The Automatic Detection of Speech Disorders in Children: Challenges, Opportunities, and Preliminary Results

Mostafa Shahin, *Member, IEEE*, Usman Zafar, and Beena Ahmed, *Member, IEEE*

***Abstract*—Given the limited accessibility to Speech and Language Pathologists (SLPs) children in need often have, pediatric Computer-Aided Speech Therapy (CAST) tools can play an important role in the early diagnosis and treatment of speech disorders. However, various challenges impede the implementation of accurate automated analysis of speech disorders in children. In this article, we first discuss three key challenges in processing child disordered speech: 1) the unreliability of low-level annotation and scarcity of speech corpora, 2) speaker diarization of therapy sessions and 3) inaccurate children's acoustic models. We next explore opportunities to overcome some of these challenges. First, we investigate the effectiveness of high-level paralinguistic features in disordered speech detection to reduce the dependency on annotated data. A binary classifier trained using paralinguistic features extracted from both typically developing children and those suffering from Speech Sound Disorders (SSD) achieved 87% subject-level classification accuracy. Second, we tackle the speech disorder detection problem as an anomaly detection problem where models are trained merely on typically developing speech, reducing the need for disordered training data. A phoneme-level *F*1 score of 0.77 was obtained from an anomaly detection-based system trained on speech attribute features to classify between typical and atypical phoneme pronunciations of children with speech disorder. Finally, we test the efficiency of an x-vector based speaker diarization technique in pediatric therapy sessions. The method successfully distinguished between therapist and child speech with a Diarization Error Rate (DER) of 10%.**

***Index Terms*—Anomaly detection, automatic assessment, paralinguistic features, speech disorder, speaker diarization.**

I. Introduction

SPEECH disorders in children affect their fluency and intelligibility. Delay in their diagnosis and treatment increases the risk of social impairment and learning disabilities [1]. With the dramatic shortage of Speech and Language Pathologists (SLPs), there is an increasing interest in Computer-Aided Speech Therapy (CAST).

There are a number of CAST tools that specifically target children with speech disorders. These include Speech Training, Assessment, and Remediation (STAR) [2], Vocaliza [3], Speech Assessment and Interactive Learning System (SAILS) [4], Phoneme Factory Sound Sorter (PFSS) [5], Tabby Talks [6], [7], Apraxiaville [8], ArtikPix [9] and Apraxia World [10]. However most of these tools focus on the automated delivery of the speech therapy stimuli, leaving therapy to be guided directly by the child [3] or the SLP [7]. Only STAR, Vocaliza and Tabby Talks utilize speech analysis to provide automated feedback and help guide the therapy exercises [11].

Of these, STAR [2] utilized phoneme-level Automatic Speech Recognition (ASR) to detect phoneme substitution errors in children with articulation disorder. While in Vocaliza [3], the ASR recognized word was compared with the expected one to accept or reject the productions of children with speech dysarthria. Tabby Talks [7] proposed a client-server platform to facilitate speech therapy for children with Childhood Apraxia of Speech (CAS). The system allowed the SLP to assign exercises to children and follow-up on their progress. Three speech processing modules were implemented to detect phonological, prosodic and groping pronunciation errors in child speech and then generate an automatic progress report to the SLP.

However, the accuracy of all of these automated disordered speech analysis tools is still not reliable enough to be used clinically. The wide variety of speech disorder types and severity levels coupled with the lack of sufficiently sized disordered speech corpora pose a great challenge towards their implementation. Moreover, child speech suffers from higher inter- and intra-speaker variability than adult speech making it more difficult to handle.

There is some recent research on the automatic detection of the different types of children speech disorders including stuttering [12]–[14], speech sound disorders [15]–[18], hypernasality [19]–[23] and dysarthria. However, due to the unavailability of public child disordered speech corpora, most of these works report results on their own proprietary datasets which makes it hard to compare the different techniques. Recently, Eshky *et al.* introduced UltraSuite [24], a freely available ultrasound and acoustic dataset of complete therapy sessions from children with Speech Sound Disorders (SSD).

Manuscript received May 16, 2019; revised October 23, 2019; accepted November 25, 2019. Date of publication December 12, 2019; date of current version April 8, 2020. This work was supported by NPRP under Grant #[8-293-2-124] from the Qatar National Research Fund (a member of Qatar Foundation). The guest editor coordinating the review of this paper and approving it for publication was Dr. Claudia Manfredi. *(Corresponding author: Mostafa Shahin.)*

M. Shahin is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: mostafa_shahin@ieee.org).

U. Zafar is with the Electrical and Computer Engineering Program, Texas A&M University at Qatar, Doha 23874, Qatar (e-mail: usman.zafar@qatar.tamu.edu).

B. Ahmed was with the Electrical and Computer Engineering Program, Texas A&M University at Qatar, Doha 23874, Qatar. She is now with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: beena.ahmed@unsw.edu.au).

Digital Object Identifier 10.1109/JSTSP.2019.2959393

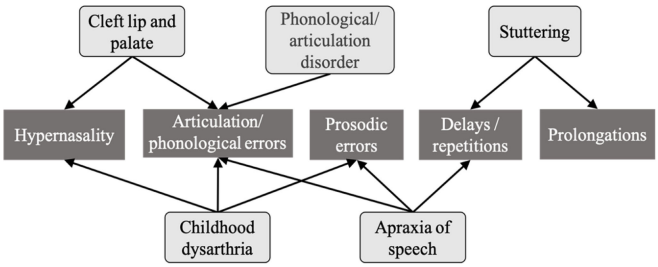


Fig. 1. Children's speech sound disorders and their associated pronunciation errors.

In this paper, we highlight three major challenges in developing a reliable tool to automatically detect childhood SSDs: namely 1) the lack of sufficiently sized disordered speech corpora, 2) the need for effective speaker diarization systems to extract child segments from recordings of speech therapy sessions and 3) limitations in the automatic speech recognition of children's speech. We then detail three proposed methods to address some of these challenges along with preliminary results. First, we design a system to detect children with speech sound disorders using segment-level paralinguistic features extracted from child speech during therapy sessions. By using high-level paralinguistic features, we overcome the need for accurate low-level annotation, e.g. phonetic annotation, of speech corpora, which suffers from high level of unreliability. Second, we utilize an anomaly detection approach to perform phoneme-level discrimination between correctly pronounced and mispronounced phonemes from children with CAS. The motivation behind using anomaly detection is that it models the typical pronunciation using only a correctly pronounced dataset with no need for a disordered speech dataset, which is hard to collect and annotate. Finally, due to the lack of research in the speaker diarization of pediatric therapy sessions, we test the effectiveness of a state-of-the-art deep learning-based speaker diarization approach on recordings of therapy sessions of children with SSD.

The rest of the paper is structured as follows. A brief background of SSD types and review of existing work on the automatic detection of pronunciation errors is given in Section I. Current challenges are discussed in Section III. Section IV details methods used in our preliminary work with results and conclusions drawn in Section V and Section VI respectively.

II. Background

A. Speech Sound Disorders

The mouth, jaw, tongue, lips, palates and other articulators all work together to produce human speech. Failure in moving these parts appropriately affects speech intelligibility leading to speech sound disorders (SSDs). There are various causes of SSDs including weakness in the oral musculature, such as dysarthria, structural based disorders such as cleft lip and palate and the inability to properly interpret and execute messages from the brain due to a neurological disorder such as apraxia. Some of the most common speech sound disorders affecting children with their speech characteristics and symptoms are given below and summarized in Fig. 1

1) Childhood Apraxia of Speech: Childhood apraxia of speech (CAS) is a neurological speech disorder that affects a child's ability to orchestrate the movement of the speech production muscles required to pronounce the desired sound without any weakness in the muscles themselves. CAS, also known as developmental verbal dyspraxia, can cause serious communication problems, if not detected and treated in its early stage [25].

Children with CAS have various pronunciation difficulties including: 1) inconsistency in pronouncing words, i.e. the same word is pronounced in different ways, 2) inappropriate prosody, i.e. producing robotic-like speech, and 3) articulatory struggle, i.e. moving articulators several times before starting to speak [26]. These difficulties lead to the following pronunciation errors: 1) phonological errors where the child tends to replace a phoneme with another phoneme, omit phoneme completely or insert unexpected phonemes, 2) prosodic errors where the child places stress on the wrong syllable in multisyllabic words and 3) groping errors where the child delays onset of sound production (Fig. 1).

2) Childhood Dysarthria: Dysarthria is a motor speech disorder that occurs due to a weakness of the muscles responsible for speech production [27]. Dysarthria is associated with diseases that cause brain damage such as brain injury, cerebral palsy, amyotrophic lateral sclerosis, etc.

Dysarthria affects the prosodic features of the speech such as rate, loudness and rhythm (Fig. 1). Children with dysarthria speak either too slow or too fast with uneven loudness and rhythm which leads to inconsistent prosody and pronunciation errors. Mild hypernasality and/or hoarse voice are also common in dysarthric speech [28].

