (SW) words. While this is a limitation to the study, their findings demonstrated that, when rank ordered, 83% of the LSR values which fell in in the upper and lower extremes of the continuum came from participants with suspected CAS, thus confirming that lexical stress errors are a valid marker of CAS. The LSR was further explored by Hosom and colleagues (2004), who demonstrated the feasibility of using ASR methods to increase the efficiency of computing LSRs. The ASR system described in this study first used forced alignment to detect vowel phoneme boundaries, then automatically extracted the same frequency, amplitude and duration variables which were analysed in Shriberg et al. (2003) to calculate the LSR. The results from the automated LSR measurement fell within the standard error of the mean LSRs reported in Shriberg et al. (2003) (Hosom et al., 2004), however, these ASR methods were not further explored with larger sample sizes nor adopted into clinical practice.

More recently, prosodic deficits emerged as a discriminant measure of CAS in two discriminant function analysis models (Murray, McCabe, Heard, & Ballard, 2015). In Model 1, Murray and colleagues (2015) demonstrated that, from the 24 quantitative measures extracted from assessment data, two measures – percent lexical stress match and presence of syllable segregation – presented 82% diagnostic accuracy against expert diagnosis of CAS and comorbid CAS (i.e. CAS plus an additional diagnosis). Greater diagnostic accuracy was obtained after removing four children with comorbid CAS and three non-CAS children with structural impairments from the dataset. Model 2 achieved 91% diagnostic accuracy with expert diagnosis using four quantitative measures including percent lexical stress match, presence of syllable segregation, percent phonemes correct and accuracy on diadochokinetic tasks (Murray et al., 2015). Model 2 achieved 100% diagnostic sensitivity and specificity for all children used to create the model with 97% sensitivity and 100% specificity when applied to the four comorbid CAS children and three non-CAS children with structural deficits

(Murray et al., 2015). While four behaviours emerged in this analysis, lexical stress was the strongest predictor in the model. It is also important to note that lexical stress operates over larger units of multiple syllables, and it is the planning/programming of syllable sequences that is particularly impaired in apraxia of speech (ASHA, 2007; Hall, Jordan, & Robin, 2007). Therefore, an accurate automated measurement for lexical stress production would allow development of a powerful diagnostic and treatment outcome tool.

ASR tools for lexical stress in CAS

In light of the demonstrated importance of lexical and phrasal stress as core features of CAS, it is critical that ASR tools developed for the evaluation and/or treatment of CAS are able to accurately determine the lexical stress patterns produced. Compared with phoneme accuracy, lexical stress production is relatively easy to measure acoustically. The three variables of vowel f0, intensity and duration are straightforward to extract, once the vowel is identified in the acoustic signal.

Despite the relative ease of automated measurement of lexical stress, the results of the systematic literature review presented in Chapter 2 indicate that only seven of the papers reviewed has designed tools to specifically analyse word level (i.e. lexical) stress. From these seven, only four studies had tested these tools using speech disordered populations, with small sample sizes (n < 20). Of the four studies exploring automated analysis of prosody in disordered speech, two had been specifically developed to evaluate lexical stress in CAS, also with small sample sizes (n < 7). Performance accuracy data presented in Chapter 2 indicate that ASR tools are able to reliably classify lexical stress patterns in typically developing speech but classification of lexical stress patterns in disordered speech continues to fall short of the accepted clinical threshold for reliability between raters when compared to expert perceptual judgment (McKechnie et al., 2018).

Chapter 4 presents a paper exploring the accuracy of a custom-designed automated lexical stress classification tool. Lexical stress analysis has been chosen both for its demonstrated importance as a key feature of CAS and potential for use as an outcome measure in intervention (Ballard et al., 2010; Miller, Plante, Ballard, & Robin, 2018). This paper extends previous work in this area by (a) applying the ASA to a larger sample of children with CAS, and (b) priming the tool's dictionary with knowledge of the specific mispronunciations made by the participants in an attempt to overcome some of the previously reported limitations in automated locating of the vowel boundaries in the acoustic signal. Through testing and refining the lexical classifier tool, the two primary aims of the paper are to determine (a) whether this tool can perform reliably for both children with typical development and with CAS, for words across a range of stress contrast patterns, and (b) ultimately, whether the tool is ready for integration into an app-based therapy program for CAS to provide children with automated feedback on performance accuracy.

Chapter 4: An automated lexical stress classification tool for assessing dysprosody in childhood apraxia of speech

Paper 2: An automated lexical stress classification tool for assessing dysprosody in childhood apraxia of speech

The paper presented in this chapter is currently under review for publication as follows:

McKechnie, J., Shahin, M., Ahmed, B., Murray, E., McCabe, P., Arciuli, J., & Ballard, K.J. (submitted). An automated lexical stress tool for assessing dysprosody in childhood apraxia of speech. *Journal of Speech, Language and Hearing Research.*

Author Contribution Statement

As co-author of the above paper and primary supervisor, I confirm that Jacqueline McKechnie made the following contributions:

- Conception of the research questions in collaboration with co-authors
- Literature reviews
- Data collection
- Data entry and data analysis/interpretation in collaboration with co-authors
- Writing of the first draft of the paper, with subsequent drafts developed in collaboration with co-authors
- Journal submission and review process

Kirrie J. Ballard

Date: 27.2.19

An automated lexical stress classification tool for assessing dysprosody in childhood apraxia of speech

Jacqueline McKechnie [1, 2], Mostafa Shahin [3], Beena Ahmed [3,4], Elizabeth Murray[1], Patricia McCabe [1], Joanne Arciuli [1], & Kirrie J. Ballard [1]

[1] Faculty of Health Sciences, The University of Sydney, Lidcombe, NSW, Australia

[2] Faculty of Health, The University of Canberra, Bruce, ACT

[3]Texas A&M University at Qatar, Doha, Qatar

[4] Faculty of Engineering, University of New South Wales, Sydney, NSW, Australia

Keywords: childhood apraxia of speech, lexical stress, automated speech analysis

Running head: Automated lexical stress classification for CAS

Address for Correspondence:
Jacqueline McKechnie
Faculty of Health Sciences
University of Sydney
Lidcombe, NSW 2141
Ph: (02) 9351 9413
Email: jacqueline.mckechnie@sydney.edu.au

<u>Abstract</u>

Purpose: Childhood apraxia of speech (CAS) is characterized by difficulty with production of lexical stress contrasts in polysyllabic words, particularly those with weak (W) – strong (S) onset (e.g. tomato: /tə’matoʊ/). Here, we explore the potential for automated classification tools to increase objectivity, accuracy and efficiency of lexical stress analysis in words with different stress patterns across the first two syllables.

Method: Speech samples from 16 typically developing (TD) children and 26 children with CAS producing 50 common polysyllabic words were input to a Deep Neural Network (DNN)-based classification tool. We extend earlier work by comparing automated classification accuracy with clinical auditory perceptual judgment using samples from both TD children and children with CAS. We also compare classification accuracy for TD speech to CAS speech; explore potential improvement to classification accuracy using a knowledge-driven analysis approach where lexical stress analysis algorithms can accommodate common syllabic speech sound errors in the sample; and explore both within-word segmental features and within-participant factors such as age and severity of speech disorder as potential sources of automated classification error.

Result: Classification accuracy for TD speech overall met the clinical threshold of > 80% agreement with human judgment, although high accuracy for strong-weak words (SW) drove this result. The threshold was not reached for CAS speech overall (76.77%) but was met for SW words (86.8%). Accuracy for CAS was moderately correlated with phonemic accuracy and, when restricted to words produced with perceptually accurate lexical stress, tool classification reached 80% accurate for SW words and for the combined set of SW and weak-strong (WS) words (strong-strong/SS words excluded). There was no significant advantage to using a knowledge-driven approach. Within-word features such as liquid or glide consonants

adjacent to the vowel and non-schwa unstressed vowel phonemes were only weakly correlated with classification error.

Conclusion: Automated speech analysis tools continue to improve in their ability to make decisions that are comparable with traditional clinical auditory perceptual judgment. The system tested here had clinically acceptable accuracy for words with SW stress for both TD and CAS speech and for words produced with perceptually accurate lexical stress. The findings represent an improvement over previous methods for lexical stress analysis in childhood speech disorders in terms of ease of use and accuracy against human perceptual judgments. Future challenges are improving the accuracy of these tools for impaired speech and, in particular, analysis of words with weak onsets that are commonly affected in childhood speech impairments.

Introduction

Difficulty with the production of lexical stress has been identified as one of the core deficits in childhood apraxia of speech (CAS) (ASHA, 2007) and has been studied for its potential as a diagnostic marker (Shriberg, Aram, & Kwiatkowski, 1997c; Murray, McCabe, Heard & Ballard, 2015). Assessment of lexical stress production is traditionally impressionistic (Peppe, 2009) and therefore vulnerable to various sources of error and bias within- and between-rater (Charter, 2003; Kent, 1996)**.** Objective acoustic measurement is advantageous for overcoming issues of perceptual bias or drift, however, manual measurement is time consuming for clinicians (Diehl & Paul, 2009). This study aims to further the work of Shahin and colleagues (Parnandi et al., 2015; Shahin, Ahmed, & Ballard, 2012; Shahin et al., 2015; Shahin, Gutierrez-Osuna, & Ahmed, 2016) in the development of an automated lexical stress classification tool for CAS. Here, we compare tool-based classification of stress patterns with expert auditory perceptual judgment. We also explore the potential for knowledge-driven systems to boost tool-based classification accuracy for mispronounced words; as well as examine classification errors for potential within-word segmental factors, which may affect tool accuracy and so guide stimulus selection for reliable assessment instruments in the future.

CAS is a congenital speech sound disorder of neurological origin which affects the accuracy and consistency of the movements and movement transitions required for speech sound production in the absence of any muscular or nerve deficits (ASHA, 2007). The primary impairment is in the programming of the temporal and spatial parameters of movement sequences, manifesting in speech sound and/or prosodic errors (ASHA, 2007). Experts in CAS have reached some level of consensus around three segmental and suprasegmental features that are consistent with deficits in programming of speech movements: “(a) inconsistent errors on consonants and vowels in repeated productions of syllables or words;

(b) lengthened and disrupted coarticulatory transitions between sounds and syllables; and (c) inappropriate prosody, especially in the realization of lexical or phrasal stress” (ASHA, 2007, pp 4, 54 and 59).

Prosodic deficits continue to demonstrate significance as a valid diagnostic feature of CAS (Hosom, Shriberg & Green, 2004; Murray, McCabe, Heard, & Ballard, 2015; Shriberg et al., 2003)). Murray and colleagues (2015) conducted a discriminant function analysis using a set of 24 quantitative measures extracted from a comprehensive clinical battery for diagnosing CAS. The gold standard comparison was expert diagnosis based on ASHA’s 3-item consensus-based feature list (described above) (ASHA, 2007) and Strand’s 10-point checklist (Shriberg, Lohmeier, Strand, & Jakielski, 2012). Perceptually-judged error in producing lexical stress contrast in polysyllabic words was the strongest predictor of CAS diagnosis in the regression models presented (Murray et al., 2015). This warrants development of an objective and efficient assessment tool for lexical stress to aid clinical diagnosis of CAS.

Lexical Stress

The English language uses lexical stress patterns in which stressed or strong syllables and unstressed or weak syllables tend to alternate both within words and across words within a phrase or sentence (Fletcher, 2010; Greenberg, 1999). Over 90% of English words are polysyllabic (contain more than one syllable) and therefore carry alternating lexical stress (Arciuli, Monaghan, & Seva, 2010). Most polysyllabic English words are classified as having either strong-weak (SW; e.g. DInosaur /ˈdaɪnəˌsɔ/) or weak-strong (WS, e.g. poTAto /pəˈteɪˌtoʊ/) over the first two syllables, with a tendency towards final syllable lengthening and medial syllable shortening (Fletcher, 2010). Vowels in stressed syllables tend to be longer (msec) (Fletcher, 2010; Greenberg, 1999), louder (dB), and higher in fundamental frequency (f0) than vowels in unstressed syllables (Kochanski, Grabe, Coleman, & Rosner, 2005). Duration and loudness make a greater contribution to listeners' perception of prominence than fundamental frequency (Kochanski et al., 2005) especially in a single word picture naming task (Ballard, Djaja, Arciuli, James, & van Doorn, 2012). Lexical stress in English can signal differences in grammatical word classes, such as noun (e.g. REcord) and verb (e.g. reCORD) and can be influential in spoken word recognition tasks (e.g. Arciuli & Slowiaczek, 2007; Cooper, Cutler & Wales, 2002). Given that 85-90% of content words in English carry initial stress, stressed syllables tend to be used by the listener to identify word boundaries within connected speech (Cutler & Norris, 1988). The influence of lexical stress on word identification and word segmentation extends beyond the local/adjacent syllabic context to more distal prosodic patterns, with manipulation of lexical stress patterns earlier in a word string having a demonstrated effect on the way in which listeners perceive and use lexical stress to determine word boundaries later in the string (e.g. Breen, Dilley, MacAuley & Sanders, 2014; Dilley, Mattys & Vinke, 2010; Morrill, Dilley & MacAuley, 2014). Difficulty with production of lexical stress contrasts impacts negatively on speech intelligibility (Field, 2005; Klopfenstein, 2009), reduces speech naturalness and can lead to negative perceptions about the social and communicative competence of the speaker (Paul et al., 2005).

Measuring lexical stress

Lexical stress is a good target for acoustic analysis as it involves manipulation of segmental or syllabic duration, fundamental frequency and intensity; all variables that are easily calculated by speech analysis software. Studies focused on acoustic analyses of lexical stress have also returned findings which support this as a key feature of apraxia of speech in both developmental (Munson, Bjorum, & Windsor, 2003; Nijland et al., 2003; Shriberg et al., 2003; Skinder, Connaghan, Strand, & Betz, 2000; Skinder, Strand, & Mignerey, 1999) and acquired forms (Ballard et al., 2014; Ballard et al., 2016; Duffy et al., 2017; Vergis et al., 2014). Many of these studies did not directly compare their acoustic measures with perceptual judgments of speech. Two of these studies reported finding no acoustic differences between typically developing and CAS groups in the execution of lexical stress contrasts, despite listeners perceiving that the speakers with CAS had achieved stress production less accurately than typically developing speakers (Munson et al., 2003; Skinder et al., 1999). Skinder and colleagues (1999) suggested that listener perception may have been influenced by segmental errors, while Munson and colleagues (2003) proposed that the acoustic differences produced by speakers with CAS may not have been consistently perceived by the listeners if the degree of difference in prominence across syllables did not match the canonical representation. This hypothesis supports the findings of Fear, Cutler and Butterfield (1995) who demonstrated that listeners have a tendency to preferentially make a binary distinction between stressed and unstressed syllables, even though acoustic analysis demonstrated that an intermediate category exists in words that contain de-stressed but unreduced vowels. Two exceptions are further explored here. First, Shriberg and colleagues (2003) developed the lexical stress ratio (LSR; a single index generated from acoustic variables of vowel duration, intensity and f0) and reported that inter-rater agreement for the global judgment of whether a child should be diagnosed as suspected CAS was higher when the child's LSR fell in either the upper or lower extremes of the distribution (Shriberg et al., 2003). Second, Ballard, Robin, McCabe & McDonald (2010) reported high agreement between auditory-perceptual judgment of lexical stress accuracy and manually calculated normalized Pairwise Variability Indices (PVI), particularly for vowel duration, peak intensity and/or peak f0. PVI (Low, Grabe & Nolan, 2000; see equation below) calculates the degree of asymmetry across two adjacent syllables in a string and provides a measure that has been normalized for speech rate, vocal intensity, or f0, respectively.

Advances in technology have made objective/acoustic analysis readily available through the use of freeware such as smartphone applications like Wavepad Audio Editor (NCH software) and speech analysis freeware such as Praat (Boersma & Weenick, 2011). However, objective manual measurements are perceived to be too time consuming for clinicians to use on a regular basis (Diehl & Paul, 2009). Many clinicians report that the analysis component of the assessment process is at least equally (McLeod & Baker, 2014), if not more time consuming (Skahan, Watson, & Lof, 2007), than the direct assessment activities.

Automated analysis of lexical stress.

Automated analysis of lexical stress has been investigated for its potential to support both foreign language learning (Delmonte, 2009; Ferrer et al., 2015; Hacker, Cincarek, Maier, HeBler, & Noth, 2007; Shahin, Epps, & Ahmed, 2016) as well as assessment and treatment of various pediatric speech disorders including CAS (Hosom, Shriberg, & Green, 2004; Parnandi et al., 2015; Shahin et al., 2015), speech impairment (Sztaho, Nagy, & Vicsi, 2010) and autism (van Santen, Prud'hommeaux, & Black, 2009). Of the tools that have been applied to disordered speech, studies have reported that automated analyses range from 10% to 77.6% agreement with human judgment (Sztaho et al., 2010, and Shahin et al., 2015, respectively); that automated measures fall within the standard error of the mean of manually calculated measures (Hosom et al., 2004); and that automated analyses demonstrate moderate to strong correlation with human judgments (Hosom et al., 2004; van Santen, Prud’hommeaux, & Black, 2009). We propose applying a threshold of 80% agreement between automated acoustic analysis and human judgment of speech as this is the threshold of clinically acceptable agreement often used between two human raters (Charter, 2003; Cucchiarini, 1996). Across both the language learning and speech disordered populations, automated lexical stress analysis tools that have been able to achieve this 80% threshold have done so for correctly pronounced words (i.e. words with no segmental substitutions, distortions, deletions or additions); these tools typically do not reach clinically acceptable standards when analyzing mispronounced words (see McKechnie et al., 2018, for a review). The best performing tools reviewed by McKechnie and colleagues (2018) that had been applied to mispronounced or disordered speech had generally used knowledge-driven methods, where the tools had been supplied with data on the types of speech errors contained within the speech samples analyzed. This type of specificity limits the wider clinical applicability of such tools and necessitates the use of confined dictionaries of words for analysis as larger dictionaries will increase the phonetic neighborhood and increase the likelihood of automated systems recognizing an erroneous word based on phonetic similarity (Rubin & Kurniawan, 2013).

Shahin, Gutierrez-Osuna & Ahmed (2016) developed software, which automatically classifies children's lexical stress patterns across each adjacent syllable pair in isolated polysyllabic word productions. This tool calculates eight acoustic features for each syllable in a word, derived from the duration, f0, intensity and spectral energy of two consecutive syllables: peak to peak Teager Energy Operator (TEO) amplitude over syllable nucleus, mean TEO energy over syllable nucleus, maximum TEO energy over syllable nucleus, nucleus duration, syllable duration, maximum f0 over syllable nucleus, mean f0 over syllable nucleus, and 27 Mel-scale energy bands over syllable nucleus. These features are combined into a single wide feature vector and input into a deep neural network (DNN) classifier. From these combined features, the tool classifies each production as having either a SW, WS, SS, or WW (weak-weak) stress pattern across adjacent syllables and assigns a confidence estimate for that classification, expressed as a proportion of one. The confidence estimate is a mathematical expression of the degree of certainty that a given word was produced with the recognized (i.e. automatically assigned) stress pattern. The tool does output pairwise comparisons across all syllables for a word but, consistent with work cited earlier, we focus here on the first two syllables. Typically developing (TD) children's productions of three and four-syllable polysyllabic words initiated with these four different stress patterns were entered into the DNN classifier with overall classification accuracy against dictionary-defined stress patterns reaching 88%. Using a binary classification (SW, WS), the tool labelled stress patterns with 93% accuracy. However, for children with CAS, accuracy with the binary classification compared to human auditory perceptual judgment was lower at 73.4%.

The DNN tool (Shahin, Gutierrez-Osuna, & Ahmed, 2016) has advantages over previous models developed by the same team (Shahin, Ahmed, & Ballard, 2012; Shahin et al., 2015). First, the tool was trained using child speech rather than adult speech. Second, the DNN classifier used raw syllable-level features rather than normalized PVI measures to learn more sophisticated relationships and so reduce errors rates compared with earlier versions (Shahin et al., 2012). Although not using PVI values to inform the lexical stress classification, these can still be extracted from the output. This is particularly useful for clinicians, given that children may have difficulty learning to control only some features to mark stress (e.g. relative vowel duration but not f0). Furthermore, these measures will be useful to compare speech-impaired children's performance to emerging normative PVI data for English and other languages (Arciuli & Ballard, 2017; Arciuli & Colombo, 2016; Arciuli, Simpson, Vogel & Ballard, 2014; Ballard, Djaja, Arciuli, James & van Doorn, 2012).

Such automated tools have the potential to increase objectivity, accuracy and efficiency of speech analysis and clinical diagnosis. These findings offer support for the use of acoustic measures to profile prosodic difficulties and monitor treatment-related change.

Purpose

This study is an extension of Shahin's work (2016), which analyzed only 15 words from 10 children with CAS and compared classification accuracy for TD speakers to a dictionary-defined canonical stress pattern rather than to human judgment of the child's actually-produced stress pattern. Here, we compare the tool's classification accuracy to human auditory perceptual judgment using speech samples from both Australian English speaking TD children and children with CAS. We extend on earlier work by including a larger number of participants with CAS, and a wider range of 3, 4 and 5-syllable polysyllabic words. We also perform deeper analysis of the tool's classification accuracy using several methods. First, we explore the effects of pre-training the tool with information about specific pronunciation errors made by the children, given the advantage for knowledge-driven methods identified in the review by McKechnie and colleagues (2018). We also explore the influence of phonetic contexts within words, given that syllabic nuclei are influenced by phonetic context and that phoneme boundaries may be more or less distinct depending on context (Peterson & Lehiste, 1960); Finally, we investigate the potential influence of the age of the speaker; and severity of speech impairment (as measured by percentage of phonemes produced correctly).

Our hypotheses were as follows:

1. An automated lexical stress classifier using acoustic features of duration, f0, intensity and spectral energy across adjacent syllables in polysyllabic words will achieve ≥80% agreement with traditional 'gold standard' auditory perceptual judgments for TD speech.

2. The automated lexical stress classifier will achieve higher classification accuracy for TD speakers than for CAS speakers, for whom the likelihood of mispronunciation is high.

3. Classification accuracy will be higher when using a knowledge-driven system trained on the segmental errors represented in the disordered speech sample.

4. Classification errors will be associated with within-word features known to reduce inter-rater reliability in perceptual and manual acoustic measurement such as equivocal stress across the first two syllables (e.g. HAMBURger/ˈhæmˈbɜgə/); short vowel phonemes in the stressed syllable, (e.g. BUTterfly /ˈbʌtəˌflaɪ); ambiguous phoneme boundaries (i.e. liquid consonants at syllable onsets or offsets such as in "elephant"); or words in which weak syllables have particularly low intensity and/or undetectable pitch (i.e. unstressed vowels adjacent to unvoiced phonemes, such as "potato").

Method

Participants

Sixteen typically developing children (seven males, nine females; M = 6 yrs, range 4 – 10 yrs, IQR = 3) and twenty-six children with CAS (twenty-two males, four females; M = 4.5 yrs, range 4 – 12 yrs, IQR = 3) participated. All children were Australian English speakers.

Typically developing children were recruited via convenience sampling from the local university community. Inclusion criteria included: aged 4 – 12 years, and parent-report of typically developing receptive and expressive language skills, age-appropriate speech sound production skills as demonstrated by percent consonants correct scores above 85% and developmentally appropriate phonology on the Single Word Test of Polysyllables (Gozzard, Baker & McCabe, 2004), no hearing deficits, no oro-muscular structural deficits, indicated by age appropriate oral structure and function scores on the Oral and Speech Motor Protocol (Robbins & Klee, 1987), and no other developmental diagnoses.

Children with CAS were drawn from cohorts recruited for studies of CAS at a large metropolitan university. All children underwent a standard test battery for differential diagnosis of CAS (Murray et al., 2015). Inclusion criteria included: aged 4-12 years; age-appropriate receptive language skills, indicated by a score of ≥ 85 on the receptive language index of the Clinical Evaluation of Language Fundamentals – Preschool – Second Edition (CELF-P2; Semel, Wiig & Secord, 2006) or Clinical Evaluation of Language Fundamentals – Fourth Edition (CELF-4; Wiig, Semel & Secord, 2006); no hearing deficits; no oro-muscular structural deficits nor evidence of dysarthria, indicated by age appropriate oral structure and function scores on the Oral and Speech Motor Protocol (Robbins & Klee, 1987); and no other developmental diagnoses as per parent report. Table 1 presents demographic information, speech production test data and statistical comparisons for the participant groups.

Table 4.1. Participant demographic and speech production data.

	TD (n = 16)		**CAS (n = 26)**		**Statistics**
Variable	**M (SD)**	**Range**	**M (SD)**	**Range**	
Demographic					
Age (years)	6.1 (2.0)	4 – 10	5.9 (2.5)	4 – 12	Z = -0.71*
Sex	7 male 9 female		22 male 4 female		
Test of Polysyllables[1]					
PPC	95.2 (4.2)	85.6 –99.3	61.8 (21.1)	23.9 – 96.7	t = 6.24**
PVC	93.9 (5.3)	82.5 – 100	67.5 (17.6)	38.5 – 94.2	t = 5.82**
PCC	95.4 (4.8)	81.4 – 100	57.5 (24.7)	13.0 – 98.6	t = 5.66**
% Lexical stress matches	88.8 (8.4)	77.3 – 100	51.0 (26.6)	6.3 – 93.8	t = 5.5**
Severity rating[2]					
Mild (n) (> 85%)	15		5		
Mild-moderate (65 – 85%)	1[3]		5		
Moderate-severe (50 – 65%)	0		5		
Severe (< 50%)	0		11		

Note. [1] Gozzard, Baker, & McCabe (2008); [2] Based on percentage of consonants correct from the Test of Polysyllables; * p < .05, ** p < .0001; [3.] Participant sp011 was the youngest participant, aged 4 years, all errors were developmentally appropriate.

Stimuli

Stimuli included 50 color pictures, each representing a common 3-5 syllable word (Gozzard et al., 2004). Twenty-eight of the words are produced with unequivocal strong-weak stress across the first two syllables in Australian English (e.g. dinosaur, motorbike), 12 words with unequivocal weak-strong stress pattern (e.g. tomato, banana), and 9 with strong-strong stress pattern (e.g. hamburger, cucumber). The latter typically involve some degree of stress contrast with primary and secondary strong stress, but are difficult to assign to the SW or WS category as neither vowel is reduced to a schwa; for this reason, they are referred to here as having an *equivocal* stress pattern. The Macquarie Dictionary Online for Australian English was used to determine stress pattern (https://www.macquariedictionary.com.au). This range of stress contrasts was included to examine how the classifier handled degree of contrastiveness across the perceptual continuum. Words of three or more syllables were used in order to avoid conflating lexical stress pattern with final syllable lengthening effects in two-syllable words (Smith & Robb, 2006).

Procedure

Each child was seated at a desk in a quiet room in the speech pathology clinic of the University or in their own home. Stimuli were presented via a Powerpoint presentation on a laptop computer, with one picture per slide. Slide advancement was controlled by the researcher and, for each slide, the child was prompted to name the picture. If the child did not produce the target word, s/he was first prompted with a forced-choice question (e.g. “Is it a watermelon or a pear?”) and finally with a cue for delayed repetition (e.g. “This is a watermelon. Now you say it”). This ensured a high response rate.

Speech samples were recorded with Audacity® (Mazzoni & Dannenberg, 2000) or Praat (Boersma & Weenink, 2011) at 44,100KHz sampling frequency using a Roland Quad-Capture UA-55 [Roland, Los Angeles, CA] or Avid Recording Studio M-Audio Fast Track

Audio Interface [Avid, Burlington, MA] connected to a Dell Latitude laptop, and an adjustable head-worn microphone (AKG C520, AKG Acoustics, Vienna, Austria) at 5cm mouth-to-microphone distance. Each word for each child was saved in a separate file, labeled with the target word (e.g. watermelon.wav), for batch processing with the lexical stress classification tool.

Prior to analysis, words that did not match the syllabic structure of the target word (e.g. productions with weak syllable deletion or with syllables added) were excluded. This was done for two main reasons: (1) the forced alignment process of the tool was unsuccessful for these words as they did not contain the required number or class of phonemes; and (2) the focus of this study was on lexical stress as defined by Iuzzini-Siegel and colleagues (2015; see above) and not on syllable production skills. Less than 1% (0.86%) of sampled words from TD speakers and 22% of words sampled from CAS speakers were excluded at this step. Next, all samples were run through the automated lexical stress classification tool. The tool took each individual wav file, linked to text information about that target word, and aligned it with the expected phoneme sequence. This sequence was extracted using a phonetic dictionary to estimate and mark phoneme boundaries within the word using a Hidden Markov Model (HMM) acoustic model pre-trained using the Australian National Database of Spoken Language (ANDOSL) corpus of Australian English speakers (Millar, et al., 1994). The tool extracted information about the phoneme sequence and time boundaries from the speech signal, and used Praat scripts to compute acoustic feature information for f0, intensity, duration and spectral energy. It then combined eight acoustic features into one wide feature vector (i.e. peak to peak TEO amplitude over syllable nucleus, mean TEO energy over syllable nucleus, maximum TEO energy over syllable nucleus, nucleus duration, syllable duration, maximum f0 over syllable nucleus, mean f0 over syllable nucleus, and 27 Mel-scale energy bands over syllable nucleus). The vector for each word was input to the DNN classifier, which then categorized each word as either SW or WS, with an associated confidence level expressed as a probability. The DNN classifier was trained using the Oregon Graduate Institute Multilanguage (OGI) corpus of American English children (Cole & Muthusamy, 1994).

