An Automated Assessment Framework for Speech Abnormalities related to Autism Spectrum Disorder

Conference Paper · June 2017

CITATIONS
2

READS
6

4 authors, including:

Ming Li
Duke Kunshan University
80 PUBLICATIONS **828** CITATIONS

SEE PROFILE

An Automated Assessment Framework for Speech Abnormalities related to Autism Spectrum Disorder

Tianyan Zhou[1], *Yixiang Xie*[12], *Xiaobing Zou*[2], *Ming Li*[1]

[1]SYSU-CMU Joint Institute of Engineering, School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China
[2]The Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

liming46@mail.sysu.edu.cn, zouxb@vip.tom.com

Abstract

Autism Spectrum Disorders (ASD), a neurodevelopmental disability, has become one of the high incidence diseases among children. Studies indicate that early diagnosis and intervention treatments help to achieving positive longitudinal outcomes. In this paper, we focus on the speech abnormalities of young children with ASD and present an automated assessment framework to assist clinicians in quantifying atypical prosody related to ASD. First, we use the openSmile toolkit to extract utterance level large 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 speech spectrogram. Third, we fuse the deep learning framework with the conventional baseline at the score level to further improve the system performance. We collect a database of spontaneous speech recorded during the Autism Diagnostic Observation Schedule (ADOS) Modules 2 tasks. This database consists of 70 children and 58 of them are diagnosed as ASD with different severity. Experimental results on this database show that our proposed methods can effectively predict the atypical prosody score for young children with the risk of ASD.

Index Terms: atypical prosody, deep learning, autism spectrum disorder, recurrent neural network

1. Introduction

Autism Spectrum Disorders (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. The latest analysis from the Centers for Disease Control and Prevention estimates that 1 in 68 children has ASD in the United States [1]. Early behavioral/educational interventions have been proved to be very successful in many clinical studies. This attaches great significance to the recognition of common ASD signs and make a diagnosis at the early stage.

In paralinguistics, prosody relates to several communicative functions such as intonation, tone, pitch, stress, and rhythm. Prosody can reflect many important elements of language including the emphasis, contrast and affective state of the speaker [2]. These are all critical information in human communication. Nevertheless, speech abnormalities, or say atypical prosody is one of the common symptoms reported for ASD. Specifically, children with ASD may speak in flat, robot-like or a sing-song voice, which is an important sign the clinicians should consider during the diagnosis.

The Autism Diagnostic Observation Schedule (ADOS) is a standard instrument to help clinicians observe children's languages and behaviors relevant to the diagnosis of autism. It consists of a series of structured and semi-structured tasks assessing social interaction, communication, play, and imaginative use of materials [3]. There are four modules designed to be performed according to subject's language capability. Moreover, speech abnormality is an observation item listed in all the modules. The ADOS screening provides a code to quantify this item on an integer scale from '0' to '2', with '0' representing appropriate prosody; '1' standing for some changes on pitch/tone, a bit flat/exaggerate intonation, slightly abnormal volume, a little slow/fast/jerky rhythm; '2' designating markedly and consistently abnormalities on the aforementioned aspects.

In the ADOS screening, therapist identifies subject's atypical prosody level and gives a code. 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 problem is the subjective disagreements between clinicians, which makes the results become ambiguous. Researchers have proposed strategies to utilize signal processing techniques to support clinicians with quantitative analysis [4, 5, 6] of ASD children's prosody. However, their experiments mainly focus on children and adolescents with higher ages and better language capabilities under the ADOS Module 3 setup. Furthermore, since pattern recognition and machine learning methods have presented promising performance in modeling behavior symptom and relationship with expert's experience [7, 8], some automated screening and evaluating tools based on the objective features extracted directly from recording are proposed [9, 10]. These automated coding tools showed potential to be scalable and assist clinicians to analyze the variation trend of a specific symptom during long term tracking assessments.

