Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 56 (2019) 80–94

www.elsevier.com/locate/csl

An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder☆

Ming Li[a], Dengke Tang[c], Junlin Zeng[c], Tianyan Zhou[c], Huilin Zhu[b], Biyuan Chen[b], Xiaobing Zou*[,b]

[a] *Data Science Research Center, Duke Kunshan University, China*
[b] *The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China*
[c] *School of Electronics and Information Technology, Sun Yat-Sen University, Guangzhou, China*

Received 17 December 2017; received in revised form 6 November 2018; accepted 7 November 2018
Available online xxx

Abstract

Autism Spectrum Disorder (ASD), a neurodevelopmental disability, has become one of the high incidence diseases among children. Studies indicate that early diagnosis and intervention treatments help to achieve positive longitudinal outcomes. In this paper, we focus on the speech and language abnormalities of young children with ASD and present an automated assessment framework in quantifying atypical prosody and stereotyped idiosyncratic phrases related to ASD. For detecting atypical prosody from speech, we propose both the hand-crafted feature based method as well as the end-to-end deep learning framework. First, we use the OpenSMILE toolkit to extract utterance level high dimensional acoustic features followed by a support vector machine (SVM) backend as the conventional baseline. Second, we propose several end-to-end deep neural network setups and configurations to model the atypical prosody label directly from the constant Q transform spectrogram of speech. Third, we apply cross-validation on the training data to perform segments selection and enhance the subject level classification performance. Fourth, we fuse the deep learning based methods with the conventional baseline at the score level to further enhance the overall system performance. For detecting the stereotyped idiosyncratic usage of words or phrases from speech transcripts, we adopt language model, dependency treebank and Term Frequency−Inverse Document Frequency (TF−IDF) in addition to Linguistic Inquiry and Word Count software (LIWC) methods to extract a set of text features followed by a standard SVM backend. We collect a database of spontaneous Mandarin speech recorded during the Autism Diagnostic Observation Schedule (ADOS) Module 2 and Module 3 sessions. The Module 2 part consists of 118 children while the Module 3 part includes 71 children. Experimental results on this database show that our proposed methods can effectively predict the atypical prosody and stereotyped idiosyncratic phrases codes for young children with the risk of ASD. On the two categories classification task, the unweighted accuracy of the aforementioned two tasks are 88.1% and 77.8%, respectively.

Keywords: Autism spectrum disorder; Atypical prosody; Stereotyped idiosyncratic phases; Recurrent neural network; Convolutional neural network

☆This paper has been recommended for acceptance by Prof. R. K. Moore.

* Corresponding author.
E-mail addresses: ming.li369@dukekunshan.edu.cn (M. Li), zouxb@vip.tom.com (X. Zou).

https://doi.org/10.1016/j.csl.2018.11.002
0885-2308/ 2018 Published by Elsevier Ltd.

1. Introduction

Autism Spectrum Disorder (ASD) refers to a group of symptoms related to social impairments and communication difficulties. It has become one of the high incidence diseases among children. A recent analysis from the Centers for Disease Control and Prevention estimates that 1 in 68 children has ASD in the United States (Christensen et al., 2016). Early behavioral and educational interventions have been proved to be very successful in many clinical studies. This attaches great significance to the recognition of common ASD behavior patterns and diagnoses at the early stage.

In paralinguistics, prosody relates to several communicative functions such as intonation, tone, pitch, stress, rhythm, etc. Prosody can reflect many important elements of language including the emphasis, contrast and affective state of the speaker (McCann and Peppé, 2003). These are critical information in human communication. Therefore, atypical prosody is one of the common symptoms related to ASD. Specifically, children with ASD may speak in flat, robot-like or a sing-song voice (Fusaroli et al., 2016).

In this work, we not only focus on the speech signal but also study the language patterns for ASD detection. For ASD children who are verbally fluent, they may have various kinds of language communication abnormalities, such as stereotyped, repetitive and idiosyncratic usage of words or phrases. Children with stereotyped idiosyncratic usage of words or phrases often use some inflexible and rigid words and expressions during the conversation. The words or phrases they uttered sometimes may be inappropriate for the context. Moreover, ASD children may create some new and weird words during the conversation.

Both the aforementioned speech and language cues are important for clinicians to perform diagnosis. The Autism Diagnostic Observation Schedule (ADOS) is a standard screening test to help clinicians observe children's language and behavior patterns relevant to the diagnosis of autism. It consists of a series of structured and semi-structured tasks assessing social interaction, communication, playing, and imaginative usage of materials (Lord et al., 2000). There are four different modules designed mainly according to the subject's age and linguistic capability. Moreover, speech and language abnormalities are included in all these four modules. The ADOS screening provides codes to quantify the items on an integer scale from "0" to "2" based on the severity of each abnormality category (Gotham et al., 2009). Taking atypical prosody as an example, "0" denotes no abnormal prosody; "1" stands for some changes on pitch/tone, a bit flat/exaggerate intonation, slightly abnormal volume, a little slow/fast/jerky rhythm and "2" implies markedly and consistently abnormalities on the aforementioned aspects (Lord et al., 2000).

In the ADOS screening, therapists need to identify multiple behavior codes related to speech and language, including atypical prosody, stereotyped idiosyncratic phrases, etc. As many research and treatment methods in the psychology field, this kind of evaluation or diagnosis requires experienced experts or clinicians with intensive specialized training. Another issue is the subjective inconsistency between clinicians, which could sometimes make the results ambiguous at some certain levels. Researchers have proposed strategies to utilize speech and language processing techniques to support clinicians with quantitative analysis of ASD children's prosody (Bone et al., 2012; Chaspari et al., 2012; Bone et al., 2015; 2017) and language patterns (Kumar et al., 2016). Furthermore, since pattern recognition and machine learning methods have demonstrated promising results in modeling behavior symptoms and relationships with expert's experience (Narayanan and Georgiou, 2013; Xiao et al., 2015), some automated screening and evaluating tools based on objective measurements directly extracted from recordings are proposed (Gong et al., 2016; Xiao et al., 2016). These automated coding tools showed great potential to be scalable and assist clinicians to analyze the variation trend of a specific symptom in long term monitoring or assessments.

In this paper, we focus on the speech and language abnormalities and present an automated assessment framework to determine the existence and severity level of atypical prosody and stereotyped idiosyncratic phrases for young children under the ADOS Module 2 and 3 setup.

On the speech side, we model the atypical prosody abnormality using both the traditional strategy and the deep learning framework. We demonstrate that the end-to-end techniques can achieve comparable performance against the baseline system even on a small-scale dataset. Since we directly model the ASD related atypical prosody code from the spectrograms in an end-to-end manner, there is no prior domain knowledge required for feature engineering. The fusion of the two systems can further improve the overall system performance at the segment level in terms of the unweighted average recall (UAR). This result shows that the end-to-end framework has great potential in the field of behavior signal processing (BSP) (Black et al., 2013). Moreover, among all the speech segments in an ADOS conversation session, not every segment reflects the atypical prosody information and therefore we adopt the

cross validation strategy on the training set to perform segment selection and improve the accuracy on both segment and person levels.