All samples were run through the classification tool twice: 1) the HMM model aligned the produced phoneme sequence against the expected sequence using a phonetic dictionary, which contained a single canonical representation of the target word (i.e. single-pronunciation HMM-based forced alignment), and 2) the HMM model aligned the produced phoneme sequence against the expected sequence using a phonetic dictionary, which contained multiple phonemic representations of the target words based on the range of actual variations/mispronunciations produced by the participants in the study (i.e. multiple-pronunciation HMM-based forced alignment). This was done on the hypothesis that mispronounced words may have generated errors in the forced alignment stage of processing which, in turn, may have affected the feature vector analysis and subsequent stress pattern classification.

All productions were randomly ordered and played back to an experienced speech-language pathologist (the first author) for perceptual rating of stress pattern using a 5-point Likert scale (i.e. 1 = unambiguously WS, 2 = somewhat WS, 3 = equal stress, 4 = somewhat SW and 5 = unambiguously SW). Following this, 48% of productions were randomly selected for independent rating by a second experienced speech-language pathologist (the last author), to establish reliability. Both raters were blinded to the output of the automated analysis at time of rating. Rater 2 was also blinded to participant group. Inter-rater reliability analysis was performed using the weighted Cohen's kappa statistic (Cohen, 1960). The resulting reliability estimate indicated substantial agreement, K = 0.695 (Landis & Koch, 1977). Prior to data analysis, perceptual ratings of stress patterns were collapsed to a 3-point scale, where 1 and 2 were combined into a single category coded 1 for WS; and 4 and 5 were combined to a single category coded 2 for SW.

Statistical Analysis

The primary dependent measure was the agreement between the tool and the primary human rater for lexical stress pattern assigned to a word, where 1 indicated a match between automated and manual classifications and 0 indicated a mismatch. Percent agreement and Cohen's kappa statistic (Cohen, 1960) were used to calculate strength agreement between the tool and human rater by group, lexical stress type, and HMM-based forced alignment method (single-pronunciation HMM model vs. multiple-pronunciation HMM model). Then the independent samples t-test and McNemar's chi-squared test (McNemar, 1947) were used to compare levels of agreement (i.e. tool accuracy) between and within groups and conditions.

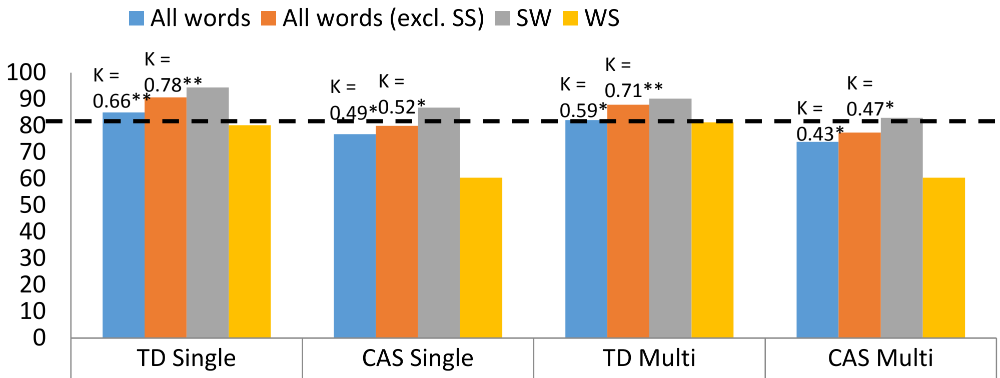
First, between-group comparisons (TD vs. CAS) tested differences in tool accuracy under (a) the single-pronunciation HMM-based forced alignment method and (b) multiple-pronunciation HMM-based forced alignment for all words together, and then for subgroups of words (i.e. excluding SS words, and considering SW or WS separately). Second, analysis of the tool's accuracy was conducted considering each group (TD, CAS) separately to explore (a) accuracy of single-pronunciation vs. multiple-pronunciation HMM-based forced alignment for all words and then the subgroups of words, (b) for the two alignment methods separately, accuracy for all words vs. all words when the SS words were excluded and then for SW vs WS words, and (c), for CAS data only, post-hoc analysis of accuracy for words perceived to have correct vs. incorrect lexical stress or correct vs. either correct or incorrect stress was used to further explore our findings. Alpha values were set at 0.01 to adjust for multiple comparisons. Effect sizes were calculated using Hedges' *g* (Hedges, 1981)

A series of correlation analyses were then run for the TD and CAS children separately. Point biserial correlation, using the nonparametric Spearman's rho statistic, was used to explore whether classification accuracy was associated with the tool's confidence estimate for the assigned classification, or with presence / absence of segmental features that may contribute to lower lexical stress contrastiveness or less reliable detection of phoneme boundaries. These features included nasal or liquid phonemes adjacent to the vowel, non-schwa unstressed vowels, or unvoiced plosives adjacent to an unstressed vowel which can lead to vowel devoicing. In addition, post-hoc analyses were conducted to further explore potential sources of classification error. We investigated whether classification accuracy was associated with age or phonemic accuracy, as measured by percent consonants correct (PCC), percent vowels correct (PVC) or percent phonemes correct (PPC). PCC, PVC and PPC are frequently used as measures of severity for children with speech sound disorders (Shriberg & Kwiatkowski, 1982). For these latter analyses, Spearman rho was used for the TD children, due to non-normally distributed data, and Pearson's correlation coefficient for the CAS group.

Results

Agreement between classifier and human judgment

Figure 1 presents the percent agreement with human auditory perceptual judgment for the automated lexical stress classification tool using single- and multiple- pronunciation HMM-based forced alignment in the TD and CAS groups. For TD children, the 80% agreement threshold was passed for both alignment methods for (i) all sampled words, (ii) all words excluding those with equivocal stress; (iii) unequivocal SW words only; and (iv) unequivocal WS words only. For CAS children, SW words reached > 80% agreement under both alignment methods with WS words at about 60% agreement.



	TD Single	CAS Single	TD Multi	CAS Multi
All words	85.03	76.77	82.06	73.91
All words (excl. SS)	90.63	80.00	87.89	77.41
SW	94.46	86.84	90.14	83.00
WS	80.19	60.37	81.25	60.40

Figure 4.1. Percent agreement and Cohen's kappa values for automated classification with single- vs. multiple-pronunciation HMM-based forced alignment compared with auditory perceptual judgment. TD = typically developing, CAS = childhood apraxia of speech, SS = strong-strong stress, SW = strong-weak stress, and WS = weak-strong stress, * = moderate effect, ** = substantial effect.

Figure 1. also presents the Cohen's kappa calculations for each group for the two alignment methods, considering (i) all the sampled words, excluding those perceptually judged as equal stress (given that the tool could only classify into either SW or WS) and (ii) the set of sampled words, excluding those perceptually judged as equal stress as well as those with typically equivocal stress (i.e. SS). Substantial agreement was achieved using single-pronunciation HMM-based forced alignment to classify all words produced by TD children (with or without the equivocal words included); and when using multiple-pronunciation HMM-based forced alignment to classify to all words excluding equivocal words from TD children. For the CAS children, all comparisons reached moderate agreement.

Between Groups Comparisons

Independent samples t-tests revealed that the tool's accuracy in classifying lexical stress patterns was significantly higher for the speech of TD children compared with CAS children on all comparisons. The effect size for WS words was medium. For all other comparisons, effect sizes were small. (see Table 2).

Table 4.2. Between group comparisons of the lexical stress classification tool's accuracy against human judgment for typically developing children (TD) vs. children with apraxia of speech (CAS), considering the single and multiple pronunciation HMM-based forced alignment methods.

	Single pronunciation			**Multiple pronunciation**		
Comparison	Statistic	*p*	*g*	Statistic	*p*	*g*
All sampled words	t = 3.53	0.0004	0.2005	t = 3.55	0.0012	0.191
All words excluding SS	t = 4.57	< 0.0001	0.304	t = 4.51	0.0001	0.2837
SW words	t = 3.38	0.0008	0.2311	t = 2.71	0.0069	0.1992
WS words	t = 3.48	0.0006	0.4378	t = 3.50	0.0005	0.463

Note. SS – Strong-Strong stress pattern (e.g. hamburger), SW = Strong-Weak (e.g. dinosaur), WS = Weak-Strong (e.g. tomato), *g* = Hedges' *g*

Within Group Comparisons

Single- vs. multiple-pronunciation HMM-based forced alignment: For the TD group, the single-pronunciation HMM-based forced alignment demonstrated significantly greater accuracy against human judgment when classifying SW words. There were no statistically significant differences between single-pronunciation HMM-based forced alignment and multiple-pronunciation HMM-based forced alignment for any of the other sample word groupings for the TD participants nor for any comparisons for the CAS group (see Table 3)

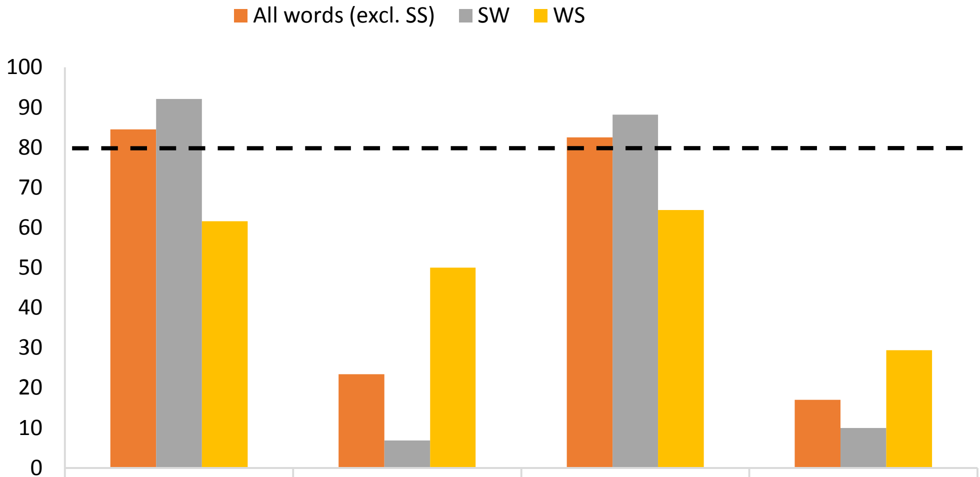
Word type: For the TD group, there was a statistically significant improvement in automated classification accuracy, for single-pronunciation HMM-based forced alignment when the equivocal words were removed from the speech sample. The effect size was small. In addition, SW words were classified significantly more accurately than WS words using single pronunciation HMM-based forced alignment, with a medium effect size. Using multiple-pronunciation HMM-based forced alignment, no significant improvement in classification accuracy was gained for the TD group by removing equivocal words, neither did the difference in classification accuracy between SW and WS words reach statistical significance for this group. For the CAS group, there was no significant increase in automated classification accuracy by removing equivocal words for either single-pronunciation HMM-based forced alignment model or multiple-pronunciation HMM-based forced alignment. Using both single- and the multiple-pronunciation HMM-based forced alignment, SW words were classified with significantly greater accuracy than WS words in the CAS group. The effect size was medium (see Table 3).

Words perceived with correct or incorrect lexical stress: Within the CAS group, for both the single- and multiple-pronunciation HMM-based forced alignment, automated classification accuracy for words perceived to have correct lexical stress met or exceeded the 80% inter-rater agreement threshold for (i) all words excluding those with equivocal stress and (ii) SW words (see Figure 2). For these word classes, there was a statistically significant difference in classification accuracy between words with perceptually accurate lexical stress and words with perceptually incorrect lexical stress and effect sizes were large in both forced alignment methods (see Table 3). In both single- and multiple-pronunciation HMM-based forced alignment, a total of 47 words from the analyzed sample were perceived as being produced with incorrect lexical stress. These 47 words were produced by 18 of the 26 participants with CAS. In the sample analyzed by the single-pronunciation HMM model, the median number of errors per participant (n = 18) was three (range 1 – 5). In the sample analyzed using multiple-pronunciation HMM-based forced alignment, the median number of errors per participant (n = 18) was 2.5 (range 1 – 5). Removing words produced with incorrect lexical stress assignment resulted in a statistically significant improvement in classification accuracy compared with results obtained from analysis of both perceptually correct and incorrectly stressed words with alpha set at 0.05 (see Table 3). Classification accuracy for WS words from children with CAS was not significantly improved when analysis was performed on words produced with perceptually accurate lexical stress (see Figure 2).

Table 4.3. Within group comparisons of the lexical stress classification tool's accuracy against human judgment for typically developing children (TD) and children with childhood apraxia of speech (CAS).

Single pronunciation vs Multiple pronunciation						
	TD			**CAS**		
Comparison	**Statistic**	***p***		**Statistic**	***p***	
All sampled words	$X^2 = 3.69$	0.0547		$X^2 = 2.68$	0.1019	
All words excluding SS	$X^2 = 3.70$	0.0543		$X^2 = 1.74$	0.1878	
SW words	$X^2 = 6.86$	0.0088		$X^2 = 3.70$	0.0545	
WS words	$X^2 = 0.17$	0.06831		$X^2 = 0.03$	0.8551	
Word type comparisons						
	TD			**CAS**		
Comparison	**Statistic**	***p***	***g***	**Statistic**	***p***	***g***
Single pronunciation						
All words vs All excluding SS	t = 2.49	0.0131	0.182	t = 1.47	0.4118	0.0729
SW vs WS words	t = 4.41	< 0.0001	0.490	t = 7.61	< 0.0001	0.7026
Multiple pronunciation						
All words vs All excluding SS	t = 2.33	0.0201	0.1677	t = 1.49	0.1369	0.0696
SW vs WS words	t = 2.32	0.0209	0.2769	t = 5.87	<0.0001	0.5611
CAS only: Lexical stress perceived as correct vs incorrect						
	Single pronunciation			***Multiple pronunciation***		
Comparison	**Statistic**	***p***	***g***	**Statistic**	***p***	***g***
All words excluding SS	t = 10.98	< 0.0001	1.6961	t = 11.34	< 0.0001	1.7368
SW words	t = 16.48	< 0.0001	3.155	t = 12.85	< 0.0001	2.4423
WS words	t = 0.95	0.3437	0.2438	t = 2.83	0.0053	0.7308
CAS only: Lexical stress perceived as correct vs either correct or incorrect						
	Single pronunciation			***Multiple pronunciation***		
Comparison	**Statistic**	***p***	***g***	**Statistic**	***p***	***g***
All words excluding SS	t = 2.07	0.0389	0.1311	t = 2.17	0.0302	0.1495
SW words	t = 2.58	0.0101	0.1622	t = 2.18	0.0296	0.1419
WS words	t = 0.23	0.82	0.0408	t = 0.687	0.4926	0.0824

Note. SS – Strong-Strong stress pattern (e.g. hamburger), SW = Strong-Weak (e.g. dinosaur), WS = Weak-Strong (e.g. tomato)



	CAS Single: Correct	CAS Single: Incorrect	CAS Multi: Correct	CAS Multi: Incorrect
🟧 All words (excl. SS)	84.52	23.40	82.52	17.02
⬜ SW	92.08	6.90	88.18	10.00
🟨 WS	61.64	50.00	64.39	29.41

Figure 4.2. Percent agreement for automated classification with single- vs. multiple-pronunciation HMM-based forced alignment compared with auditory perceptual judgment, for words produced with correct and incorrect lexical stress. CAS = childhood apraxia of speech, SS = strong-strong stress, SW = strong-weak stress, and WS = weak-strong stress.

Table 4 presents analysis of the relationship between percent agreement values, confidence estimates and within-word segmental features. For the TD samples, there was a strong positive correlation between percent agreement values and confidence estimate values using single-pronunciation HMM-based forced alignment and a weak positive correlation between percent agreement values and confidence interval values using multiple-pronunciation HMM-based forced alignment. There was a weak negative correlation between percent agreement and non-schwa unstressed vowel for both single- and multiple- pronunciation HMM-based forced alignment and a weak negative correlation between percent agreement and within-word feature of liquid or glide consonant (vs. other consonant) adjacent to the vowel only when using single-pronunciation HMM-based forced alignment. There were no significant correlations for the within-word features of long vs short stressed vowel or unvoiced plosive (vs. voiced phoneme) plus schwa in the unstressed syllable.

For the CAS samples, there was a weak positive correlation between percent agreement values and confidence intervals using single-pronunciation HMM-based forced alignment model and a moderate positive correlation between percent agreement values and confidence interval values using multiple-pronunciation HMM-based forced alignment. There was a weak negative correlation between confidence interval values and the within-word feature of liquid or glide consonant adjacent to the vowel using multiple-pronunciation HMM-based forced alignment model. There were no other significant correlations for within-word features (see Table 4).

Table 4.4. Correlation analysis (rho) exploring the relationship between classification accuracy for the single and multiple pronunciation HMM-based forced alignment methods and (a) the tool's confidence estimates in its classification and (b) within-word segmental features for typically developing children (TD) and children with apraxia of speech (CAS).

	TD		**CAS**	
	Single Pronunciation	**Multiple Pronunciation**	**Single Pronunciation**	**Multiple Pronunciation**
Confidence: Single Pronunciation	0.726**	NA	0.392**	NA
Confidence: Multiple Pronunciation	NA	0.348*	NA	0.584**
Nasal phoneme adjacent to vowel	-0.053	-0.131	-0.136	-0.116
Liquid phoneme adjacent to vowel	-0.282*	-0.266	-0.211	-0.258
Non-scwha unstressed vowel	-0.347*	-0.329*	-0.221	-0.248
Long stressed vowel	0.013	0.022	-0.205	0.006
Unvoiced plosive + schwa unstressed vowel	-0.091	-0.019	-0.202	-0.096

Table 5 presents data on the effects of age and severity on classification accuracy. For the TD samples, there were no significant correlations between age and classification accuracy across any word types for either single- or multiple- pronunciation HMM-based forced alignment. There were no significant correlations between consonant, vowel or overall phoneme accuracy and classification accuracy using single-pronunciation HMM-based forced alignment. Using multiple-pronunciation HMM-based forced alignment, there was a strong positive correlation between consonant accuracy and classification accuracy for (i) all sampled words and (ii) all words excluding those with equivocal stress. Vowel phoneme accuracy and overall phoneme accuracy were also strongly correlated with overall classification accuracy for TD samples using multiple-pronunciation HMM-based forced alignment.

For the CAS samples, classification accuracy was moderately correlated with age across all word types except SW words using both single- and multiple- pronunciation HMM-based forced alignment. Classification accuracy demonstrated moderate positive correlation with percent vowels correct for all word types except WS words and a moderate positive correlation with overall phoneme accuracy for SW words using single-pronunciation HMM-based forced alignment. Consonant, vowel and overall phoneme accuracy were each moderately correlated with classification accuracy for all word types using multiple-pronunciation HMM-based forced alignment model (see Table 5).

Table 4.5. Correlation analysis exploring the relationship between classification accuracy, considering all words or specific subsets of words, and (a) age and (b) measures of speech impairment severity (i.e. percent consonants [PCC], vowels [PVC], or phonemes correct [PPC]) for typically developing children (TD) and children with apraxia of speech (CAS).

	Single pronunciation tool				**Multiple pronunciation tool**			
	Age	**PCC**	**PVC**	**PPC**	**Age**	**PCC**	**PVC**	**PPC**
TD								
All words	0.241	0.079	0.485	0.258	0.439	0.608*	0.591*	0.518*
All excluding SS	0.179	0.040	0.111	0.082	0.472	0.583*	0.352	0.398
SW words	0.205	-0.103	0.050	0.019	0.333	0.471	0.228	0.239
WS words	0.341	0.382	0.423	0.382	0.466	0.423	0.396	0.384
CAS								
All words	0.404*	0.356	0.407*	0.384	0.447*	0.481*	0.493*	0.498*
All excluding SS	0.412*	0.334	0.392*	0.364	0.434*	0.446*	0.497*	0.476*
SW words	0.224	0.377	0.412*	0.401*	0.312	0.459*	0.491*	0.483*
WS words	0.560**	0.283	0.343	0.312	0.524**	0.403*	0.451*	0.429*

Note: SS – Strong-Strong stress pattern (e.g. hamburger), SW = Strong-Weak (e.g. dinosaur), WS = Weak-Strong (e.g. tomato),

* p < .05, ** p < .01 level (2-tailed)

Discussion

Our findings support the hypothesis that an automated lexical stress classification tool can achieve > 80% agreement with expert auditory perceptual judgments for TD speech. The tool classified stress patterns with above 80% agreement with human judgment for all word types/categories for speech samples from the TD group and these results are similar to the findings from other studies exploring automated analysis methods with typically developing speech (Kim & Beutnagel, 2011; Li, Zhang, Li, Lo, & Meng, 2011; Shahin et al., 2012; Shahin, Epps, et al., 2016; Xie, Andreae, Zhang, & Warren, 2004).

The classifier demonstrated significantly greater classification accuracy for TD speakers than for CAS speakers, satisfying our second hypothesis. Our findings also demonstrated that classification accuracy for SW words from children with CAS also met the clinical threshold of > 80% agreement between raters, whereas previous findings from disordered speech samples have not met the clinically acceptable threshold (Ferrer et al., 2015; Shahin et al., 2015; Sztaho et al., 2010). However, classification accuracy for WS words from children with CAS was well below the 80% thresholds. One possible reason for this is that producing segments of shorter duration is motorically more difficult than producing segments of longer duration (Vergis et al., 2014). Children with CAS may therefore make more significant phonemic mispronunciations as well as timing errors in their attempts at WS words and these mispronunciations contribute to poorer performance accuracy from automated tools (McKechnie et al., 2018). This hypothesis needs to be directly tested and is beyond the scope of this study. Alternatively, acoustic studies on the development of lexical stress contrastivity suggest that children's productions of WS words still may not be adult-like until the age of 12 years (Arciuli & Ballard, 2017). Such findings could also help to explain the poorer performance of the tool for WS words in both TD and CAS populations in this study, although the classifier here had been trained using child speech, which should have mitigated the influence of maturation.

Programming the dictionary of the tool’s HMM-based forced alignment module with segmental information from the range of phoneme errors produced by the participants gave no statistically significant advantage for classification accuracy. Rather, the single-pronunciation model tended to outperform the multiple-pronunciation model on measures of percent agreement with human judgment for both participant groups across most word categories. These findings are in contrast with the outcome of other research into automatic speech analysis tools which reported high accuracy and agreement with human judgment for tools trained on disordered speech using knowledge-driven recognition systems that had been specifically programmed with the types of errors the speakers were likely to produce (Chen, 2011; Duenser et al., 2016). However, these findings may also be explained with reference to the higher likelihood of error introduced by a dictionary in which there are a larger number of phonetically similar targets (Rubin & Kurniawan, 2013).

Our findings of improved classification accuracy for words produced with perceptually correct lexical stress patterns suggests that the version of the automated lexical stress classification tool that was tested in this study can determine stress patterns when productions are correct but is, as not yet able to reliably determine when stress patterns are incorrect. Although this did not hold true for WS words since removal of words produced with perceptually incorrect lexical stress gave no advantage to automated classification of WS words.

Although the spectral features extracted and filter banks used by the classifier were modeled on human speech perception and production, it is likely that there will always be differences between the human system and the modeled system. It’s possible that there are differences between the acoustic features extracted by such algorithms and the features to which the human ear is attuned when judging lexical stress accuracy. Our study implemented a tool focused on proximal prosodic contrasts (i.e. relative differences across adjacent syllables), when it is likely that the human ear can attune to, and be influenced by, prosodic patterns across the entire speech stream (e.g. Morrill et al., 2014). In addition, there can be acoustic differences in the speech signal to which the human ear does not readily attune, for example, the tendency to make binary classifications of stressed versus not stressed for words in which the de-stressed syllable contains an unreduced vowel (Fear et al., 1995). One other suggestion is that computer-driven algorithms seek to match the incoming signal to the pattern it has been trained to recognize, whereas human clinicians are trained to tune in to the incoming acoustic signal, regardless of target/expectation and are able to use contextual information, sociological factors and linguistic factors such as neighborhood density to assist with parsing and perception of spoken language. Also, the lexical stress classification system used in this study was trained only on correctly produced speech samples due to the lack of sufficiently sized databases of disordered speech data. One implication of these findings is that such tools may not yet be ready for integration into therapeutic applications until such time that they can provide accurate feedback speech production, both correct and incorrect. Until then, tools using speech recognition software are best suited to non-speech pathology applications such as education and lifestyle apps.

Our findings for TD children indicate some support for the hypothesis that classification errors are associated with more subtle lexical stress contrasts. In the TD samples, classification accuracy significantly increased when words with equivocal stress were removed. Similarly, percent agreement with human judgment tended to be lower for words in which the unstressed vowel was not fully reduced to a schwa (i.e. when the word tended more towards equivocal stress). While these syllables represent a separate and distinct acoustic category compared with stressed and unstressed syllables, the human ear has a tendency to categorise these with stressed syllables (Fear et al., 1995). In contrast, classification accuracy was not significantly improved by removing words with equivocal stress from the CAS samples, nor was there any correlation between percent agreement and the within word feature of non-schwa unstressed vowel. These findings support the hypothesis that children with CAS demonstrate reduced contrastiveness between syllables and tend towards equalized lexical stress (Ballard, Robin, McCabe, & McDonald, 2010). These findings also lend support to the hypothesis that the perception of equal or exccess stress in CAS may be a result of difficulty with control of relative timing as opposed to difficulty with the correct assignment of lexical stress (as in suggested in Vergis et al., (2014), Ballard et al., (2014) and Peter & Stoel-Gammon, (2005).

For both TD samples using the single-pronunciation model and CAS samples using the multiple-pronunciation model, classification error was weakly correlated with the within-word feature of liquid or glide phonemes adjacent to the vowel. This class of phonemes has the least distinct acoustic and spectrographic boundaries (Ballard et al., 2014; Hosom, 2009; Peterson & Lehiste, 1960) which may prove problematic for the automated/computerized phoneme alignment step in the classification process. This hypothesis is only weakly supported as it did not hold true for both pronunciation models in both participant groups.

Additional within participant factor anlaysis exploring sources of classification error only partly explained our findings. Age was correlated with classification accuracy only for the CAS group. Since the TD group did not demonstrate such a correlation between age and classification accuracy, this finding is likely to be due to the relationship between age and severity of speech impairment. Phonemic accuracy was moderately correlated with classification accuracy for some word types from the TD group using the multiple pronunciations HMM model. As might be expected, phonemic accuracy was more influential in classification accuracy for the CAS group, where the likelihood of mispronunciation was high. Consonant, vowel and overall phoneme accuracy each demonstrated moderate correlation with tool classification accuracy in all word types for multiple-pronunciation HMM-based forced alignment with vowel accuracy also correlating with classifiation accuracy for all but the WS words in single-pronunciation HMM-based forced alignment. Using percent consonants correct as a measure of severity of speech involvement (Shriberg & Kwiatkowski, 1982), classification accuracy was reduced as severity of speech impairment increased but only for multiple-pronunciation HMM-based forced alignment. Vowel accuracy was more significantly correlated with the tool's performance accuracy across the range of tool and word types. This was to be expected given that the vowel is the nucleus of the syllable and the tool performed its analysis of lexical stress at the syllable level. It is surprising that phonemic accuracy was more influential to performance accuracy of multiple-pronunciation HMM-based forced alignment than to the accuracy of single-pronunciation HMM-based forced alignment. Since the dictionary in this model of the tool had already been primed with information about the phonemic variations produced by the participants, one would expect to have a reduced likelihood that mispronunciations would affect the tool's ability to correctly classify lexical stress. From this data, this is not the case. One possible reason the multiple-pronunciation HMM-based forced alignment system did not significantly improve lexical stress classification accuracy is that the acoustic model was trained on adult Australian English speakers. This may have caused alignment problems if, instead of recognizing mispronounced words, the aligner corrupted correctly produced words where the phoneme sequence was actually matched to a sequence in the single-pronunciation forced alignment system. Another explanation may be to do with the fact that increasing the size of the dictionary resulted in higher error rates based on erroneous activation of phonetically similar targets (Rubin & Kurniawan, 2013). However, it is likely that factors other than phonemic mispronunciation and lexical stress errors are influencing automated classification accuracy, as vowel and phoneme errors accounted for approximately 26% of the variance in classification accuracy in both the single- and multiplt-pronunciation HMM-based forced alingment models.

Limitations and future directions

This research rasied as many questions as it has answered. Further research should investigate whether chidren with CAS make more significant segmental errors as well as timing errors in their productions of WS words and the potential influence this would have on autoamted lexical stress classification accuracy. Our dataset was unbalanced, with more SW words sampled than WS words. This was due to the facts that: (i) SW words are more common in the English language, particularly for nouns, while the WS pattern tends to be more common in verbs (Arciuli & Cupples, 2004, 2006); and (ii) the children were sampled using a picture naming task which accounted for the datast being comprised of nouns (i.e. picturable words) and therefore made up of more SW words than WS words. Future research should include a larger sample of WS words, particularly those produced with perceptually correct lexical stress, in order to explore potential factors related to the tool's significantly poorer performance on WS words even when words produced with perceptually inaccurate lexical stress were removed.

Further exploration of the similariteis and differences between acoustic features extracted by machine learning algorithms and those to which the human ear are attuned when judging lexical stress accuracy is warranted. This would aid in determing why the algorithm does not match human perception, particuarly for words spoken with inaccurate stress patterns.

Deeper analysis of the phonemic errors and their influence on syllable structure is required in order to further explain the finding that priming the acoustic model with specific knowledge about the types of mispronunciations present in the speech samples offered no advantage to the tool's classification accuracy.

