INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Automatically Predicting Perceived Conversation Quality in a Pediatric Sample Enriched for Autism

Yahan Yang[1], *Sunghye Cho*[1], *Maxine Covello*[2], *Azia Knox*[2], *Osbert Bastani*[1], *James Weimer*[3], *Edgar Dobriban*[1], *Robert Schultz*[1,2], *Insup Lee*[1], *Julia Parish-Morris*[1,2]

[1]University of Pennsylvania, Philadelphia, PA [2]Children's Hospital of Philadelphia, Philadelphia, PA
[3]Vanderbilt University, Nashville, TN

{yangy96@seas, csunghye@sas, obastani@seas, dobriban@wharton, lee@seas}.upenn.edu,
{covellom, knoxa2, schultzrt, parishmorrisj}@chop.edu, james.weimer@vanderbilt.edu

Abstract

Social interaction quality ratings derived from short natural conversations can differentiate children with and without autism at the group level. In this work, we explored conversations between children and an unfamiliar adult who rated their social interaction success on six dimensions. Using hand-crafted acoustic and lexical features, we built different classifiers to predict children's dimensional conversation quality. The best classifier achieved 61% accuracy, which outperformed human raters (49%). Follow-up analyses revealed that a subset of features determined communication quality scores. Additionally, we extracted acoustic features using a pretrained audio transformer and improved our prediction to 68%. This study suggests that automatically predicting conversation quality could be an inexpensive and objective way to monitor intervention progress in children with communication challenges, and could be used to identify intervention targets for improving conversational success.

Index Terms: autism spectrum disorder, conversational audio analysis, machine learning classification and interpretation

1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by social communication challenges, and the presence of repetitive behaviors and restricted interests [1]. Autism is lifelong, but early detection and intervention during childhood has been shown to improve outcomes for some individuals on the spectrum [2]. As a highly heterogeneous condition, autism can look very different from one individual to the next [3]. In some instances, autism is straightforward to identify; in other cases, co-occurring conditions and unique presentations make it more complicated to accurately diagnose [4]. Long wait lists for expert assessment are a critical barrier to speedy identification and early intervention for autistic children [5], leading to calls for inexpensive, objective automatic screening systems that could provide pre-assessments, thus shortening clinical wait lists [6]. Prior studies designed to detect autism using automated approaches have either relied on features produced by experts, or derived predictors from long structured interaction sessions ($\geq$ 30 minutes), which are difficult to scale quickly and inexpensively in local community settings [7, 8].

Autistic individuals often have subtle speech and language differences that manifest during conversations with other people [9] and may disrupt social interaction [10]. Recent work [11] demonstrates that first impressions made by children during short "get-to-know-you" conversations with non-expert adults can provide a convenient and quick way to gauge social communication differences in boys with autism - although this approach may be less accurate for girls. Providing a tool to evaluate dimensional interaction quality[1] during short conversations holds promise as a way to streamline the identification and assessment of communication difficulties in children, including autistic children. However, it is unclear which *ingredients* of a conversation predict better or worse first impressions. Pinpointing specific linguistic features that contribute to the perception of "social success" during natural conversations could help clinicians develop personalized supports for children that may otherwise experience subjective distress during social interactions due to subtle speech and language differences.

One recent study used a naturalistic conversation between a child and a young adult to build an automatic classification system to detect autism [12]. At the same time, recent developments in deep learning—such as transformer architectures—have shown promising performance in extracting contextual acoustic representations for audio classification or autism detection [13, 14, 15]. However, prior studies did not aim to predict the social quality of children's conversations, as perceived by listeners, which is critical for the goal of monitoring response to social skills interventions. The current work takes a step in this direction.

The goal of this study is to produce a tool that uses acoustic and lexical features to predict conversation quality based on brief samples from children with and without autism, and young adult conversation partners. We aim to (1) predict perceptions of conversation quality (low, medium, high) using hand-crafted features and several popular classifier types, (2) identify important hand-crafted features that contribute to the sense of "conversational success", and (3) predict conversation quality using representations extracted from pretrained transformers.