On the language side, besides n-gram language model, categorical word counts from the Linguistic Inquiry and Word Count software (LIWC) these baseline features and maximum entropy classifier (Kumar et al., 2016), we also propose dependency treebank and Term Frequency-Inverse Document Frequency (TF−IDF) these two methods to extract features that are more related to the stereotyped idiosyncratic usage of words or phrases. We concatenate all these four features together and adopt a standard SVM classifier as the backend.

Furthermore, we also investigate the cutoff boundary of the code 0/1/2 by merging 1/2 as a new code to form a binary classification task. Experimental results show that our trained models are more confident at distinguishing between normal and abnormal cases rather than estimating the detailed severity level of abnormal behaviors. In this study, our goal is not just to recognize the three-category code and use it the same way as described in the ADOS manual (adding together all the codes and compare with the cutoff threshold). This two-category code itself can serve as a quantitative measure of atypical prosody or stereotyped idiosyncratic phrases. We can use the proposed method to perform coarse screening. Besides that, we can also fuse the recognized two-category code with other automatic calculated codes from related tests, e.g. respond to name (Liu et al., 2017), response to non-social sound stimuli, joint attention, etc.

The remainder of the paper is organized as follows. Section 2 describes our database. The proposed methods are explained in Section 3 and Section 4, respectively. Experimental results and discussions are presented in Section 5 while conclusions and future works are provided in Section 6.

2. Database description

We perform experiments on the data collected from our behavior observation and analysis lab in the Third Affiliated Hospital of Sun Yat-sen University as demonstrated in Fig. 1. Our audio database is collected in the real ADOS module 2 and module 3 screening environment. As you can observe from Fig. 1, our multimodal behavior signal capture system is equipped with multiple HD cameras and Kinect sensor to capture vision data during the child-psychologist interactions. As for the audio data, every participant wears a wireless recording device (presented in Fig. 2) to collect the multi-channel audio data. By doing this, we can obtain both child's and clinician's speech with high quality and purity comparing to the single channel recording method. All the ADOS module 2 and module 3 screening sessions in this database are performed by the same doctor in the Third Affiliated Hospital of Sun Yat-sen Universit. In this case, we can reduce the inconsistency and variability due to different doctors. Moreover, this doctor is a professional and certified ADOS therapist. The data collection is approved by the children's family and the institutional review board (IRB) of the hospital.

Fig. 1. Behavior observation and analysis lab in the Third Affiliated Hospital of Sun Yat-sen University.

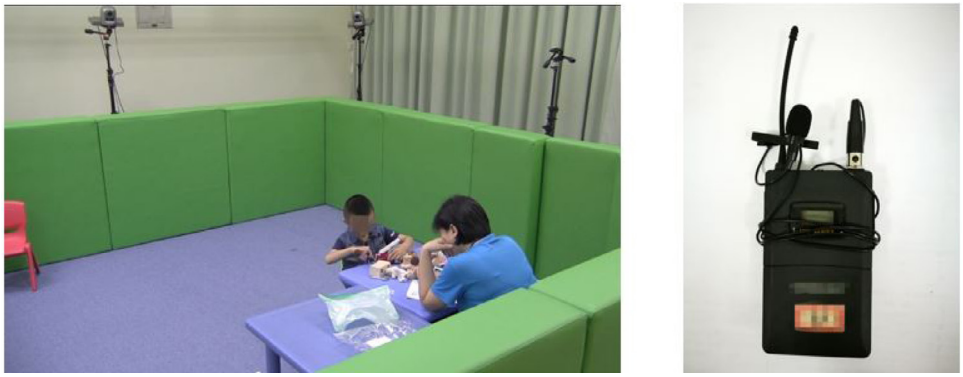


Fig. 2. Data collecting environment and the audio recording device.

2.1. ADOS module 2 data for the atypical prosody detection experiment

There are multiple spontaneous child-clinician interactions during the ADOS screening with different sub-tasks. Since we intend to focus on the speech abnormalities of ASD children, we extract a subset from the ADOS sessions. For ADOS module 2, we select speech data during the interaction of "Demonstration Task", "Description of a Picture", "Telling a Story From a Book" and "Birthday Party" these four questions. They are either designed to observe children's spontaneous and expressive language capabilities or contain a large proportion of speech conversations.

The details of our database are showed in Table 1 and Table 2, respectively. For module 2, our dataset contains 118 children, including 25 typical-developed ones (by the ADOS overall score). However, due to the heterogeneous characteristics of ASD, a few ASD children can still have normal prosody and some typical-developed children may also have atypical prosody. Here our focus is to predict the atypical prosody code rather than the ASD label.

2.2. ADOS module 3 data for the stereotyped idiosyncratic phrases detection experiment

We record both the child's and clinician's speech during the real ADOS Module 3 screening in the hospital the same way as we described in Section 2.1. All the children and clinicians in our dataset speak Mandarin. To analyze the texts of conversations, we manually transcribed the speech into texts.

Our dataset contains 71 children for Module 3. The statistical information of the dataset is shown in Table 2. For this study, we select the audio segments when children are answering those socioemotional questions: the topics are "Emotion", "Social Difficulties and Annoyance", "Friends, Relationships, and Marriage" and "Loneliness", respectively.

3. Methods for atypical prosody detection

The baseline system is implemented using the OpenSMILE feature extractor followed by a Support Vector Machine (SVM) classifier. Our end-to-end deep learning framework uses spectrograms as the input, and performs

Table 1
ADOS Module 2 database description.

Item	Statistics
Age(month)	Range: 26–142, Mean:57.77
Gender	Male:105, Female:13
Language	Mandarin:110, Cantonese:8
Atypical prosody code(subjects)	'0':19,'1':46,'2':53
ADOS Diagnosis	Autism:51, ASD:42, below ASD cutoffs:25
Randomly selected subjects	training data:93, testing data:25

Table 2
ADOS Module 3 database description.

Item	Statistics
Age(month)	Range: 24–166, Mean:91.95
Gender	Male:64, Female:7
Language	Mandarin:71
Stereotyped idiosyncratic phrases code(all subjects)	'0':11,'1':42,'2':18
ADOS Diagnosis	Autism:59, ASD:4, below ASD cutoffs:8
Randomly selected subjects	training data:50, testing data:21
Stereotyped idiosyncratic phrases code(testing subjects)	'0':3,'1':14,'2':4

supervised learning using deep neural networks. Finally we perform score level fusion by averaging the prediction scores from the aforementioned systems. Section 3.2 and Section 3.3 introduce these two methods in detail.