In this paper, we focus on the speech abnormalities and present an automated assessment framework to distinguish the severity level of atypical prosody for young children under the ADOS Module 2 setup. Specifically, we model the speech abnormalities using both the traditional strategy and the deep learning framework. We demonstrate that the end-to-end techniques can achieve similar 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 the end-to-end approach, there is no prior in-domain knowledge required for feature engineering. Moreover, the fusion of the two systems can further improve the overall system performance 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) [11]. We also investigate the cutoff boundary of the code 0/1/2 by merging 0/1 or 1/2 as a new code. Classification results indicate that our

Table 1: *Demographic statistics ADOS module 2.*

Items	Count/Statistics
Age(months)	Range: 26-125, Mean: 56.31
Gender	Male: 60, Female: 10
Language	Mandarin: 65, Cantonese: 5
Atypical code(subjects)	'0': 12, '1': 25, '2': 33
Diagnosis	Autism: 48, ASD: 9, below ASD cutoffs: 13

algorithms are more confident to identify normal prosody from abnormal prosody, which matches with the clinicians that the boundary between slightly and serious abnormal is not so clear.

The rest of this paper is organized as follows. Section 2 describes our data corpus. Section 3 presents the specific methods of our two different systems. Experimental setup and results are provided in Section 4. Section 5 presents the conclusions and our future works.

2. Dataset Description

We perform experiments on our in-house collected dataset. Our audio database is collected in the real ADOS module 2 screening environment. As you can observe from Figure 1, our multimodal behavior signal capture system is equipped with multiple HD cameras and Kinect sensor to capture vision data while child-psychologist interactions are conducting. As for audio data, every participant wears a recording device (presented in Figure 1) 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. The data collection is approved by the children's family, the institutional review board (IRB) of the hospital and the university.

There are many 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 "Demonstration Task", "Description of a Picture", "Telling a Story From a Book" and "Birthday Party" these four tasks. They are either designed to observe children's spontaneous and expressive language capabilities or contain a large proportion of speech conversations.

The demographic statistics of our database are showed in Table 1. For module 2, our dataset contains 70 children with 13 of them below the ASD cut-off point. However, a few ASD children can still have normal prosody and some normal children may also have atypical prosody. Our focus is to predict the atypical prosody code rather than the ASD label.

Figure 1: *Data collecting environment and recording device.*

3. Method

The baseline system is implemented using the OpenSmile feature extractor followed by a Support Vector Machine (SVM) paradigm. Our end-to-end deep learning framework uses the spectrograms information as the input, and does supervised learning through neural networks. Section 2.2 and 2.3 introduce these two systems in detail, respectively.

3.1. Multi-channel Speaker Diarization

Our database is collected using several recording devices carried by each participant. 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 [12]. 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 [13] to further purify the speech.

3.2. Baseline System

The baseline features are extracted using openSMILE, which is a popular open-source toolkit for extracting large-dimensional acoustic and prosodic features. We use the "IS10avic.conf" as the configuration file. This file is designed for Interspeech 2010 paralinguistics challenge [14] and the feature set contains 1584 utterance level features including pitch, loudness, jitter, MFCC, MFB, LSP and statistical functionals.

We adopted SVM with linear kernels to train the supervised classification model. SVM is efficient and can achieve satisfying performance in many applications with limited training data. As a result, we use it to represent the baseline and make a comparison with the deep learning framework.

3.3. End-to-end Framework

The existing acoustic features are basically proposed according to human perception and experience. These features may not capture the optimal discriminative information among all classification tasks. In recent years, the deep learning method has been proved to be quite successful in acoustic modeling and audio classification. As a result, we intend to apply the end-to-end framework on our ASD related atypical prosody classification task.

3.3.1. Perception Aware Spectrograms

It seems spectrograms are the standard way to represent audio for deep learning systems [15, 16]. For training the network, we extract the spectrogram from audio data as the network input.

We test two different spectrograms. The first one is the traditional short-time Fourier transform (STFT) spectrogram. The second one is the constant Q transform (CQT) spectrogram [17]. It is initially proposed in the field of music processing. Different from STFT, CQT ensures a constant Q factor across the entire spectrum. This change can bring a benefit for processing human speech, which is a higher frequency resolution at lower frequencies and a higher temporal resolution at higher frequencies.