3) Cleft Lip and Palate: Children born with cleft lip and/or palate (CLP) have difficulty in pronouncing some sounds properly due to a split in one or both sides of the upper lip (i.e. cleft lip) or a cut in the front part (hard) or back part (soft) of the palate (i.e. cleft palate). Although, in most cases, the cleft lip and/or palate can be clearly seen by doctors at birth, its effect on speech will not be known till the child starts speaking. One of each two children with CLP will need to have a speech therapy [29].

Hypernasality is the most common speech problem in children with CLP caused when the soft palate fails to block air from passing through the nose (Fig. 1). CLP may also cause articulation errors in consonants such as /s/, /z/, /sh/, in addition to distortion in stop consonants such as /p/, /b/, /t/ [22].

4) Stuttering: Stuttering is characterized by three main behaviors: 1) blocks, when the child makes a long pause before producing sound, 2) prolongations, when a child stretches out some sounds and 3) repetitions, when a child continuously repeats part of a word or sentence [30], [31] (Fig. 1).

The symptoms start at preschool age (2–4 years old) when the child learns how to form a sentence. The cause of stuttering is unknown and most likely inherited from a family member [31].

5) Articulation and Phonological Disorders: Articulation and phonological disorders are the most prevalent types of SSDs among preschool children. In some cases, they are associated with other types of speech disorders with known causes such

as motor speech disorder (dysarthria and apraxia) or structural speech disorders (cleft lip and palate) as discussed earlier, however, most articulation and phonological disorders have no known cause [32].

Articulation disorders affect the ability of a child to correctly pronounce sounds, for example a child may replace /s/ with /th/ (say /thith/ instead of /this/) or /r/ with /w/ (say /wabbit/ instead of /rabbit/). On the other hand, phonological disorders are more about sound patterns where the child is able to produce the sound but uses it inappropriately, for example says /doe/ instead of /go/, or /tat/ instead of /cat/. Children with phonological disorders also make reduction errors, where a child reduces two consecutive consonants into one, for example dropping /s/ or /p/ in /spoon/ [32].

There are four main pronunciation errors associated with articulation and phonological disorders [15]: 1) phoneme replacement, i.e. replacing certain phoneme with another, 2) phoneme distortion, i.e. altering acoustic characteristics of one or more phonemes, 3) phoneme deletion and 4) phoneme insertion (Fig. 1).

B. Automatic Detection of Speech Sound Disorders

As depicted in Fig. 1, childhood SSDs cause five main pronunciation errors. In this section we will discuss the existing work on the automatic detection of these pronunciation errors. As the automatic detection of stuttering has been summarized in some recent publications [14], [33], it has been excluded from this review. We will therefore focus on articulation and phonological errors, hypernasality and prosodic errors.

1) Phonological and Articulation Errors: Phoneme-level pronunciation errors are characterized by the insertion, deletion, substitution or distortion of phonemes. As demonstrated in the previous section, these pronunciation errors are common in most types of SSDs. As these errors are also dominant in second language learners, there is significant research work addressing their automatic detection for Computer Aided Language Learning (CALL) systems [34]–[37]. However, there has been limited research on the detection of these errors in child disordered speech.

One of the most widely used methods in detecting phonetic errors is the Extended Recognition Network (ERN) approach [34], [38]. ERN is a search lattice created to cover all possible mispronunciation errors expected in the specific domain, defined using a priori expert knowledge of expected mispronunciations. In our previous work, we designed an ERN covering pronunciation errors expected by children with CAS. We further compared two acoustic models used to decode the ERN, a Gaussian Mixture Model-Hidden Markov Model (GMM-HMM)-based and a Deep Neural Network (DNN-HMM)-based [39].

Ward *et al.* [40] constructed a lattice for the most common error patterns of children with different types of SSDs and decoded it with Hierarchical Neural Network (HNN) acoustic model. Later they improved their acoustic model using transfer learning to incorporate out-of-domain adult speech corpus [16].

Dudy *et al.* [17] incorporated an ERN in the computation of the Goodness of Pronunciation (GOP) derived from the mispronunciations of children with articulation disorders and autism. The computed GOP was then fed to a Support Vector Machine (SVM) classifier to classify each phoneme as correctly pronounced or mispronounced.

A speaker verification-based method using i-vector features was also applied to discriminate between typically developing (TD) and SSD speakers [41]–[43]. In [41], the authors trained a GMM model for TD children and another for children with SSD using Mel Frequency Cepstral Coefficients (MFCC) acoustic features. An i-vector representation of each group was then derived from their GMM models and used to discriminate between the TD and SSD groups. A similar approach was adopted in [43] to detect mispronunciations of a single Arabic phoneme at different positions in the pronounced word.

In children with CLP, a common articulation error known as Glottal Activity Error (GAE), is the distortion of unvoiced and voiced stops due to the absence of glottal vibrations in the latter and existence in the former. In a method to detect GAE proposed by Vikram et al [22], the speech was first passed through a glottal activity detector to detect regions where the glottal vibration exists, followed by two detectors for syllable vowels and voiced consonants respectively. A bandpass filter was then used to eliminate the nasalization effect on the vowel and make it more discriminable from consonants. The method reported an average detection accuracy of around 89% when compared to manual annotation of Kannada children with CLP.

2) Hypernasality: The automatic detection of hypernasality can be performed on the frame-level [44] or phoneme level [20], [45]. Most hypernasality detectors perform binary classification to determine the existence or absence of hypernasality [20], [45]–[47].

In [44], Maier *et al.* proposed a multi-level classification system to detect different types of pronunciation errors in children with CLP including hypernasality. They used frame-level MFCC features and phoneme-level Teager Energy Operator (TEO) features along with other word and speaker level features measuring pronunciation fluency. The best performance of 63% phoneme-level Unweighted Average Recall (UAR) was obtained with a combination of MFCCs and TEO features. In [45], measures of voice quality including jitter, shimmer and harmonic-to-noise ratio were utilized along with MFCCs to perform a binary classification of Spanish children vowel production as normal or hypernasal.

The detection of the degree of hypernasality is also important to diagnose the severity of CLP. In [48] a four-way k-Nearest Neighbor (kNN) classifier was proposed to classify hypernasal speech as normal, moderate, mild or severe. The classifier was trained with different types of acoustic features including pitch, amplified energy and short-time energy in frequency sub-bands and MFCC. The system was applied on 120 (male and female) Chinese children balanced over gender and hypernasality level. The author showed that MFCCs achieved the overall best performance of around 80% to detect the level of hypernasality.

In [49] authors proposed a system to estimate a continuous measure representing the degree of hypernasality in child speech. The system was composed of two modules, a glottal

activity detector to detect the regions with glottal activity, and a DNN classifier trained on MFCC features extracted from glottal activity regions to classify between oral and nasal speech. The posterior probability of the nasal class was used as a hypernasality score. The system achieved a 0.82 correlation when compared with the measures obtained from a nasometer and around 93% binary classification accuracy.

MFCCs shows the best performance in most of the existing work for hypernasality detection. However, Dubey *et al.* [20] criticized the use of MFCCs in high-pitched speech signals, like child speech, due to their inability to filter out the harmonic effects in low frequencies where nasality evidence appears. Instead, the authors employed the Pitch Adaptive MFCC (PAMFCC) features by applying pitch adaptive liftering to remove the harmonics effect. These features were first proposed in [50] to improve performance of child speech recognition systems. The PAMFCCs improved the accuracy of SVM-based hypernasality detection by around 6% absolute compared to MFCCs when applied on /a/, /i/ and /u/ vowel productions of Kannada children with CLP.

3) Prosodic Errors: Inappropriate lexical stress production is a common pronunciation problem in children with apraxia or dysarthria. The automatic detection of these errors requires a system to automatically measure the level of syllabic stress. Ferrer *et al.* [51] proposed a method for lexical stress classification as a part of EduSpeak CALL for English; a GMM model trained using both prosodic and spectral features was employed to classify each syllable as unstressed, primary or secondary. Chen *et al.* [52] utilized differential features computed from the vowels of consecutive syllables in a SVM classifier to detect the lexical stress pattern in the speech productions of Chinese learners of English. A phoneme-specific SVM model to classify each vowel as stressed or unstressed was proposed in [53] and applied on non-native English speakers. HMM [54], Maximum Entropy (MaxEnt) [55] and Deep Belief Network (DBN) [56] have also been employed to perform lexical stress classification. In our previous work, we proposed a Convolutional Neural Network (CNN) based classifier to discriminate between primary stressed, secondary stressed and unstressed syllables and applied it to English and Arabic speech [57].

In the context of child disordered speech, we trained a DNN model using both differential and raw acoustic features to detect lexical stress errors in children with CAS [58]. To the best of our knowledge, aside from this, no further work has been done on lexical stress error detection in child disordered speech.

