INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Analyzing short term dynamic speech features for understanding behavioral traits of children with autism spectrum disorder

Young-Kyung Kim[1*], *Rimita Lahiri*[1*], *Md Nasir*[2], *So Hyun Kim*[3], *Somer Bishop*[4], *Catherine Lord*[5], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory, University of Southern California, USA
[2]Microsoft AI for Good Research Lab, Redmond, Washington, USA
[3]Center for Autism and the Developing Brain, Weill Cornell Medicine, USA
[4]Department of Psychiatry, University of California, San Francisco, USA
[5]Semel Institute of Neuroscience and Human Behavior, University of California, Los Angeles, USA

(youngkyk,rlahiri,shrikann)@usc.edu, mdnasir@microsoft.com, sok2015@med.cornell.edu, bishop.somer@ucsf.edu, CLord@mednet.ucla.edu

Abstract

Computational methodologies have shown promise in advancing diagnostic and intervention research in the domain of *Autism Spectrum Disorder (ASD)*. Prior works have investigated speech features to assess disorder severity and also to differentiate between children with and without an ASD diagnosis. In this work, we explore short term dynamic functionals of speech features both within and across speakers to understand if local changes in speech provide information toward phenotyping of ASD. We compare the contributions of static and dynamic functionals representing conversational speech toward the clinical diagnosis state. Our results show that predictions obtained from a combination of dynamic and static functionals have comparable or superior performance to the predictions obtained from just static speech functionals. We also analyze the relationship between speech production and ASD diagnosis through correlation analyses between speech functionals and manually-derived behavioral codes related to autism severity. The experimental results support the notion that dynamic speech functionals capture complementary information which can facilitate enriched analysis of clinically-meaningful behavioral inference tasks.

Index Terms: dynamic functionals, diagnosis, developmental disorder

1. Introduction

ASD refers to a range of neuro-developmental disorders characterized by an early onset of significant social-communicative challenges along with restrictive, repetitive behaviors and interests. Recent studies report continual increase in prevalence of ASD in children from 1 in 59 children in 2014 to 1 in 54 children in 2020[1]. ASD diagnosis is a complex, challenging and time-consuming process as it relies on behavior symptoms in the absence of any reliable biological markers or medical tests.

While there are ongoing efforts to better understand the association between genetic and neuro-biological factors and ASD, a significant amount of research has been invested in building computational tools for domain experts and creating objective measures for early diagnostics, intervention planning and assessment. In particular, different physiological and behavioral signal-based features are being extensively studied to identify features that capture behavioral traits relevant to existing diagnostic instruments (ADOS [1], ADI-R [2]) in order to support behavioral phenotyping and stratification in the context of diagnosis and subsequent intervention.

Previous studies [3, 4] have explored computational approaches for validating behavioral markers and inferring ASD diagnosis predictions using observable behavioral information obtained from conversations involving children and interlocutors. For instance, Bone *et al.* [5] analyzed the association between objective signal-derived prosodic cues and subjective perceptions of prosodic awkwardness in settings of story retelling from adolescents with an ASD diagnosis; [6] studied lexical features to characterize the verbal behavior of children with ASD and non-ASD developmental disorders. However, since most of the literature relies on individual features computed for each speaker individually based on short-term vocal and lexical cues, they may not capture the full extent of two interlocutors' coordination and reciprocity, which is important when characterizing ASD.

Behavioral patterns in interactions are inherently dynamic in nature, and features derived from local changes reflect this behavior better when compared to those derived from global changes. In recent years, multiple works have proposed the use of various forms of conversational speech dynamics as features for downstream inference tasks. For example, emotion recognition has benefited from the use of temporal dynamics in form of autoregressive methods and spectral moments [7], hidden Markov models [8], nonlinear dynamics [9], and more recently recurrent neural networks [10, 11]. Curhan *et al.* [12] showed that measures of vocal activity level, conversational engagement, prosodic emphasis, and mirroring can help predicting negotiation trends. Deception detection [13] from vocal cues have been recently shown to improve by capturing the conversational dynamics [14, 15]. In the clinical domain, the dynamics captured by spectral energy variability have been shown to an indicator of depression [16, 17]. [18] has reported superior performance in couples therapy outcome prediction using dynamic functionals. Warlaumont *et al.* [19] has found measures of conversational dynamics, both at short and long timescales, can vary in between population with and without ASD diagnosis. Vocal arousal dynamics in child-psychologist interaction was shown to distinguish between high and low ASD severity [20]. In this work, this motivates us to capture dynamics based on the aggregated turns of each interlocutor to encode important