As a result of convenience sampling, both groups were unbalanced on sex with the TD group having a greater proportion of females than the CAS group. To address this we performed between groups analysis of classification accuracy, separating participants into male and female groups, and found no significant differences in tool performance.

The HMM-based forced alignment process of the tool was trained using adult Australian English speech samples so that the phoneme segmentation process was not affected by accent differences. This module of the tool may need to be further trained or adapted using data from Australian children. Future directions for this research includes directly testing the forced alignment component of the tool by comparing the sequence of recognized phonemes with the sequence of phonemes actually produced by the child.

While the HMM-based forced alignment process was trained using Australian English speech, the DNN-based classifier was trained using a corpus of US English speech. This introduced the potential to negatively affect classification accuracy. While the influence of accent needs to be directlly tested, US English and Australian English are dialectical variations of the same stress-timed language and therefore have similar alternating lexical stress across adjacent syllables in the majority of words.

There are some limitations inherent in using a forced alignment sytem. One is that the phonemes undergo coarticulatory adjustments so that any given phoneme will vary based on its phonetic context. Therefore, as is well-known, phoneme boundaries are rarely discrete moments in time but estimates of best fit. This is particularly the case for phonemes such as liquids/glides transitioning into or out of vowel phonemes (Hosom, 2009). Another is that such systems require a constrained vocabulary and can only match the incoming speech signal to words within the predefined dictionary. Additionally, the system requires adequate training such that it can recognize words even when produced with speaker dependent variations in the speech signal (Hosom, 2009). Constraining tasks and vocabulary to reduce the potential sources of variability in the speech signal may increase computerized analysis accuracy. However, it also has the effect of limiting the ecological validity of the speech sample and reducing the clinical utility and widespread application of computerized analysis processes if an ‘off the shelf’ tool cannot readily be applied to different populations and different word sets (Hosom, 2009). Further research could consider improving the acoustic model used in the forced alignment module of the tool. One way to achieve this would be to use a more advanced acoustic model based on deep learning (Hinton, et. al. 2012). Alternatively, using domain adaptation techniques, suitable in instances where limited data from the target population is available, to adapt an acoustic model built on adult speech to childrens speech or disordered speech (Asami, Masumura, Yamaguchi, Masataki & Aono, 2017). There is also potential for ‘unsupervised systems’, built using different automatic speech recognition technology, which do not require the same level of training, to perform as accurately as trained systems and achieve comparable accuracy compared with human judgment (Tamburini & Caini, 2005).

Further research is needed beyond the single word to explore the potential for automated/computerized analysis processes to evaluate other types of prosodic function such as sentential stress, emphatic stress etc (Peppé, 2009; van Santen et al., 2009).

Conclusions

This study has potential to guide the development of a test of lexical stress production for children, with an associated automated analysis tool for diagnosis relative to normative and other-disorder populations. Error analysis can provide guidelines for refining the tool to maximize sensitivity and specificity. Such automated analysis tools may make the analysis of lexical stress difficulties more accessible to clinicians who may have limited time and fluency with acoustic analyses. This is especially salient considering the availability of easily accessible technology to capture high quality audio recordings within the clinic using free software and smart devices.

The findings of this study are similar to the results of other studies exploring the use of automated speech analysis tools for assessment and modification of speech production skills. However, classification accuracy for disordered speech, particularly WS words, is not yet reliable enough for integration into commercial or clinical systems. These findings support the findings of earlier studies on automated speech analysis which suggest that automated systems do not function well when applied to mispronounced words (see McKechnie et al., 2018, for a review).

Automated speech analysis remains a difficult problem for clinical populations in the current state of technological development. However, the promising results from TD samples and

CAS samples of SW words in the current study suggests that, once trained on larger datasets of disordered speech and with a greater range of WS exemplars, such tools have the potential to reach clinically acceptable benchmarks of accuracy against human raters in the near future.

Acknowledgements: This research was made possible by NPRP Grant # 8-293-2-124 (Ahmed, Osuna, Ballard) from the Qatar National Research Fund (a member of the Qatar Foundation) and an Australian Postgraduate Award (McKechnie). The statements made herein are solely the responsibility of the authors.

Chapter 5: Treatment considerations and alternative service delivery methods for childhood apraxia of speech

Based on the systematic review, presented in Chapter 2, and the automated lexical stress classification study, presented in Chapter 4, apps that rely on ASA to provide feedback to children on their performance during tablet-based speech practice are not yet ready to be implemented in clinical practice. However, it is important to consider the impact of using ASA-based feedback on the therapy delivery. The results of the systematic review demonstrated that most ASA tools that have been studied aim to provide a binary decision either on (i) whether a target behaviour is recognised or not, or (ii) whether the observed behaviour correctly matched the target or not. In motor learning, this is referred to as Knowledge of Results feedback (KR). This contrasts with Knowledge of Performance feedback (KP) that describes to the learner both whether or not their movement was correct and how or why it was in/correct. In the principles of motor learning (PML) framework (Schmidt & Lee, 2011), it is proposed that KP feedback accelerates acquisition of new skills but can interfere with longer-term learning, as measured by maintenance of skills after training has ended. This is possibly due to the dependence of the learner on the teacher for identifying error and guiding how to change the movement to increase accuracy, rather than developing self-evaluation and self-correction skills (i.e. the guidance hypothesis, see (Salmoni, Schmidt, & Walter, 1984). It is possibly a more passive form of learning. In contrast, KR feedback provides no guidance on how to improve accuracy once a movement is identified as incorrect. Therefore, it is proposed that the learner is forced to contemplate what went wrong and how the movement might be changed in a trial and error fashion. It is thought to be a more active form of learning.

In Chapter 6, the influence of KP versus KR is tested in children with CAS undertaking a tablet-based speech therapy intervention. While KP and KR are important to consider, the approach to teaching new motor speech skills also needs to be considered. Murray et al. (2015) provided a systematic review of different treatment approaches for CAS, which are briefly discussed here to explain which treatment approach was selected for the study presented in Chapter 6.

Efficacy of treatment for CAS

There are six treatment approaches for CAS with preliminary evidence of efficacy from at least one randomised controlled trial (RCT) or two controlled single-case experimental design studies (Murray, McCabe, & Ballard, 2014, 2015). Five are motor-based approaches including (i) Nuffield Dyspraxia Programme – Third Edition [NDP3] (Murray et al., 2015); (ii) Rapid Syllable Transition Treatment [ReST] (Ballard, Robin, McCabe, & McDonald, 2010; McCabe, Preston, & Evans, 2016; Murray et al., 2015; Thomas, McCabe, & Ballard, 2014; Thomas, McCabe, & Ballard, 2017); (iii) Dynamic Temporal and Tactile Cueing [DTTC] (Edeal & Gildersleeve-Neumann, 2011; Maas, Butalla, & Farinella, 2012; Strand & Debertine, 2000); (iv) Motor Speech Treatment Protocol [MSTP] (Namasivayam et al., 2015a; Namasivayam et al., 2015b); and (v) Ultrasound biofeedback (McCabe et al., 2016; Preston, Brick, & Landi, 2013; Preston, Leece, & Maas, 2016). The sixth approach, Integrated Phonological Awareness, simultaneously targets phonological literacy skills and segmental speech motor skills (McNeil, Gillon, & Dodd, 2009; McNeill, Gillon, & Dodd, 2009, 2010). The majority of these six approaches explicitly incorporate PML (see Chapter 1 for an overview as well as Maas, Gildersleeve-Neumann, Jakielski, & Stoeckel, 2014; Maas et al., 2008). Currently, the highest level of evidence obtained for treatments for CAS has been Level II (NHMRC, 2009), with the first RCT in CAS published in 2015 comparing NDP3 with ReST (Murray et al., 2015).

In the RCT by Murray et al. (2015), NDP3 and ReST treatments were administered using closely distributed practice (four 50 minute sessions per week for three weeks) with a high dose within sessions (at least 100 production trials per session) (Murray, McCabe, &

Ballard, 2012). Feedback schedules adhered to the specific protocol of each treatment. NDP3 treatment prescribes high frequency immediate feedback (i.e. on 100% of production trials) incorporating both KR and KP with metaphoric, kinematic and tactile articulation cues provided as needed (Murray et al., 2012). ReST prescribes a short period of pre-practice which incorporates high frequency KR+KP with metaphoric, kinematic and tactile articulation cues provided as needed, followed by a longer period of practice during which children receive intermittent KR feedback on a fading schedule following a three second delay (Murray et al., 2012). A Cochrane review recently concluded that these two treatments demonstrated similar effectiveness for children with CAS (Morgan, Murray, & Liégeois, 2018). However, as discussed in Chapter 1, translation of these treatment conditions to the Australian clinical context remains a challenge due to the high level of contact with a clinician.

Recent research evidence has emerged supporting the efficacy of the ReST intervention delivered via alternative methods. Tele-practice delivery of clinician-led intervention, matching the recommended high dose frequency and large number of practice trials recommended by Murray et al. (2014; see also Murray et al., 2015), generated treatment and generalisation gains that were similar to the gains reported following face-to-face implementation of ReST (Thomas, McCabe, Ballard, & Lincoln, 2016). Combining clinician-delivered with parent-delivered ReST intervention demonstrated mixed results, with some children achieving treatment and generalisation gains that were similar to tradition clinician-led intervention and other children making more modest or equivocal improvements (Thomas et al., 2017). These studies included small samples but the authors suggested that smaller gains may have been the result of within-child factors or the ability of the parents to judge the accuracy of their child’s productions and to adhere faithfully to the treatment protocol (Thomas et al., 2017).

One of the critical considerations, when exploring alternate service delivery options for CAS, is how these options will affect the structure of the treatment protocol and how different PMLs can be incorporated. For example, Thomas and colleagues (2014) explicitly investigated the effect of lower dose frequency (i.e., less closely distributed practice). They reported that twice weekly ReST intervention over six weeks resulted in similar treatment gains compared with ReST four times per week (i.e. equivalent cumulative intervention intensity to Murray et al., 2015). However, children receiving the lower session frequency showed stable performance in the maintenance period rather than the ongoing improvement after treatment concluded that was noted by Murray et al. (2015).

In light of the fact that NDP3 is the most frequently used intervention for CAS in Australia (Gomez, McCabe, & Purcell, 2018)), and that the recent RCT showed similar efficacy for NDP3 and ReST, it is timely to consider the impact of alternative service delivery methods on treatment efficacy for CAS using NDP3 intervention. Furthermore, it is likely that future tablet-based speech therapy apps utilising ASR and ASA will need to use real-word stimuli; NDP3 uses real words while ReST uses pseudo-words which cannot be automatically recognised. Currently, there has been no well-controlled study published that investigates alternatives to face-to-face intensive implementation of NDP3.

The NDP3 contains over 500 picturable stimulus items presented in a hierarchy from least to most complex on the theoretical basis that motor learning is complex and children need to engage in frequent, systematic practice in order to progress from foundation levels (i.e. single sounds, simple consonant and vowel syllables, single syllable words) to more complex speech patterns (i.e. two to three syllable words, phrases and sentences) (Murray et al., 2012; Williams & Stephens, 2004). Three individual goals are selected for each child, at different levels of the hierarchy of stimulus complexity, based on a comprehensive assessment of their speech sounds, prosody, vocal and nasal quality (Williams & Stephens, 2004). This provides variable practice of speech targets. Goals include learning new speech sounds or speech patterns; combining known speech behaviours into new and more complex word shapes; as well as lexical, phrasal and sentential stress, as children move through the hierarchy from single syllables to polysyllabic words and phrases (Murray et al., 2012; Williams & Stephens, 2004). Children receive KR and KP on 100% of production attempts with the aim of achieving 90% accuracy on spontaneous productions (i.e. no clinician cues or input) across 12 trials. In this way, treatment sessions following the NDP3 protocol are most similar to the pre-practice phase of a PML approach to intervention. The program was originally recommended for use in 1-hr treatment sessions once or twice per week with a clinician, supported by 20 minutes daily home practice in between sessions (Williams & Stephens, 2004).

Chapter 6 presents the results of an experimental study manipulating feedback conditions during intensive treatment of CAS using the NDP3. Whereas previous studies exploring the feedback principle during motor learning for speech have manipulated feedback schedules (i.e. immediate versus delayed feedback, e.g. Austermann Hula, Robin, Maas, Ballard, & Schmidt, 2008) or feedback frequency (i.e. feedback on every attempt versus feedback which reduces in frequency to 50% or 10%, e.g. Maas et al., 2012), this study directly compares the response to intervention in two groups receiving high frequency feedback of different types (i.e. KR+KP versus KR only). The study is designed to explore the potential for mobile technology, such as apps with in-built ASA technology, to facilitate the high intensity practice necessary for learning new motor speech behaviours. Chapter 6 explicitly compares two groups of children receiving NDP3 treatment via tablet-based stimulus presentation using a custom designed app. One group received traditional high frequency KR+KP feedback four days per week in accordance with the published NDP3 treatment protocol (Murray et al., 2012; Williams & Stephens, 2004) and the other group receiving high frequency KR+KP feedback from the clinician one day per week and high frequency KR feedback only on the other three days per week. The latter group is a simulation of the common service delivery model of one face-to-face session with a clinician per week supported by home practice (Sugden, Baker, Munro, Williams, & Trivette, 2017). In this case, the study design simulates home practice using an app which, in the future, could be equipped with in-built ASA providing the KR feedback on whether each speech attempt is recognised or evaluated as correct against a stored exemplar. A parallel-group design, with participants matched for age and severity of CAS, was used. Stratified randomisation was employed to randomly assign one child from within each matched pair to one treatment group and the other child within a pair to the alternate group. The study contributes additional Level II evidence on the efficacy of NDP3 for treating CAS.

Chapter 6: Tablet-based delivery of intensive speech therapy in children with childhood apraxia of speech: Influence of type of feedback

Paper 3: Tablet-based delivery of intensive speech therapy in children with Childhood Apraxia of Speech: Influence of type of feedback

The paper presented in this chapter has been submitted for publication as follows:

McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Murray, E., McCabe, P. & Ballard, K.J. (submitted). Tablet-based delivery of intensive speech therapy in children with Childhood Apraxia of Speech: Influence of type of feedback. *Journal of Communication Disorders.*

Author Contribution Statement

As co-author of the above paper and primary supervisor, I confirm that Jacqueline McKechnie made the following contributions:

- Conception of the research questions in collaboration with co-authors
- Literature reviews
- Collection of data
- Data entry and data analysis/interpretation in collaboration with co-authors
- Writing of the first draft of the paper, with subsequent drafts developed in collaboration with co-authors
- Journal submission

Kirrie J. Ballard

Date: 27.2.19

Tablet-based delivery of intensive treatment in childhood apraxia of speech: Influence of type of feedback

Jacqueline McKechnie [1,2*], Beena Ahmed [3,4], Ricardo Gutierrez-Osuna[5], Elizabeth Murray[1], Patricia McCabe [1], & Kirrie J. Ballard [1]

[1] Faculty of Health Sciences, The University of Sydney, Lidcombe, NSW, Australia

[2] Faculty of Health, University of Canberra, Bruce, ACT, Australia

[3] Texas A&M University at Qatar, Doha, Qatar

[4] Faculty of Engineering, University of New South Wales, Sydney, NSW, Australia

[5] Texas A & M University, College Station, TX, USA

[*] Present address of corresponding author

Keywords: childhood apraxia of speech, mobile technology, service delivery, principles of motor learning

Running head: Tablet-based treatment for CAS: Influence of feedback

Address for Correspondence:
Jacqueline McKechnie
Faculty of Health Sciences
University of Sydney
Lidcombe, NSW 2141
Email: jacqueline.mckechnie@sydney.edu.au

Abstract.

Purpose: This randomised controlled trial explored the influence of different types of feedback on response to intervention for children with childhood apraxia of speech (CAS). This was a preliminary study investigating the feasibility and effectiveness of using mobile technology that, in the future, could be equipped with automatic speech recognition (ASR) software providing feedback on speech production accuracy. Such technology has potential to bridge the gap between recommended intervention intensity as supported by research and typical intervention intensity provided by clinicians in the community.

Method: 14 children with CAS, aged 4-10 years, participated in a parallel group design, matched for age and severity of CAS. Both groups attended a university clinic for 1-hour therapy sessions 4 days a week for 3 weeks. One group received high frequency feedback comprised of both knowledge of results (KR) and knowledge of performance (KP), in the style of traditional, face-to-face intensive intervention on all days (KP group). The other group received high frequency KR+KP feedback on 1 day per week and high frequency KR feedback only on the other 3 days per week (KR group), simulating the service delivery model of one clinic session per week supported by app-based home practice. Linear mixed effects modeling was used to test the effects of group (KP, KR), time (pre-treatment, 1-week,1-month and 4-months post-treatment) and their interaction on both treated and untreated items.

Results: Both experimental groups responded to treatment, with positive gains to treated and untreated words over time and no significant differences between groups at any time point. However, only the KP group made significant gains immediately post-treatment. Small sample size and large within group variability likely reduced statistical power to detect group differences. Survey data indicated that children and their families generally viewed app-based interventions in a positive light.

Conclusion: Mobile technology has the potential to increase motivation and engagement with therapy and to mitigate barriers associated with distance and access to speech pathology services. Further research is needed to explore the influence of type and frequency of feedback on motor learning and how these parameters interact with task, child and context-related factors.

1. Introduction.

Childhood apraxia of speech (CAS) is a disorder of speech motor control that causes substantial disruption to development of intelligible and natural sounding speech (ASHA, 2007). The speech of children with CAS is notable for substitutions and distortions of speech sounds and altered prosody. CAS often persists throughout childhood and, due to its effect on learning of speech sounds and speech prosody, it can negatively impact the acquisition of phonological awareness and literacy skills (McNeill, Gillon & Dodd, 2009; Lewis, Freebairn, Hansen, Iyengar & Taylor, 2004). As a disorder of speech motor control, it is often recommended that CAS treatment apply principles of motor learning (PML) including high frequency of treatment sessions and high numbers of practice trials per session (Maas et al., 2008; Schmidt & Lee, 2011). However, parents often report difficulty accessing, attending and affording this level of clinical care and a willingness for alternative service delivery methods to alleviate these burdens (Ruggero, McCabe, Ballard, & Munro, 2012). It is here that mobile technology can play a role in giving children with CAS access to engaging high intensity speech therapy that follows the best-practice PML. The current study explores the implications for application of motor learning principles when relying on mobile technology for service delivery.

1.1 Treatment for CAS

There are different approaches to treatment for CAS currently used around the world. These include motor-based approaches, linguistic approaches and multi-modal communication approaches. In a systematic review of the evidence on treatment for CAS, Murray and colleagues (2014) identified three treatment protocols as having the strongest levels of evidence to support their use in a clinical setting to achieve positive treatment, maintenance and generalization effects. These included Dynamic Temporal and Tactile

Cueing [DTTC], Rapid Syllable Transition Treatment [ReST], and Integrated Phonological Awareness Intervention. There was suggestive evidence for ten other treatment approaches including the Nuffield Dyspraxia Programme – Third Edition (NDP3; Williams & Stephens, 2004), commonly used across Australia as best-practice (Gomez, McCabe, & Purcell, 2018). This review then led to the first and, currently, only randomized controlled trial of treatment for CAS, comparing the NDP3 and ReST (Murray, McCabe, & Ballard, 2015). Results of the RCT indicated that both NDP3 and ReST treatments resulted in similar positive treatment outcomes, particularly for generalization to real words. The NDP3 demonstrated greater immediate gains in speech accuracy and ReST treatment lead to better maintenance of treatment gains and generalization to untreated pseudo-words (Murray et. al., 2015). However, a subsequent Cochrane review (Morgan, Murray, & Liégeois, 2018) offered a more conservative interpretation of these findings based on a re-analysis. This suggested no reliable difference existed between the two treatment groups on acquisition or maintenance of targets based on small absolute mean differences in accuracy scores between the groups and that both treatment protocols demonstrated a similar, moderate level of evidence (Morgan et al., 2018). Therefore, due to NDP3's use of real words and their potential to be analysed with ASR, NDP3 is applied in the current study.

Virtually all treatments for CAS focus on segmental aspects of speech production such as sounds, syllable and word shapes, and consistency of production over repeated attempts. Several approaches have incorporated early practice for production of suprasegmental features, particularly lexical and phrasal stress (e.g. Ballard, Robin, McCabe & McDonald, 2010; Strand, Stoekel & Baas, 2006). Some have designed stimuli that can simultaneously stimulate phonological awareness and early reading skills (e.g. McCabe, McDonald-D'Silva, van Rees, Ballard & Arciuli, 2014; Moriarty & Gillon, 2006). Regardless of the specific therapy approach, the majority of studies have advocated for incorporating one or more of the principles of motor learning (see Schmidt & Lee, 2011; Maas et al., 2008 for a review).

1.2 Principles of Motor Learning

Much of what we know about PML has come from limb movement studies in non-disordered populations or investigations involving adults with acquired apraxia of speech (AOS) or dysarthria. Limb movement studies have demonstrated that greater long-term learning occurs when practice of motor targets is variable, randomized, and frequent, with delayed feedback provided on an intermittent schedule (see Maas et al, 2008 for a review). However, investigation into adult motor speech disorders revealed that some participants benefited more from low frequency feedback and others from high frequency feedback, with similar mixed results when exploring the effects of delayed versus immediate feedback (Austermann Hula, Robin, Maas, Ballard, & Schmidt, 2008). The type of feedback received also influences acquisition and retention effects. Specific augmented feedback about how a movement was performed and what to adjust on the next trial (i.e. Knowledge of Performance, KP) enhances acquisition but potentially inhibits maintenance of skill post-treatment. In contrast, feedback on the outcome or accuracy of the motor movement (i.e. Knowledge of Results, KR) leads to greater maintenance of skill (Schmidt & Lee, 2011)**.** However, KR is most effective when the learner has some internal representation of the target movement program and some ability to self-evaluate and self-correct (see Maas et al., 2008 for a review).

Few studies have explicitly investigated the influence of specific principles of motor learning in CAS. The principles that have been studied include (a) amount of practice, where providing ~ 150 trials per session leads to greater treatment, generalization and maintenance effects than only 30-40 trials per session (Edeal & Gildersleeve-Neumann, 2011); (b) treatment intensity, where twice weekly treatment sessions led to significantly better outcomes than once per week treatment sessions (Namasivayam, Pukonen, Goshulak, Hard, Rudzicz, Rietveld, Maassen, Kroll & van Lieshout, 2015); (c) practice schedule (i.e., blocked versus random practice; Maas & Farinella, 2012), where findings were mixed across participants; (d) feedback frequency (i.e., low versus high frequency feedback; Maas et al., 2012) where findings were also mixed across participants (see Maas et al 2014 for a review); and (e) distribution of practice (i.e. closely distributed at four times weekly for three weeks versus less closely distributed at twice weekly for six weeks; Thomas, McCabe & Ballard, 2014) where findings indicated comparable outcomes between the two distribution methods.

In their RCT comparing treatment outcomes from the NDP3 and ReST treatments, Murray and colleagues (2015) suggested that the type and frequency of feedback provided to children may have influenced children's responses to intervention. Although both groups made significant improvements with treatment, the NDP3 group with KR + KP feedback on 100% of trials tended toward slightly greater improvement on treated targets immediately post-treatment (i.e. greater acquisition) than the ReST group with 50% KR feedback only. Conversely, the ReST group showed slightly greater maintenance of treatment effects than the NDP3 group. While the robustness of these differences has been questioned (Morgan et al., 2018), they are consistent with previous work arguing that high frequency KR + KP feedback confers an acquisition advantage, while low frequency KR feedback confers a maintenance advantage (e.g. Maas et al., 2008; Schmidt & Lee, 2011). This is worth further investigation given that others have reported equivocal effects for high versus reduced frequency feedback when treating CAS (Maas, Butalla & Farinella, 2012). The current study was designed to specifically investigate the influence of the type of feedback received during speech production practice when therapy was delivered using mobile technology that has potential to provide KR feedback only via automated speech recognition. To isolate the effect of feedback type, we maintained feedback frequency at 100% for both experimental groups.

1.3 Service delivery

Despite research consistently demonstrating that best practice intervention frequency for speech sound disorders, including CAS, is between 2 and 4 sessions per week (Murray, McCabe, & Ballard, 2014; Namasivayam et al., 2015; Sugden, Baker, Munro, Williams, & Trivette, 2018; Thomas, McCabe, & Ballard, 2014), these intervention frequencies are uncommon in clinical practice (Gomez et al., 2018; Ruggero et al., 2012; Sugden, Baker, Munro, Williams, & Trivette, 2017). Parent involvement and home practice activities are routinely prescribed by treating clinicians as a way to supplement face-to-face therapy sessions (Lim, McCabe, & Purcell, 2017; Sugden, Baker, Munro, & Williams, 2016; Sugden et al., 2017). Homework can also provide the frequent and regular practice of speech production targets that is needed for children to acquire new skills and habitualise these new movement skills, as well as different but related movement skills, into non-intervention contexts (Gordon-Brannan & Weiss, 2007; McLeod & Baker, 2017; Olswang & Bain, 2013). Effective home practice requires that the child be motivated to engage in their practice activities and that parents or carers can be available to supervise the practice sessions and provide feedback on the accuracy of the child’s productions. However, parents and children perceive speech practice as “work” (McAllister, McCormack, McLeod, & Harrison, 2011; Thomas, McCabe, & Ballard, 2017). It is here that computer-based or app-delivered home practice can be useful for increasing a child’s engagement and motivation to participate in speech homework (Hair, Monroe, Ahmed, Ballard, & Gutierrez-Osuna, 2018; Nordness & Beukelman, 2010; Toki & Pange, 2010).

1.4 Computer-based treatment approaches

Software packages designed to act as a virtual speech-language pathologist (SLP) can be effective for a range of paediatric and adult speech disorders (Chen et al., 2016; Furlong, Erickson, & Morris, 2017). Seven out of the 20 studies reviewed in Chen et al. (2016) and six out of the 14 studies reviewed by Furlong et al. (2017) reported on computer-based treatment programs that were designed to provide speech production feedback to the user. Most of these provided only implicit feedback on accuracy using visual cues such as waveforms or animated response-contingent reactions. When feedback on speech accuracy was explicit, it was experimenter/clinician controlled and judged. None of the included studies in either review included mobile technology.

The efficacy or effectiveness of therapy delivered via most tablet and smartphone applications, however, has not been empirically tested. This may be due in part to the risks in running time- and cost-intensive experimental trials in the fast turnover environment of the app market, along with the relative low cost and low risk of the products themselves (Edwards & Dukhovny, 2017). A recent evidence-based analysis of the quality and potential therapeutic benefit of mobile applications for children’s speech disorders found that less than 3% of more than 5000 identified apps met criteria that would warrant full evaluation (Furlong, Morris, Serry, & Erickson, 2018). Of that 3% (132 unique apps that were appraised), only 19 apps (14%) were deemed to have therapeutic potential (Furlong et al., 2018).

The majority of available computer- or app-based intervention tools offer digital stimulus presentation via engaging graphics and sound effects. They typically do not provide the child with explicit feedback on the accuracy of their productions (KP or KR) nor offer remote and/or automated assessment for the SLP to monitor. The lack of integrated, automated feedback is largely due to the challenges involved in developing ASR software which can provide decisions on speech production accuracy that are highly reliable with expert clinician judgements and delivered in a timely manner (see McKechnie et al., 2018 for a review). There has been limited research on computer-based or mobile technology approaches for CAS, perhaps due to the historical challenges in defining a relatively homogeneous group of children for testing and developing computerised approaches that treat the range of CAS features, not just segmental accuracy.

1.5 Service delivery for CAS using mobile technology

To address the mismatch between the need for children with CAS to receive intensive treatment and the reality of service delivery models in Australia and elsewhere, our group have developed *Tabby Talks*, which is a multi-tiered system for facilitating remote access to speech pathology services (Parnandi et al., 2015; Parnandi et al., 2013). *Tabby Talks* consists of three components: (1) android platform software running on mobile tablets, (2) server-based learning management software (i.e., Moodle) running a speech analysis engine to evaluate children's speech attempts offline for assessment of progress in therapy, and (3) a clinician interface allowing for the remote management and updating of clients and therapy exercises (see Table 1).

Table 6.1. Features available in the *Tabby Talks* multi-tiered system for facilitating remote access to speech pathology services.

App features (online real-time)	Server features (offline)
- Real-speech audio models - Coloured flash cards - Swipe features and simple memory game - Record and playback function - Animated cartoon cat providing motivational feedback - Star chart and medals for reaching milestones - [ASR-ready][1]	- Speech recognition software - Individual case files - Access to saved audio recordings of every production trial, for each child - Graphs of session by session accuracy - Bar charts presenting star and medal data for each practiced word or goal.

Note. [1]ASR=automatic speech recognition / analysis. At the time of this study, the tablet-based app was ASR-ready. ASR was not used here as reliability of the app-compatible algorithms was still being tested.

The first step in testing a service delivery system such as *Tabby Talks* in CAS is to examine the impact of app-delivered therapy on learning, given that some parameters of the treatment session may change when feedback is based on automated speech analysis delivering only KR feedback (i.e. right / wrong decisions). While the PML approach advocates KR feedback for best maintenance of treatment effects, a learner must first be trained in producing the target movement skills accurately through what is referred to as pre-practice. Pre-practice, unlike practice, is where the clinician/trainer provides detailed KP feedback to guide and shape performance so that the learner can experience the sensorimotor consequences of performing the targeted movement(s) correctly. Pre-practice serves to guide the learner in developing an internal reference of correctness that can be accessed later during practice, once KP is removed, to guide self-evaluation and self-correction. Therefore, we propose that *Tabby Talks* can be used to provide high intensity and frequent practice on speech behaviors that the child has begun to acquire, in between the weekly in-clinic pre-practice sessions with the speech pathologist.