2. Sample Characteristics

2.1. Dataset

Our dataset consisted of 72 five-minute "get-to-know-you" conversations (audio recordings with corresponding transcripts) between a child participant and a young adult confederate. Thirty-five participants were diagnosed with autism by an expert clinician, and thirty-seven participants were categorized as neurotypical or typically developing (TD). Autism and TD groups were matched on key demographic characteristics, including age, sex ratio, and full-scale IQ score as shown in Table 1.

For all "get-to-know-you" conversations, no instructions or topics were given beforehand, and all speakers, including autistic children, were verbally fluent native English speakers. Confederates were not aware of children's diagnostic status, and were assigned to conversations based on availability. This

[1]We interchangeably use interaction quality and conversation quality in the paper.

10.21437/Interspeech.2023-2251

study was overseen by the Institutional Review Board at Children's Hospital of Philadelphia (CHOP), and parental consent and children's verbal assent were obtained.

	Autism	TD
Number	35	37
Age (mean/std)	11.11/2.84	9.54/2.61
Sex	M: 22 F:13	M: 23 F: 14
ADOS		
Soc Aff (mean/std)	9.96/3.36	1.24/1.28
RRB (mean/std)	2.17/1.49	0.21/0.54
Overall (mean/std)	12.13/3.62	1.45/1.42

Table 1: *Demographic and clinical characteristics of participants. ADOS: Autism Diagnostic Observation Schedule - 2nd Edition, Module 3 [16], Soc Aff: Social affect subscore of the ADOS-2, RRB: repetitive behaviors and restricted interests subscore of the ADOS-2.*

2.2. Conversation Score

After the 5-minute conversation was complete, confederates filled out a modified version of the Conversation Rating Scale questionnaire [17], which included six different conversation ratings (Table 2). The sum of the first five items was computed, with questions 4 and 5 reverse-scored so that higher scores indicated better perceived conversation quality; and this was called the conversation score. The sixth score, which focused on eye contact during conversations, was excluded, since it could not be predicted using audio and transcript data only. We show the statistics of the conversation score in Figure 1 and the correlations between different scores in Figure 2.

	Description	Rating Scale
1.	The other person was interested in what I had to say.	1-7
2.	This person was warm and friendly.	1-7
3.	The conversation flowed smoothly.	1-7
4.	The other person acted bored by our conversation.	1-7
5.	The other person created a sense of distance between us.	1-7
6.	The other person made appropriate eye contact with me during our conversation.	1-7

Table 2: *Conversation Rating Scale*

We split the dataset into 3 classes given by the sum of the conversation score: low quality[2] (10-20, $n = 13$), medium quality (20-30, $n = 37$), high quality (30-35, $n = 22$). There were two reasons for considering such a split: 1) the distribution of conversation scores was positively skewed and the split balanced it, 2) it ensured a more easily interpretable clinical meaning, because only conversations that received high scores across all scales could be considered high-quality conversations.

3. Methodology

3.1. Hand-crafted Features

To extract features from conversations, we first separated audio files and transcripts by speaker. Then, we extracted two sets of features: lexical features based on expert knowledge, and

[2]There was no conversation that had a score below 10.

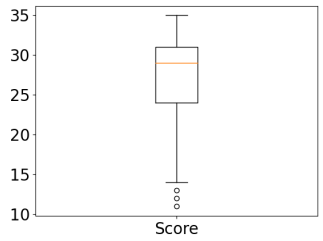


Figure 1: *Box-plot of the conversation score sum (first 5 conversation ratings).*

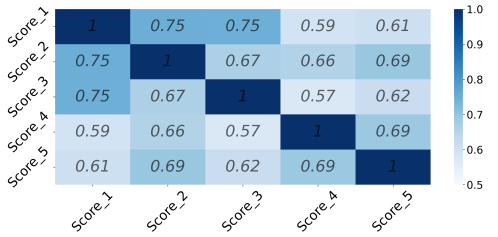


Figure 2: *Correlation heatmap of conversation scores. Note that we reverse coded questions 4 and 5.*

acoustic features used in previous studies [12, 18, 19], see Table 3. Following previous work [12, 18, 20], we computed acoustic features using openSMILE [21] with the eGeMAPSv02 configuration [22]. The 88 extracted features were high-level statistics including voice probability, mel-frequency cepstral coefficients (MFCCs), pitch, etc. The total number of hand-crafted features was 100. We further selected features based on their Pearson correlation with the outcome to mitigate over-fitting issues. Feature selection occurred after the train-test splits, and the cut-off threshold was 0.21, chosen via a hyper-parameter search. The mean number of hand-crafted features chosen was 33.

