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.