1.6 Purpose

This study aims to explicitly investigate the influence of type of feedback on response to treatment to determine the feasibility for such technology and software to provide an effective supplement to face-to-face intensive treatment. Here, *Tabby Talks* was populated with stimuli from the NDP3 (with permission from the authors, Williams & Stephens, 2004). Traditionally, NDP3 treatment is delivered face-to-face, multiple times per week, with 100% frequency of both KR and KP feedback (Williams & Stephens, 2004; Murray et. al. 2015). Here, we compared this traditional approach with a simulation of home-based app-delivered treatment where face-to-face therapy is delivered once per week with 100% frequency of KR and KP and, the remaining sessions are conducted in the style of home practice with only KR feedback provided at 100% frequency, simulating app-delivered, ASR-based feedback conditions.

To maintain experimental control, other conditions were held constant across the two groups: children in both treatment conditions attended the clinic for all therapy sessions, all sessions were delivered by trained student speech-language pathologists under the supervision of experienced clinicians, all treatment stimuli were delivered via the *Tabby Talks* app, and the student clinicians delivered all feedback verbally. The only treatment variable that we manipulated was the type of feedback received. Future studies will examine the feasibility of using our ASR algorithms for delivery of the KR feedback in home-based therapy.

1.7 Research Aims and Hypotheses

This study aimed to compare two methods of feedback during tablet-delivered NDP3 treatment, and to compare both methods to our historical data for traditional paper-based delivery of NDP3 (Murray et al., 2015). We also invited participants to complete a questionnaire exploring satisfaction with the treatment process; motivation and engagement with therapy activities; app features, likes, and dislikes; ease of use; and interest in further treatment using the app. We hypothesized that:

(i) Tablet-based delivery of NDP3 using high frequency KR+KP feedback would obtain similar treatment outcomes to Murray et. al.'s (2015) traditional paper-based delivery of NDP3.

(ii) Compared to participants in the high frequency KR+KP group and the traditional paper-based NDP3 group, participants in the high frequency KR condition may demonstrate smaller treatment gains immediately post-treatment (i.e. evidence of slower acquisition and generalization) but greater maintenance at 1- and 4-months post-treatment (i.e. evidence of more robust learning).

(iii) The experimental groups would demonstrate at least similar long-term outcomes to Murray et. al.'s (2015) traditional NDP3 delivery.

(iv) Participants would report overall satisfaction with tablet-based intervention including: high levels of child motivation, enjoyment and engagement with therapy activities; preference for tablet-based activities as compared with traditional paper-based activities; and willingness to use tablet-based intervention in the future.

2. Method

This study was approved by the Human Research Ethics Committee at the University of Sydney (Protocol number 2013/703). All parents provided written informed consent for their child to participate and children older than 6 years of age provided written assent.

2.1 Participants

Recruitment occurred via university research volunteer websites, advertisement in magazines of relevant professional associations, as well as flyers to community-based SLPs, social media forums for SLPs and special interest groups for CAS.

Inclusion criteria were (1) confirmed clinical diagnosis of CAS by the research team, as described below, (2) aged between 4 and 12 years at the time of treatment, (3) age appropriate receptive language skills, indicated by a standard score of ≥ 85 on the receptive language index of the Clinical Evaluation of Language Fundamentals – Fourth Edition (CELF-4; Semel, Wiig & Secord, 2006) or CELF-Preschool-Second Edition (CELF-P2; Wiig, Semel & Secord, 2006), (4) normal or adjusted to normal hearing and vision, (5) the child and at least one parent being native English speakers, and (6) no other diagnosed genetic, developmental or acquired diagnosis (e.g. autism spectrum disorder, dysarthria or intellectual disability).

A total of 38 children were referred. Referral sources were first interviewed by phone or via email to rule out potential contraindications to the inclusion criteria above. Comprehensive assessments were carried out in two stages. Assessments to determine eligibility for participation in the study included (1) a case history questionnaire; (2) hearing screening to exclude undiagnosed hearing impairment; (3) Peabody Picture Vocabulary Test – Fourth Edition (PPVT-4) (Dunn & Dunn, 2007) which is highly correlated with psychometric assessments of cognitive functioning and used here to exclude potential intellectual disability; (4) CELF-4 or CELF-P2 Australian Standardizations to exclude delayed receptive language skills; and (5) the Oral and Speech Motor Protocol (Robbins & Klee, 1987) to exclude oral-structural or dysarthria diagnoses. In addition, speech samples for perceptually judging the presence and severity of CAS were obtained through administration of (6) The Goldman-Fristoe Test of Articulation – Second Edition (GFTA-2) (Goldman & Fristoe, 2000); (7) the DEAP Inconsistency subtest (Dodd, Hua, Crosbie, Holm & Ozanne, 2002); (8) Single Word Test of Polysyllables (Gozzard, Baker & McCabe, 2004, 2008); and (9) NDP3 assessment (Williams & Stephens, 2004). Three experienced SLPs (first, fifth and sixth authors) independently confirmed diagnosis of CAS based on the presence of the three consensus-based features of CAS: (1) inconsistent errors on consonants and vowels, (2) difficulty transitioning between sounds and syllables; and (3) prosodic difficulties (ASHA, 2007). A flowchart demonstrating the outcome at each stage of the referral and screening/eligibility process for each of the 38 referred children is shown in Figure 1.

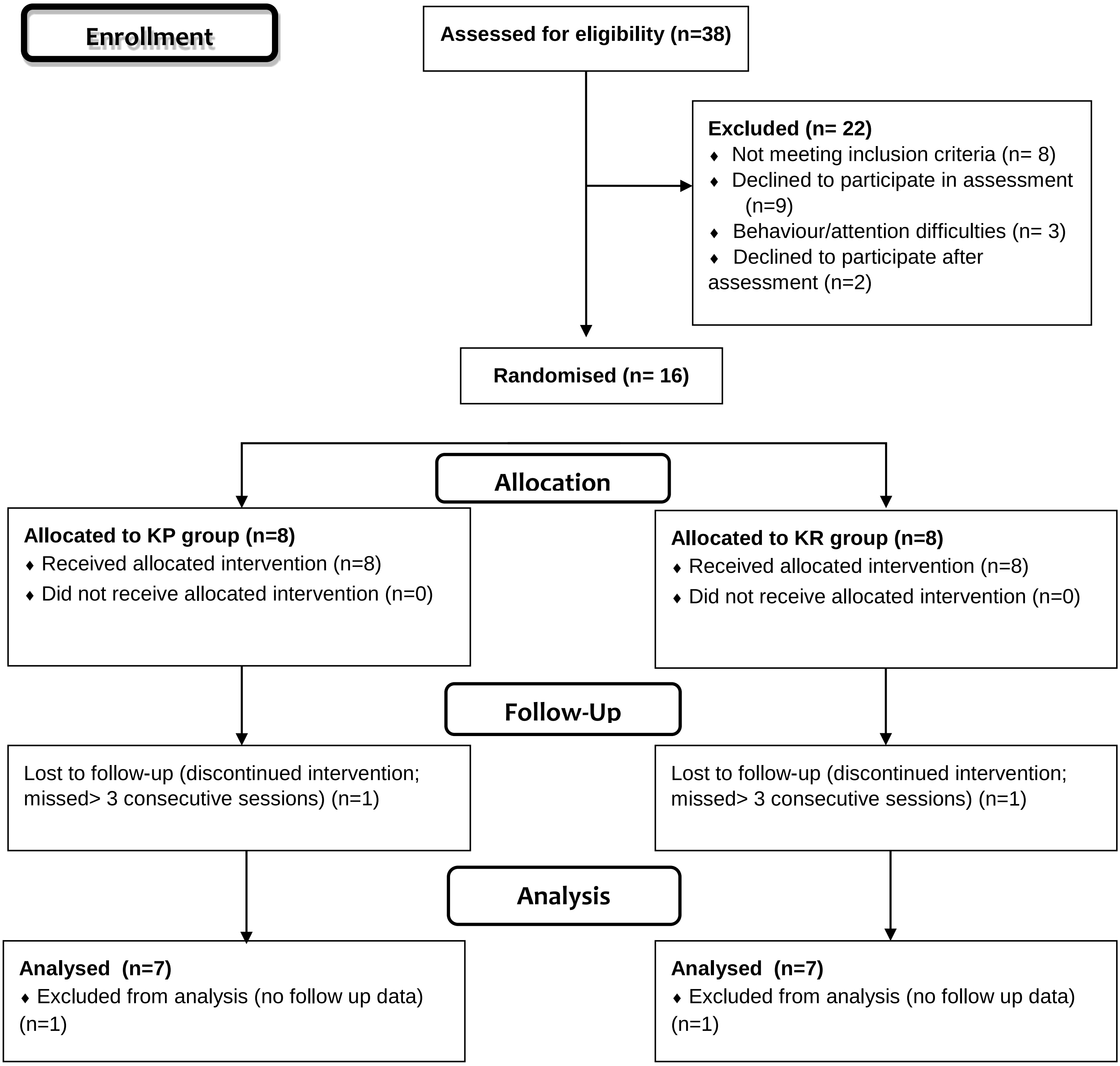

Enrollment
Assessed for eligibility (n=38)
Excluded (n= 22)
♦ Not meeting inclusion criteria (n= 8)
♦ Declined to participate in assessment (n=9)
♦ Behaviour/attention difficulties (n= 3)
♦ Declined to participate after assessment (n=2)
Randomised (n= 16)
Allocation
Allocated to KP group (n=8)
♦ Received allocated intervention (n=8)
♦ Did not receive allocated intervention (n=0)
Allocated to KR group (n=8)
♦ Received allocated intervention (n=8)
♦ Did not receive allocated intervention (n=0)
Follow-Up
Lost to follow-up (discontinued intervention; missed> 3 consecutive sessions) (n=1)
Lost to follow-up (discontinued intervention; missed> 3 consecutive sessions) (n=1)
Analysis
Analysed (n=7)
♦ Excluded from analysis (no follow up data) (n=1)
Analysed (n=7)
♦ Excluded from analysis (no follow up data) (n=1)

Figure 6.1. CONSORT flowchart of participant assignment, treatment and follow up.

Fourteen children were included in the study: 13 males and 1 female aged between 4 and 11 years, with a mean age of 6;7 years (SD = 2;5; range of 4;1 to 10;10 years). Two sets of twins participated. Severity of CAS, ranged from mild to severe, as measured by Percent Consonants Correct (PCC; Shriberg, Austin, Lewis, McSweeny & Wilson, 1997) for the Single Word Test of Polysyllables. Inter-rater reliability was > 85% for point-to-point transcription reliability on both these tests (Kratochwill et al., 2010). Demographic data are presented in Table 2. There were no significant differences between the two groups on any of the baseline variables (i.e. age, primary and secondary outcome measures or CAS severity; see Table 2).

Table 6.2. Comparison of pre-treatment variables by group for children with apraxia of speech assigned to either the Knowledge of Performance (KP) or Knowledge of Results (KR) feedback group.

	KP group (n=7)		**KR group (n=7)**			
Variable assessed	***M (SD)***	**Range**	***M (SD)***	**Range**	**t**	***p***
Demographic						
Age in months	81.7 (32.3)	50 - 129	83.6 (33.7)	54 - 131	-0.11	.92
Sex	7 Male		6 Male 1 Female			
Had previous speech treatment?	7/7		7/7			
Primary outcome measures at baseline						
Accuracy on items treated	18.6 (15.2)	0 - 42.3	20.5 (13.0)	0 - 36.4	-0.25	.81
Accuracy on items expected to generalize	55.2 (12.5)	41.8 -77.3	51.6 (20.6)	24.5 -76.1	0.40	.70
Secondary outcome measures at baseline						
DEAP Inconsistency	48.0 (20.0)	16 - 64	46.3 (15.6)	24 – 68	0.18	.86
Single Word Test of Polysyllables						
PPC	68.9 (19.9)	37 -89	62.8 (29.8)	24 - 92	0.49	.66
PVC	72.7 (16.8)	45 - 91	70.0 (23.2)	39 - 92	0.26	.80
PCC	66.0 (23.2)	32 - 96	57.4 (34.9)	12 - 93	0.54	.60
Percent lexical stress matches	62.5 (20.6)	34 - 90	55.9 (28.3)	24 – 88	0.50	.63
GFTA-2						
Standard score	73.7 (24.3)	51 -109	72.3 (22.0)	40 - 102	0.15	.88
PPC	75.0 (14.0)	52 - 91	69.8 (26.8)	36 - 97	0.45	.66
PVC	82.6 (8.9)	65 - 91	81.8 (16.2)	61 - 99	0.11	.91
PCC	70.8 (17.8)	45 - 95	62.9 (33.4)	17 - 97	0.55	.59
Speech disorder severity						
Severe (< 50%)	*n* = 2		*n* = 3			
Moderate-severe (50-65%)	*n* = 1		*n* = 1			
Mild-moderate (65-85%	*n* = 3		*n* = 0			
Mild (> 85%)	*n* = 1		*n* = 3			
CELF-P2 / CELF-4						
Receptive language score	97.3 (13.3)	82 - 121	90.1 (7.6)	81 - 106	1.23	.24
Expressive language score	84.7 (14.5)	66 - 107	85 (18.6)	63 - 112	0.03	.98

Note: t = t-test statistic; DEAP = Diagnostic Evaluation of Articulation and Phonology (Dodd, Hua, Crosbie, Holm & Ozanne, 2002); Single Word Test of Polysyllables (Gozzard, Baker & McCabe, 2004, 2008); PPC = percent phonemes correct; PVC = percent vowels correct; PCC = percent consonants correct; GFTA-2 = Goldman-Fristoe Test of Articulation – Second Edition (Goldman & Fristoe, 2000); Speech disorder severity was based on PCC from the Single Word Test of Polysyllables; CELF-P2 = Clinical Evaluation of Language Fundamentals – Preschool – Second Edition (Semel, Wiig & Secord, 2006); CELF-4 = Clinical Evaluation of Language Fundamentals – Fourth Edition (Wiig, Semel & Secord, 2006).

2.2 Design

The study used a parallel-group design with groups matched by age and severity of disorder. Stratified randomisation was employed to assign pairs of children, age- and severity- matched, to each treatment condition; that is, one child from each pair was randomly assigned to one treatment group and the matched pair assigned to the other group. In this way, each child within the sets of twins was randomised to a different group. The KP Group received KR+KP feedback throughout all four sessions per week, while the KR Group received KR+KP feedback for the first session each week and then KR feedback only for the remaining three sessions in a week. All other components of the protocol were identical across the groups. Figure 2 provides an overview of the assessment and treatment timeline of the experiment.

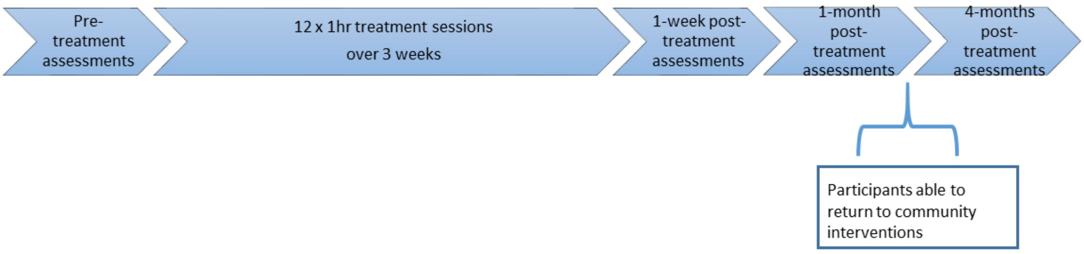

Pre-treatment assessments
12 x 1hr treatment sessions over 3 weeks
1-week post-treatment assessments
1-month post-treatment assessments
4-months post-treatment assessments
Participants able to return to community interventions

Figure 6.2 Intervention timeline

2.3 Intervention

The NDP3 was implemented as described by Williams and Stephens (2004, 2010) and operationalized by Murray and colleagues (Murray, McCabe & Ballard, 2012, 2015). Each child had three individualized speech production goals determined based on their pre-treatment assessment results. Goals were selected to include new speech sounds as single sounds or in known syllable shapes, new syllable structures using known sounds, and prosodic accuracy (i.e. lexical or phrasal stress). Five NDP3 stimulus words or phrases were selected per goal. Whereas the children in the original RCT (Murray et al., 2015) completed their speech production practice within 18-minute blocks using play-based activities, the nature of using app-delivered intervention required some adjustments to be made. Here, each goal was targeted in a 16-minute block using list-based exercises (i.e. swiping through the set of words/phrases and producing each target) and/or a memory game within the *Tabby Talks* app, with 2 minutes of free play between each goal. The total number of production trials per session was kept consistent with the protocol of Murray et al. (2012; 2015). Children needed to achieve 90% spontaneous accuracy on each target item before new stimuli were introduced into the goal. Once all five stimuli within a goal reached criterion accuracy, the child was stepped up to the next level in the NDP3 hierarchy. Immediate feedback was provided on 100% of production attempts throughout the sessions; however, the two groups differed in the type of feedback received during their treatment sessions.

2.4 Feedback Conditions

The KP Group received both KP feedback (i.e. specific, performance-based information about articulators/voicing/timing and how to adapt or change their production for next time) and KR feedback (i.e. on outcome accuracy only) on all production attempts (i.e. 100% KR+KP feedback) on all four days per week, following the protocol of Murray et. al

(2015). Teaching and cueing were provided as needed through verbal instructions, articulation placement cues, visual-tactile cues, metaphors, analogies and modeling.

The KR Group received 100% KR+KP feedback on one day per week, as described above; and 100% KR only feedback on the other three days per week. For children experiencing a high degree of difficulty (5 sequential incorrect responses), a brief period of KP feedback was introduced in order to establish a correct production before resuming with high frequency KR only (McCabe et al., 2014). While this departs from the goal of 100% KR feedback, this threshold for number of sequential errors is easily implemented in the app (Hair et al., 2018) and is necessary for duty of care. Clinicians collected data on the type of feedback provided to these children and engaged in continuous real-time monitoring of feedback type to ensure that a ratio of 80% KR to 20% KP was maintained for items on which a child was experiencing significant difficulty.

For both groups, when a production was correct, the child was instructed to repeat the response three times, with KR feedback provided by the clinician. This procedure is consistent with the NDP3 manual and the protocol developed by Murray et al. (2012). To maintain experimental control, all sessions were delivered in a University clinic. Student speech pathologists provided the treatment and delivered all feedback under the supervision of the first, fourth and last authors.

Dose was controlled across both treatment groups. Treatment was delivered over 12 1-hour sessions, four days per week for 3 weeks during school vacation periods. Children in the KP group received an average of 156.2 response trials per session (SD = 44.9) and the KR children an average of 142.5 (SD = 36.6), and these dose levels were not statistically different (t = 0.7348, p = 0.49).

The student clinicians received two days of training in providing treatment, transcription and data collection. Inter-rater reliability for point-by-point transcription after this training was ≥ 85%. To avoid potential clinician effects, each clinician was randomly allocated one child from each group and delivered two sessions per day – one child in the KP treatment condition and the other child in the KR feedback condition. The clinicians treated the same children for the 3-week block of treatment. The clinicians were, therefore, aware that treatment involved a comparison of two types of feedback, however, they remained blinded to the research hypotheses. To ensure adherence to the treatment protocol and avoid interference from one feedback condition to the other, treatment fidelity was measured in every session.

Caregivers were informed that their child would be treated using the NDP3 but were blinded to the feedback condition their child was receiving. Caregivers were able to observe treatment via one-way mirrors and could speak to other caregivers in the waiting room. Two of the participating families included twins who were paired with one another and consequently allocated to different treatment groups; therefore, the caregivers from these two families were aware that the nature of the experiment involved manipulation of the feedback conditions. All caregivers remained blinded to the experimental hypotheses and were instructed that no home practice should be done during the study. Reports containing detailed descriptions of the children's treatment condition, goals, progress, beneficial cues and strategies and recommendations for further treatment were provided to the caregivers after the 1-week post-treatment follow up assessment. No stimuli were provided to families and they were requested to refrain from practicing or resuming treatment until after the 1-month post-treatment assessment, which matched Murray et al.'s RCT (2015).

2.5 Outcomes

All children completed an individualized experimental probe immediately prior to commencing treatment. Probes varied in length from 116 to 176 items (M = 148, SD = 15.3) and consisted of (a) treated NDP3 items, to test for a treatment effect; and (b) untreated items from the NDP3 Assessment, to test for generalization of any treatment effect. The untreated items represented a range of difficulty in the NDP3 hierarchy from one level below the lowest level of treatment complexity to two levels above the highest level of treatment complexity (see Appendix A). These untreated items were analysed as a set and not by difficulty level.

Post-treatment assessments were conducted at 1-week, 1-month and 4-months post-treatment as per Murray et al. (2015). At each of these time points, the children completed their experimental probe and the DEAP Inconsistency subtest as an additional measure of generalization. In addition, each child and their caregiver completed a user-experience questionnaire at 1-week post-treatment. At the 1-month post-treatment time point, the GFTA-2 and Single Word Test of Polysyllables were also re-administered. All caregivers reported that their child had received no additional SLP input between the commencement of treatment and the 1-month post-treatment evaluation. Four children in each group reported resuming regular SLP services between 1-month and 4-months post-treatment.

2.5.1 Primary Outcome Measures. The primary dependent variable was percent accuracy of responses on experimental probe stimuli, judged perceptually. To be judged correct and scored as 1, each word or phrase was required to have: (a) all phonemes produced accurately, including no phonetic distortions, (b) smooth transitions between sounds and syllables (i.e. no syllable segregations or within word groping), and (c) accurate prosody (i.e., lexical or phrasal stress) across syllables. If any error was perceived on sounds, transitions, or prosody, the item was judged incorrect and scored as 0.

2.5.2 Secondary Outcome Measures. A secondary outcome measure to further explore generalization effects was the score on the Inconsistency subtest of the DEAP. In addition, responses on the Single Word Test of Polysyllables and GFTA-2 were analysed to explore potential changes to percent phonemes correct (PPC), percent consonants correct (PCC), percent vowels correct (PVC) and prosodic accuracy (i.e. percent lexical stress match) of untreated single words.

2.6 Recording equipment

All treatment sessions were audio- and video-recorded using the Cinde 88 audiovisual system (Cinde, Melbourne, Australia) and the Bosch Video Management System (Bosch Sicherheitssysteme GmbH, Grasbrunn, Germany). In addition, treatment sessions were audio-recorded using within-room digital voice recorders such as the Olympus VN-732PC or Sony Stereo ICD-UX200F digital voice recorder to enable off-line calculation of treatment fidelity and intra- and inter-rater reliability on the dependent variables. All pre- and post-treatment evaluations were audio- and video-recorded as above as well as audio-recorded using Roland Quad-Capture UA-55 [Roland, Los Angeles, CA] or Avid M-Track Audio [Avid, Burlington, MA] via an adjustable head-worn microphone (AKG C520, AKG Acoustics, Vienna, Austria) at 5cm mouth-to-microphone distance.

2.7 Reliability and treatment fidelity

2.7.1 Treatment sessions. Reliability for judgments of correct/incorrect on response trials was recorded for 25% of each treatment session. Mean inter-rater reliability was 88% (SD = 10.3) Treatment sessions were also closely monitored to ensure adherence to the treatment protocol. Data were collected on transcription accuracy, judgements of correct/incorrect, provision of appropriate feedback according to children’s allocated treatment group, provision of teaching/cueing where appropriate and eliciting three repetitions of a correctly produced treatment target. These data were compiled to generate an overall measure of treatment fidelity. Mean fidelity was 84.7% (SD = 9.5).

2.7.2 Experimental probes. Twenty-five percent of each probe assessment was re-rated to determine intra- and inter-rater reliability on primary outcome measures. For point-by-point transcription, mean intra-rater reliability was 89% (SD = 5.4) and mean inter-rater reliability was 84% (SD = 6.2). For judgments of correct/incorrect, mean intra-rater reliability was 92% (SD = 6.1) and mean inter-rater reliability was 87% (SD = 6.3).

Reliability for point-by-point transcription accuracy was also calculated on 20% of the secondary outcome data. This included broad transcription of the DEAP inconsistency subtest and phonetic transcription (with diacritics for errors) on the GFTA-2 and Single Word Test of Polysyllables. Mean inter-rater reliability was 85% (SD = 9.8).

2.8 Statistical Analysis

All statistical analyses were run using IBM SPSS Statistics 24 for Windows (IBM Corp, 2016)**.** A series of linear mixed effects models were run to test the effects of group (KP, KR), time (pre- and 1-week, 1-month and 4-months post-treatment) and their interaction on (a) treated items, exploring the treatment effect, (b) untreated but related items, exploring generalization of any treatment effect, and (c) the DEAP scores, also a measure of generalization. First order autoregressive and unstructured models were tested with and without the covariates of age and baseline severity (i.e. PPC score for the Single Word Test of Polysyllables), using Sidak adjustment for multiple comparisons for post hoc testing.

To assess for treatment-related changes in the secondary outcome measures from the Single Word Test of Polysyllables and GFTA-2, repeated measures analysis of variance (ANOVA) was used. This analysis included the between-subjects factor of group (KP, KR) and two-level within-subjects factor of time (pre, 1-month post) with 95% confidence intervals and alpha set at .05.

2.9 Questionnaire

A 16-item questionnaire was developed using a combination of yes/no, multiple choice, likert scale and open-ended response types (see Appendix B). Parents completed the questionnaire using pen and paper during their child's 1-week post-treatment session. Children over the age of 10 were invited to read and respond to the questions independently, with a clinician present to assist with any reading difficulties. Children under the age of 10 responded to the questions in an interview format with the assessing clinician.

Data from likert scale questions were collated to form condensed categories (e.g. 'highly motivating' and 'motivating' were combined). These data as well as binary and multiple-choice questions were analysed using descriptive statistics to report frequencies. We used qualitative content analysis (Graneheim & Lundman, 2004) to explore the responses to open-ended questions. The first author analysed each response by summarising meaning units, creating codes and identifying major themes. An independent rater conducted the same procedure for reliability. Themes were compared and potential sources of disagreement were discussed until consensus was reached.

3. Results

To assess for treatment and generalization effects, first order autoregressive and unstructured linear mixed effects models were tested with and without the covariates of age and baseline speech disorder severity (i.e. PPC score for the Single Word Test of Polysyllables). In all cases, except for age for the treated items, both covariates were significant. For all dependent variables, the unstructured model including the covariate of severity provided the best fit, with residuals being normally distributed. However, the findings were the same when either covariate was included in the model; note that age and severity were highly correlated in this sample (Pearson r = .679, p = .008). Results for the unstructured models, covarying for severity, are reported here.

3.1 Primary Outcomes

Performance on treated words across all four time points for the two experimental groups and also for the historical comparison group from Murray et al. (2015) is shown in Figure 3A. Performance on untreated but related items is shown in Figure 3B. Means and standard deviations for all measures made over four time points are presented in Table 3. Individual data for all 14 participants for change in percent correct from pre- to immediately post-treatment is also graphed in Figure 4, for transparency.

Table 6.3. Mean (SD) for treatment and generalization measures across the four test points for children with apraxia of speech assigned to either the Knowledge of Performance (KP) or Knowledge of Results (KR) feedback group.

	Pre-treatment		Post-treatment					
			1-week		1-month		4-months	
Treatment group	KP	KR	KP	KR	KP	KR	KP	KR
Primary outcomes								
Treated items [1]	18.6 (15.2)	20.5 (13.0)	43.7 (24.7)	23.8 (17.1)	45.7 (27.2)	28.1 (22.7)	59.0 (24.7)	45.1 (14.6)
Generalization items [1]	55.2 (12.5)	51.6 (20.6)	69.0 (8.7)	60.54 (16.8)	65.1 (15.9)	55.2 (20.9)	74.5 (12.6)	60.5 (24.3)
Secondary outcomes								
DEAP Inconsistency	48 (20)	46.3 (15.6)	44 (17.4)	41.7 (13.2)	39.4 (20.2)	43.4 (17.5)	33.1 (16.1)	38.3 (24.9)
Single-word Test of Polysyllables								
PPC	68.9 (19.9)	62.79 (29.8)	__	__	78.0 (11.5)	66.0 (23.3)	__	__
PVC	72.7 (16.8)	70.0 (23.2)	__	__	78.4 (9.1)	67.1 (19.2)	__	__
PCC	66.0 (23.2)	57.4 (34.9)	__	__	77.6 (14.4)	66.5 (30.1)	__	__
Percent lexical stress matches	62.5 (20.6)	55.9 (28.3)	__	__	61.7 (11.5)	46.6 (29.7)	__	__
GFTA-2								
Standard score	73.7 (24.3)	72.3 (22.0)			78.6 (25.6)	72.4 (25.8)		
PPC	75.0 (14.0)	69.8 (26.8)	__	__	79.9 (11.1)	72.2 (24.5)	__	__

PVC	82.6 (8.9)	81.8 (16.2)	__	__	83.9 (7.1)	79.6 (17.2)	__	__
PCC	70.8 (17.8)	77.7 (14.9)	__	__	62.9 (33.4)	68.1 (30.4)	__	__

Note: [1]Percent correct; DEAP = DEAP = Diagnostic Evaluation of Articulation and Phonology (Dodd et al., 2002); PPC = percent phonemes correct; PVC = percent vowels correct; PCC = percent consonants correct; GFTA-2 = Goldman-Fristoe Test of Articulation – Second Edition (Goldman & Fristoe, 2000).