3.2. Deep learning-based acoustic features

As an alternative to hand-crafted acoustic features, we also extracted deep learning-based acoustic features, and concatenated them with the hand-crafted lexical features. We extracted acoustic features using the Audio Spectrogram Transformer (AST) [13], which was pretrained on the AudioSet dataset. We preprocessed the audio waveform of the participants following [13]: the input audio was split into 25ms frames first, and then the Hamming window function was applied to each frame. We applied the Short Time Fourier Transform (STFT) and converted the resulting power spectrum to filter banks on a Mel scale. The AST had the spectrogram of participant-only audio as the input (a 30-second clip). We froze the pretrained AST model to extract deep learning-based acoustic representations. We experimented with using audio from both the participant and confederate or only the confederate, but this led to a lower performance compared to using audio from the participant only.

4. Experiments

In our experiments using hand-crafted features, we compared five different classifiers: support vector machines (SVM), k-Nearest Neighbors (kNN), Gradient Boosting (GB), Decision Trees (Tree), and Random Forests (Forest), implemented in the scikit-learn package [23]. For deep learning-based features, we fused the acoustic features and hand-crafted lexical features by

Lexical Features	Description
Filler/Pause count	Number of pause filler (e.g. 'um, hmm, uhm") (Participant)
Laugh count	Number of laugh (Participant)
No count	Number of 'no' (Participant)
Question count	Number of questions (Participant)
Short vs Whole	Ratio of one-word sentences to all sentences (Participant)
Average length	Average length (in number of words) of sentences (Both)
Word ratio	Ratio of the role's words to total number of words (Both)
Sentence ratio	Ratio of the role's turn to total number of turns (Participant)
Backchannel count	Number of backchannel utterances (Participant)
Acoustic Features	**Description**
Percent of Silence	Silence time / Total time (Participant)
openSMILE	Pitch and voice quality related features

Table 3: *Summary of hand-crafted features. Participant: indicates features are from participant side only; Confederate: indicates features are from confederate side only; Both: indicates features are from both role*

concatenation and trained a linear neural network model using the SGD optimizer [24] for 80 epochs with a learning rate of 0.005. The training ran on an NVIDIA Quadro RTX 6000 GPU with 24GB of RAM. In addition to predicting conversation quality (Sec 4.1), we also used our selected features (both hand-crafted and deep learning-based) to perform autism prediction for participants (Sec 4.4). We evaluated our classifiers using 5-fold cross-validation, and reported the mean value and standard deviation across the folds.

4.1. Conversation Score Results

4.1.1. Using Hand-crafted Features

The classification accuracy results in Table 4 show that our hand-crafted features were able to predict conversation quality (low, medium, high) with relatively high accuracy for this three-way classification problem. SVMs achieved the highest accuracy (61%) on interaction quality prediction. We also observed that both acoustic and lexical features were necessary for the best accuracy.

Acc (%)	SVM	kNN	GB	Tree	Forest
Lexical	47.67 ($\pm$11.91)	48.67 ($\pm$0.08)	43.0 ($\pm$9.91)	41.33 ($\pm$4.99)	46.33 ($\pm$11.57)
Acoustic	59.67 ($\pm$12.67)	60.00 ($\pm$10.33)	55.33 ($\pm$4.00)	40.67 ($\pm$9.98)	52.67 ($\pm$7.42)
Lexical+ Acoustic	**61.0** ($\pm$10.73)	58.33 ($\pm$8.82)	44.67 ($\pm$9.33)	42.00 ($\pm$8.33)	52.67 ($\pm$9.52)

Table 4: *Results of conversation quality prediction using hand-crafted features. We experimented with lexical features, acoustic features, or both. Acc denotes Accuracy.*

4.1.2. Using Deep-learning Features

Table 5 shows that features extracted by AST attained a higher accuracy in predicting conversation quality than hand-crafted features. The results also demonstrate that including both lexical and acoustic features was essential for optimizing conversation quality prediction.