[*] equal contribution

[1] https://www.autismspeaks.org/press-release/cdc-estimate-autism-prevalence-increases-nearly-10-percent-1-54-children-us

http://dx.doi.org/10.21437/Interspeech.2021-2111

Table 1: *Demographic details of ADOS dataset*

Category	Statistics
Age(years)	Range: 3.58-13.17 (mean, std): (8.61, 2.49)
Gender	123 male, 42 female
Non-verbal IQ	Range: 47-141 (mean, std): (96.01, 18.79)
Clinical Diagnosis	86 ASD 42 ADHD (Attention Deficit Hyperactivity Disorder) 14 mood/anxiety disorder 12 language disorder 10 intellectual disability, 1 no diagnosis

conversational and behavioral patterns of speech. More specifically, we aim to understand the contribution of dynamic functionals in characterizing behavioral patterns of children with an ASD diagnosis.

We analyze the vocal speech patterns of children – both those with and without an ASD diagnosis – engaged in interaction with clinicians. We present a correlation analysis to interpret the relationship between the extracted features and manually coded clinical ratings related to ASD diagnosis. We formulate the prediction task as a binary classification problem of differentiating between children with an ASD diagnosis and those who do not. We compare the predictions using the features derived from the static and dynamic functionals to better understand the benefit of explicitly using dynamic functionals for predicting the diagnosis state.

2. Conversational Data

The *Autism Diagnostic Observation Schedule (ADOS)-2* [21] instrument refers to a sequence of semi-structured activities between a child and a clinician to assess behavioral patterns associated with ASD. A typical ADOS-2 interaction session lasts 40-60 minutes, where a child is engaged in multiple subtasks to evoke maximum response.

In this work, we choose to focus on the *Emotions* and *Social difficulties & annoyance* subtasks from the Module 3 administration, designed for verbally fluent children. In the *Emotions* subtask, the child is asked questions related to the identification of situations and activities that elicit different emotions. During the *Social difficulties & annoyance* subtask, the child is asked to describe his/her opinion on different social issues in different circumstances (at home or school) and about their coping strategies.

For this work, we carry out data standardization for each speaker in each session after aligning the speaker's turn using manually derived transcripts (following SALT transcription guidelines [22]). We exclude sessions having fewer than 25 turns so as to enable reliable computation of 3rd and 4th order dynamic functionals. After preprocessing, our final dataset contains a total of 281 sessions from 165 children (144 ASD, 137 non-ASD). Almost every child contributed 2 interaction sessions, corresponding to the 2 subtasks mentioned above.

3. Experimental Methodology

We conduct two sets of experiments, (i) correlation based analyses between the extracted static and dynamic functionals and ranked measures of the child's ASD severity termed as *Calibrated Severity Score (CSS)* [21], and (ii) binary classification between children with and without ASD diagnosis based on different sets of static and dynamic functionals. The former is undertaken to understand the relationship between the extracted functionals and the clinically-meaningful CSS, while the latter is used to understand the additional predictive power that the dynamic functionals present over the static functionals.

3.1. Acoustic-Prosodic and Turn-Taking Feature Analysis

All the speech features are extracted using openSMILE [23]. We consider features relating to acoustics, prosody, and voice quality. All 15 dimensions of *Mel Frequency Cepstral Coefficients (MFCC)* and the 8 dimensions of *Mel Frequency Band (MFB)* features are included in the spectral set. Loudness, pitch envelope and their first order differences are considered as the prosodic set, while voicing probability, local jitter, the differential frame-to-frame jitter, local shimmer, and the first order difference of each of these features made up the voice-quality set. All these features are computed for every 10ms interval of the audio file.