3.1. Data preprocessing

Our database is collected using wearable recording devices carried by every person in the ADOS screening. Hence, the corpus contains multi-channel time synchronized audio data. In order to obtain each participant's clean speech, we need to preprocess the data using multi-channel speaker diarization techniques. In this scenario, each speaker's voice is loudest on their own microphone. As a result, the energetic difference between primary speech and secondary speech is an available characteristic (Dubey et al., 2016; Xiao et al., 2011). In practice, we use the time-synchronized energy measurements across channels to remove most secondary speech. Next, we apply the children-customized single channel speaker diarization technique (Zhou et al., 2016) to further increase the purity of the child's cluster. It is worth noting that a few children resisted in carrying the recording devices, we had to put the devices on the table. However, the child is still closer to the microphone compared to the clinician in order to utilize the multi-channel energy information for diarization.

We split each speech recording into multiple 3 seconds long segments for increasing the number of training samples. The segment window shift for recordings belong to code "0" class and code "1,2" class in the training dataset are 150 ms and 600 ms, respectively. The reason of using high overlaps between segments in class "0" is to perform data augmentation and reduce the imbalance of training data. There is no overlap on segments in the testing data. The number of these segments for both training and testing data without data selection are shown in Table 4. The energy normalization is used after the diarization at the segment level in order to standardize the input spectrum.

Finally, we perform training and classification on the segment level and use majority voting to generate the person level prediction score.

Table 3
Network structures for our end-to-end approaches.

Network	Detail
RNN	BLSTM with 500 hidden units
	BLSTM with 500 hidden units
	Fully connected layer
CNN	conv1: 16 7 × 7 kernels, 1 stride
	conv2: 32 5 × 5 kernels, 1 stride
	conv3: 64 3 × 3 kernels, 1 stride
	conv4: 128 3 × 3 kernels, 1 stride
	conv5: 128 3 × 3 kernels, 1 stride
	conv6: 256 3 × 3 kernels, 1 stride
	pooling: 3 × 3 pool, 2 × 1 stride
	Fully connected layer
CNN+RNN	conv1: 16 7 × 7 kernels, 1 stride
	conv2: 32 5 × 5 kernels, 1 stride
	conv3: 32 3 × 3 kernels, 1 stride
	conv4: 32 3 × 3 kernels, 1 stride
	pooling: 3 × 3 pool, 2 × 1 stride
	GRU: 500 hidden units or LSTM: 500 hidden units
	Fully connected layer

Table 4
Number of segments in the original and the selected ADOS Module 2 database.

Segment Selection	Partition	Score '0'	Score '1' & '2'
No	Training (93 children)	11107	13341
	Testing (25 children)	109	531
Yes	Training(93 children)	4122	6566
	Testing (25 children)	38	282

3.2. Baseline system

The baseline features are extracted using OpenSMILE (Eyben, 2010), which is a popular open-source toolkit for extracting high-dimensional acoustic and prosodic features at the utterance level. We use the "IS10avic.conf" as the configuration file. This file is designed for Interspeech 2010 paralinguistics challenge (Schuller et al., 2010) and the feature set contains 1584 utterance level features including pitch, loudness, jitter, MFCC, MFB, LSP and their statistical functionals.

We adopted a linear kernel SVM as the supervised classification model. SVM is efficient and can achieve good performance in many applications with limited training data. As a result, we use OpenSMILE features with LibSVM toolkit (Chang and Lin, 2011) as the baseline and make a comparison with our end-to-end deep learning methods.

3.3. End-to-end framework

Many current acoustic features are designed with human perception and domain knowledge. These features may not capture the optimal discriminative information for all kinds of audio classification tasks. In recent years, end-to-end deep learning methods have been shown to be quite successful in multiple audio classification tasks for its superior performance and less demand for domain knowledge. As a result, we try to apply the end-to-end framework in our ASD related atypical prosody classification task as well.

3.3.1. Perception aware spectrograms

For training the network, we extract the constant Q transform (CQT) spectrogram (Lidy and Schindler, 2016) from audio signal as the network input. CQT is initially proposed in the field of music processing (Schrkhuber and Klapuri, 2010). Different from STFT, CQT ensures a constant Q factor across the entire spectrum. CQT spectrogram has a higher frequency resolution at lower frequencies and a higher temporal resolution at higher frequencies. This property can also benefit speech processing and classification tasks, e.g. speech paralinguistics detection (Cai et al., 2017a) and spoofing detection (Cai et al., 2017b).

3.3.2. Modified group delay spectrograms

In addition to the STFT spectrogram, CQT spectrogram, Mel-function energies, Gammatone filter bank energies, these commonly used spectrograms or features, we believe that phase information could also be useful in determining the atypical prosody. As a kind of phase related representation, Modified Group Delay (MGD) spectrograms and cepstral coefficient features (MGDCC) have been used in speech recognition (Hegde et al., 2007; Zhu and Paliwal, 2004), phoneme recognition (Murthy and Gadde, 2003), voice activity detection (Wang et al., 2017b), spoofing detection (Wang et al., 2017a), speaker verification (Hegde et al., 2004; Madikeri et al., 2015; Alam et al., 2015), etc. In this study, we follow the methods in Zhu and Paliwal (2004) to extract the MGD spectrogram for the subsequent end-to-end modeling.

3.3.3. End-to-end deep neural networks

In order to learn the discriminative feature automatically, we test several network setups including convolutional neural network (CNN), recurrent neural networks (RNN) and the combination of these two. Table 3 presents the detail of our network configurations.

The 2-D spectrogram contains a sequence of column vectors along the time axis. We apply two layers of Bidirectional Long Short-Term Memory (BLSTM) cells on these sequences to learn the discriminative representation.

Given an input sequence $\boldsymbol{x} = (x_1, \ldots, x_T)$ and the hidden vector $\boldsymbol{h} = (h_1, \ldots, h_T)$, for a standard recurrent neural networks(RNNs), the output vector $\boldsymbol{y} = (y_1, \ldots, y_T)$ can be computed from $t = 1$ to T according to the following iterative equations:

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{1}$$

$$y_t = W_{ht}h_t + b_y \tag{2}$$

where H is the activation function of hidden layer, W is the weight matrix, and b is the bias vectors.

Bidirectional RNNs (BRNNs) were proposed to make full usage of the context of audio sequences in both preceding and succeeding directions. Furthermore, the Long Short-Term Memory (LSTM) structure with memory blocks was proposed to capture the long term dependencies. Every memory block contains self-connected memory cells and three adaptive and multiplicative gate units i.e. input, output,and forget gates which can respectively provide write, read, reset operations for the cells. Among them, forget gates are shown to be essential for problems involving continual or very long input strings. In this study, we adopt the BLSTM structure since it can deal with long-range context in both forward and backward directions. Furthermore, since BLSTM is just considered to build up high level representation of input features, additional fully-connected layer is needed to map the representation into binary categorical output.

Besides RNN, we also employ the standard CNN architecture with multiple convolution layers followed by the pooling and fully connected layers. This setup has been shown to be very effective in multiple speech paralinguistic attribute recognition tasks (Lozano-Diez et al., 2015; Mao et al., 2014; Milde and Biemann, 2015; Takahashi et al., 2016). Since system input is the CQT spectrogram, CNN can serve as a feature extractor to automatically learn the discriminative feature related to the labels. After pooling, a fixed-dimensional representation is learnt and feed into the subsequent fully connected layer for classification.