III. Challenges

A. Availability of Speech Corpora

One major factor constraining work on the automatic speech analysis of disordered speech is the limited availability of public disordered speech corpora, specifically for children. The collection of such data is not only costly and time-consuming, but additionally due to the large variability in acoustic characteristics across different ages, dialects, types and severity of speech disorders, it is hard to design a generic dataset. Consequently, most research groups working in this area design, collect and annotate their own tailored speech corpora. The absence of baseline results makes it difficult to compare between different algorithms presented in the literature.

TABLE I
List of Child Speech Corpora of Speech Sound Disorder (SSD), Phonological Disorder (PD), and Typically Developing (TD)

Dataset	Type of Disorders	N# Participants	Age range (years)
Cummings Corpus [60]	SSD	30	3 – 6
Torrington Eaton [61]	SSD	51	4 – 6
Chiat Corpus [62]	PD	3	5
Granada Corpus [63]	PD	19	3 – 5
Preston Corpus [64]	SSD	44	4 – 5
UltraSuite [24]	SSD	86	5 – 13
PF-STAR [65]	TD	158	4 – 14
CMU kids [66]	TD	76	6 – 11
OGI kids [67]	TD	1100	4 – 11
CU kids [68]	TD	780	4 – 10

Table I lists the publicly available speech corpora of child disordered speech in addition to some widely used speech corpora from typically developing children.

An additional issue is that the low-level manual annotation of disordered speech, such as phonetic transcription, is unreliable. Shriberg *et al.* [59] showed that the intra- and inter-rater agreement of phonetic transcription task for children with motor speech disorders reaches as low as 70% and 80% for vowels and consonant respectively. An accurate ground-truth transcription of speech corpora is crucial for the training, validation and testing of supervised models.

A number of approaches have been proposed, in the context of child disordered speech, to deal with the shortage of speech corpora and unreliability of low-level annotation including incorporating out-of-domain adult speech [16], extracting features on speaker-level such as i-vectors [41]–[43] and using robust acoustic features such as pause events [69].

Here, we propose two approaches to cope with this issue. First, we propose using paralinguistic features, which can be extracted on segment or speaker level, to discriminate between children with TD and SSD speech. Second, we propose leveraging upon anomaly detection techniques; this allows models to be built using typical speech only, with disordered speech detected by treating it as an anomaly. In Sections IV and V, we provide a detailed discussion of methods used and preliminary results for each of these approaches.

B. Speaker Diarization

Most of the available speech corpora collected from children with speech disorders are recordings of complete therapy sessions which are part of the subjective diagnosis or treatment process. Therefore, speaker diarization is first required to detect child segments that need to be automatically analyzed.

Speaker diarization is a challenging problem even for adult speech, with large diarization error rates still prevalent in literature. Given their dependency on the domain of the training dataset and amount of training data used, it is difficult to adapt diarization systems for domains where data are scarce. Current diarization systems focus on isolating speech from speakers in meetings, broadcast news and telephonic conversations.

The most common and effective methodologies for diarization involve using i-vectors and DNN embeddings [70]–[72]. [71] used i-vectors along with Probabilistic Linear Discriminant Analysis (PLDA) modelling to score the i-vectors prior to clustering. i-vectors are extracted from speaker adapted GMMs using factor analysis techniques and shown to perform well [73]. More recently, neural network embeddings were used instead of i-vectors to improve performance. [74] used a Recurrent Neural Network (RNN), pre-trained on a speaker identification task, to extract embeddings termed d-vectors, for speaker diarization and applied a non-parametric clustering method to obtain the final diarization. [75] used a Time-Delayed Neural Network (TDNN) architecture with a statistical pooling layer to capture frame level statistics. Embeddings extracted from this network (x-vectors) were shown to perform well in a range of different domains [74], [75].

Within the domain of children's speech, the limited work done to date has shown that it is a non-trivial task [76]. [76] presented a deep learning-based tool to diarize child vs female vs male speech with limited reported experimental results. [76] used an i-vector based PLDA scoring system and reported results on multiple child language environment datasets including Analyzing Child Language Experiences around the World (ACLEW) [77] and Diarization Hard (DiHard) [78]. [79] proposed a HMM based speaker diarization system with DNNs replacing GMMs for probability estimation and reported results on private data collected at a children day-care-center. [78] presented a stacked RNN architecture to separate child and adult speech by training the neural network on multiple objective functions and then jointly training the shared neural network with results.

C. Automatic Speech Recognition for Children

The quality of the acoustic model used to analyze children's disordered speech, particularly to detect phonological and articulation errors, can dramatically impact its reliability. Despite the enormous improvement in acoustic modeling of adult speech over the last few decades, less progress has been made on the acoustic modeling of child speech.

Automatic speech recognition systems trained on adult speech have shown a dramatic degradation in performance when tested on child speech due to linguistic and acoustic mismatches between adult and child speech [80]. Children have higher fundamental and formant frequencies due to their smaller vocal cords and shorter vocal tract. Consequently, in the feature extraction step, it is difficult to eliminate the speaker dependent component (fundamental frequency) and retain only the phoneme dependent components (formants). Furthermore, the shape of the vocal tract changes rapidly as children grow up and their ability to correctly pronounce speech sounds improves. This leads to wide intra- and inter-speaker variations compared to adult speech. Children's speech also suffers from a slow, highly variable speaking rate and inconsistent degree of spontaneity.

Substantial amounts of speech data are therefore needed to accurately model the large variability in children's speech. Google reported Word Error Rate (WER) of as low as 10% for large vocabulary speech recognition systems used to recognize children's queries on the YouTube kids app. Their acoustic model is based on a Convolutional Long short-term memory DNN (CLDNN) and trained with around 1000 hours of children speech selected from their Voice Search traffic [81]. Other commercial companies such as SoapBox [82] labs and KidSense [83] claim to have a highly accurate deep learning based large vocabulary speech recognition system for children trained on large amounts of proprietary speech data.

However, as mentioned in Section III.A, the publicly available speech corpora collected from children are still not large enough to train a reliable and generic acoustic model for children. As a result, several approaches have been proposed to utilize out-of-domain adult speech in enhancing children acoustic model either at the feature level or model level.

Vocal Tract Length Normalization (VTLN) is one of the most popular frequency wrapping techniques used to reduce the speaker variations in acoustic space. It has been successfully used to improve performance of child speech recognition by alleviating the acoustic mismatch between child and adult speech [84] and amongst children from different age ranges [85]. In [80], VTLN combined with model adaptation and speaker normalization techniques achieved 45% relative improvement in recognizing child speech compared to a GMM-HMM acoustic model trained on adult speech. Further, VTLN was also effective in improving the phoneme recognition accuracy of both children and adult speech when used with a DNN-HMM acoustic model [86].

In [87], Stochastic Feature Mapping (SFM) was employed to augment child speech by randomly transforming features from one of the out-of-domain (adult) speakers to one of the in-domain (child) speakers using feature space Maximum Likelihood Linear Regression (fMLLR). The transformed data were then combined with child speech to build a DNN based acoustic model and led to a relative improvement in the WER of around 6% compared to a DNN model trained merely on child speech.

PAMFCCs were proposed in [50] to reduce the effect of the increase of pitch and formant frequencies on the extracted features and produce pitch-robust MFCC features. The proposed method improved the WER of DNN-based adult acoustic model tested on child speech by 11% relative to traditional MFCC features.

Most recently, in [88] the authors examined the effectiveness of applying deep learning to learn useful features directly from the raw speech signal instead of using hand-crafted features such as MFCCs. A deep CNN was trained in an end-to-end fashion using a combination of child and adult speech and gained around 16% relative improvement in WER compared to a DNN model trained on MFCC features when tested on child speech.

Several works have studied the effect of age on the performance of child speech recognition. As expected, the performance degraded with a decrease of age when either acoustic model trained on adult speech [84] or age specific acoustic model [89] were used. Due to the limited availability of child speech corpora, it is hard to train an accurate acoustic model for each age range. Therefore, acoustic model adaptation was used to adapt

an adult acoustic model to the different age ranges [80]. An age dependent speaker normalization technique was proposed in [90] using subglottal resonances.

As data hungry, deep learning based acoustic models become the state-of-the-art in speech recognition systems, different domain adaptation algorithms have been proposed in literature to augment speech data of a source domain with out-of-domain speech data. Some of these methods showed promising results when applied on child speech recognition. In [91], a teacher-student domain adaptation approach was employed to adapt a pre-trained adult acoustic model (teacher) to a child acoustic model (student) using a set of unlabeled parallel speech corpus (adult/child). In [92], transfer learning from adult to child DNN models was performed by fine tuning pre-trained adult speech with a labeled children's corpus. A similar approach was utilized in [93], where only the lower layers of the network were fine-tuned. A multi-task learning approach with child speech as a primary task and adult speech as a secondary task was adopted in [94] to train CNN and TDNN acoustic models.

IV. Methods

In this section we present the opportunities we explored to overcome some of the challenges identified in processing child disordered speech.

A. Paralinguistic Features

Given the challenges in phoneme-level annotation of children's speech in general and disordered speech in specific, we investigated the effectiveness of higher-level features such as paralinguistic features in discriminating between typical and disordered speech.