After extracting the raw features, we calculate five static functionals (mean, standard deviation, median, mininum, and maximum) across each session in the dataset. For the static functionals, we only consider the child's turns. To calculate dynamic functionals, we first average the frames of each relevant turn, and take the first, second, third, and fourth order differences within turn-pairs as our dynamic functionals as shown in Figure 1. We define turn-pairs as consisting of consecutive turns either between the same speaker or across different speakers.

3.2. Correlation Analysis

The correlational analysis is set up to estimate the association between the functionals (static and dynamic) and manually coded behavioral ratings (CSS) related to ASD diagnosis. CSS is a metric quantifying ASD severity with relative independence from individual characteristics such as age and verbal IQ on a 10 point scale. Because of the ranked nature of CSS, we chose Spearman's rho [24] for this analysis over Pearson's correlation coefficient. The correlation metrics serve as a knowledge-driven

Table 2: *Top 5 features based on absolute correlation values for static functionals of different feature categories with CSS (the indices for MFCCs and MFBs are shown in parentheses)*

Prosodic Features			Voice Quality Features			Acoustic Features		
feature	func	corr	feature	func	corr	feature	func	corr
Pitch Envelope	Max.	-0.1875	Voicing	Min.	0.3053	MFCC(2)	Min.	0.1852
Pitch Envelope	Mean	-0.1561	Voicing Diff	Min.	0.1350	MFCC(0)	Max.	-0.1832
Pitch Diff	Mean	0.1467	Dynamic Jitter Diff	Median	-.1287	MFB(7)	Min.	-0.1694
Pitch Envelope	Min.	0.1308	Jitter	Median	0.1263	MFB(4)	Max.	-0.1620
Loudness	Mean	0.1260	Voicing Diff	Max.	0.1175	MFCC(6)	Min.	-0.1603

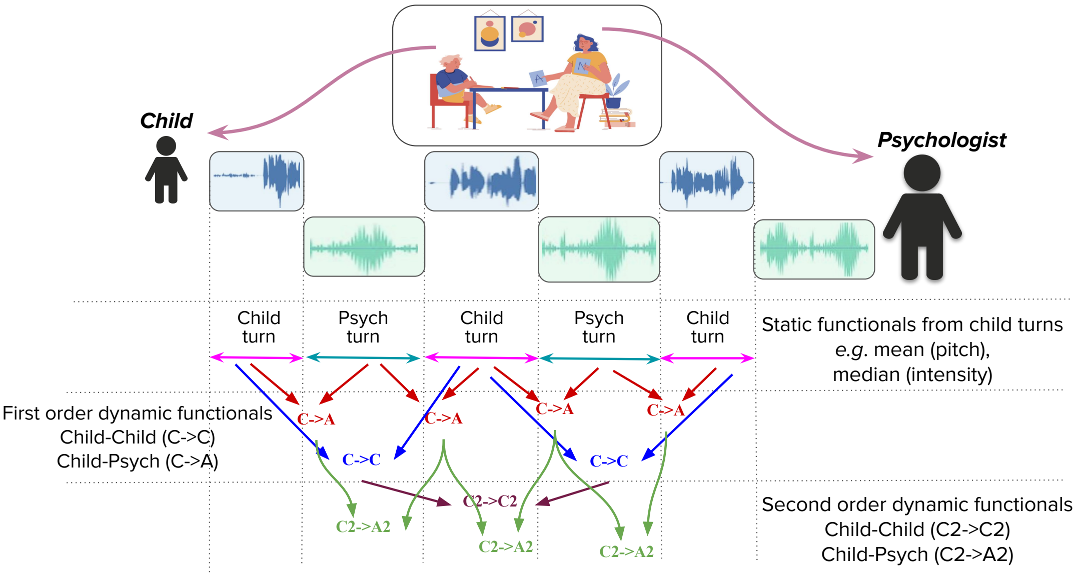


Figure 1: *Static and dynamic functionals*

Table 3: *Top 5 features based on absolute correlation values for dynamic functionals of different feature categories with CSS (the indices for MFCCs and MFBs are shown in parentheses)*