3.1.1 Treatment Effects

There was no statistically significant difference when comparing average percent improvement from baseline for the KP group here and the traditional NDP3 group (See Table 4).

Table 6.4. Comparison of average gain (i.e. percent improvement from baseline) immediately post-treatment for, (i) treated items and (ii) items expected to generalise, for the tablet-based Knowledge of Performance group (KP) and traditional NDP3 group (TRAD) from Murray et al. (2015).

	KP group (N = 7)	***TRAD group (N = 13)***	***Statistics***		
	Mean (SD)	**Mean (SD)**	**MD**	**SE**	***p***
Treated items	25.1 (21.6)	39.8 (17.3)	14.7	8.8	.113
Items expected to generalise	13.8 (5.9)	10.3 (8.6)	3.6	3.7	.346

Note: MD = mean difference, SE = standard error, alpha was set at .05.

For the two experimental groups here, adjusting for severity, the main effect of time was highly significant; however, the effect of group and the group by time interaction did not reach significance (see Table 5 and Figure 3A). Due to the exploratory nature of this study, with a relatively small participant sample, post hoc comparisons were explored (See Table 6). For the KP group, there was a significant improvement from pre- to 1-week post-treatment, the difference from pre- to 1-month post-treatment approached significance but was robust for the pre- to 4-months comparison. For the KR group, only the pre- to 4-month comparison reached significance. As shown in Figure 4, the effect for the KP group was driven by three participants who improved more than 30 percentage points from pre- to 1-week post-treatment.

Table 6.5. Type III Tests of Fixed Effects for the dependent measure of treated items produced correctly in the one pre- and three post-treatment probes.

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	12.123	2.274	.157
Group	1	11.531	1.693	.219
Time	3	12	19.267	.000
Severity	1	11	5.030	.046
Group * Time	3	12	2.084	.156

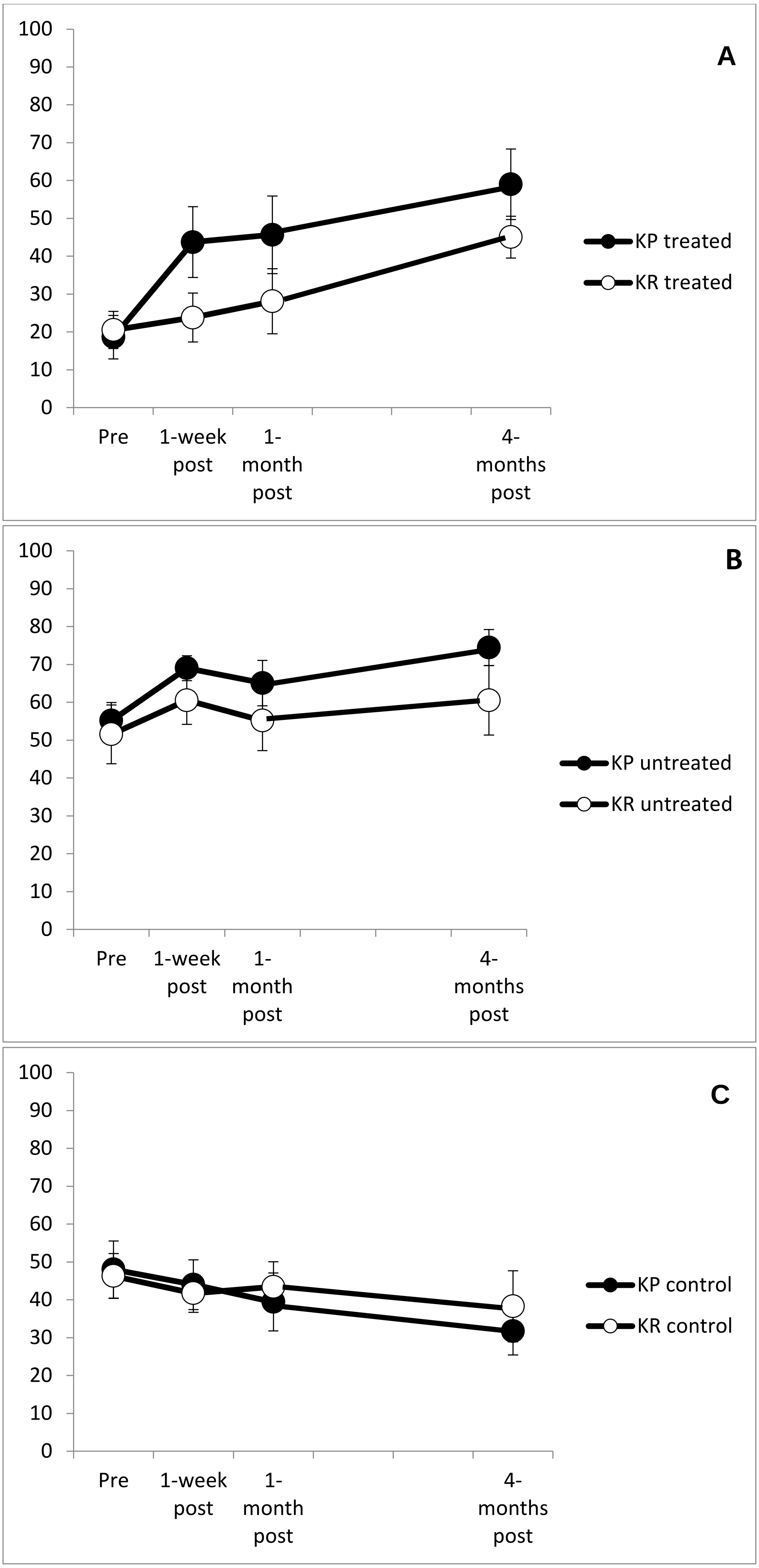

A
100
90
80
70
60
50
40
30
20
10
0
Pre
1-week post
1-month post
4-months post
KP treated
KR treated
B
100
90
80
70
60
50
40
30
20
10
0
Pre
1-week post
1-month post
4-months post
KP untreated
KR untreated
C
100
90
80
70
60
50
40
30
20
10
0
Pre
1-week post
1-month post
4-months post
KP control
KR control

Figure 6.3. Mean performance at pre-treatment, 1-week, 1-month and 4-months post-treatment for: **A.** treated items; **B.** untreated items; **C.** DEAP Inconsistency.

Note: KP = 100% knowledge of results and performance feedback for all 4 sessions each week; KR = 100% knowledge of results and performance feedback on session 1 and 100% knowledge of results feedback on sessions 2 – 4 each week. Error bars represent standard error. DEAP = Diagnostic Evaluation of Articulation and Phonology (Dodd, Hua, Crosbie, Holm & Ozanne, 2002).

Table 6.6. Post-hoc comparisons of average gain (e.g. percent improvement from baseline) for, (i) treated items and (ii) items expected to generalise, at each of the three post-treatment time points for the Knowledge of Performance (KP) group and Knowledge of Results (KR) group.

	Pre-treatment to 1-week post-treatment			***Pre-treatment to 1-month post-treatment***			***Pre-treatment to 4-months post-treatment***		
	MD	**SE**	***p***	**MD**	**SE**	***p***	**MD**	**SE**	***p***
Treated Items									
KP	25.1	6.4	.012*	27.0	9.0	.063	40.4	6.2	.000**
KR	3.3	6.4	.997	7.6	9.0	.959	24.5	6.2	.011*
Items expected to generalise									
KP	13.9	3.0	.004**	9.9	4.1	.174	19.3	4.0	.003**
KR	9.0	3.0	.070	3.6	4.1	.948	9.0	4.0	.250

Note: MD = mean difference, SE = standard error, * denotes *p* < .05, ** denotes *p* < .01.

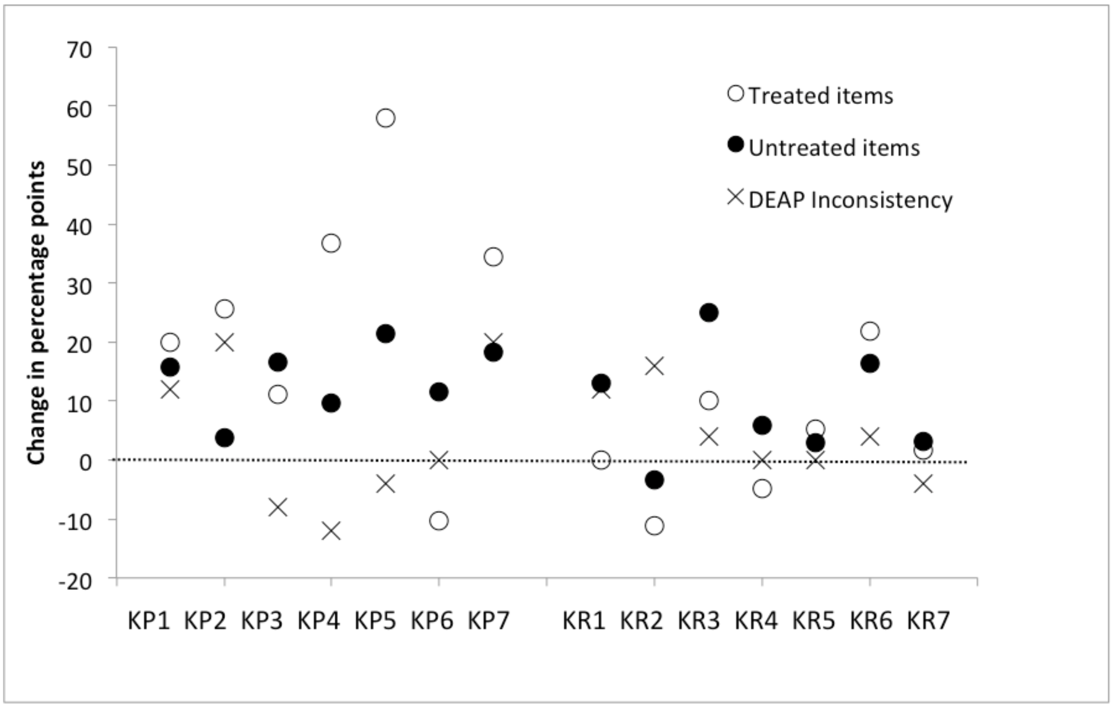


Figure 6.4 Individual percent change from pre-treatment to 1-week post-treatment for treated items, untreated items and the DEAP Inconsistency subtest.

Note: DEAP = Diagnostic Evaluation of Articulation and Phonology (Dodd, Hua, Crosbie, Holm & Ozanne, 2002).

To explore the issue of statistical power, we conducted a power analysis. First, the effect size (partial eta squared) for the group by time interaction was estimated using a traditional repeated measures ANOVA with group (KP, KR) and time (the first three time points only, free of influence from recommencement of community-based therapy). This yielded an effect size of $\eta_p^2 = 0.179$. To achieve a statistically significant interaction with this effect size, the sample size would need to be 26 per group (total sample size of 52; with alpha 0.05, power 0.8, 2 groups, 3 measurement time points, using G*Power v3.1.9.2). Conversely, with the current total sample size of 14, the effect size would have needed to be 0.36 to reach significance.

Long term outcomes for treated items between the two experimental groups in this study and the historical comparison group from Murray et al. (2015) was explored using repeated measures ANOVA with group (KP, KR, TRAD) and time. There were no significant differences between groups at the 4-months post-treatment time point (See Table 7).

Table 6.7 Comparison of average accuracy at 4-months post-treatment for, (i) treated items and (ii) items expected to generalise, for the Knowledge of Performance group (KP), Knowledge of Results (KR) group and traditional NDP3 group (TRAD) from Murray et al. (2015).

	KP group (N = 7)	***KR group (N = 7)***	***TRAD group (N = 13)***	***Statistics***	
	Mean (SD)	**Mean (SD)**	**Mean (SD)**	***F***	***p***
Treated items	59.03	45.07	64.46	3.13	0.062
Items expected to generalise	74.45	60.5086	58.9869	1.87	0.175

3.1.2 Generalisation effect

Average gain (i.e. percent improvement from baseline) on items expected to generalize was similar between the KP group here and the traditional NDP3 group (see Table 4).

Considering the two experimental groups in this study, the first analysis considered the untreated word set. Adjusting for severity, the main effect of time was highly significant; however, the effect of group and the group by time interaction did not reach significance (see Table 8 and Figure 3B). Again, due to the exploratory nature of the study, post hoc tests with Sidak adjustment for multiple comparisons were examined. Pre-treatment performance was compared to each of the three post-treatment time points for the two groups (see Figure 3B and Figure 4). For the KP group, there was a significant improvement from pre- to 1-week post-treatment, the pre- to 1-month post-treatment comparison was not significant, but the pre- to 4-months post-treatment was significant. For the KR group, the pre- to 1-week post-treatment approached significance, and no other comparisons were significant (See Table 6).

Long term outcomes for items expected to generalize were compared between the two experimental groups in this study and the historical comparison group from Murray et al. (2015) using repeated measures ANOVA with group (KP, KR, TRAD) and time. There were no significant differences between groups at the 4-months post-treatment time point (See Table 7).

Table 6.8. Type III Tests of Fixed Effects for the dependent measure of untreated related items produced correctly in the one pre- and three post-treatment probes.

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	11.611	34.659	.000
Group	1	11.362	3.383	.092
Time	3	12.000	14.233	.000
Severity	1	11.000	108.354	.000
Group * Time	3	12.000	1.292	.322

The second analysis considered performance on the DEAP inconsistency subtest. Adjusting for severity, no main effects or the interaction were significant (see Table 9, Figure 3C, and Figure 4). As expected, there were no significant post hoc comparisons across time points for either group.

Table 6.9 Type III Tests of Fixed Effects for the dependent measure of Inconsistency Score on the Diagnostic Evaluation of Articulation and Phonology (Dodd et al., 2002) at the pre- and the three post-treatment probes.

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	11.305	164.863	.000
Group	1	11.149	.179	.680
Time	3	12.000	2.915	.078
Severity	1	11.000	43.316	.000
Group * Time	3	12.000	.880	.479

3.2 Secondary Outcome Measures: Generalization Effects

Statistical analysis of the four outcome measures derived from the Single Word Test of Polysyllables (PCC, PVC, PPC and percent lexical stress match) and GFTA-2 (PCC, PVC, PPC and Standard Score) (See Table 10) demonstrated no group or interaction effect for any measure in either test. For the Single-Word Test of Polysyllables only, there was a large significant main effect of time (pre-treatment to 1-month post-treatment) for PCC and a large significant main effect of time for PPC (Cohen, 1969).

Table 6.10. Results of statistical comparisons made for the secondary outcomes measured at only two time points between the Knowledge of Performance (KP) and Knowledge of Results (KR) feedback groups.

	Pre-treatment to 1-month post-treatment											
Single word test of polysyllables	**PPC**			**PVC**			**PCC**			**% LS match**		
	F	*p*	η_p^2	*F*	*p*	η_p^2	*F*	*p*	η_p^2	*F*	*p*	η_p^2
Group	0.613	.449	.049	0.607	.451	.048	0.497	.494	.040	0.811	.386	.063
Time	5.358	**.039***	.309	0.214	.652	.018	11.423	**.005***	.488	1.769	.208	.128
Group * Time	1.235	.288	.093	1.947	.188	.140	0.180	.180	.015	1.276	.281	.096
GFTA-2	**PPC**			**PVC**			**PCC**			**Standard score**		
	F	*p*	η_p^2	*F*	*p*	η_p^2	*F*	*p*	η_p^2	*F*	*p*	η_p^2
Group	0.379	.550	.031	0.159	.697	.013	0.444	.518	.036	0.101	.756	.008
Time	1.804	.204	.131	0.025	.878	.002	3.489	.086	.225	0.459	.511	.037
Group * Time	0.216	.651	.018	0.349	.565	.028	0.076	.788	.006	0.404	.537	.033

Note. Effect (η_p^2) = partial eta squared with .01 representing a small effect, .06 representing a medium effect and .14 representing a large effect. * denotes significant at p < .05.

3.3 User experience questionnaires

Questionnaires were completed by thirteen of the children. One child declined to participate in the survey. Questionnaires from parents reporting on their perception of their child’s experience were returned by all twelve of the parents. The parents of twins completed one questionnaire per child.

3.3.1 Closed-ended questions

Frequency of responses to each of the binary and multiple-choice questions are shown in Figure 5A-G.

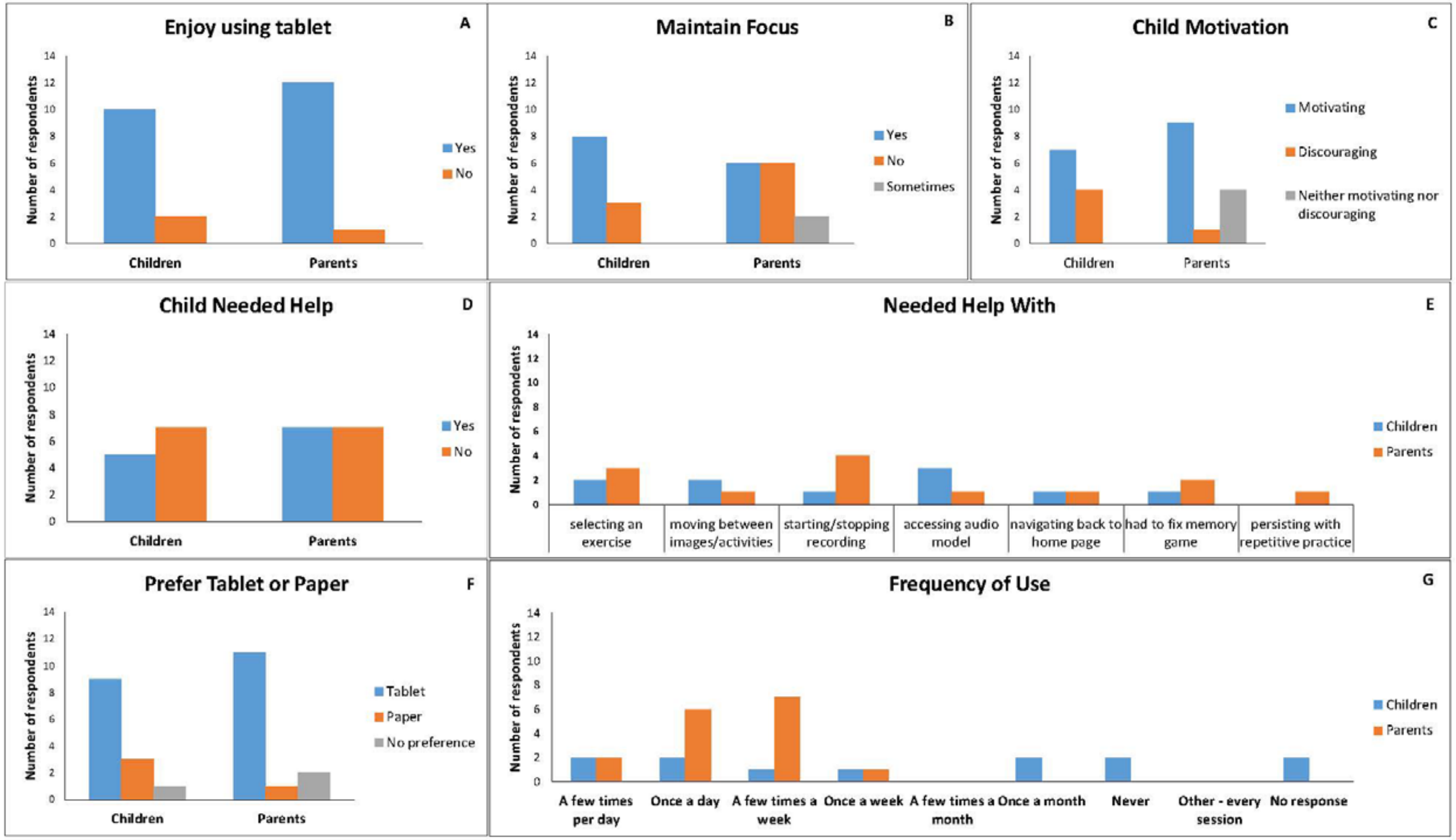


Figure 6.5A-G: Participant responses to user experience surveys

3.3.1.1 Enjoyment, engagement and motivation: The majority of children (10/13) and parents (12/14) reported that the child enjoyed using the tablet to practice their speech production targets (Figure 5A). Eight out of fourteen children reported being able to maintain focus on the therapy activities and 3/14 reported not being able to maintain focus, while parent responses were mixed (Figure 5B). Half of all children (7/13) and 9/14 parents reported finding the tablet-based activities motivating (Figure 5C). Four parents reported that their child had a neutral response to tablet-based activities.

3.3.1.2 Using the app: According to both child and parent responses, experiences were mixed as to whether the children needed help navigating the various app features during therapy (see Figure 5D). A range of features were selected for needing assistance, with no strong tendency for any one feature (See Figure 5E). Some parents (3/14) commented that their child mostly followed clinician instructions for operating the app or the clinician navigated the app for the child (3/14 parents).

3.3.1.3 Willingness to use apps in future: The majority of children (9/13) and parents (11/14) reported a preference for tablet-based practice compared with traditional paper-based activities (Figure 5F). All but one parent (13/14) reported a willingness to engage in tablet-delivered intervention more than once per week. The children were more varied in their responses: 6/14 indicated that they would use the app once a week or more, 2/14 that they would use it only once per month, 2/14 that they would not use it for speech practice, and 2/14 did not respond (Figure 5G).

3.3.2 Open-ended questions

The survey captured data on the users' likes and dislikes about the tablet-based exercises, reasons for preference toward tablet- vs paper-based exercises. Responses fell into two main themes of *the app/equipment* and *the experience.* Additional feedback was invited in a final open-ended question.

3.3.2.1 The app/equipment: Most of the app features were identified as 'likes'. Most commonly mentioned was the feature for playing back a child's audio recordings (4 children, 6 parents). Also reported as 'likes' by both children and parents were: listening to an audio model, the memory game, the reward stars, and the pictures. Three parents commented that tablet-based treatment was preferable because the app made home practice easier to set up and repeat. In contrast, the one parent who preferred paper-based activities noted their versatility: *"more adaptable…you can do more activities with paper"*.

Few children or parents mentioned aspects of the app or equipment that they disliked. Two children commented that the headset microphone was uncomfortable. Three parents wanted greater interactivity with the app and games that provide a greater range of motivators: *"it's just like flash cards"* (parent of KR7), and *"add visuals to the tablet images to show rate and emphasis within words"* (parent of KR5).

3.3.2.2 The experience: Four children commented that tablet exercises were *"fun"*. Fun, enjoyment and variety were also listed as reasons for children (2/13) and parents (5/14) preferring tablet-based activities. Three children reported that they liked learning new words and achieving goals; although, two children commented that they disliked practising difficult words.

Three children and one parent commented that the therapy program was *"too long"* and another two that there were *"lots of words to do"*. The most common dislike of parents was the repetitive nature of the activities (6/14).

3.3.2.3 Other feedback: The final survey question invited other comments or feedback. Three parents expressed gratitude for having been involved. Two parents reported that they felt their child was too young for the type of intensive treatment provided in the study. Two parents commented that their child was frustrated by KR-style feedback but one commented that they could see the benefit of using KR to encourage the child to “think more about their own speech”. Two parents expressed concern that their child’s behaviour had been negatively affected by participating in the study, and in one of these cases had also negatively impacted her child’s approach to therapy. In both cases, the child was in the KR condition.

4. Discussion

This study compared two methods of feedback during tablet-delivered NDP3 treatment. This investigation is a necessary first step towards determining whether app-delivered right/wrong (KR) feedback during intensive at-home practice of new motor speech targets can effectively facilitate acquisition and maintenance of new segmental and suprasegmental speech patterns. Such technology has the potential to bridge the gap between optimal service delivery intensity in CAS and current service delivery models in Australia.

We hypothesized that (i) tablet-based delivery of NDP3 using high frequency KP feedback would obtain similar treatment and generalization outcomes to Murray et. al.’s (2015) traditional paper-based delivery of NDP3, (ii) participants in the high frequency KR condition may demonstrate smaller gains immediately post-treatment (i.e. evidence of slower acquisition and generalization), compared with the KP group, but greater maintenance at 1- and 4- months post-treatment (i.e. evidence of more robust learning), and (iii) the experimental groups would demonstrate similar long-term maintenance of any treatment and generalization effects to Murray et. al.’s (2015) traditional NDP3 delivery.

Our first hypothesis was confirmed in that the KP group made statistically significant gains in treated and untreated word accuracy, which were similar in magnitude to the traditional NDP3 treatment group from Murray et al. (2015). Our second hypothesis was partially supported. Overall, for both treated and untreated words, no group effect was detected but the effect of time was highly significant. This suggests that children in both experimental groups responded to the treatment, with positive gains in treated and untreated words over time, regardless of the feedback condition. However, on closer examination of the data from the individual children, it was noted that 6/7 children in the KP group made substantive gains on treated words of 10 or more percentage points from pre- to 1-week post-treatment, while only 2/7 children in the KR group did; and 3/7 KP children but no KR children improved >30 percentage points. Similarly, 6/7 KP children and only 3/7 KR children showed a 10 or more percentage point gain for untreated words, indicative of generalization. It is likely that the small sample size in this study meant insufficient power to detect a significant group by time interaction effect. Our power analysis suggested that a group size of 26 was needed to achieve a significant interaction for the treated words, or else a larger effect size of 0.36. To date, this is the only study that has examined the influence of feedback type on speech intervention in CAS. These data suggest that the influence of KP vs KR feedback needs to be further explored in a larger sample, to determine whether there is indeed an effect of feedback type or whether the differences observed are driven by other factors such as age, severity of CAS, or self-evaluation ability.

The lack of significant improvement for the KR group immediately post-treatment appears consistent with the tendency for slower improvement with KR than KP. Although the KP group’s accuracy on both treated and untreated but related items at 1-week post-treatment reached significance, while the KR groups did not, there were no significant differences between the groups at any time point. This is likely because variability within groups was large, as shown in Figure 4.

Regarding the third hypothesis, both tablet-delivered treatment groups had made similar long-term gains at 4-months post-treatment that were statistically significant compared to pre-treatment performance level and similar to the gains made by the traditional NDP3 group in Murray et al. (2015). However, this also suggests that evidence of a significant treatment effect for the KP group here should be interpreted with caution. If treatment was responsible for accelerated changes in the KP group's speech production skills, one might expect that their progression over time should remain accelerated when compared with the KR group. This was not the case. Instead, the KR group demonstrated similar achievements in speech production skills at the 4-month follow up assessment. This finding may be confounded by (i) the return to community-based treatment for 4/7 children in both groups and (ii) that community clinicians were likely to have been providing KP feedback, although we do not have any evidence to support this suggestion. The lack of significant treatment effect for the KR group, also makes it difficult to attribute the improved performance at the 4-month follow up to 'maintenance'

The overall trend in improvement on both treated words and generalisation words differed between this study and the historical comparison study (Murray et al., 2015). Whereas, the traditional NDP3 group showed large improvement on treated words immediately post-treatment with a tendency towards loss of skill at follow up due to 1/13 clients having poor maintenance (Murray, McKechnie, & Williams, 2017) both groups here continued an upward trajectory during the follow up period. Reasons for this are not clear but may be due to factors to do with the use of the tablet for stimulus presentation, audio recording or self-evaluation, or the reinstatement of community-based therapy for some children. In contrast, performance on generalisation words showed the opposite effect. Where the traditional NDP3 group showed a continuous upward trend in performance accuracy, the two experimental groups here showed similar gains in untreated real words immediately post-treatment, with a trend towards deterioration of skill at 1-month post-treatment, followed by continued improvement from 1-month to 4-months post-treatment. Given that children were able to return to their regular speech pathology treatments following the 1-month post-treatment evaluation, this could explain the continued long-term improvements on all items. However, approximately half of all children did not resume treatment in this period and so it is likely additional but unidentified factors contributed to the trend for continued longer-term improvement. This is a desirable trend warranting further investigation into which child-related or treatment-related factors may have contributed to this observation.

These results echo those of previous studies in CAS and AOS that have demonstrated that responses to different feedback types and frequencies vary across participants (Maas, Butalla & Farinella, 2012; Austermann Hula, Robin, Maas, Ballard & Schmidt, 2008). Variation in response to different feedback types and frequency may be influenced by strength of internal representation of the specific speech behaviours targeted and/or pre-treatment level of proficiency. Target selection was individualised for each participant, resulting in some treatment and/or generalisation targets being relatively more difficult than others. Stimulus selection may therefore have served as a confound within and between participants (Maas et al., 2012; Wambaugh et al., 2017). This confound is almost impossible to avoid in these studies as treatment must address the sounds in error for each individual child. This is mitigated in part by limiting the sounds to those that are stimulable for a correct response and selection of three goals crossing different levels of proficiency (e.g. single sound to word level). Nonetheless, the children still vary in their ability to self-evaluate, ease in production, ability to attend and comply with the training context, and their motivation.

Feedback from participants was generally positive. The majority of respondents reported that tablet-based therapy was motivating, enjoyable and preferred compared with traditional paper-based formats. Most of the existing features of the *Tabby Talks* app were regarded favourably; however, suggestions for improvement included the need for a larger range of games and increased interactivity.

