International Journal of Speech-Language Pathology

ISSN: 1754-9507 (Print) 1754-9515 (Online) Journal homepage: https://www.tandfonline.com/loi/iasl20

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

J. McKechnie, B. Ahmed, R. Gutierrez-Osuna, P. Monroe, P. McCabe & K. J. Ballard

To cite this article: J. McKechnie, B. Ahmed, R. Gutierrez-Osuna, P. Monroe, P. McCabe & K. J. Ballard (2018) Automated speech analysis tools for children's speech production: A systematic literature review, International Journal of Speech-Language Pathology, 20:6, 583-598, DOI: 10.1080/17549507.2018.1477991

To link to this article: https://doi.org/10.1080/17549507.2018.1477991

International Journal of Speech-Language Pathology, 2018; 20: 583–598

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 typically <80% 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

Correspondence: Jacqueline McKechnie, Faculty of Health Sciences University of Sydney, PO Box 170, Lidcombe, 1825 NSW, Australia. Email: jacqueline.mckechnie@sydney.edu.au

ISSN 1754-9507 print/ISSN 1754-9515 online
Published by Informa UK Limited, trading as Taylor & Francis Group
DOI: 10.1080/17549507.2018.1477991

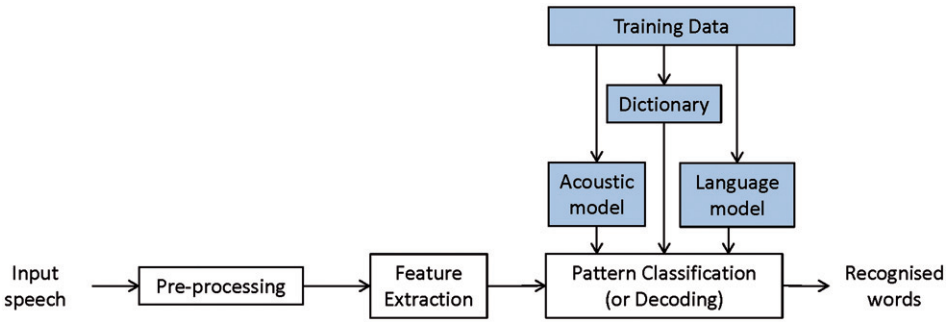

Figure 1. Basic components of a speech recognition system.

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. coarticulation 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 computer-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 management, 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 behaviour 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 pathology 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.

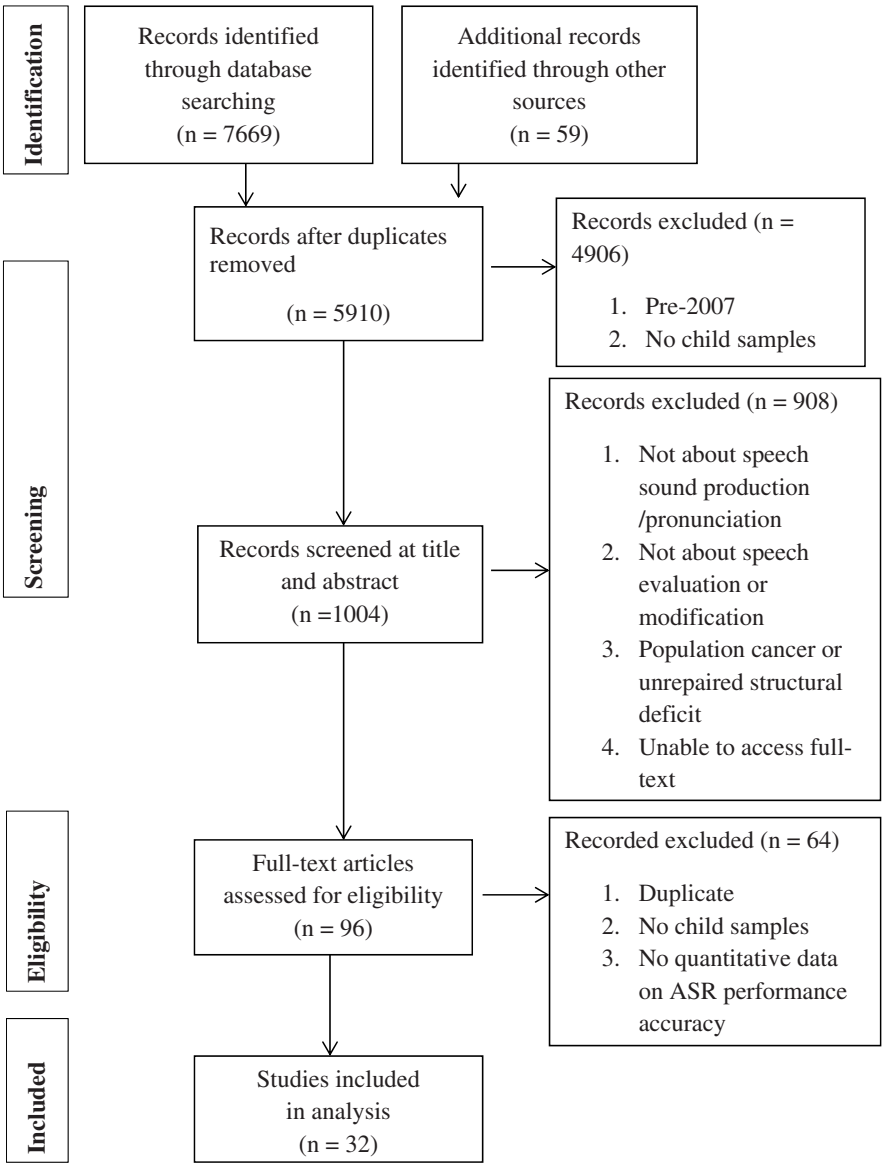

Figure 2. Systematic search and review flowchart.