Paralinguistic acoustic parameters are low-level descriptors of the prosodic, spectral and voice quality of speech, typically used to analyze emotion in speech [95] and detect autism spectrum disorder (ASD) [96], [97]. SSDs also lead to prosodic and distortion errors which impact acoustic quality making paralinguistic features a viable option to detect disordered speech.

1) Feature Extraction: The most commonly used standard paralinguistic parameter sets are the large-scale Interspeech 2013 Computational Paralinguistic ChallengE (ComParE) (6,373 parameters) [98], the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) (62 parameters) and its extended version eGeMAPS (88 parameters) [95]. Here we extracted features in the minimalistic acoustic parameter sets (GeMAPS and eGeMAPS) which include the pitch, jitter, formant frequencies, shimmer, loudness, spectral energies, first four MFCCs, and mean and standard deviation of voiced/unvoiced regions (see [95] for complete list). The frequency, energy and spectral related features, which called the Low-Level Descriptors (LLD), were computed at the frame-level and then summated over each child's speech segment using the mean and normalized standard deviation functionals. The 20th, 50th, 80th percentiles and the 20th to 80th percentile range were further computed for loudness and pitch related features along with the slope of the rising and falling signal. The temporal features, on the other hand, were extracted directly from the whole segment. The open-source OpenSmile software v2.3 [99] was used for feature extraction.

2) Classification Method: To discriminate between typically developing and disordered child speech, we trained and tested a SVM classifier, the most widely used classifier for paralingustic features [95], with Radial Basis Function (RBF) kernel at the segment-level using the 62 GeMAPS or 88 eGeMAPS segment-level parameters. We further used the Recursive Feature Elimination (RFE) feature selection method [100] to identify the most discriminative parameters.

TABLE II
List of Speech Attributes

Vowels, Stops, Affricates, Fricatives, Nasals, Liquids, Semivowels, Approximant, Coronal, High, Dental, Glottal, Labial, Low, Mid, Velar, Back, Retroflex, Anterior, Continuant, Round, Tense, Voiced, Monophthongs, Diphthongs, Silence

B. Anomaly Detection

To address the scarcity of disordered speech corpora we cast the automatic detection of disordered speech as an anomaly detection problem [101]. This also has the advantage of covering the unlimited variations in incorrect pronunciations, influenced by the speaker's native language or the type and/or degree of disorder. This is unlike systems trained using limited amounts of available mispronounced data which fail to generalize to unseen variations of the pronunciation errors.

We focus here on detecting phoneme-level pronunciation errors; a phoneme-specific anomaly detector was trained using typical pronunciations and mispronunciations detected as anomalies to the typical pronunciation.

1) Feature Extraction: To ensure the phoneme-specific anomaly detection model is sensitive to any deviation from the correct pronunciation of the modeled phoneme, we trained it with speech attribute features. Speech attribute features (manners and places of articulation) are more robust against speech variations due to speakers, environmental noise, dialect etc. compared to traditional features [102]. In addition, mispronunciations, can be defined as a change in one or more attributes of the pronounced phoneme, making them more effective in detecting pronunciation errors.

We adopted 26 speech attributes introduced by Lee et. al in [102] for bottom-up speech recognition approach as listed in Table II. These set of features are directly related to the human articulatory system and rich of linguistic information. Moreover, they have been successfully utilized in a various of speech processing domains including identification of spoken language [103], lattice rescoring for Large Vocabulary Speech Recognition (LVSR) [104] and universal phoneme recognition [105].

A binary DNN classifier was trained to determine the existence or the absence of each individual attribute. Each DNN classifier was fed by filter bank features extracted from 25 msec frame of the speech signal. All the speech attribute binary classifiers were used to evaluate the frames of each phoneme and the +ve output from each binary classifier was taken to

form a vector of 26 features that represents the probability of the presence of each specific attribute in the current frame.

2) Classification Method: The correct pronunciation of each phoneme was modeled using a One-Class SVM model (OCSVM) [106]. Unlike the multi-class SVM, here data from only one class is available and the OCSVM trained to create a decision boundary separating the data from the origin. The OCSVM operates better when there are no or less anomalies in the training data [107], as in our model where only native speech corpus were used for the training of the anomaly detection model. The OCSVM models were trained using the extracted 26 speech attribute features.

C. Speaker Diarization

We investigated the efficacy of a deep learning based speaker diarization system trained using out-of-domain, i.e. adult speech, to separate child and adult speech in speech therapy sessions. The main components of our speaker diarization system were 1) a voice activity detector (VAD), 2) a segmentation module to identify speaker changes, 3) an embedding extractor and 4) a speaker clustering module.

1) Feature Extraction and Speech Detection: We used 23 MFCCs normalized using sliding Cepstral Mean and Variance Normalization (CMVN) with a window size of 3 seconds. These settings are adopted to match the configurations of the pre-trained model as described in [75]. These normalized coefficients were then fed to a deep learning model trained on AMI meeting corpus for speech activity detection [108].

2) Segmentation and Embedding Extraction: We applied Gaussian Divergence (GD) on MFCCs extracted from 2.5 seconds windows to compare the Gaussian distributions between successive windows and detect speaker change points [109]. We then extracted the embeddings for each identified segment using a neural network system (x-vectors) pre-trained on Speaker Recognition Evaluation data from the National Institute of Standards and Technology (NIST-SRE) dataset [5]. Finally, we used k-means clustering with cosine similarity to separate the child and adult speech.

D. Speech Corpora

We used five different speech corpora to train, validate and test our proposed methods as described below.

1) TIMIT Corpus [110]: This is a standard phonetically-rich speech corpus designed for the evaluation of acoustic modeling. The corpus contains 6300 sentences produced by 630 adult speakers with 8 different American dialects. The corpus consists of $\sim$3.5 hours of speech stored in 16-bit, 16 kHz waveform files. We used this dataset to train, validate and test our anomaly detection method.

2) AMI Corpus [111]: This is a challenging meeting open-source corpus with 100 hours of speech over 58 meetings. Each meeting has between 1–4 adult speakers in a role-based setting. The AMI dataset provides recordings via various microphones. In this work, the standard headset-mix files (sampling rate 16 kHz and 16-bit resolution) were utilized for training, validation and testing. We used this dataset to validate our speaker diarization method.

3) Oregon Graduate Institute of Science & Technology (OGI) kids' Corpus [67]: The OGI kids' speech corpus consists of recordings from 1,100 children ranging from ages 4–16 years, with each child pronouncing 200 single words and 100 full sentences. Two individuals at OGI subsequently verified each utterance independently during data collection as clear/noisy and correct/incorrect. In this work, the clear and correctly pronounced data from children of ages 6–12 were utilized. All recordings were sampled at 16 kHz rate and each sample stored in 16-bit variable. We used the OGI dataset to validate and test our anomaly detection method.

4) UltraSuite Corpus [24]: This is a speech recordings and ultrasound repository collected from 58 TD children (ages: 9y 3m $\pm$ 1y 10m) and 28 children (ages: 8y 2m $\pm$ 2y) with different types of SSDs including, childhood apraxia of speech, phonological delay and phonological and articulation disorders. Speech waveforms are sampled at 22.05 kHz with sample resolution of 16-bit.

The recordings contain complete therapy sessions with speech from both the child and therapist. Recordings from SSD children were obtained from different therapy sessions, baseline (before therapy starts), therapy, mid-therapy, post-therapy and maintenance (months after therapy ends).

Each recording associated with speaker annotation file indicating the start and end of each speaker generated by automatic speaker diarization as described in [24].

We used a small set of this dataset, typically 4 speakers to test our proposed speaker diarization method. The whole dataset was used to train, validate and test our paralinguistic-based speech disorder detection system.

5) CAS Corpus: This is a disordered speech corpus collected from children with CAS. The recording and annotation processes were performed by SLPs at the University of Sydney. The corpus contained speech from 11 children pronouncing 450 single word prompts. This corpus was used to evaluate the performance of the anomaly detection method.

V. Results

The experimental results for the three methods we explored as described in section IV are as follows.

A. Paralinguistic Features

We adopted a 4-fold cross-validation approach with speech from 28 SSD children and 28 TD children (out of the total of 58 TD children), of the UltraSuite dataset, matched by age and gender to evaluate the SVM classifier. In each fold, the model was trained with segments from 51 TD and 21 SSD children and tested with segments from 7 SSD and 7 TD children. z-normalization (zero mean and unit standard deviation) was computed on each training dataset and then applied on the test dataset.