4.1 Limitations and future directions

The sample size of the study was small and within-group variability was large, thus limiting the power of our statistical analyses. CAS is relatively rare (Shriberg, Aram, & Kwiatkowski, 1997a) but much larger sample sizes may be possible with multi-centre collaboration. Our power analysis suggests that a sample size of about 26 per group is desirable and this would also allow exploration of other child-related factors that might influence or predict response to intervention. Alternatively, larger scale analyses may be possible through meta-analyses of studies which have used similar outcome measures.

Future research should explore alternative feedback type and frequency conditions and combinations. The feedback frequency and schedule used for the KR group in this study involved 100% pre-practice with KR+KP on day 1 and 100% practice with KR only on days 2 to 4 and was designed to mimic the common Australian service delivery model of once per week face-to-face with a clinician with a home-practice program with less rich feedback from an app or a parent. This model deviates from the schedule used in our previous work with PML, wherein a period of pre-practice with KR+KP is provided at the beginning of *every* session, and the child only progresses to practice with KR alone when they reach a predetermined threshold of success (Ballard, Robin, McCabe, & McDonald, 2010; Iuzzini & Forrest, 2010; McCabe et al., 2014). It is possible that the children in the KR group here did not receive sufficient pre-practice to develop a stable internal reference of correctness. This could explain why predominantly KR feedback appeared less effective than KR+KP feedback in stimulating improvement at 1-week post-treatment in this study.

It is also possible that the effects of feedback type were mediated by the frequency of feedback. High frequency feedback was used here, even though low frequency feedback has been recommended in the PML approach (Schmidt & Lee, 2011). This was in order to examine the effect of KR versus KP feedback types without the potential confound by potentially positive effects of low frequency feedback. However, high frequency feedback has been demonstrated to increase response variability if participants continually change their performance in different ways each time they are presented with feedback on error (Wulf & Shea, 2004). The within-group variability observed in this study may have been related to the high frequency feedback schedule. There was some suggestion that other aspects of the guidance hypothesis were supported, however, in that high frequency feedback guides the individual towards the correct response and that performance accuracy decreases when feedback is withdrawn (Salmoni, Schmidt, & Walter, 1984). That is, the tendency towards a drop in average accuracy at 1-month post-treatment such that performance was no longer significantly higher than pre-treatment across the entire sample may have been related to removal of feedback post-treatment. On the other hand, the finding here of an upward trajectory of improvement from1-month post-treatment to 4-months post-treatment seem to reflect the opposite effect, similar to some non-speech motor learning studies in children where higher frequency feedback has been shown to lead to greater learning and longer-term retention (Chiviacowsky, Wulf, de Medeiros, Kaefer, & Wally, 2008). Clearly, the influence of type and frequency of feedback on motor learning in children and how these principles may interact with specific task and child factors, is still not entirely clear.

Evidence from Iuzzini and Forrest (2010), who demonstrated variable reinforcement schedules only effected changes in accuracy during the third week of treatment, even after establishing a threshold of success, suggests that the KR group in this study may have benefited from a longer treatment period in order to establish acquisition of targets; or a longer period of pre-practice (Miller, Plante, Ballard, & Robin, 2018). Future research could explore whether the KR-based practice needs to be delivered for longer duration, or for more trials, to obtain a similar level of acquisition to KR+KP-based practice, and consequently to greater maintenance and generalization of these gains. Alternatively, a more gradual progression from predominantly KR+KP feedback into predominantly KR feedback (see Strand, Stoekel & Baas, 2006), gradual transition from immediate high-frequency feedback to delayed and reduced frequency feedback (Ballard et al., 2010; McCabe et al., 2014; Schmidt & Lee, 2011), or feedback fading based on successful execution of speech targets may be beneficial. One suggestion is to structure the feedback schedule beginning with three sessions of KR+KP and one session of KR in the first week, gradually progressing to one session of KR+KP and three sessions of KR in the third week of treatment. A similar gradual shift was employed by Thomas et al. (2017) who explored parent-training in ReST treatment as a method of achieving recommended intervention frequency for children with CAS, albeit with limited success.

Another factor that may have influenced the findings here was that clinicians were instructed to shift to KP feedback when children in the KR group produced five sequential incorrect productions of their selected treatment words. This was necessary in order to uphold our ethical duty of care for the children involved in the treatment, as extended intensive practice of incorrect motor plans could be harmful for learning as well as for motivation and engagement. This was monitored so that the ratio of KR and KP feedback over the study for these children was maintained at 80% KR to 20% KP trials. In clinical practice, such apps would typically be recommended for supervised use in the home environment. Clinicians would engage in progress monitoring and intervene, when required, in order to either provide coaching for the parent to assist their child to achieve more difficult speech production targets or to schedule a clinic visit in order to provide some additional pre-practice and KP-style feedback. For example, we have now implemented a threshold system where the therapy app discontinues delivery of a specific stimulus after a set number of incorrect responses (Hair et al., 2018), allowing a parent or clinician to step in and provide additional coaching with KP.

Future research is needed to explore within-participant factors in order to determine which children would be most suited to intensive practice with high frequency KR-style home practice conditions as delivered in this study.

The home practice condition was simulated in this study, as clinicians delivered all feedback. This was done for two main reasons. First, this maintained experimental control. Secondly, while automated speech analysis algorithms running offline on computers are becoming more accurate at identifying errors in children's speech (Shahin, Ji, & Ahmed, 2018), software that can run in real-time on a tablet is less sophisticated. Our speech analysis software for the tablet had not yet been sufficiently developed to meet clinically acceptable levels of reliability with human perceptual judgment and so was not incorporated into the tablet when this study was conducted. In response to the participants' feedback about the need for greater interactivity and variety of games, the research team is continuing to develop a wider range of games and alternative ASR algorithms in order to improve the gaming quality of an app designed for speech behaviour change. The team are currently trialing the effectiveness a new app using integrated ASR to determine the effectiveness of tablet-delivered treatment and ASR-generated feedback in a real home setting.

5. Conclusions

Mobile technology has the potential to increase the engagement and motivation of clients and to facilitate intensive practice of speech production targets (e.g., Hair et al., 2018). Combined with less frequent direct clinical contact via face-to-face sessions or telehealth, it can also mitigate barriers of distance and access to services for rural and remote families. With continued advancements in technology and the development and integration of accurate and reliable ASR software, mobile games are likely to become an effective supplement to face-to-face intervention. This has particular benefit for older children who can then practice independently and take greater responsibility for their remediation. It may also be helpful for some parents who find it difficult to provide reliable feedback on their child's productions (Thomas et al., 2017; Thomas, McCabe, Ballard, & Bricker-Katz, 2018). However, further research is required to understand how the parameters of therapy, and therefore the effectiveness of that therapy, can change with app-based exercises and with ASR versus parent or clinician generated feedback. Post-hoc comparisons in the current study suggest that provision of predominantly KR feedback on speech accuracy yielded small and perhaps negligible gain compared to KP feedback; however, for the 3-week block of therapy, gains under the KP feedback were not well-maintained. In building apps, it is important to build in flexibility so that practice can adhere to appropriate motor learning principles that may vary depending on the age and skill level of the child and that stimulate optimal long-term learning in a time and cost effective manner. Additional research is required to develop algorithms for prescribing these variations in practice and feedback conditions for children with CAS.

Acknowledgements: This research was made possible by NPRP Grant # 8-293-2-124 (Ahmed, Osuna, Ballard) from the Qatar National Research Fund (a member of the Qatar Foundation) and an Australian Postgraduate Award (McKechnie). During this project, Ballard was supported by an Australian Research Council Future Fellowship FT120100355. The statements made herein are solely the responsibility of the authors. The authors wish to thank the participants and families for their commitment to the treatment program during school vacation periods. Thanks also to the student clinicians who delivered the intervention and research assistants who contributed to data collection.

Appendix 6A. Generalisation items based on participants' treatment items

Participant's NDP3 treated items	**Participant's generalisation items**	
	Untreated related items of similar or lesser complexity	**Untreated related items of greater complexity**
Single sounds	Consonants and vowels in isolation	CV and VC words CVCV words
CV and VC words	Consonants and vowels in isolation Additional untreated CV and VC stimuli	CVCV words CVC words
CVCV words	CV and VC words Additional untreated CVCV stimuli	CVC words Multisyllabic words
CVC words	CVCV words Additional untreated CVC stimuli	Multisyllabic words Consonant cluster words
Multisyllabic words	CVC words Additional untreated multisyllabic stimuli	Consonant cluster words Phrases and sentences
Consonant cluster words	Multisyllabic words Additional untreated consonant cluster stimuli	Phrases and sentences
Phrases and sentences	Consonant cluster words Additional untreated phrases and sentences	

Appendix 6B: Participant satisfaction and software usability questionnaire

Participant number______________________________

Date____________________________

Child's gender (circle) **M** **F**

Person completing this form:

□ Parent

□ Child

□ Clinician

Did you/your child enjoy using the tablet for their speech therapy activities?

□ Yes

□ No

Did you/your child need any help completing the activities on the tablet?

□ Yes

□ No

If yes, please tell us what your child needed help with:

□ Selecting an exercise

□ Moving between images/activities

□ Starting the recording

□ Stopping the recording

□ Accessing the audio model

□ Navigating back to the home page

□ Internet access / connectivity for uploading recordings to the server

□ Other…. (please specify)

____

Do you want to elaborate on any items you ticked above?

Was your child able to maintain focus/attention on the exercises?

□ Yes

□ No

How motivating did your child find the therapy sessions?

|___________|_______________|______________|_____________|

Highly motivating	Motivating	Neither motivating nor Discouraging	Discouraging	Highly discouraging

What did you/your child like about the exercises on the tablet?

What did you/your child dislike about the exercises on the tablet?

In future, would you prefer to do these exercises:

□ On the tablet

□ Using paper cards/worksheets

Please elaborate on your answer…

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

□ Never
□ Once a month
□ 2 or 3 times a month
□ 4 or more times a month
□ Once a week
□ 2 or 3 times a week
□ 4 or more times a week
□ Once a day
□ 2 or 3 times a day
□ 4 or more times a day

Other (Please specify)

Any other comments/feedback….?

Chapter 7: Clinical implications and future directions for mobile technology and childhood apraxia of speech

This doctoral research was motivated by a desire to explore the potential for mobile technology to overcome some of the barriers to optimal intervention intensity that were identified in Chapter 1. The studies presented in this thesis contribute to our understanding of whether such technology is capable of effectively supplementing face-to-face clinical services by (i) providing accurate identification of errors in children's speech for diagnostic/progress monitoring purposes (i.e. feedback to clinicians); (ii) automated feedback to the child during speech production practice; (iii) engaging and motivating the child during speech production practice and (iv) facilitating changes to speech behaviours. These studies also highlight how the use of mobile technology will influence the way in which treatment protocols, based around principles of motor learning, are designed and implemented. This can support clinical decision making around which children are likely to benefit from such treatment protocols and lays a foundation for future research comparing optimal service delivery models.

Automatic speech analysis tools: Accuracy and clinical utility.

The findings from paper 1 and paper 2 (Chapters 2 and 4, respectively) suggest that automated speech analysis (ASA) tools can be effectively used to assess the intelligibility of disordered speech or to estimate the severity or degree of impairment. However, they are not yet reliable enough in their analysis of phoneme level accuracy or lexical stress patterns when applied to disordered or mispronounced speech. For tools that aim to have therapeutic benefit for children with speech sound and prosodic errors, error detection accuracy and feedback provision must be sufficient to minimise both false acceptance of error productions and false rejection of accurate productions.

The best performing tools reviewed in McKechnie et al. (2018) were those applied to phoneme level analysis that were either speaker-dependent or specifically trained on populations of disordered speakers. While this increased the performance accuracy of the specific ASA tool being investigated, this level of specificity necessarily limits the transferability of the tool to other populations and other word sets. In order for ASA tools to have wider clinical applicability, larger scale investigations using large corpi of typical and disordered children's speech are required. However, access to such large databases, especially of disordered speech, remains difficult. Until reliable models can be developed from large child speech databases, researchers have focused on testing methods that use a small amount of knowledge to guide the performance of models trained on adult speech. For example, Chapter 4 used knowledge about the phonemic errors in the sample to guide the lexical stress classification model towards correct classification of lexical stress patterns in children.

Interestingly, the results of the automated lexical stress classification study found no advantage for applying a knowledge driven approach, incorporating the specific phonemic mispronunciations made by the participants in our sample, to the preliminary phoneme segmentation and forced alignment analysis steps. One possible explanation for this is that ASA tools make their decisions based solely on acoustic information that may not be readily or consciously perceived/discernible by the listener, creating a potential mismatch between what the listener 'hears' and what the ASA tool 'hears'. This is supported by the findings of Skinder, Strand and Mignerey (1999) who reported no acoustic differences between typically developing speakers and speakers with suspected CAS, even though listeners had perceived differences in lexical stress accuracy between the two groups. There, the authors suggested that listener perception of stress might be affected by segmental errors. An alternative explanation for the mismatch between auditory perceptual judgement of stress and automated or acoustic analysis of stress is offered by Munson and colleagues (2003). Their findings indicated no group differences in the use of acoustic correlates of stress production even though speakers with suspected childhood apraxia of speech (here, referred to as CAS) were perceived to have matched the target stress contour less often than speakers with phonological disorder. The authors suggested that participants with CAS were able to produce acoustic differences between stressed and unstressed syllables but that these differences may not be consistently perceived by listeners (Munson et al., 2003). These findings suggest that listener perception of relative difference in acoustic features across adjacent syllables may be affected by the degree of difference produced. That is, the degree of acoustic contrast produced by speakers with CAS may not be sufficient for the listener to perceive that the target stress pattern has been correctly executed. This is likely to be the case if speakers with CAS produce a de-stressed but unreduced vowel within the weak syllable instead of fully reducing the vowel to a schwa. De-stressed unreduced vowels can be categorised as acoustically distinct from both stressed vowels and unstressed reduced vowels, however, human listeners preferentially make a binary distinction, and tend to categorise de-stressed unreduced vowels with stressed vowels (Fear, Cutler, & Butterfield, 1995).

Normative data on the developmental trajectory of lexical stress contrastiveness contributes further support for the theory that development of adult-like lexical stress contrasts (i.e. production of shorter syllable durations and rising intensity contours), for weak-strong (WS) words in particular, is linked to the physiological development of the speech motor system. Children as young as three years old are able to produce strong-weak (SW) patterns with adult-like acoustic contrasts (Ballard, Djaja, Arciuli, James, & van Doorn, 2012), whereas, the degree of acoustic contrast achieved during the production of weak-strong (WS) patterns continues to differ from that of adults even up to the age of 11 (Arciuli & Ballard, 2016). These acoustic differences were present for WS words despite having been deemed accurately produced based on auditory-perceptual judgment. Taken together, these findings support the theory of CAS as a disorder of speech motor control.

Future directions: Developing ASR algorithms.

Speech recognition algorithms designed for use in therapeutic applications will need to be developed and tested using larger corpi of disordered speech. This may be possible for some speech disorders such as dysarthria, which can result from a number of different aetiologies in both children and adults, for example, the TORGO database (Rudzicz, Namasivayam & Wolff, 2012. However, other disorders such as CAS are relatively rare (Shriberg, Aram, & Kwiatkowski, 1997a) and behavioural manifestation of these disorders can be heterogenous. Collating large enough databases of speech from children with speech sound disorders is a challenge that will require multi-centre research collaborations, however, if achieved, large databases and machine learning will be able to further inform researchers of the optimal features and algorithms necessary to advance the field towards successful ASR approaches to disordered speech.

Further research is also needed to explore different ASR algorithms, both for phoneme verification and lexical or phrasal stress classification, in order to increase sensitivity and specificity and reduce false acceptance and false rejection rates to within a clinically acceptable threshold. Earlier work from the collaborators on our team had developed a lattice-based pronunciation verification method using Hidden Markov Model Deep Neural Network (HMM-DNN) acoustic models specifically for disordered speech (Shahin, Ahmed, McKechnie, Ballard, & Gutierrez-Osuna, 2014). The main limitation of these methods has been that their effectiveness depends upon participants producing only errors which have been included as probable pronunciation variants in the system's search lattice (Shahin, Ji, & Ahmed, 2018). When errors are unexpected or deviate from those the model has been programmed to analyse, the performance accuracy of the ASR is decreased. Shahin and colleagues (2018) have recently developed and tested an alternative pronunciation verification approach based on a One-Class Support Vector Machine (OCSVM) model. This approach learns the place and manner of articulation and the voicing features for each phoneme and uses this to evaluate speech input. It compares the input to the learned phoneme model and decides if it is a match or a mismatch (Shahin et al., 2018). When compared with the performance of traditional Goodness of Pronunciation models, the OCSVM performed with greater phoneme verification accuracy and reduced false acceptance and false rejection rates (Shahin et al., 2018). While these types of approaches are gaining traction and more consistently approaching our clinical threshold, they are still focused on adult speech with little known about their performance with children's speech, typical or disordered.

Automatic speech analysis tools: Findings and implications for intervention.

The results from paper 3 (Chapter 6) suggest that children with CAS benefit from intensive clinician-led intervention incorporating knowledge of results (KR) and knowledge of performance (KP) feedback in order to acquire new motor speech behaviours. These findings are in line with what previous research has demonstrated in regards to optimal intervention intensity for speech sound disorders, including CAS, that higher dose frequency improves outcomes (e.g. Kaipa & Peterson, 2016; Murray, McCabe, & Ballard, 2014; Murray, McCabe, & Ballard, 2015; Namasivayam et al., 2015). Some children made substantive gains in the simulated home practice/ASR-based feedback condition receiving predominantly KR feedback on three out of four sessions per week, consistent with evidence from the motor learning field. However, many other children showed maintained improvement in speech production skills following the mixed KR+KP feedback condition, which is not entirely consistent. One possible explanation for the mixed pattern of improvement among children in the KR group is related to age and stability of the pre-treatment motor plans for treated targets. The four children who made positive gains in the KR condition were aged seven or under, while the two children who showed negative change were over the age of ten years. While it is thought that a less stable internal reference of correctness may inhibit ability to learn from simple KR feedback, these findings suggest that an overly stable, or over-practiced but erroneous movement pattern may also inhibit learning through KR feedback alone. This hypothesis could be tested in future larger-scale studies.

The session structure used in this study has been previously untested in children with CAS. Earlier research from our team included a pre-practice phase for some portion of every intervention session. Pre-practice periods typically continued until participants had produced five of their selected treatment targets correctly (Ballard, Robin, McCabe, & McDonald, 2010; McCabe et al., 2014; Murray et al., 2015; Thomas, McCabe, & Ballard, 2014; Thomas, McCabe, Ballard, & Lincoln, 2016). The protocol employed by Thomas and colleagues (2014; 2016) included 25 minutes in each of the first two 50 minute sessions in a treatment block followed by 10 minutes at the start of each subsequent session, while the protocol described by Murray and colleagues (2015) allowed for the majority of the first two sessions in a treatment block to comprise pre-practice, followed by 10-15 minutes pre-practice at the start of each subsequent session. Research from the Mayo clinic team also described a gradual decrease in feedback specificity indicative of a shift from pre-practice KP guidance to more KR, however, the authors did not disclose whether a specific threshold of production accuracy was necessary to trigger reductions in KP (Strand, Stoekel & Baas, 2006). Caution is warranted when extrapolating existing guidelines for PML from nonspeech motor literature to speech treatment for CAS. These guidelines have been challenged by studies from Maas and colleagues (Maas, Butalla, & Farinella, 2012; Maas & Farinella, 2012). Their team compared the recommended low frequency feedback during practice with high frequency feedback and found that around half of the children in their study benefited from low frequency feedback, while half benefited from high frequency feedback (Maas et al., 2012). Similarly, when comparing the effects of blocked versus random practice of speech targets, the authors found that some children benefited from random practice, as recommended in the nonspeech motor learning literature on PML, while others benefited more from blocked practice (Maas & Farinella, 2012).

As discussed in Chapter 1, the recommended motor learning principles are now found in many treatment protocols; however, relatively little investigation has been carried out comparing the effects and potential interactions of the different principles in speech motor learning, specifically child speech. Clearly some children do not respond well to the recommended principles, at least at the time of their enrolment into a particular study. It is not possible to identify the factors influencing these participant differences without evaluation of large cohorts of children that are representative of the variability in the population.

Future directions: Intervention protocols.

Once ASR algorithms have been optimised and rigorously evaluated against clinically acceptable reliability standards, additional research is warranted to determine which children might be best suited to a service provision model such as the one explored in Chapter 6 (paper 3). In that study, clinician-delivered intervention was provided once per week followed by tablet-delivered intervention using automated feedback on speech production accuracy three days per week. Participant variables such as age, baseline level of proficiency on speech targets, severity of speech disorder, consistency of speech disorder, receptive/expressive language skills, accuracy of phonological representations and speech perception abilities need to be explored in order to identify factors predictive of a positive response to intervention and specific intervention approaches. Retrospective factor analysis conducted on 20 participants across two studies (McKechnie et al., 2016; Murray et al., 2015), all of whom received clinician-led treatment four days per week for three weeks and received 100% KR+KP feedback using the Nuffield Dyspraxia Programme – Third Edition

(NDP3), found no significant correlations for speech production, mental function, or oral structure and function skills with individual difference scores immediately post-treatment (Murray, McKechnie, & Williams, 2017). However, the authors identified several factors, which predicted better maintenance of skill in the post-treatment follow up period. These factors included: greater speech inconsistency (i.e. reduced stability offers greater potential for change), speech targets at lower levels of the NDP3 goal selection hierarchy, lower expressive language skills and lower working memory skills (Murray et al., 2017). Within-participant factors that were related to greater generalisation of skills to untrained speech behaviours included younger age and greater memory (the reverse effect compared to what was observed for treated behaviours) and phonological awareness skills (Murray et al., 2017). The finding of greater inconsistency predicting better maintenance of skill may be considered further support for the findings in Chapter 6 that the children aged 10 or over were those who did not respond to the KR only feedback condition. That is, these children potentially had more stable and inflexible motor plans due to the additional years spent practising incorrect movement patterns. This theory warrants further exploration using a more rigorous measure of movement stability such as the spatiotemporal index proposed by Smith and colleagues (1995). Findings of poorer expressive language and speech targets at lower levels of complexity on the NDP3 hierarchy predicting better maintenance is in contrast to what has been previously suggested in motor learning literature where targeting more complex targets leads to greater learning (see Schmidt & Lee, 2011). However, these findings may also reflect constraints within the available stimuli of the NDP3 program; for example, basic phrase and sentence stimuli that offer little prosodic variation. Importantly, this type of factor analysis needs to be extended to children in a KR-only group to explore which factors might predict a positive response to intervention using this feedback condition. Despite revealing some significant relationships, the analysis by Murray et al. (2017) was likely underpowered with only 20 children. Likewise, the study reported in Chapter 6 was small and lacked the statistical power to perform such analyses. Power calculations indicated that, in order to detect large correlations between treatment outcomes and participant related factors, a minimum of 10 participants per experimental group was needed. To be sensitive to small correlations between treatment outcomes and within-participant factors, a sample size of 47 was needed. This type of information about predictive factors for positive treatment outcomes would be clinically useful for practising SLPs to determine which of their clients might benefit from this type of practice.

Another area that would benefit from more insight into individual variation, is the influence of participant variables in response to different motor learning principles. Our study suggests that future research could experiment with the timing of the shift from predominantly KP+KR feedback to predominantly KR feedback. First, it is possible that a longer treatment period may have seen the KR group reach a similar level of acquisition to the KP+KR group and, subsequently, achieve greater maintenance and generalisation of these gains (see Iuzzini and Forrest, 2010, where gains in accuracy emerged only during the third week of treatment or Ballard, Robin, McCabe & McDonald, 2010, where all three participants required numerous trials before demonstrating improvement). A longer period of pre-practice may also be beneficial, as was demonstrated by Miller, Plante, Ballard & Robin (2018). Another suggestion would be to gradually transition from KP+KR feedback towards KR only feedback based on successful execution of a larger number of speech targets as in Strand, Stoekel and Baas (2006); or to employ a gradual shift in feedback similar to that used by Thomas, McCabe and Ballard (2017) in their investigation of the efficacy of parent-implemented intervention. In that study, the authors commenced with three clinician-delivered and one parent-delivered session per week, moving to two clinician-delivered and two parent-delivered sessions in the second week and finally one clinician-delivered and one parent-delivered session in the third week of treatment (Thomas, McCabe, & Ballard, 2017). A similar gradual progression of type of feedback could be designed and explored.

User perspectives on tablet-based intervention

The use of tablets to engage in speech pathology intervention was generally received favourably. The majority of the 14 participants reported a preference for tablet-based therapy over traditional paper-based formats and willingness to use tablet-based intervention in the future as a motivating and enjoyable method of speech production practice. These findings echoed previous reports of clinicians' and consumers' desire for alternative service delivery formats as a means to overcoming some of the service delivery barriers discussed in Chapter 1 (e.g. McAllister, McCormack, McLeod, & Harrison, 2011; Ruggero, McCabe, Ballard, & Munro, 2012).

Literature on the use of apps in treatment of paediatric speech sound disorders is limited. Of the studies available, most focus on development stages of the app but have reported positive findings for overall level of enjoyment, engagement and participation from the children (Ahmed et al., 2018; Anjos et al., 2018; Byun et al., 2017; Hair, Monroe, Ahmed, Ballard, & Gutierrez-Osuna, 2018; Tommy & Minoi, 2016). Investigations into computer-based interventions for children with speech disorders have also demonstrated that children generally enjoy and prefer computer-based approaches over traditional table-top approaches (Lan, Aryal, Ahmed, Ballard, & Gutierez-Osuna, 2014; Nordness & Beukelman, 2010; Tan, Johnston, Ballard, Ferguson, & Perera-Schulz, 2013; Toki & Pange, 2010; Wren & Roulstone, 2008). The use of apps in communication interventions for children with autism has been found to increase time on-task and decrease challenging behaviours; plus, children tend to demonstrate a preference for app-based approaches compared with traditional approaches (Flores et al., 2012; Ganz, Hong, & Goodwyn, 2013; Lee et al., 2015). Emerging results regarding positive improvements to language (Cumming & Draper Rodriguez, 2013), emergent literacy (Brouwer et al., 2017) and prosody (Simmons, Paul, & Shic, 2016) have also been reported, though none of these apps used automated feedback on performance accuracy. In acquired neurogenic disorders, apps are successfully engaging adult clients with home practice in between clinic visits and achieving positive treatment outcomes for clients (e.g. Des Roches, Balachandran, Ascenso, Tripodis, & Kiran, 2015; Des Roches & Kiran, 2017; Kurland, Liu, & Stokes, 2018; Kurland, Wilkins, & Stokes, 2014).

In 2018, 81% of 228 SLPs surveyed in the USA reported using apps in sessions (Benedon, 2018). 90% of these SLPs indicated that their primary purpose for using apps was for direct therapy and skills development. Speech sound production skills were the second most commonly reported skill for which apps were used in therapy (Benedon, 2018). Despite thousands of available apps in mobile stores, which purport to be useful for speech and language disorders, the majority are experimentally untested and little is known of their quality, effectiveness and efficiency. The lack of experimental investigation is likely related to the high turnover of the market as well as the cost and time involved in running experimental trials (Edwards & Dukhovny, 2017). A recent quality analysis of mobile applications for speech disorders found that less than 3% (132) of the more than 5000 apps identified warranted full evaluation and, of those that were subjected to full quality evaluation, only 19 (14%) were deemed to have potential for therapeutic benefit (Furlong, Morris, Serry, & Erickson, 2018).

There is a powerful move in the computer science and eHealth fields toward co-design and this is also beginning to happen in the field of app development for childhood speech disorders (e.g. Ahmed et al., 2018; Hair et al., 2018). This work aims to ensure that apps on the market adhere to best practice in terms of theories of learning, engagement and nature of speech disorders. Sensitively designed tools can increase the likelihood that the apps are facilitating meaningful behavioural change with no adverse effects such as developing maladaptive behaviours or therapy burn-out that could complicate the therapeutic process for clinicians. Considerable research is still required in this area to optimally serve the consumers.

Future directions: Diversification of included games and activities.

In response to participants' desire for more games, greater interactivity and less repetitive practice when using apps for speech production practice (see Chapter 6 and also Ahmed et al., 2018), recent research has explored user perspectives on a suite of prototype speech-controlled games using open-source games modified to be speech-controlled incorporating freely available speech recognition software, PocketSphinx (Ahmed et al., 2018; Hair et al., 2018). In addition to the memory game that was included in the *Tabby Talks* app, Ahmed et al. (2018) examined the following four games: asteroids, where correctly pronounced words broke up asteroids that threatened to hit the spaceship; a game similar to Whack-a-Mole, where players were required to tap electronic 'cards' that flipped at random; a word search game where points accrue for words that the ASR deemed correctly produced; and a word pop game where correctly produced words caused bubbles to pop. Children reported enjoying the speech-controlled nature of the games and earning points and rewards, however, they also reported that the games quickly became boring and that they wanted games which were fast paced, more challenging and had multiple levels of difficulty. Expert evaluation of ASR accuracy demonstrated that the ASR recognised fewer productions as correct compared with the researchers, with the highest accuracy being for adult productions, followed by typically developing children, and lowest accuracy scores for children with CAS (Ahmed et al., 2018). These findings are unsurprising in light of the demonstrated difficulties with accurate recognition of both typical and disordered child speech reported in the literature (Gerosa, Giuliani & Brugnara, 2007; Yeung & Alwan, 2018; see also McKechnie et al., 2018 (Chapter 2 in this thesis) for a review).

