Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 70 (2015) 49–64

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Tabby Talks: An automated tool for the assessment of childhood apraxia of speech

Mostafa Shahin [a,*], Beena Ahmed [a], Avinash Parnandi [b], Virendra Karappa [b], Jacqueline McKechnie [c], Kirrie J. Ballard [c], Ricardo Gutierrez-Osuna [b]

[a] Department of Electrical and Computer Engineering, Texas A&M University, Doha 23874, Qatar
[b] Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843, USA
[c] Faculty of Health Sciences, The University of Sydney, Sydney, NSW 2141, Australia

Received 3 June 2014; received in revised form 26 February 2015; accepted 2 April 2015
Available online 9 April 2015

Abstract

Children with developmental disabilities such as childhood apraxia of speech (CAS) require repeated intervention sessions with a speech therapist, sometimes extending over several years. Technology-based therapy tools offer the potential to reduce the demanding workload of speech therapists as well as time and cost for families. In response to this need, we have developed "Tabby Talks," a multi-tier system for remote administration of speech therapy. This paper describes the speech processing pipeline to automatically detect common errors associated with CAS. The pipeline contains modules for voice activity detection, pronunciation verification, and lexical stress verification. The voice activity detector evaluates the intensity contour of an utterance and compares it against an adaptive threshold to detect silence segments and measure voicing delays and total production time. The pronunciation verification module uses a generic search lattice structure with multiple internal paths that covers all possible pronunciation errors (substitutions, insertions and deletions) in the child's production. Finally, the lexical stress verification module classifies the lexical stress across consecutive syllables into strong–weak or weak-strong patterns using a combination of prosodic and spectral measures. These error measures can be provided to the therapist through a web interface, to enable them to adapt the child's therapy program remotely. When evaluated on a dataset of typically developing and disordered speech from children ages 4–16 years, the system achieves a pronunciation verification accuracy of 88.2% at the phoneme level and 80.7% at the utterance level, and lexical stress classification rate of 83.3%.

Keywords: Speech therapy; Automatic speech recognition; Pronunciation verification; Computer aided pronunciation learning; Prosody

1. Introduction

Language production and speech articulation can be delayed in children due to developmental disabilities and neuromotor disorders such as childhood apraxia of speech (CAS) (Dodd, 2005). Treatment for CAS involves extended one-on-one therapy with a speech language pathologist (SLP), which can be difficult to manage due to time constraints and expenses (Adhoc Committee on CAS, 2007b). Children often have difficulty monitoring their own speech and self-correcting their errors; for this reason, they benefit from repeated practice with producing the sounds as well as listening and evaluating their attempts (Ballard et al., 2010). Early intervention can reduce the negative effects of childhood speech-language disorders

* Corresponding author. Tel.: +974 3365 8817.
E-mail addresses: mostafa.shahin@qatar.tamu.edu (M. Shahin), beena.ahmed@qatar.tamu.edu (B. Ahmed), parnandi@tamu.edu (A. Parnandi), vkarappa1@tamu.edu (V. Karappa), jacqueline.mckechnie@sydney.edu.au (J. McKechnie), kirrie.ballard@sydney.edu.au (K.J. Ballard), rgutier@tamu.edu (R. Gutierrez-Osuna).

http://dx.doi.org/10.1016/j.specom.2015.04.002
0167-6393/

such as academic difficulties (Adhoc Committee on CAS, 2007b). Unfortunately publicly-funded services are often under-resourced. This leads to long wait periods for sessions, which rarely are comprehensive, more often than not are cursory and provide limited interaction with the therapist (Theodoros and Russell, 2008). Private services are expensive, forcing parents to budget the amount of therapy sessions delivered to the child. Children with speech disorders in rural and remote areas or underdeveloped countries may be at a disadvantage because of poor access to speech therapy services, which tend to be concentrated in major cities (Theodoros, 2008). Children with CAS benefit from both phonetic- and linguistic-based treatment approaches (Ballard et al., 2010; Gillon and Moriarty, 2007; Strand et al., 2006). As these children generally require intensive treatment that starts early and continues throughout childhood (Forrest, 2003), their treatment protocol benefits significantly from technology aids. Interactive and automatic speech monitoring tools, which can be used remotely at the child's home, offer a practical, adaptive and cost-effective alternative to face-to-face intervention sessions for children with CAS.

In previous work (Parnandi et al., 2013) we described the system architecture of an automated therapy tool for CAS. The proposed system, "Tabby Talks," consists of (1) a clinician interface where the therapist can create and assign exercises to different children and monitor each child's progress, (2) a tablet-based mobile application which prompts the child with the assigned exercises and records the child speech; and (3) a speech recognition engine running on a server that receives the recorded speech, analyzes it and provides the assessment results to the clinician. This paper describes the speech processing engine within "Tabby Talks," which was designed to identify the three main types of errors commonly associated with CAS: groping errors (delay in sound production), articulation errors (incorrect pronunciation of phones) and prosodic errors (inconsistent lexical stress) (Crary et al., 1984; Nijland et al., 2002; Stackhouse, 1992). The module consists of three components, Voice Activity Detection (VAD), Pronunciation Verification (PV), and Lexical Stress pattern Verification (LSV) (Shahin et al., 2012). VAD uses an energy-based algorithm with a silence threshold to identify non-speech frames at the start of the recording and determine delays in production. The PV algorithm generates a search lattice for each prompted utterance with alternative paths for likely insertion, deletion or substitution errors. A speech recognizer uses the generated lattice for decoding. Finally, the LSV algorithm classifies lexical stress patterns in multisyllabic words into two categories: strong–weak (SW) and weak–strong (WS), and compares them against the expected pattern.

The main contributions in this paper include: (1) the application of automatic speech recognition (ASR) tools to assess errors occurring in pediatric speech sound disorders, (2) a detailed modeling of errors associated with CAS using speech processing modules and algorithms, (3) a generic phoneme level lattice structure for use in identifying pronunciation errors and (4) a speaker independent, multi-syllabic lexical stress classifier.

The remainder of this paper is structured as follows. Section 2 provides background material on childhood apraxia of speech and reviews previous work on speech recognition based speech therapy tools. Section 3 describes the system architecture of 'Tabby Talks'. The speech corpus used is presented in Section 4. Sections 5–7 illustrate the method, experiments and evaluation of the three main components of the system (VAD, PV and LSV). Finally, Section 8 draws conclusions from the study and provides directions for future work.

2. Background

2.1. Childhood apraxia of speech

Developmental communication disorders, including speech sound disorders, are one of the most common reasons for pediatric referrals (Harel et al., 1996). These disorders are difficult to diagnose since they are highly co-morbid, with many children not falling within a single diagnostic cluster (Newbury and Monaco, 2010). Among these disorders, childhood apraxia of speech (CAS), also known as developmental verbal dyspraxia, can lead to a serious communicative disability (Adhoc Committee on CAS, 2007b). Current estimates of children suffering from CAS range from 3.4% to 4.3% in the US (Delaney and Kent, 2004). Starting appropriate intervention at an early age is critical to develop intelligible speech and lay the foundations for the development of language and literacy (Forrest, 2003).

CAS is a neurological disorder that interferes with an individual's ability to correctly pronounce sounds, syllables and words; the area of the brain responsible for sending motor commands is damaged or not fully developed, which affects the planning or specification of movements for accurate speech production. CAS represents a loss in the ability to consistently position and coordinate speech articulators (face, tongue, lips, jaw) and sequence those sounds into syllables or words (Shriberg et al., 1997). In a 2007 position statement (Adhoc Committee on CAS, 2007a), the American Speech Language Hearing Association (ASHA) specified three key behaviors associated with CAS:

(1) inconsistency in production of speech sounds in words across repeated attempts,
(2) difficulty transitioning between sounds and syllables to form a fluently and accurately produced word (articulatory struggle), and
(3) inappropriate prosody (lexical stress patterns) resulting in robotic-like speech, with each syllable produced with equal stress.

2.2. Speech technology tools in disordered voice and speech therapy

To the best of our knowledge, despite the extensive work on ASR, limited work has been reported on the development of speech therapy tools incorporating ASR capabilities for use in pediatric speech sound disorders such as CAS. This can be partly attributed to the fact that ASR still exhibits higher error rates for typically developing children due to variations in vocal tract length, formant frequency, pronunciation and grammar. The limited work with CAS includes an ASR system for the automatic assessment of CAS proposed in Hosom et al. (2004), which was tested on two participants. The system automatically computes two diagnostic markers of suspected CAS: (1) Lexical Stress Ratio, a weighted composite of amplitude area, frequency area, and duration in the stressed compared to the unstressed vowel, and (2) Coefficient of Variation Ratio, the average normalized variability of durations of pause and speech events in conversational speech. There has however been interest in the use of ASR to assist with a range of other difficulties including (1) pathological voice, (2) dysarthria and recently (3) learning and/or hearing difficulties.

Automated systems have been used to detect the presence of pathological voice, specifically vocal fold (i.e., laryngeal) disorders (Arias-Londoño et al., 2010; Fraile et al., 2009; Szaleniec et al., 2007; Wielgat et al., 2008), and assess the intelligibility of subjects with dysglossia and dysphonia to assist in rehabilitation (Maier et al., 2010). In these disorders, the speech production organs are affected, which results in atypicalities in the voice. Thus, the focus has been on automatically detecting voice alterations by averaging local perturbations. Jitter (perturbations in pitch) and shimmer (perturbations in amplitude) are used to assess the severity of voice pathologies (Gelzinis et al., 2008; Manfredi et al., 2000). Good results have been obtained using the differential Teager energy operator (Hansen et al., 1998) and Mel frequency cepstral coefficients (Dibazar et al., 2006; Godino-Llorente and Gomez-Vilda, 2004) as well as time–frequency decompositions (Umapathy et al., 2005) and nonlinear dynamics (Henriquez et al., 2009). The autocorrelation method was found to perform the best in tracking pitch perturbations in different pathological voices as it resulted in the least amount of errors (Seung-Jin et al., 2007).