Table III provides the distribution of the samples across the different cross-validation folds. As seen, despite balancing the

TABLE III
THE DISTRIBUTION OF SEGMENTS OVER DIFFERENT CROSS-VALIDATION (CV) FOLDS. NOTE THAT SSD HAS SIGNIFICANTLY MORE SEGMENTS THAN TD. A WEIGHTED MISCLASSIFICATION PENALTY PARAMETER (C) WAS USED TO BALANCE THE SVM MODEL

CV	N# Training Segments		N# Test Segments	
	TD	SSD	TD	SSD
Fold1	4287	9758	482	3349
Fold2	3995	9895	774	3212
Fold3	4365	10298	404	2809
Fold4	4066	9370	703	3737
Average	4178	9830	590	3276

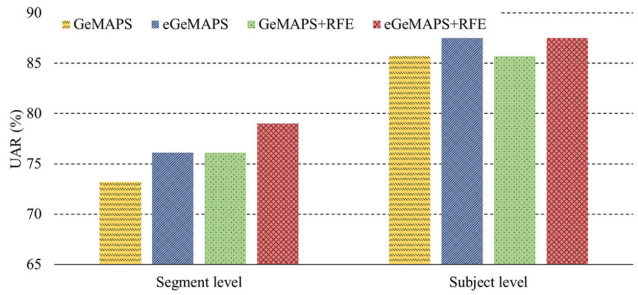


Fig. 2. The segment and subject level UAR classification accuracy of both GeMAPS and eGeMAPS and the effect of RFE feature selection.

number of children in the TD and SSD groups, the number of SSD samples far exceeded the TD segments. This was because 1) the SSD speech recordings were collected from two baseline sessions unlike TD recordings which were collected from only one session and 2) the SSD children repeated utterances at a much higher rate than the TD children. To alleviate this imbalance in the SVM model, we weighted the C hyperparameter with a higher penalty to minor class misclassification during training. Consequently, the evaluation of the system performed using the Unweighted Average Recall (UAR), i.e. the average recall of the TD and SSD classes.

As shown in Fig. 2, results with eGeMAPS feature set were significantly better than those using the GeMAPS in all folds with an average UAR of 76.1 $\pm$ 3.6 compared to 73.2 $\pm$ 1.9 for the latter. The GeMAPS average UAR increased by around 3% absolute to 76% when 55 RFE identified features were used instead of the total 62 features; similarly, the eGeMAPS UAR increased to 78.9% with 58 RFE identified features.

We obtained subject-level classification by identifying each speaker as belonging to the TD or SSD class based on the classification of the majority of their segments. Fig. 2 shows the subject level classification accuracy using the RFE identified GeMAPS and eGeMAPS features. As the number of TD and SSD subjects were equal, we calculated the traditional accuracy score. As seen, here too the RFE identified 58 eGeMAPS features led to a higher accuracy score of 87.5% compared to 85.7% obtained with the RFE identified 55 GeMAPS features.

The confusion matrix in Table IV of the best performing SVM model (RFE identified eGeMAPS features) shows that even though 23% of the TD segments were misclassified as

TABLE IV
THE SUBJECT AND SEGMENT LEVEL CONFUSION MATRIX OF THE EGEMAPS WITH RFE FEATURE SELECTION. ALL TD SUBJECTS WERE CORRECTLY CLASSIFIED WHILE 6 SSD SUBJECTS WERE MISCLASSIFIED AS TD

	Segment-level		Subject-level	
	TD	SSD	TD	SSD
TD	1830(77%)	533(23%)	**28(100%)**	0(0%)
SSD	2325(30%)	5460(70%)	6(21%)	**22(79%)**

SSD, all 28 TD subjects were correctly classified as TD subjects given they had a significantly larger number of TD segments than SSD segments. This indicates that annotating segments as typical or atypical can help improve the model sensitivity as it is possible that children with SSD may produce some typical speech segments, and conversely, TD children may produce some atypical segments.

B. Anomaly Detection

Here we first used the standard TIMIT data to build a phoneme-specific OCSVM model. Of the 630 speakers, 462 speakers were used to train the model while the other 168 speakers were split equally between the validation and test sets.

The validation set was used to tune the parameters of the OCSVM to achieve the lowest frame-level false-acceptance (*FA*) and false-rejection (*FR*) rates. 30% of the validation set was selected from the same phoneme (+ve samples) and 70% randomly selected from the frames of the other phonemes (−ve ones). Because of the imbalance between the (+ve) and (−ve) samples in the validation set, optimal parameters were selected to maximize the *F*1 score instead of the overall accuracy to consider both the *FA* and *FR* rates.

The OCSVM model of each phoneme was tested against samples from the same phoneme to estimate the model *FR* rate and against samples from other phones to estimate the *FA* rate. In testing, all the frames of each phoneme being tested were evaluated and the phoneme acceptance/rejection decision made based on the ratio between the in-class and out-of-class frames. All samples extracted from the test set were force aligned to the correct phoneme sequence of each sentence along with its corresponding speech signal.

Table V demonstrates the phoneme-level *FR* and *FA* rates for each phoneme and the number of occurrences in the test set. In this experiment, the phoneme is considered in-class (accepted) if the ratio between the in-class frames to the out-of-class frames is greater than 1 otherwise it is rejected which means that the decision threshold is equal 1. As shown in the table, most of the phonemes had both *FA* and *FR* rates less than 10% with some extremes such as /ih/, /uh/, /eh/ and /ah/.

We then tested the OCSVM models in detecting pronunciation errors using the typically developing OGI and CAS disordered speech test sets. As the OGI dataset contains correct pronunciations only, we manipulated its phonetic transcription to generate artificial pronunciation errors simulating what are expected from CAS children.

To demonstrate the effectiveness of the method we compared it with the DNN-based GOP algorithm [37]. We first force

TABLE V
THE PHONEME-LEVEL FALSE-ACCEPTANCE (FA) AND FALSE-REJECTION (FR) RATES OF THE OCSVM MODEL

Ph	N#	FA (%)	FR (%)	Ph	N#	FA (%)	FR (%)
aa	588	4.93	6.59	g	367	5.99	4.55
ae	743	5.25	5.79	hh	177	4.52	4.22
ah	436	7.8	11.29	jh	180	5	3.54
ao	617	7.29	5.01	k	822	4.62	2.63
aw	118	9.32	8.97	l	1126	6.13	7.64
ay	433	6	4.46	m	644	5.75	5.09
eh	732	11.07	9.12	n	977	6.45	5.65
er	401	5.24	6.99	ng	179	5.03	6.18
ey	395	6.84	6.32	p	456	5.92	4.33
ih	824	7.52	16.6	r	1174	4.94	5.08
iy	1378	5.95	7.79	s	1397	4.01	4.85
ow	413	7.26	8.17	sh	402	7.71	6.7
oy	131	6.11	5.36	t	755	6.49	7.66
uh	105	13.33	12.29	th	131	5.34	10.98
uw	95	5.26	6.18	v	295	6.78	9.64
b	360	4.72	4.99	w	595	2.86	4.39
ch	146	2.05	3.31	y	260	6.92	5.65
d	441	9.52	8.89	z	638	4.86	6.02
dh	279	5.73	5.73	zh	38	7.89	8.53
f	478	4.18	3.54				

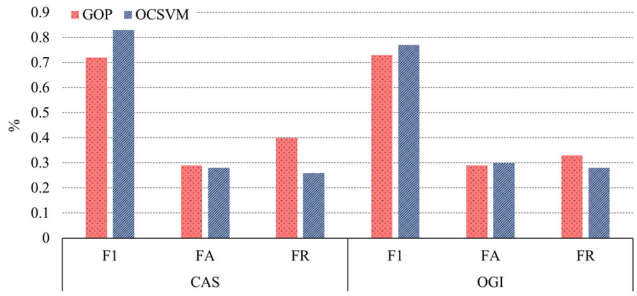


Fig. 3. The F1 scores and the false acceptance (FA) and false rejection (FR) rates of the OCSVM and GOP algorithms applied on the OGI test set and the CAS corpus.

aligned the speech signal with the manipulated version of the phonetic transcription for the OGI corpus and the expected phoneme sequence of the prompt word for the CAS corpus. Forced alignment was performed using DNN-HMM acoustic models trained on the TIMIT training set. Both the OCSVM and GOP methods were then applied on each phoneme and the phoneme accepted if the score exceeded a predefined decision threshold, otherwise rejected. The OCSVM score is the ratio between the number of in-class and out-of-class frames while the GOP score is the normalized estimated log posterior probability. We tuned the phoneme-specific decision threshold for both algorithms to achieve a maximum $F1$ score for each phoneme.

Fig. 3. shows the performance of the two algorithms against the OGI data set with artificial errors and the CAS data set. The results show that our method slightly outperformed the GOP in the artificial error task with $F1$ score and *FA* and *FR* rates of 0.77, 30% and 28% respectively compared to 0.73, 29% and 33% obtained from the GOP algorithm. On the other hand, our approach showed a significant improvement over the conventional GOP algorithm when applied on data with real pronunciation errors. Our method had a significantly higher $F1$ score of 0.83 and lower *FR* of around 26% compared to the GOP 0.72 and 40%, respectively but similar FA rates.