Moreover, the combination of CNN and RNN shows great potential recently in speech application such as speech recognition (Sainath et al., 2015), speech verification (Heigold et al., 2016), language identification (Cai et al., 2018) and paralinguistic detection(Schuller et al., 2017). In our system, the 2-D spectrogram is extended to a 3-D tensor with multiple channels (also known as feature maps) after several convolution and pooling layers. Then we run a single LSTM or gated recurrent unit (GRU) layer on 2D slices of that 3-D tensor along the time axis (Harutyunyan and Khachatrian, 2016). After that, the outputs of LSTM or GRU layer are fed to a fully connected layer. The CNN +RNN network structure is illustrated in Fig. 3.

3.4. Segment selection

Although we have multiple 3 seconds long segments for each child, it is not guaranteed that every segment will necessarily carry salient atypical prosody information. In order to improve the fitness of the neural network module with our segment level training samples, we perform leave one subject out cross validation using the CNN method on our segment level training data to filter out non-informative or ”invalid” training segments. The reason to adopt

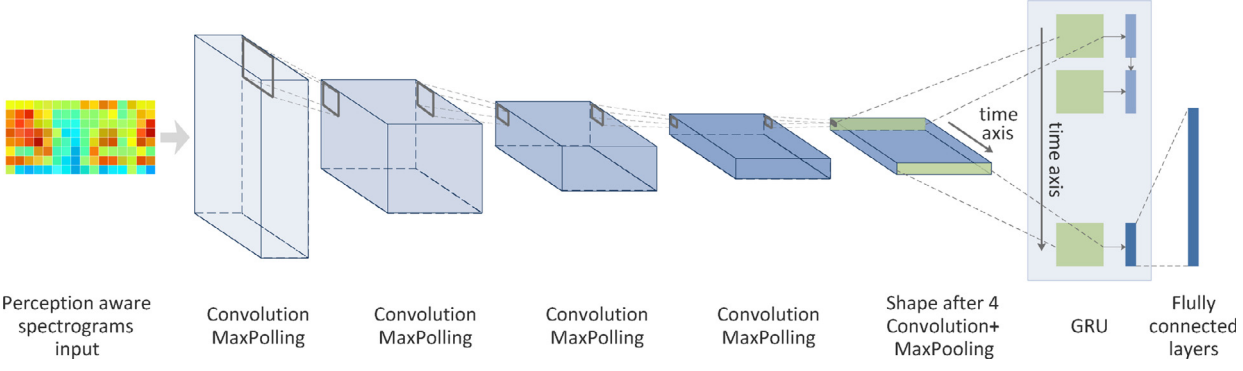


Fig. 3. The CNN+RNN(GRU) network architecture with perception aware spectrograms input. This deep learning network consists of 4 convolution layers, 1 GRU layer and 1 fully connected layer.

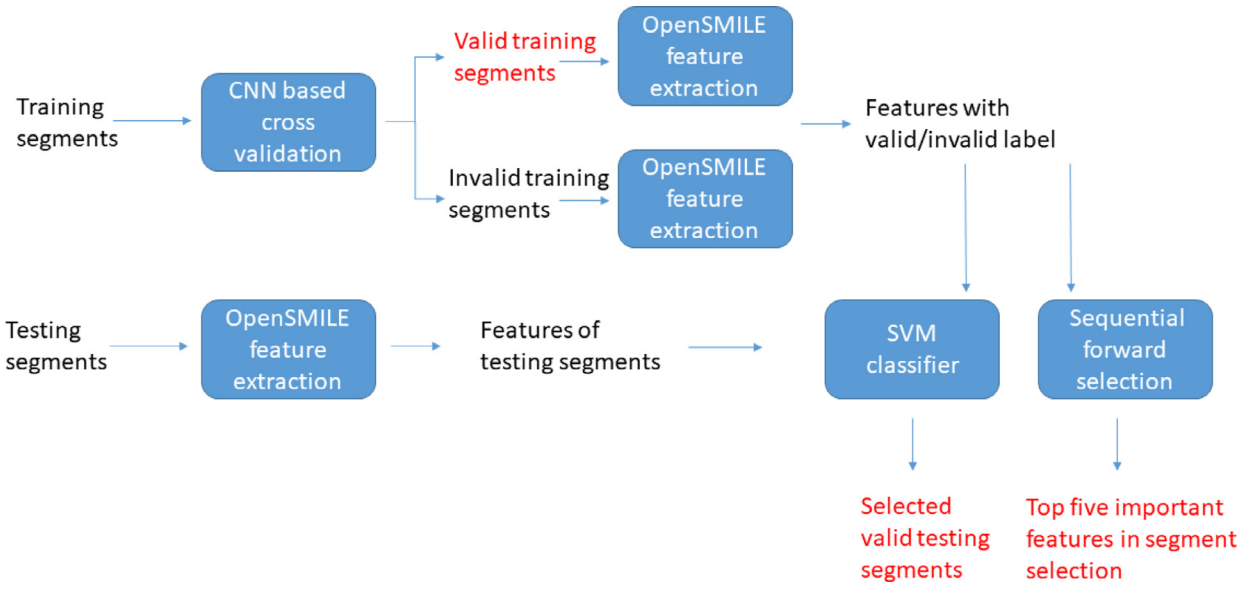


Fig. 4. The overview diagram of proposed segment selection method.

CNN method as the first stage cross validation is because CNN performs the best at the segment level without segment selection. We further train a binary SVM classifier with OpenSMILE features using the overall valid and invalid speech segments of the training set to select the good segments in the testing set. The number of segments in the original and the selected ADOS Module 2 database is shown in Table 4. And the overview diagram of the proposed segment selection method is illustrated in Fig. 4.

In order to find out the reason that some speech segments are more informative than others in the atypical prosody detection task, we use the sequential forward selection (SFS) method on OpenSMILE features to highlight those relevant factors in determining which segment may contain more discriminative information for the atypical prosody detection. The labels for this feature selection is the binary valid/invalid segment tags which are different with the 0–2 atypical prosody codes. Therefore, the selected features are informative in terms of finding valid segments rather than classifying the atypical prosody codes.

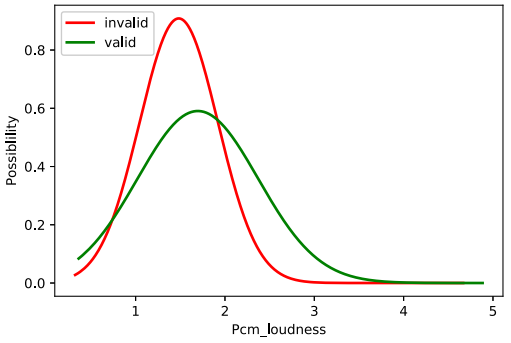
Within those 1582 dimensional features generated by the OpenSMILE toolkit, there are 76 attributes and each attribute has 19 or 21 statistical functionals (Schuller et al., 2010). After we perform SFS on these 76 attributes, the histogram of top five attributes are demonstrated in Fig. 5. We can observe that those selected ("valid") speech segments tend to have higher loudness, Signal to Noise Ratio (SNR) and energies at low frequency filter banks. Therefore, we try to only use these "valid" speech segments with more clear, loud and informative voices for the segment level training and testing, and finally generate the subject level decision.