ASR systems have been developed for dysarthria, a motor speech disorder characterized by weakness, paralysis, or poor coordination of the muscles responsible for producing speech resulting from neurological injury (Morales and Cox, 2009). Speech technology has been implemented to detect the disorder (DiCicco and Patel, 2008), assess speech intelligibility (Falk et al., 2012; Middag et al., 2011) and control assistive technology (Hasegawa-Johnson et al., 2006; Rudzicz, 2010). In DiCicco and Patel (2008), the authors present a system to automatically detect different acoustic landmarks in dysarthric speech to differentiate it from unimpaired speech. Kim et al. (2015) proposes sentence-level features that capture abnormal variation in the prosodic, voice quality and pronunciation aspects of pathological speech to improve intelligibility classification. In DiCicco and Patel (2010), 3 machine learning classifiers were used to recognize the prosodic manipulations in dysarthria to control assistive technology using voice. Ortho-Logo-Paedia (OLP) (Öster et al., 2002) uses ASR to provide immediate word-error-rate feedback and assist in identifying any confusability patterns in its training program to allow severely dysarthric people with motor disabilities to control assistive technology (Green et al., 2003; Hawley et al., 2003).

Due to factors such as low intelligibility, limited phonemic repertoire and high variability in speech within and across patients, conventional speaker adaptation algorithms perform poorly on dysarthric speakers. Baseline word accuracy rates for the ASR used in the STARDUST project dropped from 100% with typically developing speech down to 87% when used with speech from severely dysarthric adults (Green et al., 2003). Similar (if not worse) findings were reported with Vocaliza, a Spanish speech therapy application, that uses ASR to decode user utterances, decide which word sequence has been pronounced, and provide feedback to the user (Rodríguez Dueñas et al., 2008; Saz et al., 2009). When tested on children and young adults with different levels of dysarthria, Vocaliza's word accuracy rate drops from 96.7% for typically developing speech to 66.8% for disordered productions (Saz et al., 2009). The Speech Training, Assessment, and Remediation system (STAR) uses ASR to identify single letter utterances to assist in treating children with articulation problems (Bunnell et al., 2000). Testing this system on Consonant–Vowel–Consonant (CVC) words from dysarthric speakers showed a strong correlation between perceptual and ASR ratings for utterances containing substitution errors, but low correlation for correctly articulated utterances (Van Nuffelen et al., 2009). To improve speech recognition rates, given the difficulties in collecting sufficient amounts of impaired speech recordings, speaker adaptation techniques targeting dysarthric speech are being employed. Sharma and Hasegawa-Johnson (2013) extracts a 'background' model of the dysarthric speaker's general speech characteristics whereas (Rudzicz, 2012) implements acoustic-to-articulatory inversion to estimate positions of the vocal tract in dysarthric speakers to adapt unimpaired speech used in the ASR training.

Learning tools for children are also now using ASR in a similar approach to that used in the speech therapy systems described above. An interactive literacy tutor developed at the University of Colorado uses an ASR to recognize children's speech during oral reading. The tool exploits the reduced language uncertainty in the read-aloud task (i.e., the prompts are known) to improve the language modeling (Hagen et al., 2007). ASR has also found application for use in a teaching system (SPECO) for hearing-impaired children (Vicsi et al., 1999).

3. System description

As illustrated in Fig. 1, Tabby Talks consists of three major components: a fleet of mobile clients running on tablets, a clinician's therapy management interface, and a server running the speech analysis engine. The system (1) prompts the child to complete predetermined exercises on the tablet; (2) records speech productions and uploads them into the server; (3) applies speech analysis algorithms to identify production errors; (4) provides feedback to the child and reports to the therapist detailing the child's progress, and (5) allows the therapist to remotely create or modify exercises individually based on each child's performance. Details of the system implementation have been published elsewhere (Parnandi et al., 2013); in this paper we focus on the speech analysis engine.

The goal of the speech analysis module is to identify errors in the child's speech associated with the key behaviors observed in CAS: articulatory struggle, inconsistent productions and inappropriate prosody; see Table 1. For this purpose, the speech analysis module contains three main components, a Voice Activity Detector (VAD), a Pronunciation Verification (PV) component and a Lexical Stress pattern Verification component (LSV).

Fig. 2 presents a block diagram of the complete speech analysis process. The speech signal is first passed through the VAD module to detect silence segments and then calculate voicing delays and total production time. The PV component compares the speech signal against the prompt to verify the correctness of the production. Only correct pronounced words are passed to the LSV component to assess if the child's stress pattern matches the desired pattern for the prompted word. A report generator collates outputs of all these components and generates a report to be sent back to the therapist.

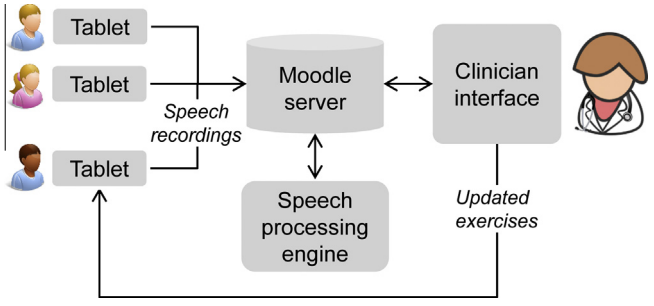

Fig. 1. General overview of the remote speech therapy system showing the server, mobile clients, and remote therapy management system.

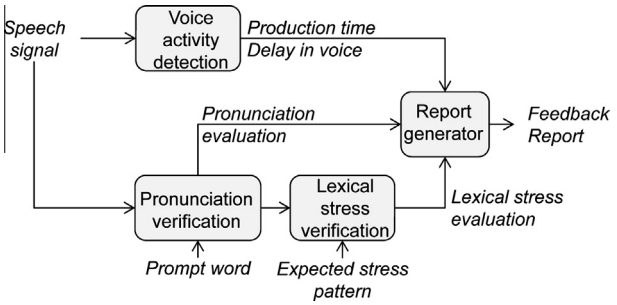

Fig. 2. Description of the speech analysis process and its three main assessment blocks: voice activity detection, pronunciation verification, and lexical stress verification.

4. Speech corpora

We trained and evaluated the algorithms in our speech processing engine using two separate speech datasets. Our training and development corpus (OGI) was collected at the Oregon Graduate Institute of Science & Technology. This marked corpus contains prompted speech for 205 isolated words, 100 general sentences, and 10 digit strings from 1100 typically developing children ranging from ages 4 to 16 years (around 60 h). Two individuals at the Oregon Graduate Institute of Science & Technology subsequently verified each utterance independently during data collection (Shobaki et al., 2000).

Our testing dataset (DOH) was collected from a speech therapy clinic in Doha, Qatar from 4 typically developing children and 2 children with disordered speech. Each child pronounced the 205 isolated words from the OGI corpus; each pronounced word was then marked as correct or incorrect by a speech therapist. The DOH data set was used to evaluate all three speech processing modules.

5. Voice Activity Detection (VAD)

5.1. Method

The goal of this module is to calculate two important measures used in the assessment of CAS: delay in voice, and absence of voice, both of which indicate the presence of "articulatory struggle". The VAD module is used to discriminate between speech and silence segments in the child's production for use in computing both of these measures. Outputs from the VAD module are also used to compute the total speech production time.

A significant amount of work has been done on VAD, including algorithms based on zero crossing rate (Rabiner

Table 1
Summary of errors associated with CAS behavior.

CAS behavior	Observed errors	Speech module component
Articulatory struggle	Delay in onset of sound production, absence of production	Voice activity detector
Inconsistent production	Number of incorrect productions in utterance	Pronunciation verification
Inappropriate prosody	Mismatch in lexical stress pattern	Lexical stress pattern verification

and Sambur, 1975), pitch estimation (Tucker, 1992), autocorrelation (Wu and Wang, 2006) and full-band and sub-band energies (Woo et al., 2000). Most of these techniques focus on identifying speech in the presence of high levels of background noise. In our context, however, the child records speech either in the clinic or at home, so we can assume a high signal–noise ratio. Under these conditions, intensity-based VADs have been found to achieve good results (Kristjansson et al., 2005). We therefore implemented a simple intensity-based VAD algorithm controlled by three parameters: the minimum silence duration, the minimum speech duration and the silence threshold. The minimum silence duration is used to eliminate short silence segments that can naturally occur during speech, e.g. closure duration in the production of plosive phonemes (Crystal and House, 1988). The minimum speech duration is used to eliminate short noise bursts during silence intervals (e.g. microphone noise, door closing, knocking), which would otherwise be misidentified as speech. Finally, the silence threshold is used to perform the silence/speech classification of each segment. The minimum silence duration and minimum speech duration are fixed and have typical values of 0.3 s and 0.1 s, respectively. In contrast, the silence threshold differs for each recording and is estimated individually for each speech file. The silence threshold (ST) is calculated by adding a percentage of the intensity dynamic range (to account for any unexpected increase in background noise) to the minimum intensity value (which represents the minimum silence intensity value) over the speech file. ST is thus given by

$$\mathrm{ST} = P(05) + r \times [P(95) - P(05)] \quad (1)$$

where r is a percentage while $P(05)$ and $P(95)$ are the 5th and 95th percentile intensity values, respectively. The 5th and 95th percentile intensity values were used instead of the minimum and maximum intensity values to reduce sensitivity to outliers. The value of r was determined experimentally using a development dataset.