TABLE VI
RESULTS ON A SUBSET OF ULTRASUITE AND COMPLETE AMI TEST/EVALUATION CORPUS

Ultrasuite Dataset					
Speaker label	01F	02F	03F	04M	Overall
DER	0.13	0.08	0.11	0.07	0.1
AMI DATASET					
Corpus		AMI-evalset		AMI-testset	
DER		0.15		0.15	

C. Speaker Diarization

For reference purposes, we first validated the performance of our speaker diarization system on the AMI corpus. We tested our system on the evaluation and test sets which consists of 11 and 6 meetings respectively as specified in [111].

We then evaluated our system on recordings of the baseline speech therapy sessions from 1 TD child and 3 children with SSD in the UltraSuite dataset, each around 10 minutes long. The UltraSuite dataset has been down sampled from 22.05 kHz to 16 kHz to match the configuration of the pre-trained x-vector DNN model.

Given the absence of an annotated reference for speaker diarization of UltraSuite dataset, we manually annotated the data. While annotating the data, we followed the protocol mentioned in [111]. We used an ASR in forced alignment mode to produce word level timings of transcripts. We provided these transcipts to the transcriber. The transcriber followed a 2-pass process to annotate the data. In the first pass, only the boundary of the automatically identified speech segments are refined and noted. In the second pass a more detailed analysis of the silent regions along with the speech regions is carried out for identifying segments and speakers.

The average speaker duration was 0.8 second for the speech pathologist and 1 second for the child. The average speaker turn was similar to the average speaker duration indicating negligible overlapping regions.

We evaluated the performance of our diarization system using the standard Diarization Error Rate (DER) metric [16] defined as the sum of three errors namely, the False Alarm Rate (FAR), the Missed Speech Rate (MSR) and Speaker Error Rate (SER) as follow:

$$DER = FAR + MSR + SER \tag{1}$$

Where FAR refers to non-speech segments being labelled as speech. MSR refers to speech segments being labelled as non-speech and SER refers to incorrect labelling of speaker segments (that is mapped speaker id is not same as reference speaker id).

Table VI presents the average DER results with both the AMI dataset and the UltraSuite dataset. As seen, for both sets we obtained a DER of 0.15. We achieved comparable performance to the state-of-the-art work in [19] which reported a DER of 0.13 and better performance to that of [20] which reported a DER of 0.23 on the AMI dataset. It is important to note that [19]

trains a full model on the AMI training data while our model is pre-trained with out-of-domain data. Across the four therapy sessions in the UltraSuite dataset, we obtained DERs ranging from 0.07 to 0.13, with an average DER of 0.1.

VI. Conclusion

In this paper, we presented current automatic assessment methods used to analyze speech from children with various types of speech sound disorders. We highlighted the most prominent techniques used to detect the three common pronunciation errors made by children with speech sound disorders, namely articulation and phonological errors, hypernasality and prosodic errors.

As phonological and prosodic errors are common also in second-language learners, most of the methods proposed in the literature are for that domain. However, there has been limited application of these techniques to child disordered speech. As an example, our group has conducted the only work on the automatic detection of lexical stress errors in child disordered speech [57].

We further discussed three key challenges hindering the development of automatic analysis of child disordered speech: 1) the lack of publicly available speech corpora collected from children with speech disorders, 2) the need for a reliable speaker diarization system to segment child speech from speech therapy sessions and 3) the slow progress in building an accurate acoustic model for children.

The collection and annotation of child disordered speech is not only costly but also time-consuming and often unreliable. This limits the number publicly available standard datasets and makes it difficult to compare the effectiveness of different algorithms. It also means that the use of deep learning techniques on disordered speech has been limited due to the substantial amount of data required to train the large number of learnable parameters in deep learning structures. Accurate speaker diarization of speech therapy sessions has also been constrained by the lack of available recording sessions of speech therapy sessions with both child and adult speech.

Currently the accuracy of child acoustic models is lagging far behind its adult counterpart. In addition to the scarcity of child speech data compared to adult, there are huge variations in the acoustic characteristics of child speech across different age groups. Various techniques have been proposed to overcome this problem, mainly by incorporating adult training data. The mismatch between the adult and child speech was handled on the feature level by using speaker normalization techniques such as VTLN [84] or at the model level with domain adaptation approaches such as transfer learning [92]. Still, the most effective way to improve child acoustic model is to use large amounts of child training data [81].

We also introduced preliminary work conducted by our research group to tackle some of the identified challenges. Firstly, we investigated the use of paralinguistic features as high-level features extracted over segment or speaker level to alleviate the need for low-level annotation of disordered speech. We showed that paralinguistic features can correctly classify 87% of child speakers as typically developing or disordered.

Secondly, to alleviate the need for large amounts of disordered training speech, we cast the disordered speech detection problem in the anomaly detection framework. Therefore, only the correct pronunciation of each phoneme was modeled and mispronunciations detected as an anomaly. The speech attribute features, namely manners and places of articulation, were chosen to train the anomaly detection method due to their robustness against speech variations. Our anomaly detection based method outperformed the DNN GOP based method with an $F1$ score of 0.83 compared to 0.72 obtained with the GOP based method.

Finally, we demonstrated an approach that efficiently uses out-of-domain resources to create a speaker diarization system for speech therapy sessions with diarization error rates comparable to current state-of-the-art methods. We showed that it is possible to use deep learning networks pre-trained with adult speech to extract embeddings from segments speech therapy session and characterize them as adult or child speech.

Acknowledgment

The statements made herein are solely the responsibility of the authors.

References

[1] B. A. Lewis, L. A. Freebairn, and H. G. Taylor, "Follow-up of children with early expressive phonology disorders," *J. Learn. Disabilities*, vol. 33, no. 5, pp. 433–444, 2000.

[2] H. T. Bunnell, D. M. Yarrington, and J. B. Polikoff, "STAR: Articulation training for young children," in *Proc. 6th Int. Conf. Spoken Lang. Process.*, vol. 4, 2000, pp. 85–88.

[3] O. Saz, S.-C. Yin, E. Lleida, R. Rose, C. Vaquero, and W. R. Rodríguez, "Tools and technologies for computer-aided speech and language therapy," *Speech Commun.*, vol. 51, no. 10, pp. 948–967, 2009.

[4] S. Rvachew and F. Brosseau-Lapré, "Speech perception intervention," in *Interventions for Speech Sound Disorders in Children*. Baltimore, MD, USA: Brookes Publishing, 2010, pp. 295–314.

[5] Y. Wren, S. Roulstone, and A. L. Williams, "Computer-based interventions," in *Interventions for Speech Sound Disorders in Children.* Baltimore, MD, USA: Brookes Publishing, 2010, pp. 275–294.

[6] A. Parnandi *et al.*, "Development of a remote therapy tool for childhood apraxia of speech," *ACM Trans. Accessible Comput.*, vol. 7, no. 3, 2015, Art. no. 10.

[7] M. Shahin *et al.*, "Tabby talks: An automated tool for the assessment of childhood apraxia of speech," *Speech Commun.*, vol. 70, pp. 49–64, 2015.

[8] S. E. Apps, Apraxiaville, 2017. [Online]. Available: http://smartyears-apps.com/apraxia-ville/

[9] E. Solutions, ArtikPix, 2018. [Online]. Available: http://expressive-solutions.com/artikpix/

[10] A. Hair, P. Monroe, B. Ahmed, K. J. Ballard, and R. Gutierrez-Osuna, "Apraxia world: A speech therapy game for children with speech sound disorders," in *Proc17th ACM Conf. Interact. Des. Children*, 2018, pp. 119–131.

[11] B. Ahmed, P. Monroe, A. Hair, C. T. Tan, R. Gutierrez-Osuna, and K. Ballard, "Speech-driven mobile games for speech therapy: User experiences and feasibility," *Int. J. Speech-Lang. Pathol.*, vol. 20, no. 6, pp. 644–658, 2018.

[12] S. Alharbi, M. Hasan, A. J. Simons, S. Brumfitt, and P. Green, "Detecting stuttering events in transcripts of children's speech," in *Proc. Int. Conf. Statist. Lang. Speech Process.*, 2017, pp. 217–228.

[13] S. Alharbi, M. Hasan, A. J. Simons, S. Brumfitt, and P. Green, "A lightly supervised approach to detect stuttering in children's speech," in *Proc. Interspeech*, 2018, pp. 3433–3437.

[14] P. Arbajian, *Disfluent Speech Segments Detection and Remediation.* Charlotte, NC, USA: The University of North Carolina at Charlotte, 2019.

[15] S. Dudy, M. Asgari, and A. Kain, " Pronunciation analysis for children with speech sound disorders," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2015, pp. 5573–5576.

[16] D. Smith *et al.*, "Improving child speech disorder assessment by incorporating out-of-domain adult speech," in *Proc. Interspeech*, 2017, pp. 2690–2694.

[17] S. Dudy, S. Bedrick, M. Asgari, and A. Kain, "Automatic analysis of pronunciations for children with speech sound disorders," *Comput. Speech Lang.*, vol. 50, pp. 62–84, 2018.