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 of <8 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 ($n = 29$ 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% ($n = 9$) of articles and to disordered speech in 71.9% ($n = 23$).

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 pronunciation (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 intelligibility 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,

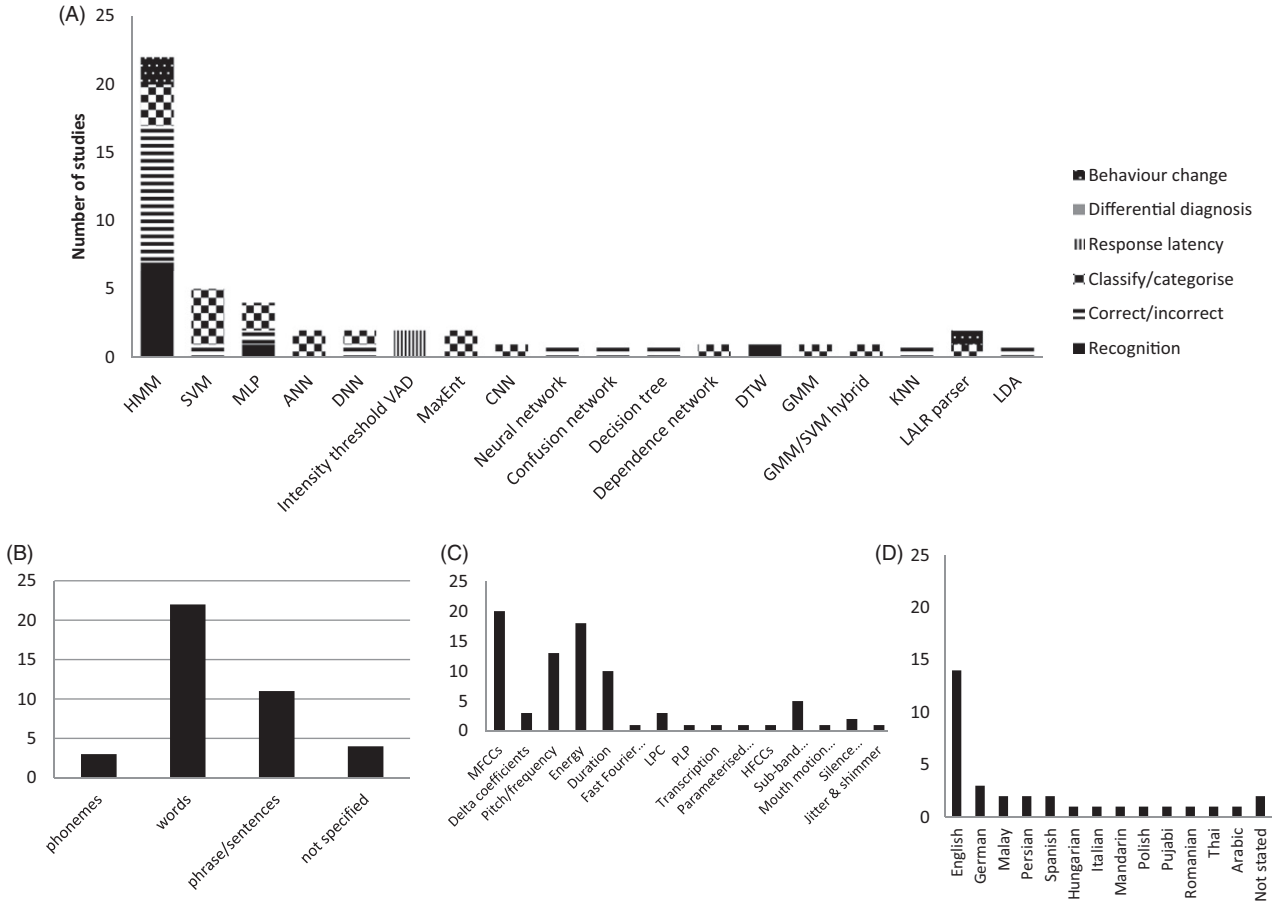

Figure 3. Frequency across the 32 studies of A. each automated technology used and proportion of tools addressing each analysis focus (HMM = Hidden Markov Models; SVM = Support Vector Machine; MLP = MultiLayer Perceptron; ANN = Artificial Neural Network; DNN = Deep Neural Network VAD = Voice Activity Detector; MaxEnt = Maximum Entropy; CNN = Convolutional Neural Network; DTW = Dynamic Time Warping; GMM = Gaussian Mixture Models; KNN = k-nearest neighbour algorithm; LALR parser was not defined in the study; LDA = Linear Discriminant Analysis); B. each type of speech sample elicited; C. use for each feature extraction method (MFCCs = mel-frequency cepstral coefficients; LPC = linear predictive coding coefficients; PLP = perceptual linear prediction coding coefficients; HFCCs = 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 information 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 ($n = 22$ 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 ($n = 24$ 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 phoneme 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 performance 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].

ORCID

J. McKechnie http://orcid.org/0000-0001-5747-0888
B. Ahmed http://orcid.org/0000-0002-1240-6572
P. McCabe http://orcid.org/0000-0002-5182-1007
K. J. Ballard http://orcid.org/0000-0002-9917-5390

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 kokyb÷s vertinimas pagalbinio kalb÷jimo 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 & 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