The VAD algorithm works as follows. First, the speech file is divided into 10 ms non-overlapped frames and the intensity of each frame is calculated to estimate the silence threshold value according to Eq. (1). The intensity of each frame is then compared against the silence threshold. Frames above and below this threshold are marked as speech and silence, respectively. Speech intervals shorter than the minimum speech duration are removed and their neighboring silence intervals merged. In a final step, silence intervals shorter than the minimum silence duration are removed and their neighboring speech intervals joined together. The delay in voicing is given by the time elapsed between the start of the utterance and the start of the first speech interval (Boersma, 2001).

5.2. Speech datasets

10% of the DOH files were randomly selected as a development set to tune the parameter r. The system was then evaluated using the rest 90% of the DOH data set. Each file in both the development and test sets was manually labeled into silence and speech segments to generate ground truth of voicing delay and accurate speech production time.

5.3. Experiments and evaluation

As the VAD is used to assess voicing delay and the total production time in the child's speech, we chose evaluation criteria for the VAD that corresponded to these two measures. The calculated voicing delay and total production time were thus marked as correct if the difference with ground truth was less than 0.1 s to account for human error in perceptually marking the data (Reichardt and Niese, 1970). We first used the OGI development set to determine the value of r in (1) that resulted in the highest accuracy in voicing delay and the total production time. As shown in Fig. 3 the highest accuracy of both the delay in voice and total production time measurements was obtained at r of 0.2. The system was then evaluated with the DOH test set where the system correctly calculated voicing delays in 96.6% of the test files and total production time correctly in 94.8% of the test files.

6. Pronunciation Verification (PV)

6.1. Method

The PV module determines whether the child's utterance matches the prompt given in the therapy exercise and estimates the number of incorrect productions (i.e., phones). Errors made by the child are marked on the phoneme level. There are two main approaches for phone-level pronunciation verification. The first approach, the posterior probability PV, computes a confidence score for each phone in the expected phone sequence and accepts or rejects it according to certain threshold. The other method, the lattice-based PV creates a mispronunciation lattice and/or

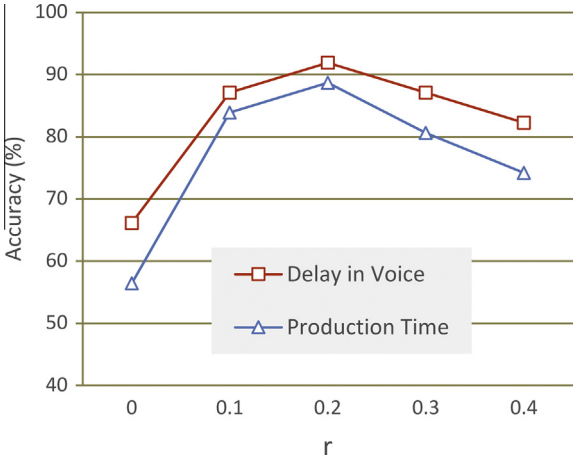

Fig. 3. The accuracy of delay in voice and total production time as a function of r value.

dictionary and aligns the produced speech to the best phoneme sequence in the lattice. Most lattice based PV algorithms need pre-defined mispronunciations rules that depend on the application (Duan et al., 2014; Harrison et al., 2009; Meng et al., 2010). Although this method has been used successfully in different applications especially for second language learning, creating similar rules in speech therapy applications is non-trivial. The development of these rules requires analysis of extensive recordings to develop a list of expected mispronunciations; currently there are limited recordings of the required duration for disorder-specific impaired speech. Hence, the generic posterior probability is still used widely, specifically in measuring quality and mispronunciation detection in impaired speech (Bone et al., 2013; Le and Provost, 2014; Pellegrini et al., 2014). In our algorithm we modify the lattice-based phone-level pronunciation verification to make it generic, thus precluding the need for any predefined rules of the custom language model. For comparison, we also implemented the posterior-based PV algorithm employed in Vocaliza (Rodríguez Dueñas et al., 2008), which computes the posterior probability for each phoneme and compares it against a decision threshold. Details of both algorithms are provided next.

6.1.1. Posterior based PV (PPV)

Several confidence measure techniques are used in ASR systems (Jiang, 2005), with those based on posterior probability being the most widely used for pronunciation verification (Franco and Neumeyer, 1996; Jing et al., 2007; Saz et al., 2009; Witt and Young, 2000). There are different methods to estimate the posterior probability, in Witt and Young (2000) the posterior probability is approximated from the likelihood ratio between the target phoneme obtained from the force alignment and the maximum likelihood phoneme obtained from phoneme loop recognition. For an accurate estimation of the posterior probability we implemented the method proposed by Wessel et al. (2001), which is similar to the method later used in Vocaliza (Yin et al., 2009). Here the posteriors are estimated using a word graph generated at recognition time; in our case, the word graph is replaced with a phone graph χ constructed from a bi-gram phone language model. The phone graph is a directed, acyclic, weighted graph where each arc a_i is defined by the tuple (p_i, s_i, e_i, A_i), p_i is the hypothesized phone attached to a_i, s_i and e_i are the starting and ending time of a_i, respectively, and A_i is the acoustic score calculated from the decoder. Each graph has two nodes to denote the start (SENT-START) and the end (SENT-END) of the utterance. The complete path Q is defined between the SENT-START and SENT-END nodes, and consists of n connected arcs as $Q = \{[p_1]_{s_1}^{e_1}, [p_2]_{s_2}^{e_2}, \ldots, [p_n]_{s_n}^{e_n}\}$. The probability of any complete path Q given the phone graph χ can be calculated as:

$$P(Q|\chi) = \prod_{i=1}^{n} A_i \cdot p(p_i|p_{i-1}) \tag{2}$$

where $p(p_i|p_{i-1})$ is the score from the bi-gram language model. Given the phone graph χ, the posterior probability of any arc a can be calculated as:

$$P(a|\chi) = \frac{\sum_{\forall Q, a \supset Q} P(Q|\chi)}{\sum_{\forall Q} P(Q|\chi)} \tag{3}$$

where the numerator represents the total probability of all complete paths passing through a, and the denominator represents the total probability of all complete paths in the graph.

Fig. 4 illustrates how the confidence measure is used to verify the correct pronunciation of any phone in the prompt word. First, the phone sequence of the prompt word is extracted using the CMU pronunciation dictionary, and then forced alignment is performed on the speech signal. The output is a set of segments, each segment labeled with the phone symbol and the starting and ending time. A bi-gram phone decoder is then used to generate a phone graph. The speech signal is divided into frames of 25 ms with 15 ms overlap and 12 Mel Frequency Cepstral Coefficients (MFCC) plus energy component along with delta and delta-delta features extracted for each frame to produce a 39-dimensional feature vector per frame. The acoustic models are Context Dependent (CD) HMMs consisting of multi-mixture tied-state tri-phones and were used in both the forced alignment and phone decoding.

To compute the confidence score of each phone, the algorithm accumulates the posterior probability of all the arcs in the phone graph with the same phone symbol p that intersect with the median time frame of the given segment. The confidence score is then compared to an empirically-determined decision threshold, and all the phones with a score below the threshold are rejected. If one or more of the utterance phones are rejected, the whole utterance is considered to be incorrect.

Though this method works efficiently for substitution errors, it is unable to detect insertion or deletion errors since the speech signal is aligned to the expected phone sequence of the prompt. As an example, if the child inserts a phone between two correctly pronounced phones, the aligner will assign the inserted phone frames to one of

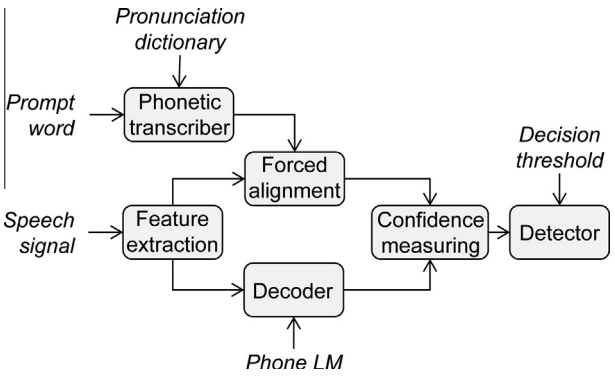

Fig. 4. Block diagram of the posterior-based Pronunciation Verification module (PPV).

the neighboring phones or share it between them; these inserted frames may lower the confidence score of one or both of the neighboring phones below the decision threshold and lead to a false rejection. On the contrary, if the confidence score of these phones is still above the decision threshold, the utterance will result in false acceptance. Likewise, if the child does not pronounce one of the expected phones, the aligner will assign some of the neighboring pronounced phone frames to the deleted phone; this also may affect the confidence score of these phones and cause correctly produced phones to be rejected.

