SPEECH PROCESSING MODULES FOR THE AUTOMATIC ASSESSMENT OF CHILDREN WITH APRAXIA OF SPEECH

By
Mostafa Ali Abdallah Shahin

A Thesis Submitted to the
Faculty of Engineering, Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
ELECTRONICS & COMMUNICATIONS ENGINEERING

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2015

SPEECH PROCESSING MODULES FOR THE AUTOMATIC ASSESSMENT OF CHILDREN WITH APRAXIA OF SPEECH

By
Mostafa Ali Abdallah Shahin

A Thesis Submitted to the
Faculty of Engineering, Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
ELECTRONICS & COMMUNICATIONS ENGINEERING

Under the Supervision of

Prof. Dr. Mohsen Abdul Raziq Ali Rashwan

.....

Professor of
Electronics & Communications
Engineering Department
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2015

SPEECH PROCESSING MODULES FOR THE AUTOMATIC ASSESSMENT OF CHILDREN WITH APRAXIA OF SPEECH

By
Mostafa Ali Abdallah Shahin

A Thesis Submitted to the
Faculty of Engineering, Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
ELECTRONICS & COMMUNICATIONS ENGINEERING

Approved by the Examining Committee

_____
Prof. Dr. Mohsen Abdul Raziq Ali Rashwan, Main Supervisor

_____
Prof. Dr. Sherif Mahdy Abdou

_____
Prof. Dr. Mohamed Waleed Talaat Fakhr
(Professor at College of Computing, Arab Academy for Science and Technology)

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2015

Engineer: **Mostafa Ali Abdallah Shahin**
Date of Birth : **02 / 08 / 1981**
Nationality : **Egyptian**
E-mail : **Mostafa_shahin@ieee.org**
Phone. : **01098881457**
Address : **Hadaek El koba, Cairo, Egypt**
Registration Date : **12 / 10 / 2010**
Awarding Date : **/ /**
Degree : **Master of science**
Department : **ELECTRONICS & COMMUNICATIONS ENGINEERING**

Supervisors : **Prof. Dr. Mohsen Abdul Raziq Rashwan**

Examiners : **Prof. Dr. Mohsen Abdul Raziq Rashwan**
Prof. Dr. Sherif Mahdy Abdou, Information Technology, Faculty of Computers and Information System, Cairo University
Prof. Dr. Mohamed Waleed Talaat Fakhr, College of Computing, Arab Academy for Science and Technology

Title of Thesis : **Speech processing modules for the automatic assessment of children with apraxia of speech**

Key Words: **Speech therapy; Automatic speech recognition; Pronunciation verification; Computer aided pronunciation learning; Prosody**

Summary :
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 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.

Acknowledgments

All praise is due to Allah.

First, I would like to express my deepest appreciation to my supervisor Prof. Dr. Mohsen Rashwan whom I owe all my technical knowledge while my work in RDI.

Beside my supervisor I would like to thank Dr. Sherief Mahdy who teaching me a lot in the speech technology filed.

A very special thanks to my colleagues, Alaa Badr, Mamdoh Ragheb and Ali Salah for their help and support.

Grateful to my mother for her prayers for me.

Finally, words are not enough to thank my wife, Hager, for her support and encouragement.

Table of Contents

List of Tables

List of Figures

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 a multi-tier system for remote administration of speech therapy. This thesis 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 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, weak-strong, strong-strong or weak-weak 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 88.7%.

Chapter 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) [1]. 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 [2]. 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 [3]. Early intervention can reduce the negative effects of childhood speech-language disorders such as academic difficulties [2]. 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 [4]. 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 [5]. Children with CAS benefit from both phonetic- and linguistic-based treatment approaches [3, 6, 7]. As these children generally require intensive treatment that starts early and continues throughout childhood [8], 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.

The proposed system, 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 thesis describes the speech processing engine within the system 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) [9-11]. The module consists of three components [12], Voice Activity Detection (VAD), Pronunciation Verification (PV) [13], and Lexical Stress pattern Verification (LSV) [14]. 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 bisyllabic lexical stress patterns in multisyllabic words into: strong-weak (SW), weak-strong (WS), strong-strong (SS) or weak-weak (WW) and compares them against the expected pattern.

The main contributions in this work 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 phoneme level lattice structure for use in identifying pronunciation errors and 4) a bisyllabic lexical stress pattern classifier.

Outline of the thesis

The thesis is structured, in nine chapters, as follows:

Chapter 1: This Introduction chapter which provided a brief description of the whole work.

Chapter 2 gives an overview of the Childhood Apraxia of Speech disorder and the assessments and treatment ways followed by some related work of using the Automatic Speech Recognition (ASR) in the automatic assessment/treatment.

A detailed description of two different approaches of ASR can be found in **chapter 3**. Section 3.1 illustrated the conventional GMM-HMM system while the hybrid DNN-HMM is illustrated in section 3.2.

Chapter 4 describes the design of the system tool, the speech processing modules and the speech corpora used in the training and evaluation of the whole system.