Child - Child				Psych - Psych				Child - Psych			
order	feature	func	corr	order	feature	func	corr	order	feature	func	corr
1	MFCC(6)	Max.	0.1999	1	Loudness	Std. Dev.	-0.3190	2	Loudness	Max.	-0.3026
3	MFCC(6)	Min.	-0.1893	1	MFB(0)	Std. Dev.	-0.3014	3	Loudness	Max.	-0.2977
1	Pitch	Median	0.1748	4	MFB(0)	Min.	0.2951	4	Loudness	Min.	0.2856
1	MFCC(6)	Std. Dev.	0.1715	1	MFB(0)	Min.	0.2919	1	MFB(0)	Std. Dev.	-0.2745
3	MFCC(7)	Std. Dev.	0.1700	2	Loudness	Std. Dev.	-0.2799	1	MFB(0)	Min.	0.2712

way to select features then used to infer ASD diagnosis in the next experiment.

3.3. Classification and Feature Selection

To understand the role of dynamic functionals in differentiating between children with and without ASD diagnosis, we set up a binary classification experiment to predict the output labels as ASD or non-ASD. We consider five different classifiers for this experiment, logistic regression, *Support Vector Machine (SVM)*, random forest, k-nearest neighbours and naive bayes classifier. For each of the classifiers, we carry out 5-fold cross-validation 5 times to avoid overfitting the data.

We consider different combinations (feature-level fusion) of static and dynamic functionals in our classification analysis to investigate the extent of combined predictive power of both static and dynamic functionals. Moreover, we report the classification F1 score of different order functionals individually and along with static functionals to study the contributions of using dynamic functionals over static ones.

To reduce the number of features used, we use a feature selection strategy based on correlation-based feature ranking. We calculate the Spearman's correlation coefficients for each feature from each set of functionals (dynamic or static) with respect to the variable of interest (CSS in this case), and rank them in descending order of correlation. In this process, we exclude the features that are not statistically significant.

4. Results

In this section we report the findings based on the experiments we conduct.

4.1. Correlation Analysis

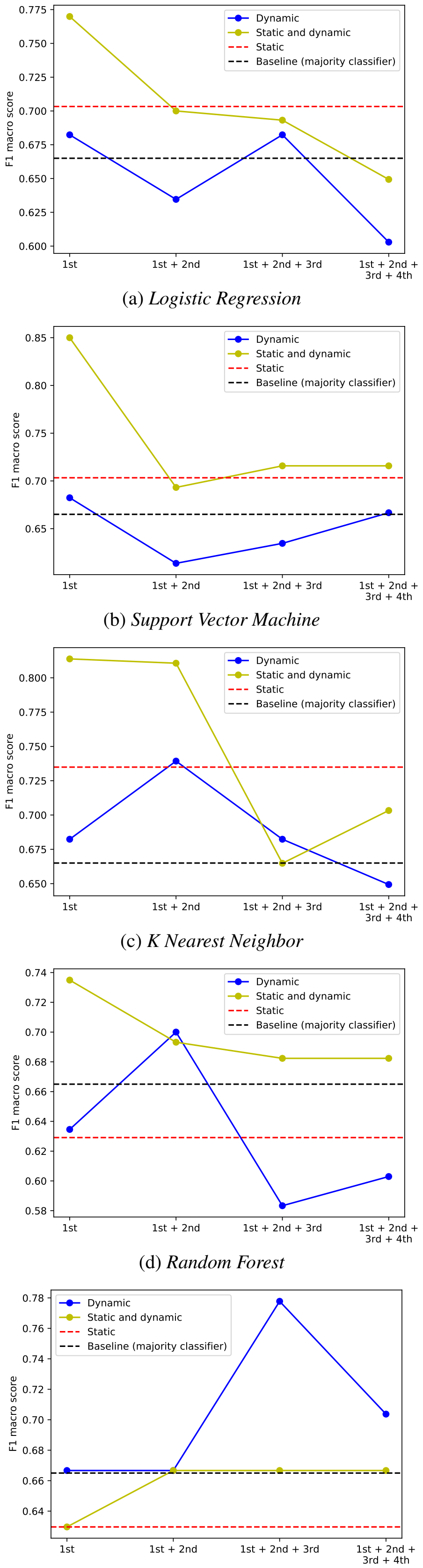
Here, we perform correlation analysis to investigate details regarding the static and dynamic functionals capturing information that can be used to make inferences related to behavioral patterns in ASD. For this experiment, we compute Spearman's correlation coefficients between the CSS and mutually exclusive sets of static and dynamic functionals.