Using principles of participatory design, researchers from the same team developed a new game, similar in style to a Mario game, where game play is the goal but speech production practice is integrated, customisable for task difficulty and session dose, and speech practice earns coins for players to spend on customising their avatar in the game store (Hair et al., 2018). User feedback was largely promising, children reported high levels of engagement and motivation to keep playing (Hair et al., 2018). However, mobile ASR software continues to require optimisation. Studies are underway investigating the use of speaker-dependent, template-based models which are trained using a child's own productions, both correctly and incorrectly pronounced, with the aim of facilitating more accurate judgments of accuracy for productions by each child. Also, domain-guided adaptation methods are being developed and explored to improve ASR accuracy in "low-resource" contexts where large training databases are not available, although currently this work is focused on ASR for recognising accented speech in adults (e.g.Juan, Besacer, Lecouteux, & Tan, 2015).

Conclusions.

The studies contained in this thesis offer foundational information about the capacity for technology to motivate the learner to engage in intensive practice of speech targets; the adequacy of automated feedback on accuracy; and the potential to facilitate changes in speech behaviours. These results provide a necessary first step in evaluating optimal service delivery models for CAS. Once adequately trained for disordered speech, either through training with large databases or through new low-resource domain-guided adaptation of models trained previously with typical adult speakers, automated speech analysis tools may be able to offer several advantages to clinicians. Such tools hold potential to expedite the objective assessment of lexical stress errors and/or speech sound production errors, which may address some of the reported challenges around caseload management and workload demands. Automated speech analysis systems, particularly when embedded into mobile applications, may also provide a method of bridging the gap between evidence-based intervention intensity and current clinical practice. Based on the results of the studies included in this thesis, these applications would be best suited for children who have already acquired an internal reference of correctness for each of their speech production targets and who can self-evaluate their productions and experiment with changing productions to be more accurate. Such tools can then be useful for facilitating intensive, engaging and motivating home practice that includes accurate feedback on production attempts. This, in turn, could facilitate positive and efficient treatment gains for children with speech sound disorders, including CAS.

References

Adams, S. G., & Page, A. D. (2000). Effects of selected practice and feedback variables on speech motor learning. *Journal of Medical Speech-Language Pathology, 8*(4), 215-220.

Aguilar-Mediavilla, E. M., Sanz-Torrent, M., & Serra-Raventos, M. (2002). A comparative study of the phonology of pre-school children with specific language impairment (SLI), language delay (LD) and normal acquisition (E). *Clinical Linguistics & Phonetics*, 16, 573–596.

Ahmed, B., Monroe, P., Hair, A., Tan, C. T., Gutierrez-Osuna, R., & Ballard, K. J. (2018). Speech-driven mobile games for speech therapy: User experiences and feasibility. *International Journal of Speech-Language Pathology, 20*(6), 644-658. doi:10.1080/17549507.2018.1513562

Allen, M. M. (2013). Intervention Efficacy and Intensity for Children With Speech Sound Disorder. *Journal of Speech, Language & Hearing Research*, *56*(3), 865-877. doi:1092-4388(2012/11-0076)

American Speech Language and Hearing Association (ASHA) (2007a). *Childhood Apraxia of Speech [Position Statement]*. Retrieved from Rockville, MD.: https://www.asha.org/policy/PS2007-00277/

American Speech Language and Hearing Association (ASHA) (2007b). *Childhood Apraxia of Speech [Technical Report]*. Retrieved from Rockville (MD): http://www.asha.org/policy/TR2007-00278

American Speech Language and Hearing Association (ASHA) (n.d.). *Telepractice.* (Practice Portal). Retrieved November 28th, 2018 from www.asha.org/Practice-Portal/Professional-Issues/Telepractice

Anjos, I., Grilo, M., Ascensão, M., Guimarães, I., Magalhães, J., & Cavaco, S. (2018, 2018//). *A Serious Mobile Game with Visual Feedback for Training Sibilant Consonants.* Paper presented at the Advances in Computer Entertainment Technology, Cham.

Arciuli, J. & Bailey, B. (in press). An acoustic study of lexical stress contrastivity in children with and without Autism Spectrum Disorders. *Journal of Child Language.* Accepted June 2018.

Arciuli, J. & Colombo, L. (2016). An acoustic investigation of the developmental trajectory of lexical stress contrastivity in Italian. *Speech Communication, 80, 22-33.*

Arciuli, J. & Slowiaczek, L. (2007). The where and when of lingiustic word-level prosody. *Neuropsychologia, 45*, 2638-2642.

Arciuli, J., & Ballard, K. J. (2016). Still not adult-like: lexical stress contrastivity in word productions of eight- to eleven-year-olds. *Journal of Child Language, 44*(5), 1274-1288. doi:10.1017/S0305000916000489

Arciuli, J., & Ballard, K. J. (2017). Still not adult-like: lexical stress contrastivity in word productions of eight- to eleven-year-olds. *Journal of Child Language, 44*(5), 1274-1288. doi:10.1017/S0305000916000489

Arciuli, J., & Cupples, L. (2004). Effects of stress typicality during spoken word recognition by native and nonnative speakers of English: Evidence from onset gating. *Memory & Cognition, 32*(1), 21-30. doi:10.3758/bf03195817

Arciuli, J., & Cupples, L. (2006). The processing of lexical stress during visual word recognition: Typicality effects and orthographic correlates. *The Quarterly Journal of Experimental Psychology, 59*(05), 920-948. doi:10.1080/02724980443000782

Arciuli, J., & Cupples, L. (2007). Would you rather “embert a cudsert” or “cudsert an embert”? How spelling patterns at the beginning of English bisyllables can cue grammatical category. In A. Schalley &D. Khlentzos (Eds.), Mental states: Language and cognitive structure (Vol. 2, pp. 213–237). Amsterdam, the Netherlands: John Benjamins

Arciuli, J., & Slowiaczek, L. M. (2007). The where and when of linguistic word-level prosody. *Neuropsychologia, 45,* 2638-2642.

Arciuli, J., Monaghan, P., & Seva, N. (2010). Learning to assign lexical stress during reading aloud: Corpus, behavioral, and computational investigations. *Journal of Memory and Language, 63*, 180-196.

Arciuli, J., Simpson, B., Vogel, A., & Ballard, K.J. (2014). Acoustic changes in the production of lexical stress during Lombard speech. *Language and Speech 57:2*, 149-162.

Asami, T., Masumura, R., Yamaguchi, Y., Masataki, H., & Aono, Y. (2017) *Domain adaptation of DNN models using knowledge distillation.* Paper presented at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2017.

Austermann Hula, S. N., Robin, D. A., Maas, E., Ballard, K. J., & Schmidt, R. A. (2008). Effects of feedback frequency and timing on acquisition, retention and transfer of speech skills in acquired apraxia of speech. *Journal of Speech, Language and Hearing Research, 51,* 1088-1113.

Baker, E. (2012). Optimal intervention intensity in speech-language pathology: Discoveries, challenges, and unchartered territories. *International Journal of Speech-Language Pathology, 14*(5), 478-485. doi:10.3109/17549507.2012.717967

Ballard, K. J., Azizi, L., Duffy, J. R., McNeil, M. R., Halaki, M., O'Dwyer, N., . . . Robin, D. A. (2016). A predictive model for diagnosing stroke-related apraxia of speech.

Neuropsychologia, 81, 129-139. doi:https://doi.org/10.1016/j.neuropsychologia.2015.12.010

Ballard, K. J., Djaja, D., Arciuli, J., James, D. G. H., & van Doorn, J. (2012). Developmental trajectory for production of prosody: Lexical stress contrastivity in children ages 3 to 7 years and in adults. *Journal of Speech Language and Hearing Research, 55*, 1822-1835.

Ballard, K. J., Robin, D. A., McCabe, P., & McDonald, J. (2010). A Treatment for Dysprosody in Childhood Apraxia of Speech. *Journal of Speech, Language and Hearing Research, 53*(3), 1227-1245.

Ballard, K. J., Savage, S., Leyton, C. E., Vogel, A. P., Hornberger, M., & Hodges, J. R. (2014). Logopenic and Nonfluent Variants of Primary Progressive Aphasia Are Differentiated by Acoustic Measures of Speech Production. *PLoS ONE, 9*(2), e89864. doi:10.1371/journal.pone.0089864

Ballard, K.J., Granier, J.P, & Robin, D.A. (2000). Understanding the nature of apraxia of speech: Theory, analysis, and treatment. *Aphasiology, 14*(10): 969-995.

Ballard, K.J., Maas, E., & Robin, D.A. (2007). Treating control of voicing in apraxia of speech with variable practice. *Aphasiology, 21*(12): 1195-1217.

Benedon, T. A. (2018). *Speech-language pathologists' practices and attitudes towards app use in therapy.* (Master of Science), University of Wisconsin-Milwaukee, UWM Digital Commons. Retrieved from https://dc.uwm.edu/etd/1748

Boersma, P., & Weenink, D. (2011). Praat: doing phonetics by computer. (Version 5.3). Retrieved from http://www.praat.org

Breen, M., Dilley, L.C., MacAuley, J.D. & Sanders, L.D. (2014) Auditory evoked potentials reveal early perceptual effects of distal prosody on speech segmentation. *Language, Cognition and Neuroscience, 29*(9): 1132-1146

Brouwer, K., Downing, H., Westhoff, S., Wait, R., Entwisle, L. K., Messersmith, J. J., & Hanson, E. K. (2017). Effects of Clinician-Guided Emergent Literacy Intervention Using Interactive Tablet Technology for Preschool Children With Cochlear Implants. *Communication Disorders Quarterly, 38*(4), 195-205. doi:10.1177/1525740116666040

Brown, T., Murray, E., & McCabe, P. (2018). The boundaries of auditory perception for within-word syllable segregation in untrained and trained adult listeners. *Clinical Linguistics & Phonetics, 32*(11), 979-996. doi:10.1080/02699206.2018.1463395

Brumbaugh, K. M., & Smit, A. B. (2013). Treating Children Ages 3–6 Who Have Speech Sound Disorder: A Survey. *Language, Speech, and Hearing Services in Schools, 44*(3), 306-319. doi:10.1044/0161-1461(2013/12-0029)

Brunnegård, K., Lohmander, A., & van Doorn, J. (2009). Untrained listeners' ratings of speech disorders in a group with cleft palate: a comparison with speech and language pathologists' ratings. *International Journal of Language & Communication Disorders, 44*(5), 656-674. doi:10.1080/13682820802295203

Byun, T. M., Campbell, H., Carey, H., Liang, W., Park, T. H., & Svirsky, M. (2017). Enhancing Intervention for Residual Rhotic Errors Via App-Delivered Biofeedback: A Case Study. *Journal of Speech, Language, and Hearing Research, 60*(6S), 1810-1817. doi:doi:10.1044/2017_JSLHR-S-16-0248

Carrigg, B., Baker, E., Parry, L., & Ballard, K. J. (2015). Persistent speech sound disorder in a 22-year-old male: Communication, educational, socio-emotional, and vocational outcomes. *SIG 16 Perspectives on School-Based Issues, 16*, 37-49.

Carrigg, B., Parry, L., Baker, E., Shriberg, L. D., & Ballard, K. J. (2016). Cognitive, linguistic and motor abilities in a multigenerational family with childhood apraxia of speech. *Archives of Clinical Neuropsychology*, 1-20.

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of General Psychology, 130*(3), 290-304.

Chen, Y. J. (2011). Identification of articulation error patterns using a novel dependence network. *IEEE Transactions on Biomedical Engineering, 58*(11), 3061-3068.

Chen, Y.-P. P., Johnson, C., Lalbakhsh, P., Caelli, T., Deng, G., Tay, D., . . . Morris, M. (2016). Systematic review of virtual speech therapists for speech disorders. *Computer Speech and Language, 37*, 98-128.

Chiviacowsky, S., Wulf, G., de Medeiros, F. L., Kaefer, A., & Wally, R. (2008). Self-controlled feedback in 10-year-old children: Higher feedback frequencies enhance learning. *Research Quarterly for Exercise and Sport*, *79*(1), 122-127.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, XX*(1), 37-46.

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences.* . New York; USA: Academic Press.

Cole, R., & Muthusamy, Y. (1994). *OGI Multilanguage Corpus LDC94S17*. Retrieved from: https://catalog.ldc.upenn.edu/LDC94S17

Cooper, N., Cutler, A. & Wales, R. (2002). Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners. *Language and Speech, 45,* 207-228.

Cucchiarini, C. (1996). Assessing transcription agreement: Methodological aspects. *Clinical Linguistics & Phonetics, 10*(2), 131-155. doi:10.3109/02699209608985167

Cumming, T. M., & Draper Rodriguez, C. (2013). Integrating the ipad into language arts instruction for students with disabilities: Engaegment and perspectives. *Journal of Special Education Technology, 28*(4), 43-52.

Cutler, N. & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance 14* (1): 113-121.

Davis, B. L., Jakielski, K. J., & Marquardt, T. P. (1998). Developmental apraxia of speech: Determiners of differential diagnosis. *Clinical Linguistics & Phonetics*, *12*(1), 25-45.

Delmonte, R. (2009). Prosodic tools for language learning. *International Journal of Speech Technology, 12*(4), 161-184. doi:10.1007/s10772-010-9065-1

Des Roches, C. A., & Kiran, S. (2017). Technology-Based Rehabilitation to Improve Communication after Acquired Brain Injury. *Frontiers in Neuroscience, 11*, 382.

Des Roches, C. A., Balachandran, I., Ascenso, E. M., Tripodis, Y., & Kiran, S. (2015). Effectiveness of an impairment-based individualized rehabilitation program using an iPad-based software platform. *Frontiers in Human Neuroscience, 8*(1015). doi:10.3389/fnhum.2014.01015

Diehl, J. J., & Paul, R. (2009). The assessment and treatment of prosodic disorders and neurological theories of prosody. *International Journal of Speech-Language Pathology, 11*(4), 287-292.

Dilley, L.C., Mattys, S., & Vinke, L. (2010). Potent prosody: Comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation. *Journal of Memory and Language, 63*, 274-294.

Dodd, B.L., Hua, Z., Crosbie, S., Holm, A., & Ozanne, A. (2002) *Diagnostic Evaluation of Articulation and Phonology.* The Psychological Corporation, London: England.

Duenser, A., Ward, L., Stefania, A., Smith, D., Freyne, J., Morgan, A., & Dodd, B. (2016). Feasibility of Technology Enabled Speech Disorder Screening. In A. Georgiou, L. K. Schaper, & S. Whetton (Eds.), *Digital Health Innovation for Consumers, Clinicians, Connectivity and Community* (Vol. 227, pp. 21-27).

Duffy, J. R., Hanley, H., Utianski, R., Clark, H., Strand, E., Josephs, K. A., & Whitwell, J. L. (2017). Temporal acoustic measures distinguish primary progressive apraxia of speech from primary progressive aphasia. *Brain and Language, 168*, 84-94. doi:10.1016/j.bandl.2017.01.012

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test.* (4th. ed.). New York, USA.: Pearson Inc.

Edeal, D. M., & Gildersleeve-Neumann, C. E. (2011). The importance of production frequency in therapy for childhood apraxia of speech. *American Journal of Speech-Language Pathology, 20,* 95-110.

Edgar, D. L., & Rosa-Lugo, L. I. (2007). The critical shortage of speech-language pathologists in the public school setting: features of the work environment that affect recruitment and retention. *Language, Speech and Hearing Services in Schools, 38*(January), 31-46.

Edwards, J., & Dukhovny, E. (2017). Technology training in Speech-Language Pathology: A focus on tablets and apps. *Perspectives of the ASHA Special Interest Groups, SIG 10, 2*(Part 1), 33-48.

Fear, B. D., Cutler, A., & Butterfield, S. (1995). The strong/weak syllable distinction in English. *Journal of the Acoustical Society of America, 97*(3), 1983-1904.

Ferrer, L., Bratt, H., Richey, C., Franco, H., Abrash, V., & Precoda, K. (2015). Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. *Speech Communication, 69,* 31-45. doi:10.1016/j.specom.2015.02.002

Field, J. (2005). Intelligibility and the Listener: The Role of Lexical Stress. *TESOL Quarterly, 39*(3), 399-423. doi:10.2307/3588487

Fletcher, J. (2010). The Prosody of Speech: Timing and Rhythm. In *The Handbook of Phonetic Sciences* (pp. 521-602): Blackwell Publishing Ltd.

Flores, M., Musgrove, K., Renner, S., Hinton, V., Strozier, S., Franklin, S., & Hil, D. (2012). A Comparison of Communication Using the Apple iPad and a Picture-based System. *Augmentative and Alternative Communication, 28*(2), 74-84. doi:10.3109/07434618.2011.644579

Forrest, K. (2003). Diagnostic Criteria of Developmental Apraxia of Speech Used by Clinical Speech-Language Pathologists. *American Journal of Speech-Language Pathology, 12*(3), 376.

Furlong, L., Erickson, S., & Morris, M. (2017). Computer-based speech therapy for childhood speech sound disorders. *Journal of Communication Disorders, 68*, 50-69.

Furlong, L., Morris, M., Serry, T., & Erickson, S. (2018). Mobile apps for treatment of speech disorders in children: An evidence-based analysis of quality and efficacy. *PLoS ONE, 13*(8), e0201513. doi:10.1371/journal.pone.0201513

Ganz, J. B., Hong, E. R., & Goodwyn, F. D. (2013). Effectiveness of the PECS Phase III app and choice between the app and traditional PECS among preschoolers with ASD. *Research in Autism Spectrum Disorders, 7*, 973-983.

Gerosa, M. G., D. Brugnara. F. (2007). Acoustic variability and automatic recognition of children's speech. *Speech Communication, 49*, 847-860.

Gillon, G. T., & Moriarty, B. C. (2007). Childhood apraxia of speech: Children at risk for persitent reading and spelling disorder. *Seminars in Speech and Language, 28*(1), 48-57.

Glogowska, M., & Campbell, R. (2000). Investigating parental views of involvement in pre-school speech and language therapy. *International Journal of Language & Communication Disorders, 35*(3), 391-405. doi:10.1080/136828200410645

Goldman, R., & Fristoe, M. (2000). *Goldman-Fristoe test of articulation* (Second ed.). Minneapolis, MN.: Pearson.

Gomez, M., McCabe, P., & Purcell, A. (2018). *Clinical management of childhood apraxia of speech: A survey of speech-language pathologists.* Paper presented at the Speech Pathology Australia National Conference, Adelaide, Australia.

Gordon-Brannan, M. E., & Weiss, C. E. (2007). *Clinical Management of Articulatory and Phonologic Disorders.* Baltimore, MD: Lippincott Williams & Wilkins.

Gozzard, H., Baker, E., & McCabe, P. (2004) *Single Word Test of Polysyllables.* Unpublished work.

Gozzard, H., Baker, E., & McCabe, P. (2008). Requests for clarification and children's speech responses: Changing “pasghetti” to “spaghetti”. *Child Language Teaching and Therapy, 24,* 249-263.

Graneheim, U.H. & Lundman, B. (2004) Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today, 2,* 105-112.

Greenberg, S. (1999). Speaking in shorthand - a syllable centric perspective for understanding pronunciation variation. *Speech Communication, 29*, 159-176.

Guadagnoli, M. A., & Lee, T. D. (2004). Challenge point: A framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of Motor Behavior, 36*(2), 212-224.

Guadagnoli, M.A., & Kohl, R.M. (2001). Knowledge of results for motor learning: Relationship between error estimation and knowledge of results frequency. *Journal of Motor Behaviour, 33,* 217-224.

Hacker, C., Cincarek, T., Maier, A., HeBler, A., & Noth, E. (2007, 15-20 April 2007). *Boosting of prosodic and pronunciation features to detect mispronunciations of non-*

native children. Paper presented at the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07.

Hair, A., Monroe, P., Ahmed, B., Ballard, K. J., & Gutierrez-Osuna, R. (2018). *Apraxia world: a speech therapy game for children with speech sound disorders*. Paper presented at the Proceedings of the 17th ACM Conference on Interaction Design and Children, Trondheim, Norway.

Hall, P. K., Jordan, L. S., & Robin, D. A. (1993). *Developmental apraxia of speech: Theory and clinical practice.* Austin, TX.: Pro-Ed.

Hearnshaw, S., Baker, E., & Munro, N. (2014). The speech perception skills of children with and without speech dsound disorder. *Journal of Communication Disorders, 71,* 61-71.

Hedges, L.V. (1981) Distribution theory for Glass'estimator of effect size and related estimators. *Journal of Educational Statistics.* Vol 6. No.2 (Summer 1981): 107-128.

Hegarty, N., Titterington, J., McLeod, S., & Taggart, L. (2018). Intervention for children with phonological impairment: Knowledge, practices and intervention intensity in the UK. *International Journal of Language & Communication Disorders, 53*(5), 995-1006.

Hinton, G., Deng, L., Dong, Y., Dahl, G.E., Abdel-rahman, M., Navdeep, J., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared view of four research groups. *IEEE Signal Processing Magazine, 29:6,* 82-97.

Hosom, J. P., Shriberg, L., & Green, J. R. (2004). Diagnostic assessment of childhood apraxia of speech using automatic speech recognition (ASR) methods. *Journal of Medical Speech-Language Pathology, 12*(4), 167-171.

Hosom, J.P. (2009). Computer processing for analysis of speech disorders. In R. Paul & P. Flipsen (Eds.), *Speech Sound Disorders in Children: In Honour of Laurence D. Shriberg* (pp. 115-140). San Diego, USA: Plural Publishing.

IBM Corp. (2016). IBM SPSS Statistics for Windows (Version 24.0). Armonk, NY: IBM Corp.

Iuzzini, J., & Forrest, K. (2010). Evaluation of a combined treatment approach for childhood apraxia of speech. *Clinical Linguistics & Phonetics, 24*(4-5), 335-345.

Iuzzini-Seigel, J., Hogan, T. P., Guarino, A. J., & Green, J. R. (2015). Reliance on auditory feedback in children with childhood apraxia of speech. *Journal of Communication Disorders, 54*, 32-42.

Juan, S. S., Besacer, L., Lecouteux, B., & Tan, T.-P. (2015). *Merging native and non-native speech for low-resource accented ASR.* Paper presented at the 3rd International Conference on Statistical Language and Speech Processing., Budapest, Hungary.

Justice, L. M. (2018). Conceptualising “dose” in paediatric language interventions: Current findings and future directions. *International Journal of Speech-Language Pathology, 20*(3), 318-323. doi:10.1080/17549507.2018.1454985

Kaipa, R., & Peterson, A. M. (2016). A systematic review of treatment intensity in speech disorders. *International Journal of Speech-Language Pathology, 18*(6), 507-520.

Keilmann, A., Braun, L., & Napiontek, U. (2004). Emotional satisfaction of parents and speech-language therapists with outcome of training intervention in children with speech annd language disorders. *Folia Phoniatrica et Logopaedica, 56*(1), 51-61.

Kenny, B., & Lincoln, M. (2012). Sport, scales or war? Metaphors speech-language pathologists use to describe caseload management. *International Journal of Speech-Language Pathology, 14*(3), 247-259.

Kent, R. D. (1996). Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology, 5*(3), 7-23.

Kent, R.D. & Kim, Y-J. (2003). Towards an acoustic typology of motor speech disorders. *Clinical Linguistics & Phonetics, 17*(6): 427-445.

Kim, I.-S., LaPointe, L., & Stierwalt, J. A. G. (2012). The effect of feedback and practice on the acquisition of novel speech behaviors. *American Journal of Speech-Language Pathology, 21*, 89-100.

Kim, Y.J., & Beutnagel, M. C. (2011). *Automatic assessment of American English lexical stress using maching learning algorithms*. Paper presented at the SLaTE, Speech and Language Technology in Education, Venice, Italy.

Klopfenstein, M. (2009). Interaction between prosody and intelligibility. *International Journal of Speech-Language Pathology, 11*(4), 326-331. doi:10.1080/17549500903003094

Knock, T. R., Ballard, K. J., Robin, D. A., & Schmidt, R. A. (2000). Influence of order of stimulus presentation on speech motor learning: A principled approach to treatment for apraxia of speech. *Aphasiology, 14*(5-6), 653-668.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: fundamental frequency lends little. *Journal of the Acoustical Society of America, 118*(2), 1038-1054.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation.* Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kurland, J., Liu, A., & Stokes, P. (2018). Effects of a Tablet-Based Home Practice Program With Telepractice on Treatment Outcomes in Chronic Aphasia. *Journal of Speech, Language, and Hearing Research, 61*(5), 1140-1156. doi:doi:10.1044/2018_JSLHR-L-17-0277

Kurland, J., Wilkins, A. R., & Stokes, P. (2014). iPractice: piloting the effectiveness of a tablet-based home practice program in aphasia treatment. *Seminars in Speech and Language, 35*(1), 51-63.

Lai, Q., Shea, C.H, Wulf, G., & Wright, D.L. (2000). Optimizing generalized motor program and parameter learning. *Resaerch Quarterly for Exercise and Sport, 71*(1): 10-24.

Lan, T., Aryal, S., Ahmed, B., Ballard, K., & Gutierez-Osuna, R. (2014). *Flappy voice: An interactive game for childhood apraxia of speech therapy*. Paper presented at the CHI PLAY 2014 - Proceedings of the 2014 Annual Symposium on Computer-Human Interaction in Play.

Lancaster, G., Keusch, S., Levin, A., Pring, T., & Martin, S. (2010). Treating children with phonological problems: does an eclectic approach to therapy work? *International Journal of Language & Communication Disorders, 45*(2), 174-181. doi:10.3109/13682820902818888

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159-174. doi:10.2307/2529310

Lawler, K., Taylor, N. F., & Shields, N. (2013). Outcomes After Caregiver-Provided Speech and Language or Other Allied Health Therapy: A Systematic Review. *Archives of Physical Medicine and Rehabilitation, 94*(6), 1139-1160. doi:https://doi.org/10.1016/j.apmr.2012.11.022

Lee, A., Lang, R., Davenport, K., Moore, M., Rispoli, M., van der Meer, L., . . . Chung, C. (2015). Comparison of therapist implemented and iPad-assisted interventions for children with autism. *Developmental Neurorehabilitation, 18*(2), 97-103.

Leitão, S., Hogben, J., & Fletcher, J. (1997). Phonological processing skills in speech and language impaired children. *European Journal of Disorders of Communication*, 32, 91–113.

Lewis, B. A., Freebairn, L. A., Hansen, A. J., Iyengar, S. K., & Taylor, H. G. (2004). School-age follow-up of children with childhood apraxia of speech. *Language, Speech and Hearing Services in Schools, 35*, 122-140.

Li, K., Zhang, S., Li, M., Lo, W.-K., & Meng, H. (2011). *Prominence model for prosodic features in automatic lexical stress and pitch accent detection.* Paper presented at the INTERSPEECH, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31.

Lim, J. M., McCabe, P., & Purcell, A. (2017). Challenges and solutions in speech-language pathology service delivery across Australia and Canada. *European Journal for Person Centred Healthcare, 5*(1), 120-128.

Lim, J. M., McCabe, P., & Purcell, A. (2017). *Look at Mummy: Challenges in training parents to deliver a home treatment program for childhood apraxia of speech*. Paper presented at the Speech Pathology Australia National Conference, May, 2017; Sydney, Australia.

Maas, E. & Farinella, K.A. (2012). Random versus blocked practice in treatment for childhood apraxia of speech. *Journal of Speech Language and Hearing Research, 55*(2): 561-578.

Maas, E., Butalla, C. E., & Farinella, K. A. (2012). Feedback frequency in treatment for childhood apraxia of speech. *American Journal of Speech-Language Pathology, 21*, 239-247.

Maas, E., Gildersleeve-Neumann, C. E., Jakielski, K. J., & Stoeckel, R. (2014). Motor-Based Intervention Protocols in Treatment of Childhood Apraxia of Speech (CAS). *Current Developmental Disorders Reports, 1*(3), 197-206. doi:10.1007/s40474-014-0016-4

Maas, E., Robin, D. A., Austerman Hula, S. N., Freedman, S. E., Wulf, G., Ballard, K. J., & Schmidt, R. A. (2008). Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology, 17,* 277-298.

Markham, D., & Hazan, V. (2004). The effect of talker- and listener-related factors on intelligibility for a real-word, open set perecption test. *Journal of Speech, Language & Hearing Research, 47*(4), 725-737.

Mattys, S. (1997). The use of time during lexical processing and segmentation: A review. *Psychonomic Bulletin & Review, 4*(3): 310-329.

Mazzoni, D., & Dannenberg, R. (2000). Audacity (Version 1.3.9). Retrieved from http://www.audacityteam.org

McAllister, L., McCormack, J., McLeod, S., & Harrison, L. J. (2011). Expectations and experiences of accessing and participating in services for childhood speech impairment. *International Journal of Speech-Language Pathology, 13*(3), 251-267. doi:10.3109/17549507.2011.535565

McCabe, P., Macdonald-D'Silva, A, van Rees, L., Ballard, K.J. & Arciuli, J. (2014). Orthographically sensitive treatment for dysprosody in children with Childhood Apraxia of Speech using ReST intervention. *Developmental Neurorehabilitation, 17*(2), 137-146.

McCabe, P., Preston, J., & Evans, P. (2016). *Comparing treatments: An RCT contrasting ultrasound and ReST therapy for CAS.* Paper presented at the Childhood Apraxia of Speech Association of North America, Chicago, IL.