6.1.2. Lattice-based PV (LPV)

To address the limitation of posterior-based PV and detect insertions and deletions of phones, our proposed PV uses a search lattice that considers different competing paths. Pronunciation verification by decoding a generated lattice has been used in some previous works such as in Hafss (Abdou et al., 2006) where a search lattice was generated according to previously determined probable mispronunciation rules to identify errors in Quranic Arabic recitation. In Black et al. (2007), a disfluency detector, a word-level search lattice was created based on a pronunciation variant dictionary constructed by expert linguists that used a garbage model at the start and end of the whole lattice to collect any insertions made before or after the word. *Our proposed system differs from these existing systems due to three enhancements. First, both of these approaches require prior information about the expected pronunciation errors whereas our proposed lattice structure is generic and no prior information is needed. Second, we introduced the garbage model between and parallel to phonemes to detect insertion or substitution errors made by the child at the phone level, instead of adding it at the word level. Lastly, we used penalty values to control the acceptance/rejection rate of the garbage model in the system.*

Our proposed approach is illustrated in Fig. 5. In the first step, the prompted word is phonetically transcribed to obtain the expected phone sequence. Next, the lattice generator uses the phone sequence to generate a search lattice, which is then fed to the ASR engine. The generated lattice is then fed to the speech recognizer together with the extracted feature vector from the utterance. Finally the speech recognizer output is matched against the expected phone sequence to give a decision on each phone and on the whole utterance as well.

The generated lattice is flexible enough to cover all possible pronunciation errors (insertion, deletion and substitution) by adding alternative paths to the correct path for each of the expected errors. The deletion path can be represented by a null arc to allow the recognizer to skip phone nodes during decoding, while the garbage node is used as an alternative to collect phones other than the expected one (substitution errors). A garbage loop is also added between two consecutive phones to collect inserted phones frames. Fig. 6(a) shows an example of the lattice for the word "chair." The terms PG and PD denote the penalties attached to the garbage and deletion arcs, respectively; these penalties are added to prevent the recognizer from skipping phones or aligning speech to the garbage node unless the fit is better than the correct path. The garbage node is composed of all the phones, connected in parallel, as shown in Fig. 6(b).

Here also the 39-dimension MFCC feature vector is extracted for each frame. The feature vector is then fed to the ASR engine along with the created lattice and the acoustic models to generate a sequence of phones. The acoustic models used are typical to the ones used in the PPV method, whereas the garbage model consists of single mixture mono-phones to reduce the complexity and speed up the recognition process. We used a phone-loop instead of an n-gram phone language model trained on correct words as the children with disorder most probably producing inconsistent phone sequence. The output phone sequence is then compared to the expected phone sequence: if matched, the utterance is marked as correct; otherwise it is marked as incorrect. Thus an utterance is marked as incorrect even if one phoneme is recognized as being incorrect.

6.2. Speech datasets

The PV algorithms were trained and developed using 705 speakers of age range 4–10 years (the target age of our application) from the OGI dataset. The PV acoustic models were trained using a training set of correctly-pronounced utterances (isolated words and full sentences) from 670 speakers (around 30 h). A development set of

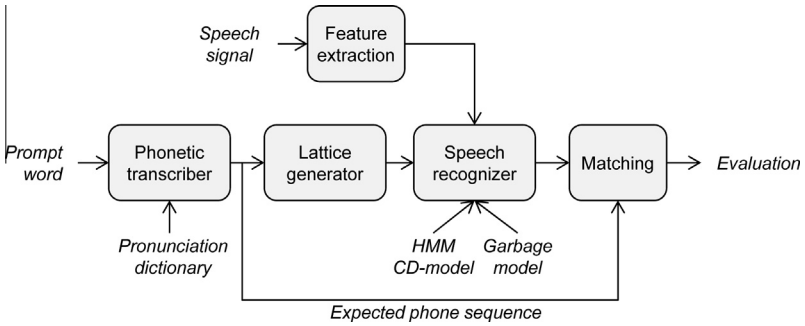

Fig. 5. Block diagram of the Lattice-based Pronunciation Verification algorithm (LPV).

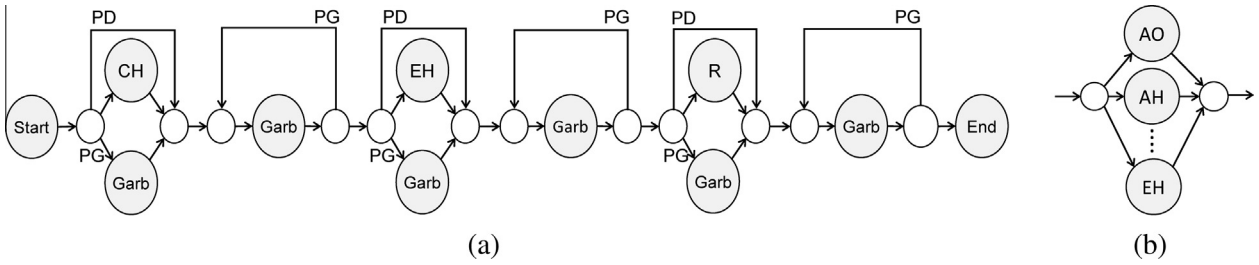

Fig. 6. (a) Example of the search lattice for the word "chair" /CH/ /EH/ /R/, where Garb denotes the garbage node, and PD and PG are the penalties attached to the deletion and garbage arcs consecutively. (b) Construction of the garbage node.

only correctly-pronounced isolated words from 35 different speakers was used to tune and determine the different parameters (the decision threshold used in the PPV model, the insertion and deletion penalties in the LPV model). The system was then evaluated using the DOH test set.

6.3. Experiments and evaluation

We performed three different experiments to compare the two PV algorithms (posterior-based and lattice-based) on a battery of performance measures. Verification labels were categorized into 4 types.

- True Positive (TP): when correctly pronounced phones/utterances were labeled as correct,
- True Negative (TN): when mispronounced phones/utterances were labeled as incorrect,
- False Positive (FP): when mispronounced phones/utterances were labeled as correct, and
- False Negatives (FN): when correctly pronounced phones/utterances were labeled as incorrect.

Using these rates, we computed four different performance measures: the overall classification rate (CR), precision, recall, and the *F*-measure:

$$\mathrm{CR} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{4}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{5}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{6}$$

$$F = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{7}$$

6.3.1. Experiment 1

In the first experiment, we compared the performance of both pronunciation-verification algorithms using the OGI development set. All the words in this dataset had been correctly pronounced, so we generated different error types (insertion/deletion/substitution) by randomly changing 30% of the expected phone sequence produced by the phonetic transcriptor (the front end of both systems). Fig. 7 illustrates the error generation process. We generated errors from the words in the OGI development set and tested the two PV algorithms using only generated substitution errors (SUB) and then using generated substitution, deletion and insertion errors (SDI). Results for both algorithms are summarized in Table 2. The decision threshold for PPV and the penalties PG and PD in the lattice of LPV were selected to balance TP and TN rates in the OGI development set. The influence of these values will be illustrated in the next experiment.

The results in Table 2 show that when only substitution errors are considered, the performance of both algorithms is comparable, with a marginal advantage for PPV. However, when insertion and deletion errors are also considered, the accuracy of PPV decreases significantly for the four performance measures; in contrast, the performance

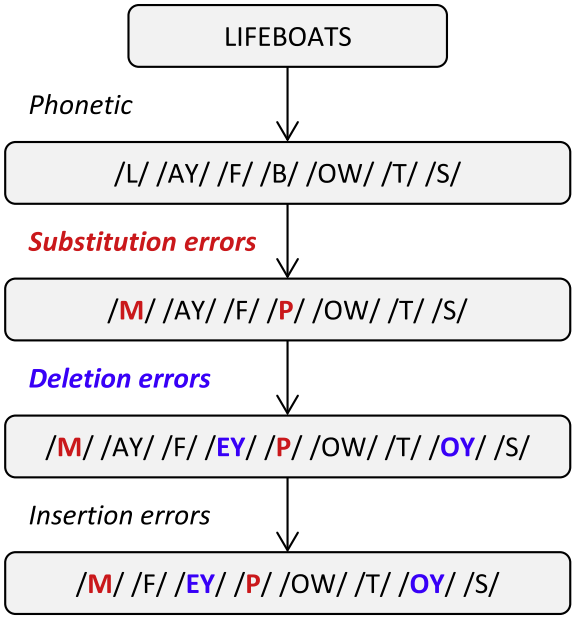

Fig. 7. Example of error generation process for the word 'Lifeboats.' First, we generate a correct (expected) phone sequence through transcription. In a second step, we amend the expected phone sequence by introducing substitution errors with /L/ and /B/ replaced with /M/ and /P/, respectively. Likewise, we insert /EY/ and /OY/ to generate deletion errors; neither /EY/ nor /OY/ were present in the speech production but are now added to the amended expected phone sequence. Finally, we remove the pronounced phone /AY/ from the transcription and so its production is now considered an insertion error (i.e. not expected but pronounced).

Table 2
Performance of LPV and PPV using OGI development set with substitution-only generated errors (SUB) and substitution, deletion and insertion generated errors (SDI).

	SUB (%)		SDI (%)	
	LPV	PPV	LPV	PPV
Classification rate	87.4	89.8	88.2	79.9
Precision	92.1	92.5	92.4	85.2
Recall	92.0	94.7	90.4	85.7
F-measure	92.0	93.6	91.4	85.5

of LPV remains stable when allowing insertion and deletion errors. The superior performance of LPV can be attributed to the presence of alternate arcs in the lattice.

6.3.2. Experiment 2

In our second experiment, we analyzed the effect of the PPV decision threshold and the LPV penalty terms (PG and PD) on system performance, measured at the utterance level instead of the phone level. In a TP utterance, a correct utterance is marked correct if all the pronounced phonemes are evaluated as correct, whereas in a FN utterance, a correct utterance is marked incorrect if one or more of the pronounced phonemes were evaluated as incorrect. Similarly in a TN utterance, an incorrect utterance is marked incorrect if one or more of the pronounced phonemes of utterance were rejected, whereas in a FP utterance, an incorrect utterance is marked correct if all the pronounced phonemes were evaluated as correct. As the whole utterance is rejected even if only one phoneme in it is rejected, it is possible that in a FN utterance, the utterance is marked as incorrect because only one phoneme in the whole utterance is misrecognized as incorrect; this may lead to a higher error on the utterance level than the phone level.