4. Methods for stereotyped idiosyncratic phrases detection

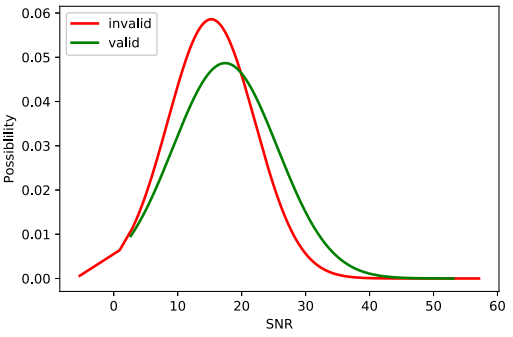
Generally, the stereotyped idiosyncratic phrases detection task can be considered as a supervised text classification problem. Given that the scale of our transcript database is quite small, we adopt several feature extractors that could match with the definition of 'stereotyped/idiosyncratic usage of words or phrases' and the domain expert knowledge from clinicians. After the features are extracted, we use the LibSVM toolkit (Chang and Lin, 2011) with a linear kernel to perform leave one subject out cross validation. In this section, we present the proposed feature extraction methods.

4.1. Language model

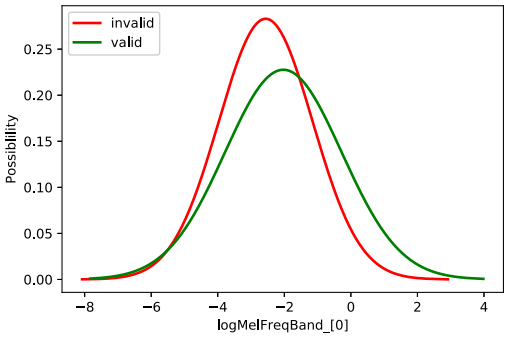
In this work, we build a 3-gram language model on Chinese texts from discussion forums, short messages and instant messaging softwares. This language model is used to examine whether the input utterance transcript is



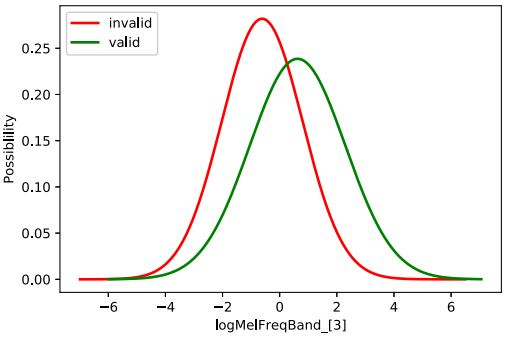
(a) Loudness



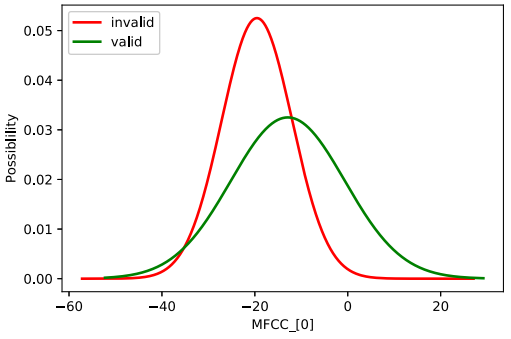
(b) SNR



(c) LMFB[0]



(d)LMFB[3]



(e)MFCC[0]

Fig. 5. Distribution of the top five attributes for distinguishing the quality of speech segment for atypical prosody detection, namely loudness, Signal to Noise Ratio (SNR), Log Mel Frequency Band (LMFB) energies, and Mel-Frequency Cepstral Coefficients (MFCC).

well-organized and whether the words or phrases are appropriate in the context. Also, language model can help us determine whether the children create new words or mutter something random to themselves. We extract the perplexity (PPL) for each training/testing text.

4.2. Dependency treebank

Besides the word trigrams, we also adopt a dependency treebank based parser to examine whether the words and phrases in the text are suitable to their context. Here, we focus more on the phenomenon that words or phrases in the

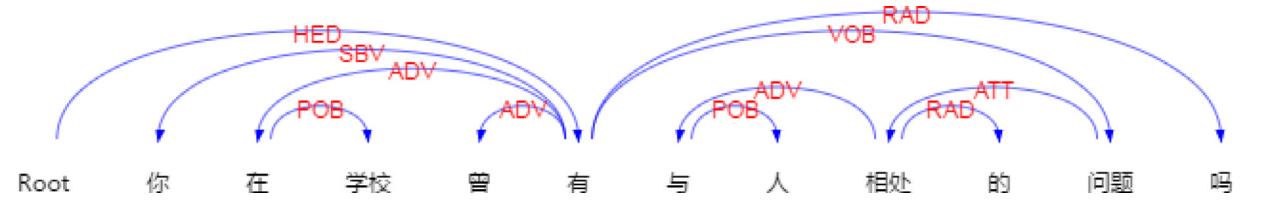


Fig. 6. An example using the mandarin syntactic parser. The English translation of this Chinese sentence is 'Have you ever had any problems with others in school?'.

speech have an improper collocation with other words, which is considered as an important clue in ADOS Module 3. Furthermore, sematic dependency relationship can also be utilized to detect semantic errors (Li et al., 2013).

Compared to n-gram language model, dependency relation could help us understand more about the structure of a sentence. In Chinese, sometimes two related words are not close to each other in the sentence. Therefore some relations can not be directly captured by the the n-gram language model if n is not high. By analyzing the dependency relations between language components, we can potentially capture some complementary information to the n-gram language model.

In this work, we use a large-scale Chinese dependency treebank based syntactic parser on the Language Technology Platform Cloud (Che et al., 2010), developed by Harbin Institute of Technology. Using this syntactic parser, we can get word pairs that are related to each other and the type of their relation. For example, for the sentence ('Have you ever had any problems with others in school?'), the parsing result is shown in Fig. 6.

For an arbitrary input text, we check the number of detected dependency relations that exist in the dependency treebank. A dependency relation exists in the dependency treebank if both the words and the type of the relation are the same as a relation in the treebank and this relation must appear more than α times in the treebank. α serves as a threshold to reduce the noise.

4.3. TF−IDF and cosine similarity

Another common symptom for ASD children is the echolalia. It is defined as the unsolicited repetition of vocalizations made by another person. The echolalia is categorized as immediate echolalia and delayed echolalia. Delayed echolalia is an important factor in evaluating stereotyped idiosyncratic usage of words or phrases (Lord et al., 2000; Grzadzinski et al., 2016). In the ADOS Module 3, the A4 code is "Stereotyped /Idiosyncratic Use of Words or Phrases". This coding item includes delayed echolalia and other highly repetitive phrases with consisted tone patterns, and these phrases are used in an inappropriate formal way. Therefore in our study, we also check the similarity between each sentence and its context.