Table 2: *Deep learning network configurations (conv: convolutional layer).*

Network	Detail
Combine CNN and RNN	conv1: 16 7×7 kernels, 1 stride conv2: 32 5×5 kernels, 1 stride conv3: 32 3×3 kernels, 1 stride pooling: 3×3 pool, 2×1 stride GRU: 500 hidden units Classification layer
RNN	GRU with 500 hidden units GRU with 500 hidden units Classification layer

3.3.2. 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 them. However, experimental results show that a single CNN cannot work well on our dataset which might due to the dynamic nature of prosody, so we only describe the other two setups.

The combination of CNN and RNN shows great potential recently in speech application such as speech recognition [18] and speech verification [19]. Since the system input is STFT/CQT spectrogram, CNN can serve as the feature extraction tool, and the output is then fed forward into a RNN. 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 gated recurrent unit (GRU) layer on 2D slices of that 3-D tensor along the time axis [20]. We also attempt to run separate GRUs on each channel, and these GRUs may share or not share weights. Experimental results demonstrate that a single GRU works better than multiple GRUs on our dataset which might due to the lack of training data . After that, the outputs of GRU are fed to a fully connected layer.

Another setup with good performance is the RNN itself. The 2-D spectrogram contains a sequence of column vectors along the time axis. We apply two layers of GRU cells on these sequence to learn discriminative information. Table 2 presents the detail of our network configurations.

4. Experimental Results

In this part, we will compare the classification results between the openSMILE+SVM baseline system and our proposed end-to-end framework. Besides the 0/1/2 three categories classification, we also investigate the cutoff boundary of the code by performing two categories classification. Both STFT/CQT spectrograms are evaluated as the network input.

For module 2, our dataset contains 70 children. In order to train the classification model and test the system performance, we separate them with 45 in training set and 25 in testing set (no person overlap). Both set have the similar 0/1/2 code distribution.

4.1. Evaluation Measure

The evaluation measure for our classification task is the unweighted average recall (UAR). UAR is defined as follows where n is the number of classes.

$$UAR = \frac{1}{n}\Sigma_{i=1}^{n}\frac{N_{prediction=i}}{N_{label=i}} \quad (1)$$

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 reliable and meaningful for this kind of tasks. We use the 0/1/2 code given by clinician as the ground truth label.

4.2. Three Categories Classification

There are 45 children in the training set, and the average audio length for each child after voice activity detection and multi-channel diarization is 135 seconds. In order to increase the number of training instances, we split each recording to a group of 3 seconds long short segments. The time shift between each segment can be tuned to further augment training set and balance the category distribution. For the three categories classification, we have 8832 segments in training set and each has 80% overlaps with the previous segment. This means five times data augmentation. Testing set contains 629 segments without inner overlap.

The baseline system takes the 1584 dimensional feature vector as the input and predict the category for each utterance. Neural network processes the 256186 STFT spectrogram or 863352 CQT spectrogram and predict categories as well. As you can see from Figure 3, SVM achieves the highest UAR 50% (by chance 33%), which is still far away from satisfaction. This might be due to the unclear and subjective boundary between code 1 and 2 in terms of the severity levels. Another possible reason could be the unbalanced distribution among categories. The number of code 0,1 and 2 in the training set is 1309,4595 and 2928, respectively. Neural networks may not learn enough information for code 0. We also notice that RNN works better than the combination of CNN and RNN. The reason might be the over-fitting problem. Without large-scale training data, we cannot drive CNN to learn a proper feature representation. Under the circumstances, complex network architecture can easily become over-fitting. Moreover, it seems that CQT spectrograms is not as effective than STFT spectrograms in our task which is different as reported in [21]. This might be because the dynamic nature of prosody information and the low time domain resolution for low frequencies may degrade the performance.