As in this experiment we wanted to study the effect of changing the decision threshold on both the acceptance and rejection rate of the system, we calculated the recall of both TP and TN separately. Varying the TP recall value led to a change in the system's true acceptance rate while varying the TN recall value resulted in a corresponding change in the true rejection rate. The recall of TP was calculated using Eq. (6) while the recall of TN was calculated as:

$$\text{Recall of TN} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{8}$$

We performed the experiment on the OGI development set using two separate scenarios. In the first scenario, each utterance was marked as being correctly pronounced and matched with the corresponding PV output label. Thus all utterances accepted by the PV system resulted in a count of TP and all utterances rejected resulted in a count of FN. In the second scenario, the marked labels of the OGI development set were modified to generate different error types (insertion, deletion and substitution) as explained in Fig. 7 and re-matched with the corresponding PV output label. Thus all the simulated wrong utterances rejected by the PV system provided a count of TN and all accepted utterances a count of FP.

Fig. 8 summarizes the results for the PPV method in terms of recall of TP and TN and CR for different values of the PPV decision threshold. As shown, increasing the decision threshold makes the system more restrictive (phonemes need a high confidence score for acceptance) and leads to a decrease in TP rate and an increase in TN rate.

Fig. 9 summarizes the results for the LPV algorithm as a function of the deletion and garbage penalties PD and PG. Decreasing both penalty terms increases the probability of the garbage and deletion paths in the search lattice and decreases the acceptance rate of the system, thus increasing TP and decreasing TN.

These results illustrate how the decision threshold (PPV method) or the PG and PD parameters (LPV method) can be used by the therapist to determine how strictly the system should labels utterances. As an example, if the system was to be used with a child who had severe speech impairment, the therapist could reduce the threshold to provide positive reinforcement instead of accurate verification. In contrast, in the case of a child who has been performing the exercises for a significant time, the therapist could adjust the threshold to provide accurate verification to track the child's performance over time. This feature can be controlled better in LPV than in PPV as we have 2 parameters which can be adjusted together to reach the desired behavior without significant loss in CR.

6.3.3. Experiment 3

In a third experiment, we evaluated both PV algorithms on the DOH test set. Here too the whole utterance was evaluated and TP, FP, TN and FN labels were identified. The decision threshold of PPV and PG and PD of the LPV are fixed to the values that gave the best balance between TP and TN in experiment 2. Results are summarized in Table 3. These results indicate that LPV outperforms PPV across the four measures.

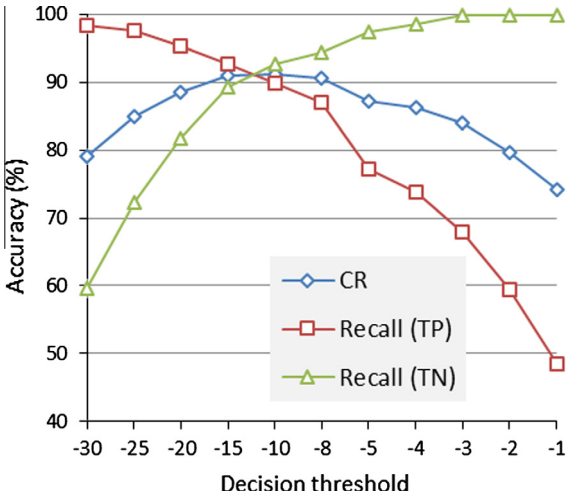

Fig. 8. The effect of changing the decision threshold on the CR and the Recall of both TP and TN in the PPV method.

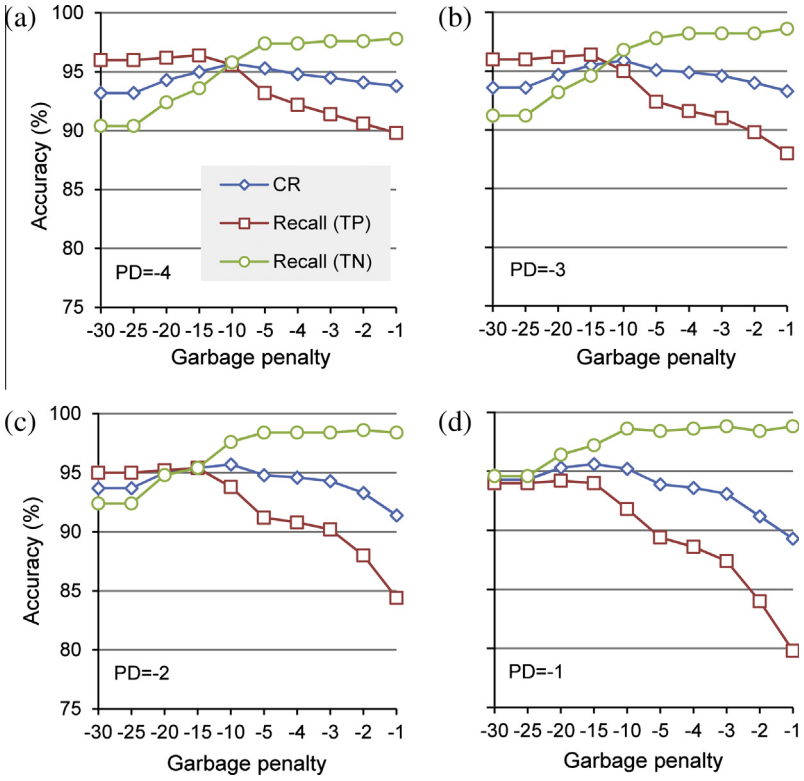

Fig. 9. The effect of changing the garbage penalty (PG) and deletion penalty (PD) on the CR and the recall of both the TP and TN where (a) PD = −4, (b) PD = −3, (c) PD = −2 and (d) PD = −1.

Table 3
Utterance level accuracy of the LPV and PPV using DOH test set.

	LPV (%)	PPV (%)
TC	80.7	76.5
Precision	91.1	89.3
Recall	76.9	70.9
F-measure	83.4	79.1

7. Lexical Stress Verification (LSV)

7.1. Method

The third major component in our architecture is responsible for evaluating lexical stress patterns and comparing them against the correct stress pattern of the prompt. Classifying lexical stress is important in second language learning applications since it can affect intelligibility. Tepperman and Narayanan (2005), identified stressed syllables in the speech of non-native English speakers using prosodic features and features related to the fundamental frequency slope and root mean square (RMS) energy range, resulting in an 87–89% accuracy compared with human-tagged stress labels. Zhao et al. (2011) presented separate support vector machine (SVM) classifiers for each vowel to classify each vowel as either stressed or unstressed; they report an accuracy of around 88% when tested with English speech from Taiwanese speakers. A deep belief network (DBN) used for the assessment of English speech recordings from both Mandarin and Cantonese speakers in Li et al. (2013) had an accuracy of around 80% when classifying each syllable as primary stressed, secondary stressed or no stressed, which increased to around 87% when classifying it as either primary stressed or no stressed. Work done at AT&T Labs with one adult female on classifying different lexical stress patterns resulted in an accuracy of 83.3% for 3-syllable words and 88.7% for 4-syllable words using 20 h of speech (Kim and Beutnagel, 2011). Hosom et al. (2004) validate the automatic measuring of the lexical stress ratio, a weighted composite of amplitude area, frequency area, and duration in the stressed compared to the unstressed vowel, in the production of eight trochaic word forms, but do not use it to detect stress patterns. *Unlike second language learning systems where each syllable is evaluated individually, in our system we evaluate the relative variation in stress patterns between two successive syllables to identify prosodic errors. Moreover our work has been performed using children's speech instead of adult speech which can result in high errors due to variations in vocal tract length, formant frequency and pronunciation.*

In this work only multi-syllabic (two or more) words marked as correctly pronounced by the PV module are passed to the LSV component. Fig. 10 shows the block diagram of the lexical stress verification process; the various time and frequency based measures used to evaluate the stress pattern are summarized in Table 4.

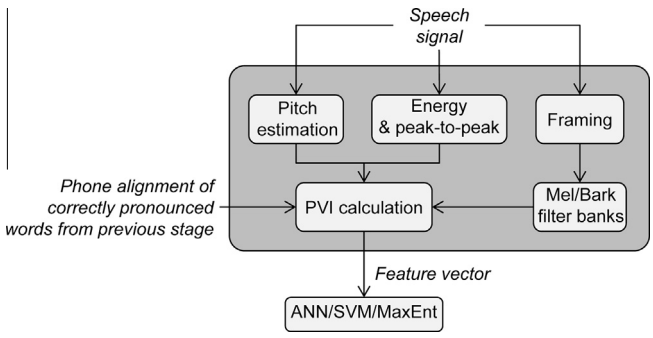

Fig. 10. Block diagram of the lexical stress pattern verification process.

Table 4
Prosodic features used for syllable stress classification (nucleus denotes vowel in syllable).