In Table 2, we report the five most correlated static functionals for each of the feature categories and in Table 3 we report the five most correlated dynamic functionals for each of the categories involving either same speaker or different speakers. In each case, we consider the significantly correlated functionals only ($p < 0.05$).

4.2. Classification Experiment

The goal of the classification experiment is to investigate the possibility of predicting ASD diagnosis based on static and dynamic functionals (both individually and in combination). As mentioned in previous sections, the classification experiment is formulated as a 2-class problem of predicting either ASD or non-ASD.

For all the experiments, the first n ranked statistically significant static and dynamic functionals are considered as input to classifiers. The value of n is chosen separately for each fea-



(a) *Logistic Regression*

(b) *Support Vector Machine*

(c) *K Nearest Neighbor*

(d) *Random Forest*

(e) *Naive Bayes*

Figure 2: *Classification experiment results: macro-averaged F1 scores vs. different orders of dynamic functionals used*

ture set from 2 to 35 based on the classification performance. We consider only child-child and child-psychologist dynamic functionals as input to the classifiers, as our primary focus is on analyzing the behavioral dynamics of the children. It is important to understand that for each order of dynamic functionals, we also consider the preceding order differences cumulatively. For example, the dynamic functional of 3rd order also includes the 1st and 2nd order dynamic functionals. Since we build each dynamic functional set in a cumulative way, the resulting selected features do not always include equal proportions of each contributing dynamic or static functional subset.

For each of the mentioned classifiers, we report the performance considering only static functionals, cumulative dynamic functionals, and also static and cumulative dynamic functionals together in terms of classification F1 score as shown in Figure 2. We consider the majority classifier (every sample is assigned to whichever is the majority class in the training set) as the baseline and report its performance in the same plot to better understand the improvement in classification performance after incorporating static and dynamic functionals.

4.3. Summary of observations

While Table 2 shows pitch envelope provides maximum correlation for static functionals, Table 3 reveals MFCC and loudness showing greater absolute correlation for dynamic functionals. Quite interestingly, within psychologist (i.e., psych - psych) correlations are found to be higher followed by child - psych dynamics; this is consistent with the observation that the psychologist adjusts their dynamics according to the clinical state of the child [25]. Results from Figure 2 suggest a combination of static and dynamic functionals offers the best performance in the majority of the cases. Amongst the dynamic functionals, the 1st order differences work the best, indicating that higher order functionals viz., 2nd, 3rd and 4th order differences are not contributing as much to the classification problem; it may also be the case that the feature selection is perhaps overfitting or ineffective.

5. Conclusion

Speech features are being extensively studied to understand and characterize ASD in the context of behavioral analysis and phenotyping. In this work, we report the relevance of different combination of static and dynamic speech functionals with respect to clinically-determined disorder severity in terms of the correlation metric. We examine the role of dynamic functionals individually and in combination with static functionals to predict ASD diagnosis. Furthermore, we show the top ranked static and dynamic functionals (based on correlation metric) carry meaningful insights to classify the behavioral patterns of children with and without ASD diagnosis.

In the future we plan to extend this work with lexical features in order to gain a comprehensive understanding about behavioral traits of children during such interactions. We also plan to explore other feature selection techniques in order to improve classification performance based on static and dynamic functionals.