The combination of Term Frequency−Inverse Document Frequency (TF−IDF) and cosine similarity is a common way to detect the similarity of two texts. We use the TF−IDF method to convert the sentences from the input text into vectors. Then the cosine similarity is applied to calculate the score for every sentence pair. We further extract two features, namely the mean of the similarity scores and the percentage of the sentence pairs whose scores are above a certain threshold.

5. Experimental results

5.1. Results for the atypical prosody detection task

In this section, we compare the classification results between the OpenSMILE+SVM baseline and our proposed end-to-end approaches. Besides the 0/1/2 three categories classification, we also perform binary classification by merging the code 1 and 2 together as a new class to enhance the practical usability. Moreover, we also show the results with segment selection and score level fusion. The details of the database and evaluation protocol are presented in Section 2 and Table 4.

5.1.1. Evaluation measure

The evaluation measure for our classification task is the unweighted average recall (UAR). UAR is defined as follows:

$$UAR = \frac{1}{n}\Sigma_{i=1}^{n}\frac{N_i^{'}}{N_i} \tag{3}$$

where n, N_i and $N_i^{'}$ denote the number of classes, the number of samples that belong to class i and the number of correctly classified samples that belong to class i, respectively.

The reason to apply unweighted rather than weighted average recall (i.e. accuracy) is that our dataset has relatively unbalanced distribution among different classes. UAR is more reasonable and meaningful for this kind of tasks. We use the 0/1/2 codes provided by clinicians as the ground truth labels.

5.1.2. Three categories classification

The baseline system takes the 1584 dimensional OpenSMILE feature vector as the input and predict the category for each utterance using a SVM classifier. Neural network takes CQT spectrogram as the input and predict categories with the softmax layer. As shown in Table 5, our proposed CNN method outperforms the SVM baseline and RNN network at the segment level (from 52.3% to 57.4%) while SVM baseline achieves the highest UAR 65.8% (by chance 33%) at the person level. After score level fusion, the segment level UAR improves to 58.4%. However, this is still far away from practical usage. This might be due to the blured and subjective boundary between code 1 and 2 in terms of the severity levels. Another possible reason could be the unbalanced data distribution among different categories.

From Table 5, we can also observe that CNN performs better than RNN here at the segment level. This might be because there are less parameters in the CNN structure and therefore less likely to be overfitted with small scale training data.

Furthermore, human annotators sometimes exhibit subjective behavior in the coding, which makes our ground truth labels less consistent. Generally, clinicians are less confident to distinguish between slightly and seriously abnormal compare to determine the existence of abnormality. This motivates us to investigate the two categories binary classification.

5.1.3. Two categories classification

We merge the instances with code 1 and 2 together to perform a binary classification of atypical prosody (code 0 vs union(1,2)). Experimental results are shown in Table 6. We evaluate SVM, CNN, RNN and CNN+RNN these four systems. At the segment level, both RNN and CNN methods outperform the SVM baseline. We also find out that CNN and RNN networks perform better than the hybrid system (CNN+RNN in Table 3 and Fig. 3). The reason might be over-fitting. Without large-scale training data, complex network architecture can easily become over-fitted.

As our goal in real application is to detect the speech abnormalities for each child, we further calculate the UAR at the person level. As you can see from Table 6, the UAR(per) of the SVM baseline and the RNN network are very close. When we fuse these two systems together, we achieve 79.3% UAR at the segment level, however, there is no additional gain at the person level which might be due to the limited number of children (person level decisions) as well as the unbalanced number of segments among different children in our database.

Table 5
Three categories classification results on the testing set for the atypical prosody detection task(UAR(seg) stands for calculating UAR with respect to segment, UAR(per) stands for calculating UAR with respect to person).

ID	Model	Inputs	UAR (seg)	UAR (per)
1	SVM	OpenSMILE features	52.3%	**65.8%**
2	RNN	CQT spectrogram	49.1%	60.6%
3	CNN	CQT spectrogram	**57.4%**	62.1%
4	Score level Fusion(1+2)		56.3%	59.1%
5	Score level Fusion(1+3)		**58.4%**	59.4%

Table 6
Two categories (binary) classification results on testing set for the atypical prosody detection task.

ID	Methods	Inputs	UAR (seg)	UAR (per)
1	SVM	OpenSMILE features	76.9%	**85.7%**
2	RNN	CQT spectrogram	78.4%	83.3%
3	CNN	CQT spectrogram	78.2%	**85.7%**
6	CNN+RNN(LSTM)	CQT spectrogram	77.4%	83.3%
7	CNN+RNN(GRU)	CQT spectrogram	69.8%	72.0%
4	Score level Fusion(1+2)		**79.3%**	83.3%

Moreover, from Table 7, we can observe that although phase information related MGD spectrogram is not as good as the CQT spectrogram with the RNN modeling, their score level fusion could improve the overall system performance at the segment level (85.7% UAR after segment selection).

5.1.4. Segment selection

Table 7 shows the results of SVM baseline and our proposed RNN system at both with and without segment selection conditions. We can observe that our proposed segment selection method improves the UAR performance dramatically at both segment level and person level. Therefore, although there are only around 50% of segments remaining in the training and testing data after segment selection, they carry more discriminative information about the atypical prosody and make the person level decision more accurate. Furthermore, our RNN system achieves 84.4% and 88.1% UAR at the segment and person level which outperforms the SVM baseline before segment selection by 7.5% and 2.7%, respectively (Table 8).

5.2. Results for the stereotyped idiosyncratic phrases detection task

As we mentioned in Section 2.2 and Table 2, our ADOS Module 3 dataset contains conversational transcripts from 71 children. After extracting the text features, we adopt the linear kernel SVM as the classifier due to the limited number of training samples. The results of each proposed feature set as well as the baseline in Kumar et al.

Table 7
Two categories (binary) classification results on testing set for the atypical prosody detection task.

		Without segment selection		With segment selection	
ID	Methods	UAR (seg)	UAR (per)	UAR (seg)	UAR (per)
1	SVM	76.9%	85.7%	81.2%	88.1%
2	RNN (CQT spectrogram)	78.4%	83.3%	84.4%	**88.1%**
8	RNN (MGD spectrogram)	67.3%	72.0%	74.9%	77.9%
4	Score level Fusion(1+2)	**79.3%**	83.3%	84.7%	85.7%
9	Score level Fusion(2+8)	78.03%	83.3%	**85.7%**	88.1%

Table 8
The segment level confusion matrix of the fused system (CQT-RNN + MGD-RNN, ID 9 in Table 7) with segment selection on two categories classification for the atypical prosody detection task.

		Predicted code		
		0	1	Total
Ground truth diagnosis code	0	38	0	38
	1	77	192	269
	Total	115	192	307

Table 9
Three Categories classification results for the stereotyped idiosyncratic phrases detection task.