As mentioned earlier in Section 1, human tagging usually exhibits subjective variation among each other, which makes our ground truth labels become ambiguous. We also discussed with the clinicians about how to distinguish from code 0/1/2. The answer is actually they are less confident to distinguish between slightly/serious abnormal than with/without abnormality. This leads us to investigate the two categories classification.

4.3. Two Categories Classification

We test two partition manners towards the code 0/1/2.

First, we merge the instances with code 1 and 2 in the training set to perform a two categories classification between with/without atypical prosody (0 vs union(1,2)). Experimental results are shown in Table 4. We also evaluate SVM, CNN+RNN and RNN these three systems. To our surprise, RNN achieves 78.75% UAR with respect to segment, which is higher than the baseline system. Given the fact that as a automated assessment framework, our goal is provide a speech abnormalities code for each subject. We further calculate the UAR with respect to person, which is much meaningful and valuable

Table 3: *Three categories classification results on testing set (UAR(seg) stands for calculating UAR with respect to segment).*

Model	Inputs	UAR (seg)
SVM	OpenSmile features	50.5%
CNN + RNN	STFT spectrogram	34.62%
	CQT spectrogram	35.48%
RNN	STFT spectrogram	45.62%
	CQT spectrogram	36.56%

Table 4: *Two categories classification results on testing set (UAR(per) stands for calculating UAR with respect to person).*

Model	Partition	UAR (seg)	UAR (per)
SVM	0 and 1/2	**74.7%**	**88.1%**
	0/1 and 2	64.6%	58.44%
CNN + RNN	0 and 1/2	69.85%	72%
	0/1 and 2	53.56%	59.41%
RNN	0 and 1/2	**78.75%**	**85.71%**
	0/1 and 2	59.98%	68.58%
fuse SVM&RNN	0 and 1/2	**81%**	**90%**
	0/1 and 2	61.79%	65.58%

in real scenario. As you can see from Table 4, the UAR(per) of SVM and RNN are very close. When we fuse these two systems, we get the highest UAR both with respect to segment and person (i.e. 81% and 90%). This result demonstrates the end-to-end framework has great potential in our task.

Then we merge the instances with code 0 and 1 in the training set to perform a two categories classification as well (union(0,1) vs 2). The results are also showed in Table 4. Comparing to the first partition manner, none of these systems can achieve satisfying performance.

The results of these two partition manners probably reveal the phenomenon that the boundary between code 1/2 are less clear than code 0/1 in our dataset.

5. Conclusions and Future Work

In this paper, we focus on the speech abnormalities of young children with ASD. We propose an automated assessment framework to predict the severity level of atypical prosody. We model the training data using both the conventional SVM model and deep neural networks. For the three categories classification task, all systems cannot achieve satisfying performance. Then we perform two different partition on the 0/1/2 code, and attempt to run two categories classification. Experimental results indicate that boundary between code 1/2 are less clear than code 0/1 in our dataset. Even the clinicians are less confident to distinguish between slightly/serious abnormal than with/without abnormality. Particularly, the results of classification on with/without atypical prosody (0 vs union(1,2)) are quite surprising. RNN achieves very close performance comparing to SVM baseline and the fusion of the two system achieves the best performance. This result demonstrates the end-to-end framework has great potential in our task.

Future work includes three parts. First, in order to utilize the strong learning ability of the deep neural networks, we need to collect more ADOS recordings, especially the subjects with normal prosody (i.e. code '0'), to augment our training set as well as balance the distribution of different classes. Second, since human tagging usually exhibits subjective disagreements among each other, we intend to obtain labels from multiple evaluators and utilize some modeling methods to estimate the ground truth [22, 23, 24]. Finally, we will use regression method to investigate the coding strategy of the clinicians and construct a better model. Considering the speech abnormalities code, the demographic information as well as the language acquisition questionnaire results, maybe we can help clinicians predicting a subject's language development score or even the overall social communication score under the ADOS module 2 scenario.