6. Acknowledgements

This work is supported by funds from NIH and Simons foundation.

7. References

[1] C. Lord, M. Rutter, P. C. DiLavore *et al.*, "Autism diagnostic observation schedule–generic," *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 1999.

[2] C. Lord, M. Rutter, and A. Le Couteur, "Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders," *Journal of autism and developmental disorders*, vol. 24, no. 5, pp. 659–685, 1994.

[3] D. Bone, M. S. Goodwin, M. P. Black, C.-C. Lee, K. Audhkhasi, and S. Narayanan, "Applying machine learning to facilitate autism diagnostics: pitfalls and promises," *Journal of autism and developmental disorders*, vol. 45, no. 5, pp. 1121–1136, 2015.

[4] D. Bone, C.-C. Lee, T. Chaspari, J. Gibson, and S. Narayanan, "Signal processing and machine learning for mental health research and clinical applications," *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 189–196, September 2017.

[5] D. Bone, M. P. Black, A. Ramakrishna, R. Grossman, and S. S. Narayanan, "Acoustic-prosodic correlates of 'awkward' prosody in story retellings from adolescents with autism," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] M. Kumar, R. Gupta, D. Bone, N. Malandrakis, S. Bishop, and S. S. Narayanan, "Objective language feature analysis in children with neurodevelopmental disorders during autism assessment." in *Interspeech*, 2016, pp. 2721–2725.

[7] S. Ntalampiras and N. Fakotakis, "Modeling the temporal evolution of acoustic parameters for speech emotion recognition," *IEEE Transactions on affective computing*, vol. 3, no. 1, pp. 116–125, 2011.

[8] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 2. IEEE, 2003, pp. II–1.

[9] E. Tzinis, G. Paraskevopoulos, C. Baziotis, and A. Potamianos, "Integrating recurrence dynamics for speech emotion recognition," *arXiv preprint arXiv:1811.04133*, 2018.

[10] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6685–6689.

[11] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[12] J. R. Curhan and A. Pentland, "Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes." *Journal of Applied Psychology*, vol. 92, no. 3, p. 802, 2007.

[13] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception." *Psychological bulletin*, vol. 129, no. 1, p. 74, 2003.

[14] Y. Zhou, H. Zhao, X. Pan, and L. Shang, "Deception detecting from speech signal using relevance vector machine and non-linear dynamics features," *Neurocomputing*, vol. 151, pp. 1042–1052, 2015.

[15] H.-C. Chou, Y.-W. Liu, and C.-C. Lee, "Joint learning of conversational temporal dynamics and acoustic features for speech deception detection in dialog games," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1044–1050.

[16] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[17] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling spectral variability for the classification of depressed speech." in *Interspeech*, 2013, pp. 857–861.

[18] M. Nasir, B. R. Baucom, P. Georgiou, and S. Narayanan, "Predicting couple therapy outcomes based on speech acoustic features," *PloS one*, vol. 12, no. 9, p. e0185123, 2017.

[19] A. S. Warlaumont, D. K. Oller, R. Dale, J. A. Richards, J. Gilkerson, and D. Xu, "Vocal interaction dynamics of children with and without autism," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 32, no. 32, 2010.

[20] D. Bone, C.-C. Lee, A. Potamianos, and S. S. Narayanan, "An investigation of vocal arousal dynamics in child-psychologist interactions using synchrony measures and a conversation-based model," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[21] K. Gotham, A. Pickles, and C. Lord, "Standardizing ados scores for a measure of severity in autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 39, no. 5, pp. 693–705, 2009.

[22] J. J. Heilmann, J. F. Miller, and A. Nockerts, "Using language sample databases," 2010.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[24] L. Myers and M. J. Sirois, "S pearman correlation coefficients, differences between," *Encyclopedia of statistical sciences*, 2004.

[25] D. Bone, M. P. Black, C.-C. Lee, M. Williams, P. Levitt, S. Lee, and S. S. Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *Journal of Speech, Language, and Hearing Research*, vol. 57, p. 1162â€"1177, aug 2014.