Feature	Description
f_1	Peak-to-peak amplitude integral over syllable nucleus
f_2	Energy mean over nucleus
f_3	Maximum energy over nucleus
f_4	TEO peak-to-peak amplitude integral over syllable nucleus
f_5	TEO energy mean over nucleus
f_6	Maximum TEO energy over nucleus
f_7	Nucleus duration
f_8	Syllable duration
f_9	Maximum pitch over nucleus
f_{10}	Mean pitch over nucleus
f_{11}	21 Bark-scale filter banks
f_{12}	27 Mel-scale filter banks

Lexical stress is produced through variations in syllable duration, intensity and fundamental frequency (Fletcher, 2010). In any English multisyllabic word the stressed syllable has longer duration as well as higher intensity and pitch values. As in our prior work (Shahin et al., 2012), the speech signal is divided into 10 ms windows, from which energy and pitch features are computed. The energy value is computed by integrating the square of the amplitudes over the frame duration, pitch value is estimated using the auto correlation method (Boersma, 1993), and peak-to-peak amplitude is computed as the difference between maximum and minimum amplitude over the nucleus duration. We also utilized amplitude and energy features from the Teager energy operator (TEO) version of the speech signal (Teager, 1980); the TEO signal has been shown to reduce the effect of noise and tracks rapid energy changes within a glottal cycle. Syllable and nucleus duration measures are obtained from the pronunciation-verification module described earlier.

We also compute the signal energy in different frequency sub-bands on the Mel-scale (Davis and Mermelstein, 1980) and Bark-scale (Zwicker, 1961). These sub-band energies are also computed for each 10 ms frame in the syllable nucleus (vowel), and then averaged over the syllable nucleus. We used 21 Bark-scale and 27 Mel-scale filter banks.

In a final step, and in order to reduce speaker dependencies, we measure the variability between two consecutive syllables in multisyllabic words with unequal stress patterns by computing the pairwise variability index (PVI) (Ling et al., 2000) for each computed prosodic measure. This provides the degree of asymmetry across pairs of neighboring syllables. The PVI for any acoustic feature f_i is given by:

$$\mathrm{PVI}_i = \frac{f_i^{(1)} - f_i^{(2)}}{(f_i^{(1)} + f_i^{(2)})/2} \tag{9}$$

where $f_i^{(1)}, f_i^{(2)}$ are the acoustic features of the first and second of two consecutive syllables. PVIs tend to be positive for words with a SW stress pattern, and negative for words with a WS stress pattern. The computed feature vector is then classified as SW or WS stress pattern using a machine learning algorithm.

7.2. Classifiers

Existing work on stress detection and prosodic labeling has used different types of classifiers, e.g. Support Vector Machines (SVM) (Chen and Wang, 2010; Zhao et al., 2011), decision tree (Deshmukh and Verma, 2009; Xie et al., 2004), Hidden Markov Model (HMM) (Li et al., 2007), Neural Network (Wagner, 2009), Maximum Entropy (MaxEnt) (Rangarajan et al., 2007). In Kim and Beutnagel (2011), the authors implemented 4 different machine learning algorithms to classify different lexical stress patterns for the CAPL system and obtained the highest accuracies with the SVM and MaxEnt classifiers. We therefore implemented the SVM and MaxEnt as they achieved better accuracies in a similar classification problem to ours, and compared them to the standard Neural Network classifier.

Based on these previous works, we used three different classifiers to classify the stress patterns in the words in the database and compared their performance. These included:

(1) A multi-layer perceptron feed forward artificial neural network (MLP) containing input, hidden and output layers. The output layer had 2 neurons, one for the SW class and another for the WS class. The number of neurons in the input layer depended on the size of the feature vector, one for each feature and the hidden layer size was empirically determined.
(2) A two-class support vector machine (SVM) containing a Gaussian kernel.
(3) A Maximum Entropy classifier (MaxEnt) that is based upon the multinomial logistic regression model.

All three classifiers were implemented using inbuilt MATLAB toolboxes. To implement the SVM and ANN we used the MatLab toolboxes svmtrain and NN toolbox. For MaxEnt we used the implementation of Jerod Weinman (Weinman).

7.3. Speech datasets

For the training and development of the LSV component, we used data from children of ages 4 to 16 years in the OGI corpus. Only multi-syllabic words and full sentences manually marked as correct were used. PVI values were computed for the features of consecutive syllables; a two-syllable word resulted in one PVI value/feature, whereas a three- and four-syllable word gave two and three PVI values/features respectively. As we computed a relative measure between each consecutive syllable, the output features became speaker independent. To ensure that the classifiers were not biased toward one of the two classes (SW or WS), we used an equal number of samples for both classes in the training and developing processes. The system performance was then evaluated using the DOH test set. The counts of speakers and stress patterns for the different data sets are summarized in Table 5.

Table 5
The counts of stress patterns in each class for the different data sets.

Data	Speech corpus	NO # speakers	WS	SW
Training	OGI	904	8304	8304
Development	OGI	169	1664	1664
Testing	DOH	6	287	188

Table 6
Components of each feature vector.

Intensity (I)	Mean energy + maximum energy + peak-to-peak amplitude
TEO intensity (I_TEO)	Intensity features derived from TEO version of the speech signal
Duration (D)	Nucleus duration + syllable duration
Pitch (P)	Maximum pitch + mean pitch
Tradition features (T)	Combination of (I_TEO + D + P)
Mel sub-band (Mel)	27 Mel-scale sub-band energies
Bark sub-band (Bark)	21 Bark-scale sub-band energies

7.4. Experiments

The PVI of all extracted acoustic measures listed in Table 4 were calculated for all data sets. The three different classifiers MLP, SVM and MaxEnt were then trained using the computed PVI values of the data in the OGI training set. Despite the extensive work on prosodic labeling, limited work has been performed on classifying lexical stress patterns in speech, making it difficult to compare our results to existing algorithms.

7.4.1. Experiment 1

The accuracy of these three classifiers using various combinations of different feature subsets (as described in Table 6) from the OGI development data set were compared in Fig. 11.

The results show that classifiers using a combination of all the traditional features perform better than those using individual subsets and classifiers using any of the set of sub-band energy features perform better than those using the traditional features alone. The results show that the best overall accuracy (83.3%) is obtained using a 20 hidden units MLP with traditional features and Bark-scale energies as inputs.

Individual SW and WS stress pattern accuracies and overall classification accuracies for the best-performing MLP are shown in Table 7. The system classifies correctly 83.6% of the SW and 83% of the WS lexical stress patterns when using the Bark-scale sub-band energies and traditional feature set.

7.4.2. Experiment 2

We asked two SLPs to independently mark the lexical stress patterns in productions from both typically developing children and children with CAS. They perceptually judged data consisting of 50 multi-syllabic words each

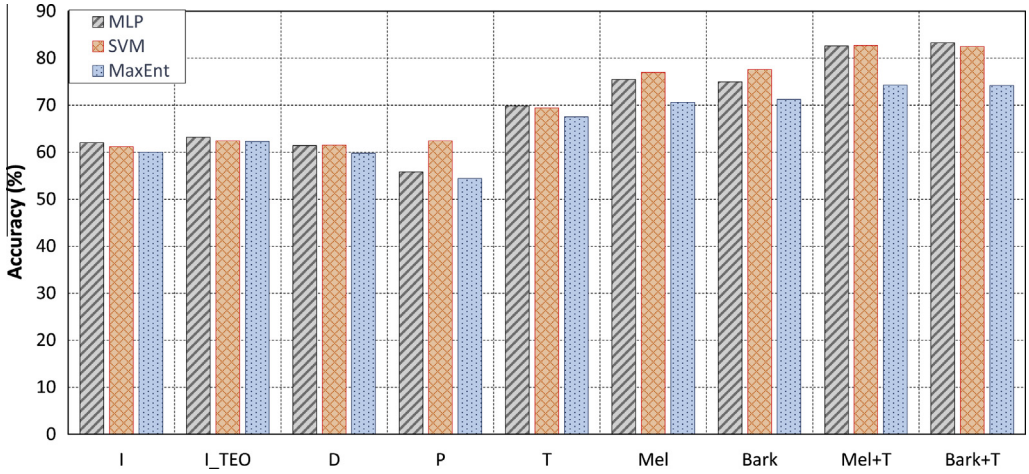

Fig. 11. The classification accuracy of the LSV for different feature sets and different classifiers where I is the intensity features, I_TEO is intensity features of the TEO signal, D is the duration features, P is the pitch features and T is the combinations of the previous traditional feature sets (I_TEO+D+P). Mel, Bark, Mel+T and Bark+T are the Mel-scale and Bark-scale sub-band energies and the combination of the Mel-scale and Bark-scale with the traditional sets of features.

Table 7
Classification accuracy for different feature sets using MLP classifier. The best accuracy is obtained using Bark-scale sub-band energies and traditional features (highlighted in bold with underline).

	SW (%)	WS (%)	Overall (%)
I	73.2	50.4	61.7
I_TEO	74.3	51.1	62.8
D	76	47.4	61.5
P	63	48.5	55.7
T	74.1	65.7	70
Mel	75.2	76	75.6
Bark	75.6	74.4	75
Mel+T	81.5	83	82.2
Bark+T	**83.6**	**83**	**83.3**

collected from 10 typically developing children and 10 children with CAS. The resulting inter-rater reliability was 98% for typically developing children while it dropped down to 82% for children with CAS, indicating the non-triviality of perceptually assessing disordered speech.