The method, experiments and results of the Voice Activity Detector (VAD) are illustrated in **chapter 5**.

Chapters 6 and 7 provide two different implementation of the Pronunciation Verification (PV) module. The first one used to evaluate each phoneme as correct or incorrect while the second one used to specify the produced phoneme.

The method used in verifying the lexical stress in the produced speech is presented in **chapter 8.**

Finally, the thesis closes with **Chapter 9**, which explicitly delineates the contributions of this research and outlines some directions for future research.

Chapter 2 : Background and Related Work

2.1. Childhood Araxia of Speech (CAS)

Developmental communication disorders, including speech sound disorders, are one of the most common reasons for pediatric referrals [15]. These disorders are difficult to diagnose since they are highly co-morbid, with many children not falling within a single diagnostic cluster [16]. Among these disorders, childhood apraxia of speech (CAS), also known as developmental verbal dyspraxia, can lead to a serious communicative disability [17]. Current estimates of children suffering from CAS range from 3.4% - 4.3% in the US [18]. Starting appropriate intervention at an early age is critical to develop intelligible speech and lay the foundations for the development of language and literacy [8].

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 [19]. In a 2007 position statement [17], 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.

Prior to language acquisition, children learn to control their own speech production with skills such as breathing, control of tone and intensity and vocalization. Developmental disabilities and neuromuscular disorders can delay the acquisition of these abilities making the children unable to start articulating the first sounds and words. Speech therapists use game-like activities to train children in these skills. Once these skills are developed, the therapist focuses on acquiring the necessary phonetic system needed to speak a language. This is done by repeating different sounds and words and evaluating the correctness of the pronunciation, training the child's phonological abilities. Once the child can articulate sounds and words, the child's phonological level of the language needs to improve to enable using the language to interact with the world; this is commonly done through speech therapy using activities with images and text involving dialog with the educator. Children with CAS benefit from both phonetic- and linguistic-based treatment approaches. **Automatic monitoring tools can thus provide an effective, adaptive and practical tool for children with CAS**. **They can also provide an objective determination of the impairment based on the value of extracted acoustic parameters.** The analysis of the disordered speech signal is usually performed by the extraction of acoustic parameters using digital signal processing techniques.

2.2. Treatment and Assessment Approach of CAS

Treatment and assessment for children with CAS is performed using a 'bottom up' approach in the Nuffield Centre Dyspraxia Program (NDP3) system, which is a complete resource - from pre-speech to connected speech - and includes a range of enhanced and unique materials for assessment and therapy [20], as shown in Figure 2.1 [21]. Initially the activities focus on the child developing a sound system, initially including all vowels and a basic range of consonants but eventually a full range of sounds. As speech is organized into syllables, with consonants and vowels being their basic building blocks, the next progression is to sequences of a single consonant (C) followed by a single vowel (V), i.e. CV-syllables such as "pa" and a single vowel (V) followed by a single consonant (C), i.e. VC-syllables such as "up".

In the NDP3 program, speech skills are conceptualized as a "brick wall", with pre-speech oro-motor skills and single consonant and vowel sounds seen as the foundations, and word level skills built up in layers of bricks on top of the foundations. Simple CV and VC syllables are the first layer, moving up in layers through CVCV (e.g. puppy), CVC (e.g. cup), CVCVC (e.g. button) and multi-syllabic words (e.g. caterpillar), clusters (e.g. star), word combinations of phrases and sentences and finally reaching the top layer of connected speech. The approach provides a multi-layered, multi-target treatment approach involving working on several layers at the same time (e.g. oro-motor skills, single sounds and CV word level), but with different speech sounds targeted at each level. The approach works from the child's strengths (what s/he can already do) and builds skills in small achievable steps from this point. It is a cumulative approach in that practice exercises from the early stages continue to be included in the therapy program as the child moves up the layers.

The assessment procedure for NDP3 is also based on a multi-layer therapy strategy, and relies on the production of:

• all the single consonants and vowels

• a set of 20 single words at each phonotactics structure (CV/VC, CVCV, CVC, CCV and multisyllabics) through picture naming, and

• phrases and sentences through imitation with pictures

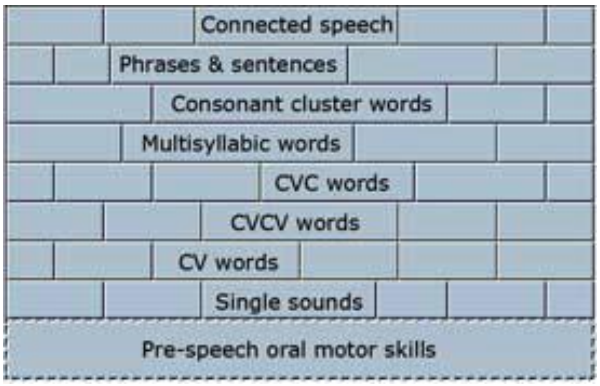

Figure 2.1: The brick wall: NDP3 focuses on establishing (as fully as possible) a set of motor programs at each level of the wall, as well as supporting the development of a full range of psycholinguistic processing skills [21].