[18] P. V. Kothalkar, J. Rudolph, C. Dollaghan, J. McGlothlin, T. F. Campbell, and J. H. Hansen, "Automatic screening to detect 'At Risk' Child speech samples using a clinical group verification framework," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 4909–4913.

[19] J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, J. C. Vásquez-Correa, C. G. Castellanos-Domínguez, and E. Nöth, "Automatic detection of hypernasal speech of children with cleft lip and palate from Spanish vowels and words using classical measures and nonlinear analysis," *Revista Facultad de Ingeniería Universidad de Antioquia*, vol. 80, pp. 109–123, 2016.

[20] A. K. Dubey, S. R. M. Prasanna, and S. Dandapat, "Pitch-adaptive front-end feature for hypernasality detection," in *Proc. Interspeech*, 2018, pp. 372–376.

[21] S. Kalita, S. R. M. Prasanna, and S. Dandapat, "Self-similarity matrix based intelligibility assessment of cleft lip and palate speech," in *Proc. Interspeech*, 2018, pp. 367–371.

[22] C. M. Vikram, S. R. M. Prasanna, A. K. Abraham, M. Pushpavathi, and S. GirishK., "Detection of glottal activity errors in production of stop consonants in children with cleft lip and palate," in *Proc. Interspeech*, 2018, pp. 382–386.

[23] M. C. VikramC., A. Tripathi, S. Kalita, and S. R. M. Prasanna, "Estimation of hypernasality scores from cleft lip and palate speech," in *Proc. Interspeech*, 2018, pp. 1701–1705.

[24] A. Eshky *et al.*, "UltraSuite: A repository of ultrasound and acoustic data from child speech therapy sessions," in *Proc. INTERSPEECH: 19th Annu. Conf. Int. Speech Commun. Assoc.*, Hyderabad, India, 2018, pp. 1888–1892.

[25] L. D. Shriberg, D. M. Aram, and J. Kwiatkowski, "Developmental apraxia of speech: I. Descriptive and theoretical perspectives," *J. Speech, Lang. Hearing Res.*, vol. 40, no. 2, pp. 273–285, 1997.

[26] Adhoc Committee on CAS, *Childhood Apraxia of Speech [Position Statement]*. Rockville, MD, USA: American Speech-Language-Hearing Association, 2007.

[27] A. J. Caruso and E. A. Strand, *Clinical Management of Motor Speech Disorders in Children*. New York, NY, USA: Thieme, 1999.

[28] A. T. Morgan and A. P. Vogel, "Intervention for dysarthria associated with acquired brain injury in children and adolescents," *Cochrane Database Syst. Rev.*, vol. 3, 2008, Art. no. CD006279.

[29] [Online]. Available: https://www.clapa.com/treatment/early-years-1-4/speech/

[30] J. S. Yaruss, "Clinical measurement of stuttering behaviors," *Contemporary Issues Commun. Sci. Disorders*, vol. 24, no. 24, pp. 33–44, 1997.

[31] H. R. Perez and J. H. Stoeckle, "Stuttering: Clinical and research update," (in eng), *Can. Family Physician Medecin de famille Canadien*, vol. 62, no. 6, pp. 479–484, 2016.

[32] J. E. Bernthal, N. W. Bankson, and P. Flipsen, *Articulation and Phonological Disorders: Speech Sound Disorders in Children*. Boston, MA, USA: Pearson, 2009.

[33] L. S. Chee, O. C. Ai, and S. Yaacob, "Overview of automatic stuttering recognition system," in *Proc. Int. Conf. Man-Mach. Syst.*, 2009, pp. 1–6.

[34] A. M. Harrison, W.-K. Lo, X.-j. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *Proc. Int. Workshop Speech Lang. Technol. Educ.*, 2009, pp. 45–48.

[35] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, no. 2–3, pp. 95–108, 2000.

[36] H. Al-Barhamtoshy, K. Jambi, W. Al-Jedaibi, D. Motaweh, S. Abdou, and M. Rashwan, "Speak correct: Phonetic editor approach," *Life Sci. J.*, vol. 11, no. 8, 2014, pp. 626–640.

[37] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)," in *Proc. Interspeech*, 2013, pp. 1886–1890.

[38] S. M. Abdou *et al.*, "Computer aided pronunciation learning system using speech recognition techniques," in *Proc. 9th Int. Conf. Spoken Lang. Process.*, 2006.

[39] M. Shahin, B. Ahmed, J. McKechnie, K. Ballard, and R. Gutierrez-Osuna, "A comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 1583–1587.

[40] L. Ward *et al.*, "Automated screening of speech development issues in children by identifying phonological error patterns," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 2661–2665.

[41] P. Kothalkar, J. Rudolph, C. Dollaghan, J. McGlothlin, T. Campbell, and J. H. Hansen, "Fusing text-dependent word-level i-Vector models to screen 'at risk' child speech," *Age (Months)*, vol. 51, no. 11, pp. 36–78, 2018.

[42] N. Ramou and M. Guerti, "Automatic detection of articulations disorders from children's speech preliminary study," *J. Commun. Technol. Electron.*, vol. 59, no. 11, pp. 1274–1279, 2014.

[43] A. Hanani, M. Attari, A. Farakhna, M. Hussein, A. Joma'a, and S. Taylor, "Automatic identification of articulation disorders for arabic children speakers," *presented at the Workshop on Child Computer Interact.*, San Francisco, CA, USA, 2016.

[44] A. Maier *et al.*, "Automatic detection of articulation disorders in children with cleft lip and palate," *J. Acoust. Soc. Amer.*, vol. 126, no. 5, pp. 2589–2602, 2009.

[45] S. M. Rendón, J. O. Arroyave, J. V. Bonilla, J. A. Londoño, and C. C. Domínguez, "Automatic detection of hypernasality in children," in *Proc. Int. Work-Conf. Interplay Between Natural Artif. Comput.*, 2011, pp. 167–174.

[46] M. Golabbakhsh *et al.*, "Automatic identification of hypernasality in normal and cleft lip and palate patients with acoustic analysis of speech," *J. Acoust. Soc. Amer.*, vol. 141, no. 2, pp. 929–935, 2017.

[47] J. R. Orozco-Arroyave *et al.*, "Automatic selection of acoustic and nonlinear dynamic features in voice signals for hypernasality detection," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 529–532.

[48] L. He, J. Zhang, Q. Liu, H. Yin, M. Lech, and Y. Huang, "Automatic evaluation of hypernasality based on a cleft palate speech database," *J. Med. Syst.*, vol. 39, no. 5, 2015, Art. no. 61.

[49] C. Vikram, A. Tripathi, S. Kalita, and S. M. Prasanna, "Estimation of hypernasality scores from cleft lip and palate speech," in *Proc. Interspeech*, 2018, pp. 1701–1705.

[50] S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-adaptive front-end features for robust children's ASR," in *Proc. Interspeech*, 2016, pp. 3459–3463.

[51] L. Ferrer, H. Bratt, C. Richey, H. Franco, V. Abrash, and K. Precoda, "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems," *Speech Commun.*, vol. 69, pp. 31–45, 2015.

[52] J.-Y. Chen and L. Wang, "Automatic lexical stress detection for Chinese learners' of English," in *Proc. 7th Int. Symp. Chin. Spoken Lang. Process.*, 2010, pp. 407–411.

[53] J. Zhao, H. Yuan, J. Liu, and S. Xia, "Automatic lexical stress detection using acoustic features for computer assisted language learning," in *Proc. APSIPA ASC*, 2011, pp. 247–251.

[54] C. Li, J. Liu, and S. Xia, "English sentence stress detection system based on HMM framework," *Appl. Math. Comput.*, vol. 185, no. 2, pp. 759–768, 2007.

[55] V. K. R. Sridhar, S. Bangalore, and S. Narayanan, "Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework," in *Proc. Conf. North Amer. Chapt. Assoc. Comput. Linguist.*, 2007, pp. 1–8.

[56] K. Li, X. Qian, S. Kang, and H. Meng, "Lexical stress detection for L2 English speech using deep belief networks," in *Proc. Interspeech*, 2013, pp. 1811–1815.

[57] M. Shahin, R. Gutierrez-Osuna, and B. Ahmed, "Classification of bisyllabic lexical stress patterns in disordered speech using deep learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 6480–6484.

[58] M. A. Shahin, B. Ahmed, and K. J. Ballard, "Classification of lexical stress patterns using deep neural network architecture," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2014, pp. 478–482.

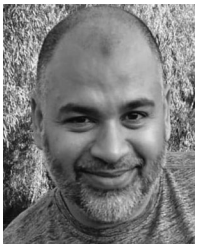
[59] L. D. Shriberg *et al.*, "Perceptual and acoustic reliability estimates for the speech disorders classification system," *Clin. Linguist. Phonet.*, vol. 24, no. 10, pp. 825–846, 2010.