McCabe, P., Preston, J., Murray, E., Bricker, G., & Morgan, A. (2017). *What happens when they grow up? Experiences of adults who were diagnosed with childhood apraxia of speech as children.* Paper presented at the Speech Pathology Australia National Conference, Sydney.

McCabe, P., Rosenthal, J. B., & McLeod, S. (1998). Features of developmental dyspraxia in the general speech-impaired population? *Clinical Linguistics & Phonetics, 12*(2), 105-126. doi:10.3109/02699209808985216

McCann, J., & Peppe, S. (2003). Prosody in autism spectrum disorders: A critical review. *International Journal of Language and Communication Disorders*, 38, 325–350.

McCormack, J., McAllister, L., McLeod, S., & Harrison, L. (2012). Knowing, having, doing: The battles of childhood speech impairment. *Child Language Teaching and Therapy, 28*(2), 141-157. doi:10.1177/0265659011417313

McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Monroe, P., McCabe, P., & Ballard, K. J. (2018). Automated speech analysis tools for children's speech production: A systematic literature review. *International Journal of Speech-Language Pathology,* 1-17. doi:10.1080/17549507.2018.1477991

McKechnie, J., Ballard, K. J., McCabe, P., Murray, E., Lan, T., Gutierrez-Osuna, R., & Ahmed, B. (2016). *Tablet-based delivery of intensive speech therapy in children with childhood apraxia of speech: Influence of type of feedback.* Paper presented at the Speech Pathology Australia National Conference, Perth, Australia.

McKechnie, J., Ballard, K.J., Robin, D.A., Jacks, A. Palethorpe, S., & Rosen, K.M. (2008) *An acoustic typology of apraxic speech – toward reliable diagnosis*. Poster presented at: INTERSPEECH, September 2008, Brisbane, Australia.

McLeod, S., & Baker, E. (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), 508-531.

McLeod, S., & Baker, E. (2017). *Children's Speech: An Evidence-Based Approach to Assessment and Intervention.* . Boston, MA: Pearson.

McNeil, M.R., Robin, D.A., & Schmidt, R.A. (1997). Apraxia of speech: Definition, differentiation and treatment. In *Clinical management of sensorimotor speech disorders.* McNeil, M.R. (Ed). New York: Thieme.

McNeill, B. C., Gillon, G. T., & Dodd, B. (2009). Phonological awareness and early reading development in childhood apraxia of speech. *International Journal of Language & Communication Disorders, 44*(2), 175-192.

McNeill, B. C., Gillon, G. T., & Dodd, B. (2009). Effectiveness of an integrated phonological awareness approach for children with childhood apraxia of speech (CAS). *Child Language Teaching and Therapy, 25*(3), 341-366. doi:10.1177/0265659009339823

McNeill, B. C., Gillon, G. T., & Dodd, B. (2010). The Longer Term Effects of an Integrated Phonological Awareness Intervention for Children With Childhood Apraxia of Speech. *Asia Pacific Journal of Speech, Language and Hearing, 13*(3), 145-161. doi:10.1179/136132810805335074

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153-157. doi:10.1007/BF02295996

Miller, H., Plante, A., Ballard, K. J., & Robin, D. (2018). *Treatment of lexical stress, segmentation, and sound distortions in childhood apraxia of speech.* . Paper presented at the International Conference on Motor Speech, Savannah, GA.

Morgan, A. T., Murray, E., & Liégeois, F. J. (2018). Interventions for childhood apraxia of speech. *Cochrane Database of Systematic Reviews*(5). doi:10.1002/14651858.CD006278.pub3

Moriarty, B.C. & Gillon, G.T. (2006) Phonological awareness intervention for children with childhood apraxia of speech. *International Journal of Language & Communication Disorders, 41*(6): 713-734.

Morrill, T.H., Dilley, L.C., & MacAuley, J.D. (2014) Prosodic patterning in distal speech context: Effect of list intonation and f0 downtrend on perception of proximal prosodic structure. *Journal of Phonetics, 46,* 68-85.

Munnoch, M., Baker, E., Munro, N., & Hearnshaw, S. (2018). *Did you say 'rat' or 'wat'? An examination of adults' ability to judge the accuracy of children's speech.* Paper presented at the Speech Pathology Australia National Conference, Adelaide, Australia.

Munson, B., Bjorum, E. M., & Windsor, J. (2003). Acoustic and perceptual correlates of stress in nonwords produced by children with suspected developmental apraxia of speech and children with phonological disorder. *Journal of Speech Language and Hearing Research, 46*, 189-202.

Munson, B., Johnson, J. M., & Edwards, J. (2012). The Role of Experience in the Perception of Phonetic Detail in Children’s Speech: A Comparison Between Speech-Language Pathologists and Clinically Untrained Listeners. *American Journal of Speech-Language Pathology, 21*(2), 124-139. doi:10.1044/1058-0360(2011/11-0009)

Murray, E., McCabe, P., & Ballard, K. J. (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*, 112-120.

Murray, E., McCabe, P., & Ballard, K. J. (2014). A systematic review of treatment outcomes for children with childhood apraxia of speech. *American Journal of Speech-Language Pathology, 23*(3), 486-504. doi:10.1044/2014_AJSLP-13-0035

Murray, E., McCabe, P., & Ballard, K. J. (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*, 669-686.

Murray, E., McCabe, P., Heard, R., & Ballard, K. J. (2015). Differential diagnosis of children with suspected childhood apraxia of speech. *Journal of Speech, Language & Hearing Research, 58*(1), 43-60. doi:10.1044/2014_JSLHR-S-12-0358

Murray, E., McKechnie, J., & Williams, P. (2017). *Exploring factors for treatment success in Childhood Apraxia of Speech using the Nuffield Dyspraxia Programme - 3rd Edition*. Paper presented at the Speech Pathology Australia National Conference, Sydney, Australia.

Murray, E., Thomas, D. C., & McKechnie, J. (2018). Comorbid morphological disorder apparent in some children aged 4-5 years with childhood apraxia of speech: findings from standardised testing. *Clinical Linguistics & Phonetics,* 1-18. doi:10.1080/02699206.2018.1513565

Namasivayam, A. K., Pukonen, M., Goshulak, D., Hard, J., Rudzicz, F., Rietveld, T., . . . Lieshout, P. (2015). Treatment intensity and childhood apraxia of speech. *International Journal of Language & Communication Disorders, 50*(4), 529-546. doi:10.1111/1460-6984.12154

Namasivayam, A., Pukonen, M., Hard, J., Jahnke, R., Kearney, E., Kroll, R., & van Lieshout, P. (2015a). Motor speech treatment protocol for developmental motor speech disorders. *Developmental Neurorehabilitation, 18*(5), 296-303. doi:10.3109/17518423.2013.832431

National Health and Medical Research Council. (2009). *NHMRC: A guide to the development, implementation and evaluation of clinical practice guidelines.* Canberra, Australia: National Health and Medical Research Council Retrieved from https://www.nhmrc.gov.au/_files_nhmrc/file/guidelines/developers/nhmrc_levels_grades_evidence_120423.pdf.

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

Nijland, L., Maasen, B., van der Muelen, S., Gabreels, F., Kraaimaat, F. W., & Schreuder, R. (2003). Planning of syllables in children with developmental apraxia of speech. *Clinical Linguistics & Phonetics, 17*(1), 1-24.

Nijland, L., Terband, H., & Maassen, B. (2015). Cognitive Functions in Childhood Apraxia of Speech. *Journal of Speech, Language, and Hearing Research, 58*(3), 550-565. doi:10.1044/2015_JSLHR-S-14-0084

Nittrouer, S., & Miller, M. E. (1997). Developmental weighting shifts for noise components of fricative-vowel syllables. *The Journal of the Acoustical Society of America*, *102*(1), 572-580.

Nordness, A. S., & Beukelman, D. R. (2010). Speech practice patterns of children with speech sound disorders: The impact of parental record keeping and computer-led practice. (Report). *Journal of Medical Speech-Language Pathology, 18*(4), 104-108.

O'Callaghan, A. M., McCallister, L., & Wilson, L. (2005). Consumers' proposed solutions to barriers to access of rural and remote speech pathology services. *Advances in Speech Language Pathology, 7*(2), 58-64. doi:10.1080/14417040500125277

O'Callaghan, C., McAllister, L., & Wilson, L. (2005). Barriers to accessing rural paediatric speech pathology services: Health care consumers' perspectives. *Australian Journal of Rural Health, 13*, 162-171.

Oliveira, C., Lousada, M., & Jesus, L. M. T. (2015). The clinical practice of speech and language therapists with children with phonologically based speech sound disorders. *Child Language Teaching & Therapy, 31*(2), 173-194.

Olswang, L. B., & Bain, B. A. (2013). Treatment Research. In L. A. C. Golper & C. Frattali (Eds.), *Measuring Outcomes in Speech-Language Pathology* (2nd ed., pp. 245-264). New York,NY: Thieme Medical.

Pappas, N. W., McLeod, S., McAllister, L., & McKinnon, D. H. (2008). Parental involvement in speech intervention: A national survey. *Clinical Linguistics & Phonetics, 22*(4-5), 335-344. doi:10.1080/02699200801919737

Parnandi, A., Karappa, V., Lan, T., Shahin, M., McKechnie, J., Ballard, K., . . . Gutierrez-Osuna, R. (2015). Development of a remote therapy tool for childhood apraxia of speech. *ACM Transactions on Accessible Computing, 7*(3). doi:10.1145/2776895

Parnandi, A., Karappa, V., Son, Y., Shahin, M., McKechnie, J., Ballard, K., . . . Gutierrez-Osuna, R. (2013). *Architecture of an automated therapy tool for childhood apraxia of speech.* Paper presented at the Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2013.

Paul, R., Augustyn, A., Klin, A., & Volkmar, F. R. (2005). Perception and production of prosody by speakers with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *33*, 205–220.

Paul, R., Shriberg, L. D., McSweeny, J., Cicchetti, D., Klin, A., & Volkmar, F. (2005). Brief Report: Relations between Prosodic Performance and Communication and Socialization Ratings in High Functioning Speakers with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders, 35*(6), 861. doi:10.1007/s10803-005-0031-8

Peppé, S. J. E. (2009). Why is prosody in speech-language pathology so difficult? *International Journal of Speech-Language Pathology, 11*(4), 258-271. doi:10.1080/17549500902906339

Peterson, G.E. & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America, 32,* 693-703.

Preston, J. L., Brick, N., & Landi, N. (2013). Ultrasound Biofeedback Treatment for Persisting Childhood Apraxia of Speech. *American Journal of Speech-Language Pathology, 22*(4), 627-643. doi:10.1044/1058-0360(2013/12-0139)

Preston, J. L., Leece, M. C., & Maas, E. (2016). Intensive Treatment with Ultrasound Visual Feedback for Speech Sound Errors in Childhood Apraxia. *Frontiers in Human Neuroscience, 10*(440). doi:10.3389/fnhum.2016.00440

production treatment for apraxia of speech: Overgeneralization and maintenance effects. *Aphasiology*, 13, 821–837.

Pye, C., Wilcox, K. A., & Siren, K. A. (1988). Refining transcriptions: the significance of transcriber 'errors'. *Journal of Child Language, 15*, 17-37.

Robbins, J. & Klee, T. (1987) Clinical assessment of oropharyngeal motor development in young children. *Journal of Speech and Hearing Disorders, 52,* 271-277.

Rubin, Z. & Kurniawan, S. (2013). Speech Adventure: Using speech recognition for cleft speech therapy. *Proceedings of the 6th International Conference on Pervasive Technologies related to Assistive Environments,* Rhodes, Greece.

Rudzicz, F., Namasivayam, A.K., & Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation, 46*(4): 523-541.

Ruggero, L., McCabe, P., Ballard, K. J., & Munro, N. (2012). Paediatric speech-language pathology service delivery: An exploratory survey of Australian parents. *International Journal of Speech-Language Pathology, 14*(4), 338-350.

Ruscello, D. M., Cartwright, L. R., Haines, K. B., & Shuster, L. I. (1993). The use of different service delivery models for children with phonological disorders. *Journal of communication disorders, 26*(3), 193-203.

Salmoni, A., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: A review and critical reappraisal. *Psychological Bulletin, 95,* 355-386.

Schellinger, S. K., Munson, B., & Edwards, J. (2017). Gradient perception of children's productions of /s/ and /θ/: A comparative study of rating methods. *Clinical Linguistics & Phonetics, 31*(1), 80-103. doi:10.1080/02699206.2016.1205665

Schmidt, R. A., & Lee, T. D. (2011). *Motor Control and Learning: A Behavioural Emphasis* (Fifth ed.). Champaign, IL: Human Kinetics.

Schneider, S., & Frens, R. (2005). Training four-syllable CV patterns in individuals with acquired apraxia of speech: Theoretical implications. *Aphasiology, 19*(3-5), 451-471.

Semel, E., Wiig, E., & Secord, W. (2006) *Clinical Evaluation of Language Fundamentals, Australian standardised* (4th ed.) Sydney, Australia: Pearson.

Shahin, M., Ahmed, B., & Ballard, K. J. (2012). *Automatic classification of unequal lexical stress patterns using machine learning algorithms.* Paper presented at the 14th Australasian Conference on Speech Science and Technology, Sydney, Australia.

Shahin, M., Ahmed, B., McKechnie, J., Ballard, K., & Gutierrez-Osuna, R. (2014). *Comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech.* Paper presented at the Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.

Shahin, M., Ahmed, B., Parnandi, A., Karappa, V., McKechnie, J., Ballard, K. J., & Gutierrez-Osuna, R. (2015). Tabby Talks: An automated tool for the assessment of childhood apraxia of speech. *Speech Communication, 70*, 49-64. doi:10.1016/j.specom.2015.04.002

Shahin, M., Epps, J. & Ahmed, B. (2016) *Automatic classification of lexical stress patterns in English and Arabic languages using deep learning.* Paper presented at the Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.

Shahin, M., Gutierrez-Osuna, R., & Ahmed, B. (2016). *Classification of bisyllabic lexical stress patterns in disordered speech using deep learning.* Paper presented at the The 41st IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China.

Shahin, M., Ji, J. X., & Ahmed, B. (2018). *One-class SVMs based pronunciation verification approach.* Paper presented at the International Conference on Pattern Recognition, Beijing, China.

Shriberg, L. D. (1993). Four New Speech and Prosody-Voice Measures for Genetics Research and Other Studies in Developmental Phonological Disorders. *36*(1), 105-140. doi:doi:10.1044/jshr.3601.105

Shriberg, L. D., & Lof, L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics & Phonetics, 5*, 225-279.

Shriberg, L. D., & Widder, C. J. (1990). Speech and prosody characteristics of adults with mental retardation. *Journal of Speech, Language, and Hearing Research, 33,* 627–653.

Shriberg, L. D., Aram, D. M., & Kwiatkowski, J. (1997a). Developmental apraxia of speech: I. Descriptive and theoretical perspectives. *Journal of Speech, Language, and Hearing Research, 40*(2), 273-285.

Shriberg, L. D., Aram, D. M., & Kwiatkowski, J. (1997b). Developmental apraxia of speech: II. toward a diagnostic marker. *Journal of Speech, Language & Hearing Research, 40*(2), 286-312.

Shriberg, L. D., Aram, D. M., & Kwiatkowski, J. (1997c). Developmental apraxia of speech: III. a subtype marked by inappropriate stress. *Journal of Speech, Language & Hearing Research, 40*(2), 313-337.

Shriberg, L. D., Campbell, T. F., Karlsson, H. B., Brown, R. L., McSweeny, J. L., & Nadler, C. J. (2003). A diagnostic marker for childhood apraxia of speech: the lexical stress ratio. *Clinical Linguistics & Phonetics, 17*(7), 549-574. doi:http://dx.doi.org/10.1080/0269920031000138123

Shriberg, L. D., Fourakis, M., Hall, S. D., Karlsson, H. B., Lohmeier, H. L., McSweeny, J. L., . . . Wilson, D. L. (2010). Extensions to the Speech Disorders Classification System (SDCS). *Clinical Linguistics & Phonetics, 24*(10), 795-824. doi:http://dx.doi.org/10.3109/02699206.2010.503006

Shriberg, L. D., Kwiatkowski, J. and Rasmussen, C., 1990, The Prosody-Voice Screening ProWle (Tucson, AZ: Communication Skill Builders).

Shriberg, L. D., Lohmeier, H. L., Strand, E. A., & Jakielski, K. J. (2012). Encoding, memory, and transcoding deficits in Childhood Apraxia of Speech. *Clinical Linguistics & Phonetics, 26*(5), 445-482. doi:10.3109/02699206.2012.655841

Shriberg, L. D., Strand, E. A., Fourakis, M., Jakielski, K. J., Hall, S. D., Karlsson, H. B., . . . Wilson, D. L. (2017a). A Diagnostic Marker to Discriminate Childhood Apraxia of Speech From Speech Delay: I. Development and Description of the Pause Marker. *Journal of Speech, Language, and Hearing Research, 60*(4), S1096-S1117. doi:10.1044/2016_JSLHR-S-15-0296

Shriberg, L. D., Strand, E. A., Fourakis, M., Jakielski, K. J., Hall, S. D., Karlsson, H. B., . . . Wilson, D. L. (2017b). A Diagnostic Marker to Discriminate Childhood Apraxia of Speech From Speech Delay: II. Validity Studies of the Pause Marker. *Journal of Speech, Language, and Hearing Research, 60*(4), S1118-S1134. doi:10.1044/2016_JSLHR-S-15-0297

Shriberg, L.D., Austin, D., Lewis, B.A., McSweeny, J.L, & Wilson, D.L. (1997). The Percentage Consonants Correct (PCC) Metric: Extensions and reliability data. *Journal of Speech, Language and Hearing Research, 40*(4): 708-722.

Simmons, E. S., Paul, R., & Shic, F. (2016). Brief report: A mobile application to treat prosodic deficits in autism spectrum disorder and other communication impairments: A pilot study. *Journal of Autism & Developmental Disorders, 46*(1), 320-327.

Skahan, S. M., Watson, M., & Lof, G. L. (2007). Speech-language pathologists' assessment practices for children with suspected speech sound disorders: results of a national survey. *American Journal of Speech-Language Pathology, 16*(3), 246-259.

Skinder, A., Connaghan, K., Strand, E. A., & Betz, S. (2000). Acoustic correlates of perceived lexical stress errors in children with developmental apraxia of speech. *Journal of Medical Speech-Language Pathology, 8*(4), 279-284.

Skinder, A., Strand, E. A., & Mignerey, M. (1999). Perceptual and acoustic analysis of lexical and sentential stress in children with developmental apraxia of speech. *Journal of Medical Speech-Language Pathology, 7*(2), 133-144.

Skinder, A., Strand, E.A., Stoel-Gammon, C., Mignerey, M., & Betz, S. (1999, November). *Lexical stress errors in children with developmental apraxia of speech versus typically developing peers.* Poster presented to the American Speech-Language and Hearing Association, San Francisco, CA.

Slowiaczek, L. (1990). Effects of lexical stress in auditory word recognition. *Language and Speech, 33*, 47–68.

Smith, A. B., & Robb, M. P. (2006). The influence of utterance position on children's production of lexical stress. *Folia Phoniatrica et Logopaedica, 58*(3), 199-206. doi:10.1159/000091733

Smith, A. B., Goffman, L., Zelaznik, H. N., Ying, G., & McGillem, C. (1995). Spatiotemporal stability and the patterning of speech moement sequences. *Experimental Brain Research, 104*, 493-501.

Snowling, M., & Stackhouse, J. (1983). Spelling performance of children with developmental verbal dyspraxia. *Developmental Medicine & Child Neurology, 25*(4), 430-437. doi:doi:10.1111/j.1469-8749.1983.tb13787.x

Spielman, J., Ramig, L.O., Mahler, L., Halpern, A., & Gavin, W.J. (2007). Effects of an extended version of Lee Silverman Voice Treatment on voice and speech in Parkinson’s Disease. *American Journal of Speech-Language Pathology, 16*(2): 95-107.

Stewart Keck, C. & Doarn, C.R. (2014). Telehealth Technology Applications in Speech-Language Pathology. *Telemedicine and e-Health, 20*(7), 653-659. doi:10.1089/tmj.2013.0295

Strand, E. A., & Debertine, P. (2000). The efficacy of integral stimulation intervention with developmental apraxia of speech. *Journal of Medical Speech-Language Pathology, 8*(4), 295-300.

Strand, E. A., Stoeckel, R. & Baas, B. (2006). Treatment of severe childhood apraxia of speech: A treatment efficacy study. *Journal of Medical Speech-Language Pathology, 14*(4), 297-307.

Sugden, E., Baker, E., Munro, N., & Williams, A. L. (2016). Involvement of parents in intervention for childhood speech sound disorders: a review of the evidence. *International Journal of Language & Communication Disorders, 51*(6), 597-625. doi:10.1111/1460-6984.12247

Sugden, E., Baker, E., Munro, N., Williams, A. L., & Trivette, C. M. (2017). An Australian survey of parent involvement in intervention for childhood speech sound disorders. *International Journal of Speech-Language Pathology,* 1-13. doi:10.1080/17549507.2017.1356936

Sugden, E., Baker, E., Munro, N., Williams, A. L., & Trivette, C. M. (2018). Service delivery and intervention intensity for phonology-based speech sound disorders. *International Journal of Language & Communication Disorders, 53*(4), 718-734. doi:10.1111/1460-6984.12399

Sullivan, K. J., Kantak, S. S., & Burtner, P. A. (2008). Motor learning in children: Feedback effects on skill acquisition. *Physical Therapy, 88*, 720-732.

Swinnen, S.P., Lee, T.D., Verschueren, S., Serrien, D.J., & Bogaerds, H. (1997). Interlimb coordination: Learning and transfer under different feedback conditions. *Human Movement Science, 16,* 749-785.

Swinnen, S.P., Walter, C.B., Lee, T.D., & Serrien, D.J. (1993). Acquiring bimanual skills: Contrasting forms of information feedback for interlimb decoupling. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1328-1344.

Sztaho, D., Nagy, K., & Vicsi, K. (2010). *Subjective tests and automatic sentence modality recognition with recordings of speech impaired children*. Paper presented at the Proceedings of the Second International Conference on Development of Multimodal Interfaces: Active Listening and Synchrony, Dublin, Ireland.

Tamburini, F., & Caini, C. (2005). An automatic sytem for detecting prosodic prominence in American English continuous speech. *International Journal of Speech Technology, 8*(1), 33-44.

Tan, C. T., Johnston, A., Ballard, K., Ferguson, S., & Perera-Schulz, D. (2013). *sPeAK-MAN: towards popular gameplay for speech therapy*. Paper presented at the Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death, Melbourne, Australia.

Terband, H., van Brenk, F., & van Doornik-van der Zee, A. (2014). Auditory feedback perturbation in children with developmental speech sound disorders. *Journal of Communication Disorders, 51*, 64-77.

Theodoros, D. (2012). A new era in speech-language pathology practice: Innovation and diversification. *International Journal of Speech-Language Pathology, 14*(3), 189-199.

Thomas, D. C., McCabe, P., & Ballard, K. J. (2014). Rapid Syllable Transitions (ReST) treatment for Childhood Apraxia of Speech: The effect off lower dose-Frequency. *Journal of Communication Disorders, 51*, 29-42.

Thomas, D. C., McCabe, P., & Ballard, K. J. (2017). Combined clinician-parent delivery of rapid syllable transition (ReST) treatment for childhood apraxia of speech. *International Journal of Speech-Language Pathology*, 1-16. doi:10.1080/17549507.2017.1316423

Thomas, D. C., McCabe, P., Ballard, K. J., & Bricker-Katz, G. (2018). Parent experiences of variations in service delivery of Rapid Syllable Transition (ReST) treatment for childhood apraxia of speech. *Developmental Neurorehabilitation, 21*(6), 391-401. doi:10.1080/17518423.2017.1323971

Thomas, D. C., McCabe, P., Ballard, K. J., & Lincoln, M. (2016). Telehealth delivery of Rapid Syllable Transitions (ReST) treatment for childhood apraxia of speech.

International Journal of Language & Communication Disorders, 51(6), 654-671. doi:10.1111/1460-6984.12238

To, C. K., Law, T., & Cheung, P. S. P. (2012). Treatment intensity in everyday clinical mangement of speech sound disorders in Hong Kong. *International Journal of Speech-Language Pathology, 45*(5), 462-466.

Toki, E. I., & Pange, J. (2010). E-learning activities for articulation in speech language therapy and learning for preschool children. *Procedia Social and Behavioural Sciences, 2*, 4274-4278.

Tommy, C. A., & Minoi, J. L. (2016, 4-8 Dec. 2016). *Speech therapy mobile application for speech and language impairment children.* Paper presented at the 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES).

treatment for sound errors in apraxia of speech and aphasia. *Journal of Speech, Language, and Hearing Research,* 41, 725–743.

Turk, A. E., & Sawusch, J. (1997). The domain of accentual lengthening in American English. *Journal of Phonetics, 25,* 25–41.

Van Kuijk, D., & Boves, L. (1999). Acoustic characteristics of lexical stress in continuous telephone speech. *Speech Communication, 27,* 95–111.

van Santen, J. P. H., Prud'hommeaux, E. T., & Black, L. M. (2009). Automated assessment of prosody production. *Speech Communication, 51*(11), 1082-1097. doi:10.1016/j.specom.2009.04.007

Verdon, S., Wilson, L., Smith-Tamaray, M., & McAllister, L. (2011). An investigation of equity of rural speech-language pathology services for children: a geographical perspective. *International Journal of Speech-Language Pathology, 13*(3), 239-250.

Vergis, M. K., Ballard, K. J., Duffy, J. R., McNeil, M. R., Scholl, D., & Layfield, C. (2014). An acoustic measure of lexical stress differentiates aphasia and aphasia plus apraxia of speech after stroke. *Aphasiology, 28*(5), 554-575. doi:10.1080/02687038.2014.889275

Wambaugh, J. L., Kalinyak-Fliszar, M. M., West, J. E., & Doyle, P. J. (1998). Effects of treatment for sound errors in apraxia of speech and aphasia. *Journal of Speech, Language, and Hearing Research,* 41, 725–743.

Wambaugh, J. L., Martinez, A. L., McNeil, M. R., & Rogers, M. A. (1999). Sound production treatment for apraxia of speech: Overgeneralization and maintenance effects. *Aphasiology*, 13, 821–837.

Wambaugh, J. L., Nessler, C., Wright, S., Mauszycki, S. C., DeLong, C., Berggren, K., & Bailey, D. J. (2017). Effects of blocked and random practice schedule on outcomes of Sound Production Treatment for Acquired Apraxia of Speech: Results of a group investigation. *Journal of Speech, Language & Hearing Research, 60,* 1739-1751.

Warren, S. F., Fey, M.E. & Yoder, P.J. (2007). Differential treatment intensity research: a missing link to creating effective communication interventions. *Mental Retardation and Developmental Disabilities Research Reviews, 13*, 70-77.

Watts Pappas, N., McAllister, L., & McLeod, S. (2015). Parental beliefs and experiences regarding involvement in intervention for their child with speech sound disorder. *Child Language Teaching and Therapy, 32*(2), 223-239. doi:10.1177/0265659015615925

Wiig, E., Semel, E. Secord, W. (2006). *Clinical evaluation of language fundamentals preschool, Australian and New Zealand standardised* (2nd ed.). Sydney, Australia: Pearson.

Williams, L. A. (2012). Intensity in phonological intervention: Is there a prescribed amount? *International Journal of Speech-Language Pathology, 14*(5), 456-461.

Williams, P. & Stephens, H. (2004). *The Nuffield Dyspraxia Programme – Third Edition.* Windsor, England: The Miracle Factory.

Wilson, L., Lincoln, M., & Onslow, M. (2002). Availability, access and quality of care: Inequities in rural speech pathology services for children and a model for redress. *Advances in Speech-Language Pathology, 4*(1), 9-22.

Wolfe, V., Martin, D., Borton, T., & Youngblood, H. C. (2003). The effect of clinical experience on cue trading for the /r - w/ contrast. *American Journal of Speech-Language Pathology, 12*(2), 221-228.

Wren, Y., & Roulstone, S. (2008). A comparison between computer and tabletop delivery of phonology therapy. *International Journal of Speech-Language Pathology, 10*(5), 346-363. doi:10.1080/17549500701873920

Wulf, G., & Shea, C. (2004). Understanding the role of augmented feedback: The good, the bad and the ugly. In M. Williams, N. Hodges, & M. Scott (Eds.), *Skill Acquisition in Sport: Research, Theory and Practice.* Florence, KY, USA: Routledge.

Wulf, G., Shea, C., & Lewthwaite, R. (2010). Motor skill learning and performance: a review od influential factors. *Medical Education, 44*, 75-84.

Xie, H., Andreae, P., Zhang, M., & Warren, P. (2004, January 2004). *Detecting stress in spoken English using Decision Trees and Support Vector Machines.* Paper presented at the ACSW Frontiers '04 Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation, Dunedin, New Zealand.

Yeung, G., & Alwan, A. (2018). *On the difficulties of automatic speech recognition for kingergarten-aged children.* Paper presented at the INTERSPEECH, Hyderabad, India.

Young, D.E. & Schmidt, R.A. (1992). Augmented kinematic feedback for motor learning. *Journal of Motor Behavior, 24,* 261-273.

Zeng, B., Law, J., & Lindsay, G. (2012). Characterizing optimal intervention intensity: The relationship between dosage and effect size in interventions for children with developmental speech and language difficulties. *International Journal of Speech-Language Pathology, 14*(5), 471-477. doi:10.3109/17549507.2012.720281