7.4.3. Experiment 3

Based on these results (Shahin et al., 2012), we then evaluated the performance of the Bark+T MLP-based LSV on the DOH test set. For this purpose, ground-truth lexical stress patterns (SW or WS) in every word in the test set were manually judged by a speech therapist. The LSV classifier achieves an overall classification accuracy of 77.6%, with 78.3% accuracy for SW patterns and 76.5% for WS patterns. As expected, the disordered nature of the speech in DOH test set led to a decrease in the performance of the LSV, however these results are comparable to the inter-rater reliability scores presented in Section 7.4.2. As our LSV performs a bi-syllabic assessment of stress, it is difficult to compare our results to those obtained with lexical stress detectors used in second language learning system where stress is evaluated for individual syllables.

8. Conclusions

We have presented a speech processing engine for the delivery of automated speech therapy in childhood apraxia of speech. The engine is able to identify voicing delays in children's productions as well as pronunciation and prosodic errors, the main error types associated with CAS. All three components of the engine (VAD, PV and LSV) were tested with a corpus of disordered speech from children with CAS. The developed intensity based VAD correctly determined 94.3% of the voicing delays in the speech production.

Compared to previous pronunciation verification methods, which rely on posterior probability confidence measures, we use a generic search lattice with different alternative paths to allow for detection of insertion, deletion and substitution errors. The use of additional arcs in the search lattice based on expected mispronunciations increased PV accuracy to 88.2% compared to the accuracy of 79.9% obtained with a PV based on posterior confidence measures. This validates the need to use information about disorder specific errors in the design of an accurate PV system for therapy purposes.

A MLP classifier trained on PVI values of intensity, pitch and duration measures and Bark-scale sub-band energies in the LSV component was able to classify both SW and WS errors in the produced stress patterns with an overall accuracy of 83.3% with typically developing children and 77.6% for children with CAS. The resulting accuracy with CAS speech is comparable to our reported perceptual inter-rater reliability of 82%. Using PVI values enabled detection of the subtle variations in the production of adjacent phonemes that result in SW and WS stress patterns. The addition of sub-band energies to the traditional features derived from intensity, pitch and duration used in prosodic assessment led to further improvement in the performance of the LSV indicating that aside from intensity and pitch, stress patterns are also impacted by variations in the frequency structure of the productions.

The proposed engine may be used as part of a comprehensive, speech assessment tool for CAS that can provide automated feedback about the speech produced in a remote, tablet-based therapy system, a capability currently lacking in existing technology-based therapy systems. The tool targets the disorder specific difficulties that children with CAS face to provide therapists with quantitative speech assessment results about the exercises completed by the child at home. This allows therapists to remotely monitor the child's progress and adapt the prescribed therapy regimen to meet the individual needs of the child.

Despite extensive work, ASR systems have found limited application in the area of speech therapy due the difficulties in working with children's speech. Our work shows that by tailoring the system to identify disorder specific errors, it is possible to use ASR techniques to obtain meaningful feedback about speech productions. The task is significantly facilitated due to (1) the structured nature of the speech exercises, (2) restricted domain size and (3) prior knowledge of the expected stress pattern, phone sequence and particular errors produced by children with CAS.

Further work with the speech processing module is ongoing. Our next immediate step is to improve the acoustic HMMs in the PV component by means of discriminative training (i.e., Maximum Mutual Information) and adaptation techniques to adjust the acoustic models to each individual user. Work is underway to trial the speech analysis engine with a large dataset collected with the tablet-based speech therapy tool as part of a clinical study.

Acknowledgements

This work was made possible by NPRP Grant # [4-638-2-236] from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Abdou, S.M., Hamid, S.E., Rashwan, M., Samir, A., Abdel-Hamid, O., Shahin, M., Nazih, W., 2006. Computer Aided Pronunciation Learning System Using Speech Recognition Techniques. Interspeech, Pittsburgh, PA, USA.

Adhoc Committee on CAS, 2007a. Childhood Apraxia of Speech [Position Statement]. American Speech-Language-Hearing Association.

Adhoc Committee on CAS, 2007b. Childhood Apraxia of Speech: Technical Report. American Speech-Language-Hearing Association.

Arias-Londoño, J.D., Godino-Llorente, J.I., Osma-Ruiz, V., Castellanos-Domínguez, G., 2010. An improved method for voice pathology detection by means of a HMM-based feature space transformation. Pattern Recogn. 43, 3100–3112.

Ballard, K.J., Robin, D.A., McCabe, P., McDonald, J., 2010. A treatment for dysprosody in childhood apraxia of speech. J. Speech, Lang., Hear. Res. 53, 1227–1245.

Black, M., Tepperman, J., Lee, S., Price, P., Narayanan, S.S., 2007. Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment. In: INTERSPEECH. ISCA, pp. 206–209.

Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proc. of the Institute of Phonetic Sciences, U. of Amsterdam, pp. 97–110.

Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot Int. 5, 341–345.

Bone, D., Chaspari, T., Gibson, J., Tsiartas, A., Van Segbroeck, M., Li, M., Lee, S., Narayanan, S., 2013. Classifying language-related developmental disorders from speech cues: the promise and the potential confounds. In: INTERSPEECH.

Bunnell, H.T., Yarrington, D.M., Polikoff, J.B., 2000. STAR: Articulation Training for Young Children. Interspeech, Beijing, China, pp. 85–88.

Chen, J.-Y., Wang, L., 2010. Automatic lexical stress detection for Chinese learners' of English. In: Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on. IEEE, pp. 407–411.

Crary, M.A., Landess, S., Towne, R., 1984. Phonological error patterns in developmental verbal dyspraxia. J. Clin. Neuropsychol. 6, 157–170.

Crystal, T.H., House, A.S., 1988. The duration of American-English stop consonants: an overview. J. Phonet. 16, 285–294.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust., Speech Signal Process. 28, 357–366.

Delaney, A.L., Kent, R.D., 2004. Developmental profiles of children diagnosed with apraxia of speech. In: Annual convention of American-Speech-Language-Hearing Association, Philadelphia.

Deshmukh, O.D., Verma, A., 2009. Nucleus-level clustering for word-independent syllable stress classification. Speech Commun. 51, 1224–1233.

Dibazar, A.A., Berger, T.W., Narayanan, S.S., 2006. Pathological voice assessment. In: Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 1, pp. 1669–1673.

DiCicco, T., Patel, R., 2008. Automatic landmark analysis of dysarthric speech. J. Med. Speech-Lang. Pathol. 16, 213–219.

DiCicco, T.M., Patel, R., 2010. Machine classification of prosodic control in dysarthria. J. Med. Speech-Lang. Pathol. 18, 35.

Dodd, B., 2005. Differential Diagnosis and Treatment of Children with Speech Disorder.

Duan, R., Zhang, J., Cao, W., Xie, Y., 2014. A Preliminary Study on ASR-based Detection of Chinese Mispronunciation by Japanese Learners.

Falk, T.H., Chan, W.-Y., Shein, F., 2012. Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. Speech Commun. 54, 622–631.

Fletcher, J., 2010. The prosody of speech: timing and rhythm. In: Hardcastle, W.J., Laver, J., Gibbon, F.E. (Eds.), The Handbook of Phonetic Sciences, second ed. Blackwell Publishing Ltd., Oxford, UK.

Forrest, K., 2003. Diagnostic criteria of developmental apraxia of speech used by clinical speech-language pathologists. Am. J. Speech-Lang. Pathol./Am. Speech-Lang.-Hearing Assoc. 12, 376–380.

Fraile, R., Godino-Llorente, J., Sáenz-Lechón, N., Osma-Ruiz, V., Gómez-Vilda, P., 2009. Automatic detection of laryngeal pathology on sustained vowels using short-term cepstral parameters: analysis of performance and theoretical justification. In: Fred, A., Filipe, J., Gamboa, H. (Eds.), Biomedical Engineering Systems and Technologies. Springer, Berlin Heidelberg, pp. 228–241.

Franco, H., Neumeyer, L., 1996. Automatic scoring of pronunciation quality for language instruction. J. Acoust. Soc. Am. 100.

Gelzinis, A., Verikas, A., Bacauskiene, M., 2008. Automated speech analysis applied to laryngeal disease categorization. Comput. Methods Prog. Biomed. 91, 36–47.

Gillon, G.T., Moriarty, B.C., 2007. Childhood apraxia of speech: children at risk for persistent reading and spelling disorder. Semin. Speech Lang. 28, 48–57.

Godino-Llorente, J.I., Gomez-Vilda, P., 2004. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. IEEE Trans. Bio-med. Eng. 51, 380–384.

Green, P., Carmichael, J., Hatzis, A., Enderby, P., Hawley, M.S., Parker, M., 2003. Automatic Speech Recognition with Sparse Training Data for Dysarthric Speakers. Interspeech, Geneva, Switzerland.

Hagen, A., Pellom, B., Cole, R., 2007. Highly accurate children's speech recognition for interactive reading tutors using subword units. Speech Commun. 49, 861–873.

Hansen, J.H., Gavidia-Ceballos, L., Kaiser, J.F., 1998. A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment. IEEE Trans. Bio-med. Eng. 45, 300–313.

Harel, S., Greenstein, Y., Kramer, U., Yifat, R., Samuel, E., Nevo, Y., Leitner, Y., Kutai, M., Fattal, A., Shinnar, S., 1996. Clinical characteristics of children referred to a child development center for evaluation of speech, language, and communication disorders. Pediatr. Neurol. 15, 305–311.

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

Hasegawa-Johnson, M., Gunderson, J., Penman, A., Huang, T., 2006. HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria. In: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. IEEE, pp. III–III.

Hawley, M., Enderby, P., Green, P., Brownsell, S., Hatzis, A., Parker, M., Carmichael, J., Cunningham, S., O'Neill, P., Palmer, R., 2003. STARDUST – speech training and recognition for dysarthric users of assistive technology. In: 7th European Conference for the Advancement of Assistive Technology in Europe, Dublin, Ireland.