[60] A. E. Cummings and J. A. Barlow, "A comparison of word lexicality in the treatment of speech sound disorders," *Clin. Linguist. Phonet.*, vol. 25, no. 4, pp. 265–286, 2011.

[61] C. Torrington Eaton and N. B. Ratner, "An exploration of the role of executive functions in preschoolers' phonological development," *Clin. Linguist. Phonet.*, vol. 30, no. 9, pp. 679–695, 2016.

[62] S. Chiat and M. Yavas, "From lexical access to lexical output: What is the problem for children with impaired phonology," *First Second Lang. Pathol.*, pp. 107–133, 1994.

[63] B. May Bernhardt *et al.*, "Word structures of Granada Spanish-speaking preschoolers with typical versus protracted phonological development," *Int. J. Lang. Commun. Disorders*, vol. 50, no. 3, pp. 298–311, 2015.
[64] J. L. Preston, M. Hull, and M. L. Edwards, "Preschool speech error patterns predict articulation and phonological awareness outcomes in children with histories of speech sound disorders," *Amer. J. Speech-Lang. Pathol.*, vol. 22, no. 2, pp. 173–84. 2013.
[65] A. Batliner *et al.*, "The PF_STAR children's speech corpus," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 2761–2764.
[66] M. Eskenazi, J. Mostow, and D. Graff, *The CMU Kids Speech Corpus*. Philadelphia, PA, USA: University of Pennsylvania, 1997.
[67] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The OGI kids' speech corpus and recognizers," in *Proc. 6th Int. Conf. Spoken Lang. Process.*, 2000, pp. 564–567.
[68] R. Cole, P. Hosom, and B. Pellom, "University of Colorado prompted and read children's speech corpus," Univ. Colorado, Boulder, CO, USA,Tech. Rep. TR-CSLR-2006-022006.
[69] J. J. Gong, M. Gong, D. Levy-Lambert, J. R. Green, T. P. Hogan, and J. V. Guttag, "Towards an automated screening tool for developmental speech and language impairments," in *Proc. Interspeech*, 2016, pp. 112–116.
[70] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.
[71] G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2014, pp. 413–417.
[72] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "I-vector based speaker recognition on short utterances," in *Proc. 12th Annu. Conf. Int, Speech Commun. Assoc.*, 2011, pp. 2341–2344.
[73] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.
[74] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with LSTM," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5239–5243.
[75] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5329–5333.
[76] A. Le Franc *et al.*, "The ACLEW DiViMe: An easy-to-use diarization tool," in *Proc. Interspeech*, 2018, pp. 1383–1387.
[77] E. Bergelson *et al.*, *Starter-ACLEW*. Databrary, 2017, doi: 10.17910/B7.390.
[78] X. Wang, J. Du, L. Sun, Q. Wang, and C.-H. Lee, "A progressive deep learning approach to child speech separation," in *Proc. IEEE 11th Int. Symp. Chin. Spoken Lang. Process.*, 2018, pp. 76–80.
[79] M. Najafian and J. H. Hansen, "Speaker independent diarization for child language environment analysis using deep neural networks," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2016, pp. 114–120.
[80] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 603–616, Nov. 2003.
[81] H. Liao *et al.*, "Large vocabulary automatic speech recognition for children," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1611–1615.
[82] S. Labs, Soap Box Labs, 2018. [Online]. Available: https://www.soapboxlabs.com/
[83] Kadho, Kid Sense, 2019. [Online]. Available: https://kidsense.ai/
[84] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Proc. 5th Eur. Conf. Speech Commun. Technol.*, 1997, pp. 2371–2374.
[85] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive books and tutors," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2003, pp. 186–191.
[86] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2014, pp. 135–140.
[87] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving children's speech recognition through out-of-domain data augmentation," in *Proc. Interspeech*, 2016, pp. 1598–1602.
[88] S. P. Dubagunta, S. H. Kabil, and M. M. Doss, "Improving children speech recognition through feature learning from raw speech signal," presented at the IEEE Int. Conf. Acoust., Speech Signal Process., Brighton, U.K., 2019.
[89] G. Yeung and A. Alwan, "On the difficulties of automatic speech recognition for Kindergarten-Aged children," in *Proc. Interspeech*, 2018, pp. pp. 1661–1665.
[90] J. Guo, R. Paturi, G. Yeung, S. M. Lulich, H. Arsikere, and A. Alwan, "Age-dependent height estimation and speaker normalization for children's speech using the first three subglottal resonances," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1665–1669.
[91] J. Li, M. L. Seltzer, X. Wang, R. Zhao, and Y. Gong, "Large-scale domain adaptation via teacher-student learning," *Interspeech*, pp. 2386–2390, 2017.
[92] R. Serizel and D. Giuliani, "Deep neural network adaptation for children's and adults' speech recognition," in *Proc. Italian Comput. Linguist. Conf.*, 2014, pp. 1–26.
[93] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," 2018, *arXiv:1805.03322*.
[94] J. Wang, S. I. Ng, D. Tao, W. Y. Ng, and T. Lee, "A study on acoustic modeling for child speech based on multi-task learning," in *Proc. IEEE 11th Int. Symp. Chin. Spoken Lang. Process.*, 2018, pp. 389–393.
[95] F. Eyben *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr./Jun. 2016.
[96] M. Schmitt, E. Marchi, F. Ringeval, and B. Schuller, "Towards cross-lingual automatic diagnosis of autism spectrum condition in children's voices," in *Proc. 12 ITG Symp. Speech Commun.*, 2016, pp. 1–5.
[97] A. Baird *et al.*, "Automatic classification of autistic child vocalisations: A novel database and results," in *Proc. Int. Speech Commun. Assoc.*, 2017, pp. 849–853.
[98] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*, 2013, pp. 849–853.
[99] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 835–838.
[100] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, 2002.
[101] M. Shahin, B. Ahmed, J. X. Ji, and K. Ballard, "Anomaly detection approach for pronunciation verification of disordered speech using speech attribute features," in *Proc. Interspeech*, 2018, pp. 1671–1675.
[102] C.-H. Lee and S. M. Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification, and recognition," in *Proc. IEEE*, vol. 101, no. 5, pp. 1089–1115, May 2013.
[103] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 209–227, 2013.
[104] I.-F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Attribute based lattice rescoring in spontaneous speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 3325–3329.
[105] S. M. Siniscalchi, D.-C. Lyu, T. Svendsen, and C.-H. Lee, "Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 875–887, Mar. 2012.
[106] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Proc Adv. Neural Inf. Process. Syst.*, 2000, pp. 582–588.
[107] M. Amer, M. Goldstein, and S. Abdennadher, "Enhancing one-class support vector machines for unsupervised anomaly detection," in *Proc. ACM SIGKDD Workshop Outlier Detection Description*, 2013, pp. 8–15.
[108] R. Yin, H. Bredin, and C. Barras, "Speaker change detection in broadcast TV using bidirectional long short-term memory networks," in *Proc. Interspeech*, 2017, pp. 3827–3831.
[109] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Speech Recognit. Workshop*, 1997, vol. 1997, pp. 97–99.
[110] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM," NIST speech disc 1-1.1," NASA STI/Recon technical report n, vol. 93, 1993.
[111] J. Carletta *et al.*, "The AMI meeting corpus: A pre-announcement," in *Proc. Int. Workshop Mach. Learn. Multimodal Interact.*, 2005, pp. 28–39.



Mostafa Shahin (M'12) graduated from the Department of Electrical Engineering, Ain Shams University, Giza, Egypt, in 2003 and the M.Sc. degree in electronics and communications engineering from Cairo University, Cairo, Egypt, in 2015. From 2004, he was in the speech research group at RDI Company, Giza. From 2011 to 2018, he was a Research Associate with the Electrical and Computer Engineering Department, Texas A&M University at Qatar, Doha, Qatar. He is currently working toward the Ph.D. degree with the School of Electrical and Telecommunication Engineering, UNSW, Sydney, Australia. His research focuses on applying machine learning techniques for speech and biomedical signal processing.

Usman Zafar received the bachelor's degree in electronics engineering from the National University of Sciences and Technology, Islamabad, Pakistan and the master's degree in computer science from the Lahore University of Management Sciences, Lahore, Pakistan. He is currently working as a Research Associate with the Qatar Environment and Energy Research Institute, Qatar. He was previously with the Electrical and Computer Engineering Program, Texas A&M University at Qatar, Doha, Qatar. Previously, He has worked on different problems in text mining (Identifying lexical variation in low-resource languages) and speaker recognition(Speaker diarization). His current research interests include applying artificial intelligence techniques to novel use-cases for the energy industry.

Beena Ahmed (M'04) received the Ph.D. degree in electrical engineering from the University of New South Wales, Sydney, NSW, Australia in 2004. She is currently a Senior Lecturer with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW, Australia. She was previously an Assistant Professor with the Electrical and Computer Engineering Program, Texas A&M University at Qatar, Doha, Qatar. Her research focuses on developing speech processing based solutions for low-resource domains such as children's speech and disordered speech.