6. Acknowledgements

This research was funded in part by the National Natural Science Foundation of China (61401524), Natural Science Foundation of Guangdong Province (2014A030313123), Natural Science Foundation of Guangzhou City (201707010363), the Fundamental Research Funds for the Central Universities(15lgjc12), Science and technology development foundation of Guangdong Province, National Key Research and Development Program (2016YFC0103905) and IBM Faculty Award.

7. References

[1] D. Christensen, J. Baio, and K. Braun, "Prevalence and characteristics of autism spectrum disorder among children aged 8 years autism and developmental disabilities monitoring network, 11 sites, united states, 2012," *MMWR Surveill Summ 2016*, vol. 65, no. SS-3, pp. 1–23, 2016.

[2] J. McCann and S. Peppé, "Prosody in autism spectrum disorders: a critical review," *International Journal of Language & Communication Disorders*, vol. 38, no. 4, pp. 325–350, 2003.

[3] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedulegeneric: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.

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

[5] T. Chaspari, E. M. Provost, A. Katsamanis, and S. Narayanan, "An acoustic analysis of shared enjoyment in eca interactions of children with autism," in *Proceeding of ICASSP*, 2012, pp. 4485–4488.

[6] D. Bone, M. P. Black, A. Ramakrishna, R. B. Grossman, and S. S. Narayanan, "Acoustic-prosodic correlates ofawkward'prosody in story retellings from adolescents with autism." in *Proceeding of Interspeech*, 2015, pp. 1616–1620.

[7] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[8] B. Xiao, Z. E. Imel, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, "" rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing," *PloS one*, vol. 10, no. 12, p. e0143055, 2015.

[9] 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," *Proceeding of Interspeech*, pp. 112–116, 2016.

[10] B. Xiao, D. Can, J. Gibson, Z. E. Imel, D. C. Atkins, P. Georgiou, and S. Narayanan, "Behavioral coding of therapist language in addiction counseling using recurrent neural networks," *Proceeding of Interspeech*, pp. 908–912, 2016.

[11] M. P. Black, A. Katsamanis, B. R. Baucom, C.-C. Lee, A. C. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Communication*, vol. 55, no. 1, pp. 1–21, 2013.

[12] H. Dubey, L. Kaushik, A. Sangwan, and J. H. Hansen, "A speaker diarization system for studying peer-led team learning groups," in *Proceeding of Interspeech*, 2016, pp. 2180–2184.

[13] T. Zhou, W. Cai, X. Chen, X. Zou, S. Zhang, and M. Li, "Speaker diarization system for autism children's real-life audio data," in *Proceeding of ISCSLP*, 2016, pp. 1–5.

[14] B. W. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, S. S. Narayanan, *et al.*, "The interspeech 2010 paralinguistic challenge." in *Proceeding of Interspeech*, vol. 2010, 2010, pp. 2795–2798.

[15] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," *Proceeding of ICASSP*, 2016.

[16] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[17] T. Lidy and A. Schindler, "Cqt-based convolutional neural networks for audio scene classification and domestic audio tagging," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2016)*, 2016.

[18] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proceeding of ICASSP*, 2015, pp. 4580–4584.

[19] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proceeding of ICASSP*, 2016, pp. 5115–5119.

[20] Combining cnn and rnn for spoken language identification. [Online]. Available: http://yerevann.github.io/2016/06/26/combining-cnn-and-rnn-for-spoken-language-identification/

[21] D. Cai, Z. Ni, W. Liu, W. Cai, G. Li, and M. Li, "End-to-end deep learning framework for speech paralinguistics detection based on perception aware spectrum," in *Proceeding of Interspeech*, 2017.

[22] K. Audhkhasi and S. S. Narayanan, "Data-dependent evaluator modeling and its application to emotional valence classification from speech." in *Proceeding of Interspeech*, 2010, pp. 2366–2369.

[23] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: whom to trust when everyone lies a bit," in *Proceedings of the 26th Annual international conference on machine learning*. ACM, 2009, pp. 889–896.

[24] P. Smyth, U. M. Fayyad, M. C. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of venus images," in *Advances in neural information processing systems*, 1995, pp. 1085–1092.