Henriquez, P., Alonso, J.B., Ferrer, M.A., Travieso, C.M., Godino-Llorente, J.I., Diaz-de-Maria, F., 2009. Characterization of healthy and pathological voice through measures based on nonlinear dynamics. IEEE Trans. Audio, Speech, Lang. Process. 17, 1186–1195.

Hosom, J.-P., Shriberg, L., Green, J.R., 2004. Diagnostic assessment of childhood apraxia of speech using automatic speech recognition (ASR) methods. J. Med. Speech-Lang. Pathol. 12, 167.

Jiang, H., 2005. Confidence measures for speech recognition: a survey. Speech Commun. 45, 455–470.

Jing, Z., Chao, H., Chu, M., Soong, F.K., Wei-Ping, Y., 2007. Generalized segment posterior probability for automatic mandarin pronunciation evaluation. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, pp. IV-201–IV-204.

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

Kim, J., Kumar, N., Tsiartas, A., Li, M., Narayanan, S.S., 2015. Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech Lang. 29, 132–144.

Kristjansson, T., Deligne, S., Olsen, P., 2005. Voicing features for robust speech detection. INTERSPEECH 2005, 3.

Le, D., Provost, E.M., 2014. Modeling pronunciation, rhythm, and intonation for automatic assessment of speech quality in aphasia rehabilitation. In: Fifteenth Annual Conference of the International Speech Communication Association.

Li, C., Liu, J., Xia, S., 2007. English sentence stress detection system based on HMM framework. Appl. Math. Comput. 185, 759–768.

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

Ling, L.E., Grabe, E., Nolan, F., 2000. Quantitative characterizations of speech rhythm: syllable-timing in Singapore English. Lang. Speech 43, 377–401.

Maier, A., Haderlein, T., Stelzle, F., Nöth, E., Nkenke, E., Rosanowski, F., Schützenberger, A., Schuster, M., 2010. Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. EURASIP J. Audio Speech Music Process. 2010, 1–7.

Manfredi, C., D'Aniello, M., Bruscaglioni, P., Ismaelli, A., 2000. A comparative analysis of fundamental frequency estimation methods with application to pathological voices. Med. Eng. Phys. 22, 135–147.

Meng, H., Lo, W.-K., Harrison, A.M., Lee, P., Wong, K.-H., Leung, W.-K., Meng, F., 2010. Development of Automatic Speech Recognition and Synthesis Technologies to Support Chinese Learners of English: The CUHK Experience.

Middag, C., Bocklet, T., Martens, J.-P., Nöth, E., 2011. Combining Phonological and Acoustic ASR-Free Features for Pathological Speech Intelligibility Assessment. Interspeech.

Morales, S.O.C., Cox, S.J., 2009. Modelling errors in automatic speech recognition for dysarthric speakers. EURASIP J. Adv. Signal Process 2009, 1–14.

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

Nijland, L., Maassen, B., Van der Meulen, S., Gabreels, F., Kraaimaat, F.W., Schreuder, R., 2002. Coarticulation patterns in children with developmental apraxia of speech. Clin. Linguist. Phonet. 16, 461–483.

Öster, A.-M., House, D., Protopapas, A., Hatzis, A., 2002. Presentation of a New EU Project for Speech Therapy: OLP (Ortho-Logo-Paedia). TMH-QPSR, Fonetik.

Parnandi, A., Karappa, V., Son, Y., Shahin, M., McKechnie, J., Ballard, K., Ahmed, B., Gutierrez-Osuna, R., 2013. Architecture of an automated therapy tool for childhood apraxia of speech. In: Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility. ACM, Bellevue, Washington, pp. 1–8.

Pellegrini, T., Fontan, L., Mauclair, J., Farinas, J., Robert, M., 2014. The goodness of pronunciation algorithm applied to disordered speech. In: Fifteenth Annual Conference of the International Speech Communication Association.

Rabiner, L.R., Sambur, M.R., 1975. An algorithm for determining the endpoints of isolated utterances. Bell Syst. Tech. J. 54, 297–315.

Rangarajan, V., Narayanan, S., Bangalore, S., 2007. Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework. Proc. NAACL HLT, 1–8.

Reichardt, W., Niese, H., 1970. Choice of sound duration and silent intervals for test and comparison signals in the subjective measurement of loudness level. J. Acoust. Soc. Am. 47, 1083–1090.

Rodríguez Dueñas, W., Vaquero, C., Saz, O., Lleida, E., 2008. Speech technology applied to children with speech disorders. 4th Kuala Lumpur International Conference on Biomedical Engineering 2008. Springer, Berlin Heidelberg, pp. 247–250.

Rudzicz, F., 2010. Towards a noisy-channel model of dysarthria in speech recognition. Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies. Association for Computational Linguistics, pp. 80–88.

Rudzicz, F., 2012. Using articulatory likelihoods in the recognition of dysarthric speech. Speech Commun. 54, 430–444.

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 Commun. 51, 948–967.

Seung-Jin, J., Seong-Hee, C., Hyo-Min, K., Hong-Shik, C., Young-Ro, Y., 2007. Evaluation of performance of several established pitch detection algorithms in pathological voices. In: Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE, pp. 620–623.

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

Sharma, H.V., Hasegawa-Johnson, M., 2013. Acoustic model adaptation using in-domain background models for dysarthric speech recognition. Comput. Speech Lang. 27, 1147–1162.

Shobaki, K., Hosom, J.-P., Cole, R.A., 2000. The OGI kids^2 speech corpus and recognizers. In: The International Conference on Spoken Language Processing (ICSLP), Beijing, China, pp. 564–567.

Shriberg, L.D., Aram, D.M., Kwiatkowski, J., 1997. Developmental apraxia of speech: I. Descriptive and theoretical perspectives. J. Speech, Lang., Hearing Res.: JSLHR 40, 273–285.

Stackhouse, J., 1992. Developmental verbal dyspraxia. I: A review and critique. Eur. J. Disorders Commun.: J. College Speech Lang. Therapists, London 27, 19–34.

Strand, E.A., Stoeckel, R., Baas, B., 2006. Treatment of severe childhood apraxia of speech: a treatment efficacy study. J. Med. Speech-Lang. Pathol. 14, 297–307.

Szaleniec, J., Modrzejewski, M., Szaleniec, M., WszoŁek, W., 2007. Application of New Acoustic Parameters in ANN-aided Pathological Speech Diagnosis.

Teager, H., 1980. Some observations on oral air flow during phonation. IEEE Trans. Acoust., Speech Signal Process. 28, 599–601.

Tepperman, J., Narayanan, S., 2005. Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners. In: ICASSP (1). Citeseer, pp. 937–940.

Theodoros, D., 2008. Telerehabilitation for service delivery in speech-language pathology. J. Telemed. Telecare 14, 221–224.

Theodoros, D., Russell, T., 2008. Telerehabilitation: current perspectives. Stud. Health Technol. Inform. 131, 191–209.

Tucker, R., 1992. Voice activity detection using a periodicity measure. IEE Proc. I (Commun., Speech Vision), 377–380.

Umapathy, K., Krishnan, S., Parsa, V., Jamieson, D.G., 2005. Discrimination of pathological voices using a time-frequency approach. IEEE Trans. Bio-med. Eng. 52, 421–430.

Van Nuffelen, G., Middag, C., De Bodt, M., Martens, J.P., 2009. Speech technology-based assessment of phoneme intelligibility in dysarthria. Int. J. Lang. Commun. Disorders/Roy. College Speech Lang. Therapists 44, 716–730.

Vicsi, K., Roach, P., Öster, A.-M., Kacic, Z., Barczikay, P., Sinka, I., 1999. SPECO – a multimedia multilingual teaching and training system for speech handicapped children. In: Sixth European Conference on Speech Communication and Technology, Budapest.

Wagner, A., 2009. Analysis and recognition of accentual patterns. INTERSPEECH, 2427–2430.

Weinman, J., MaxEnt for Matlab.

Wessel, F., Schluter, R., Macherey, K., Ney, H., 2001. Confidence measures for large vocabulary continuous speech recognition. IEEE Trans. Speech Audio Process. 9, 288–298.

Wielgat, R., Zielinski, T.P., Wozniak, T., Grabias, S., Krol, D., 2008. Automatic recognition of pathological phoneme production. Folia Phonia. logo.: Off. Organ Int. Assoc. Logopedics Phoniatrics 60, 323–331.

Witt, S.M., Young, S.J., 2000. Phone-level pronunciation scoring and assessment for interactive language learning. Speech Commun. 30, 95–108.

Woo, K.-H., Yang, T.-Y., Park, K.-J., Lee, C., 2000. Robust voice activity detection algorithm for estimating noise spectrum. Electron. Lett. 36, 180–181.

Wu, B.-F., Wang, K.-C., 2006. Voice activity detection based on auto-correlation function using wavelet transform and teager energy operator. Comput. Linguist. Chin. Lang. Process. 11, 87–100.

Xie, H., Andreae, P., Zhang, M., Warren, P., 2004. Detecting stress in spoken English using decision trees and support vector machines. In: Proceedings of the Second Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation, vol. 32. Australian Computer Society Inc, pp. 145–150.

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

Zhao, J., Yuan, H., Liu, J., Xia, S., 2011. Automatic lexical stress detection using acoustic features for computer assisted language learning. Proc. APSIPA ASC, 247–251.

Zwicker, E., 1961. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). J. Acoust. Soc. Am. 33, 248.