		Three Categories code 0 vs 1 vs 2			
ID	Features	Accuracy	UAR	Precision	F1-score
1	n-gram language model	66.7%	50.8%	40.2%	44.9%
2	dependency treebank	66.7%	42.1%	39.5%	40.7%
3	TF–IDF	57.2%	46.0%	34.9%	39.7%
4	LIWC (Kumar et al., 2016)	42.8%	44.8%	42.5%	43.6%
5	Maximum Entropy (Kumar et al., 2016)	65.1%	35.5%	33.3%	34.4%
6	n-gram model+dependency treebank +TF–IDF + LIWC	**52.4%**	**55.6%**	**47.8%**	**51.4%**

(2016) are presented in Table 9 and Table 10, respectively. We can observe that combining multiple features together enhances the system performance dramatically.

In the three categories classification task, our proposed method (ID 6) achieves 52.4% accuracy and 55.6% UAR, respectively. In the two categories classification task, our proposed method (ID 6) achieves 85.7% accuracy and 77.8% UAR which outperforms the n-gram language model, maximum entropy classifier and Linguistic Inquiry and Word Count (LIWC) based baseline (Kumar et al., 2016). The confusion matrix is shown in Table 11. In future, we will combine multiple quantitative scores from different speech and language abnormity detection tasks with video or eye tracking based objective measures together to distinguish between ASD and typical developed children.

6. Conclusions and future works

In this paper, we present an automated assessment framework in quantifying atypical prosody and stereotyped idiosyncratic phrases related to ASD. We collected an audio database during the ADOS screening sessions, the Module 2 part consists of 118 children while the Module 3 part includes 71 children. For detecting the atypical prosody, the proposed end-to-end deep learning methods achieve superior performance at the segment level, but not at the person-level. The cross validation based segment selection method improves the accuracy at both segment and person level. Score level fusion further enhance the segment level system performance. For detecting the stereotyped idiosyncratic phrases, we adopt language model, dependency treebank, term frequency–inverse document frequency and LIWC these methods to extract a set of text features followed by a standard SVM backend. Experimental results

Table 10
Two Categories classification results for the stereotyped idiosyncratic phrases detection task.

		Two Categories code 0 vs union(1,2)			
ID	Features	Accuracy	UAR	Precision	F1-score
1	n-gram language model	81.0%	47.2%	42.5%	44.7%
2	Dependency treebank	76.2%	72.2%	63.3%	67.5%
3	TF–IDF	81.0%	75.0%	66.9%	70.7%
4	LIWC(Kumar et al., 2016)	76.2%	58.3%	56.6%	57.5%
5	Maximum Entropy(Kumar et al., 2016)	80.1%	47.2%	42.5%	44.7%
6	n-gram model+dependency treebank +TF–IDF + LIWC	**85.7%**	**77.8%**	**72.1%**	**74.8%**

Table 11
The confusion matrix of the fusion system (n-gram model+dependency treebank+TF-IDF+LIWC) on two categories classification for the stereotyped idiosyncratic phrases detection task.

		Predicted code		
		0	1	Total
Ground truth diagnosis code	0	2	1	3
	1	2	16	18
	Total	4	17	21

show that the proposed feature sets are effective and better results are achieved by merging code 1 and 2 together as a two categories binary classification task.

In future works, we definitely need to collect more data at the person level. Furthermore, we will pay more attention to the child-therapist interaction at the controlled environment (e.g. ADOS) as well as the child-parents interaction in free living condition at home. Last but not the least, the 'ground truth' codes of our current ADOS data come from the ADOS screening results performed by one doctor, which may be subjective. If we can obtain labels from multiple clinicians for each recording, it could possibly reduce the subjective variabilities, improve the robustness and compare the systems performance with the human agreement level.

Acknowledgments

This research was funded in part by the National Natural Science Foundation of China (61773413,81873801,81601533), Natural Science Foundation of Guangzhou City (201707010363), Guangdong Science and Technology Program for Industrial Development (20160914), Six talent peaks project in Jiangsu Province (JY-074) and National Key Research and Development Program (2016YFC0103905).

References

Alam, M.J., Kenny, P., Stafylakis, T., 2015. Combining amplitude and phase-based features for speaker verification with short duration utterances. In: Proceeding of Interspeech.

Black, M.P., Katsamanis, A., Baucom, B.R., Lee, C.-C., Lammert, A.C., Christensen, A., Georgiou, P.G., Narayanan, S.S., 2013. Toward automating a human behavioral coding system for married couples interactions using speech acoustic features. Speech Commun. 55 (1), 1–21.

Bone, D., Black, M.P., Lee, C.-C., Williams, M.E., Levitt, P., Lee, S., Narayanan, S., 2012. Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist. In: Proceeding of Interspeech, pp. 1043–1046.

Bone, D., Black, M.P., Ramakrishna, A., Grossman, R.B., Narayanan, S.S., 2015. Acoustic-prosodic correlates of 'awkward' prosody in story retellings from adolescents with autism. In: Proceeding of Interspeech, pp. 1616–1620.

Bone, D., Chaspari, T., Narayanan, S., 2017. Chapter 15 behavioral signal processing and autism: Learning from multimodal behavioral signals. Autism Imaging and Devices.

Cai, D., Ni, Z., Liu, W., Cai, W., Li, G., Li, M., Cai, D., Ni, Z., Liu, W., Cai, W., 2017a. End-to-end deep learning framework for speech paralinguistics detection based on perception aware spectrum. In: Proceeding of Interspeech, pp. 3452–3456.

Cai, W., Cai, D., Liu, W., Li, G., Li, M., 2017b. Countermeasures for automatic speaker verification replay spoofing attack : On data augmentation, feature representation, classification and fusion. In: Proceeding of Interspeech, pp. 17–21.

Cai, W., Cai, Z., Liu, W., Wang, X., Li, M., 2018. Insights into end-to-end learning scheme for language identification. In: Proceeding of ICASSP.

Chang, C.C., Lin, C.J., 2011. LIBSVM: A Library for Support Vector Machines. ACM.

Chaspari, T., Provost, E.M., Katsamanis, A., Narayanan, S., 2012. An acoustic analysis of shared enjoyment in eca interactions of children with autism. In: Proceeding of ICASSP, pp. 4485–4488.

Che, W., Li, Z., Liu, T., 2010. Ltp: A chinese language technology platform. J. Chin. Inform. Process. 2 (6), 13–16.

Christensen, D.L., Bilder, D.A., Zahorodny, W., Pettygrove, S., Durkin, M.S., Fitzgerald, R.T., Rice, C., Kurzius-Spencer, M., Baio, J., Yeargin-Allsopp, M., 2016. Prevalence and characteristics of autism spectrum disorder among 4-year-old children in the autism and developmental disabilities monitoring network. J. Develop. Behav. Pediat. 37 (1), 1–8.

Dubey, H., Kaushik, L., Sangwan, A., Hansen, J.H., 2016. A speaker diarization system for studying peer-led team learning groups. In: Proceeding of Interspeech, pp. 2180–2184.

Eyben, F., 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceeding of ACM International Conference on Multimedia, pp. 1459–1462.

Fusaroli, R., Lambrechts, A., Bang, D., Bowler, D.M., Gaigg, S.B., 2016. "is voice a marker for autism spectrum disorder? a systematic review and meta-analysis". Autism Res. 10 (3).

Gong, J.J., Gong, M., Levy-Lambert, D., Green, J.R., Hogan, T.P., Guttag, J.V., 2016. Towards an automated screening tool for developmental speech and language impairments. In: Proceeding of Interspeech, pp. 112–116.

Gotham, K., Pickles, A., Lord, C., 2009. Standardizing ados scores for a measure of severity in autism spectrum disorders. J. Aut. Develop. Disorders 39 (5), 693–705.

Grzadzinski, R., Dick, C., Lord, C., Bishop, S., 2016. Parent-reported and clinician-observed autism spectrum disorder (asd) symptoms in children with attention deficit/hyperactivity disorder (adhd): implications for practice under dsm-5. Molecular Aut. 7 (1), 7.

Harutyunyan, H., Khachatrian, H., 2016. Combining cnn and rnn for spoken language identification. In: http://yerevann.github.io/2016/06/26/combining-cnn-and-rnn-for-spoken-language-identification/.

Hegde, R.M., Murthy, H.A., Gadde, V.R.R., 2007. Significance of the modified group delay feature in speech recognition. IEEE Trans. Audio, Speech, Lang. Process. 15 (1), 190–202.

Hegde, R.M., Murthy, H.A., Rao, G.R., 2004. Application of the modified group delay function to speaker identification and discrimination. In: Proceeding of ICASSP, 1, pp. I–517.

Heigold, G., Moreno, I., Bengio, S., Shazeer, N., 2016. End-to-end text-dependent speaker verification. In: Proceeding of ICASSP, pp. 5115–5119.
Kumar, M., Gupta, R., Bone, D., Malandrakis, N., Bishop, S., Narayanan, S.S., 2016. Objective language feature analysis in children with neurodevelopmental disorders during autism assessment. In: Proceeding of Interspeech, pp. 2721–2725.
Li, J., Zhang, Y., Zhu, J., Zhang, Z., 2013. Semantic automatic error-detecting for chinese text based on semantic dependency relationship. In: Proceeding of Workshop on Chinese Lexical Semantics, pp. 406–415.
Lidy, T., Schindler, A., 2016. Cqt-based convolutional neural networks for audio scene classification and domestic audio tagging. In: Proceeding of 2016 Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2016).
Liu, W., Zhou, T., Zhang, C., Zou, X., Li, M., 2017. Response to name: A dataset and a multimodal machine learning framework towards autism study. In: Proceeding of Affective Computing and Intelligent Interaction (ACII), pp. 178–183.
Lord, C., Risi, S., Lambrecht, L., Cook, E.J., Leventhal, B., DiLavore, P., Pickles, A., Rutter, M., 2000. The autism diagnostic observation schedułełgeneric: A standard measure of social and communication deficits associated with the spectrum of autism. J. Autism Develop. Disorders 30 (3), 205.
Lozano-Diez, A., Zazo-Candil, R., Gonzalez-Dominguez, J., Toledano, D.T., Gonzalez-Rodriguez, J., 2015. An end-to-end approach to language identification in short utterances using convolutional neural networks. In: Proceeding of Interspeech.
Madikeri, S.R., Talambedu, A., Murthy, H.A., 2015. Modified group delay feature based total variability space modelling for speaker recognition. Int. J. Speech Technol. 18 (1), 17–23.
Mao, Q., Dong, M., Huang, Z., Zhan, Y., 2014. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans. Multim. 16 (8), 2203–2213.
McCann, J., Peppé, S., 2003. Prosody in autism spectrum disorders: a critical review. Int. J. Lang. Commun. Disorders 38 (4), 325–350.
Milde, B., Biemann, C., 2015. Using representation learning and out-of-domain data for a paralinguistic speech task. In: Proceeding of Interspeech, pp. 904–908.
Murthy, H.A., Gadde, V., 2003. The modified group delay function and its application to phoneme recognition. In: Proceeding of ICASSP, 1. IEEE, pp. I–68.
Narayanan, S., Georgiou, P.G., 2013. Behavioral signal processing: Deriving human behavioral informatics from speech and language. Proc. IEEE 101 (5), 1203–1233.
Sainath, T.N., Vinyals, O., Senior, A., Sak, H., 2015. Convolutional, long short-term memory, fully connected deep neural networks. In: Proceeding of ICASSP, pp. 4580–4584.
Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., et al., 2017. The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring. In: Proceeding of Interspeech, pp. 3442–3446.
Schuller, B.W., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C.A., Narayanan, S.S., et al., 2010. The interspeech 2010 paralinguistic challenge. In: Proceeding of Interspeech, 2010, pp. 2795–2798.
Schrkhuber, C., Klapuri, A., 2010. Constant-q transform toolbox for music processing. In: Proceeding of 7th Sound and Music Computing Conference.
Takahashi, N., Gygli, M., Pfister, B., Van Gool, L., 2016. Deep convolutional neural networks and data augmentation for acoustic event detection. Proc. Interspeech 2982–2986.
Wang, L., Nakagawa, S., Zhang, Z., Yoshida, Y., Kawakami, Y., 2017a. Spoofing speech detection using modified relative phase information. IEEE J. Select. Topics Signal Process. 11 (4), 660–670.
Wang, L., Phapatanaburi, K., Go, Z., Nakagawa, S., Iwahashi, M., Dang, J., 2017b. Phase aware deep neural network for noise robust voice activity detection. In: Proceeding of ICME, pp. 1087–1092.
Xiao, B., Can, D., Gibson, J., Imel, Z.E., Atkins, D.C., Georgiou, P., Narayanan, S., 2016. Behavioral coding of therapist language in addiction counseling using recurrent neural networks. In: Proceeding of Interspeech, pp. 908–912.
Xiao, B., Ghosh, P.K., Georgiou, P., Narayanan, S.S., 2011. Overlapped speech detection using long-term spectro-temporal similarity in stereo recording. In: Proceeding of ICASSP, pp. 5216–5219.
Xiao, B., Imel, Z.E., Georgiou, P.G., Atkins, D.C., Narayanan, S.S., 2015. “rate my therapist”: Automated detection of empathy in drug and alcohol counseling via speech and language processing. PloS one 10 (12), e0143055.
Zhou, T., Cai, W., Chen, X., Zou, X., Zhang, S., Li, M., 2016. Speaker diarization system for autism children’s real-life audio data. In: Proceeding of ISCSLP, pp. 1–5.
Zhu, D., Paliwal, K.K., 2004. Product of power spectrum and group delay function for speech recognition. In: Proceeding of ICASSP, 1, pp. I–125.