Class	Accuracy(%)
Lexical	54.0 ($\pm$13.06)
Acoustic	62.67 ($\pm$11.62)
Lexical + Acoustic	**68.0** ($\pm$8.84)

Table 5: *Prediction of conversation quality using lexical features and acoustic features extracted using an AST transformer.*

4.2. Human Evaluation

We performed human evaluation as a baseline to evaluate our automatic conversation predictor. Three non-expert undergraduate student raters listened to the conversations and filled in the conversation score survey. The sixth score was omitted. We averaged the scores collected from the three raters and reported human performance in Table 6. Our automatic conversation predictor's accuracy was approximately 12% and 19% higher than human raters, using hand-crafted and deep learning features, respectively. This demonstrates the feasibility of using an automatic tool for assessing conversation quality in children with and without autism. We also observed that the recall of low-quality conversations by human raters was low, which was similar to the tendency observed when using automatic conversation predictors. We hypothesize that this happened because the number of low-quality conversations is small compared to the other two categories. Therefore, increasing our sample size could be beneficial for improving our understanding of communication difficulties in children.

	Class	Precision	Recall	F1-score
Automated Prediction	Low	60.00	46.15	52.17
	Medium	70.45	83.78	76.54
	High	66.67	54.55	60.00
	Acc (%)		**68.06**	
Human Raters	Low	100.00	23.08	37.50
	Medium	50.00	48.65	49.32
	High	42.42	63.64	50.91
	Acc (%)		48.61	

Table 6: *Comparison of machine learning method and human raters for predicting conversation quality (low, medium, high).*

4.3. Feature Analysis

To identify which features contributed the most to perceived conversation quality, we computed the contributions of the hand-crafted features in the SVM model using SHAP (SHapley Additive exPlanations) [25], a technique for understanding the behavior of machine learning models. We did not apply SHAP to the transformer model despite its higher accuracy, because analyzing the spectrogram feature-by-feature is not easily interpretable. The feature analysis in Figure 3 suggests that the most important features for predicting social interaction quality were the participants' loudness, pitch range in semitones, the number of questions produced by the participant, the ratio of short vs. long sentences produced by the participant, and the number of backchannel words produced during the participant's turn. These selected features could prove useful as targets in personalized interventions aimed at improving perceptions of social interaction quality.

We also compared which features predict perceptions of conversation quality in autistic children vs. TD children separately, by plotting the impact of hand-crafted features on conversation quality in Figures 4a and 4b. We observed that for both groups, loudness, the number of questions asked by participants, short sentence ratio, and F0 semitone were important, but with different feature importance scores. For autistic children,

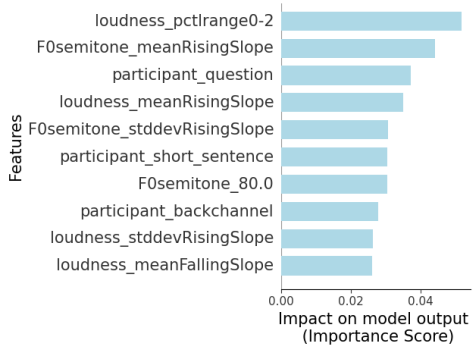


Figure 3: *Feature importance by SHAP. The variables of the selected acoustic features are from eGeMAPSv02 configuration. The variables of selected lexical features are from Table 3.*

the number of backchannel words produced was also an important feature that impacted perceptions of conversation quality.

4.4. Autism Prediction

We used the extracted features designed for assessing conversation scores to predict participants' diagnostic status (autism or TD). The results of autism prediction using hand-crafted features are shown in Tables 7 and 8. We observed that the accuracy achieved using deep-learning extracted features was comparable to the performance from previous work [12]. This further showed that our automatic conversation predictor was a reasonable candidate to assist with autism prediction.

Acc (%)	SVM	kNN	GB	Tree	Forest
Lexical	59.67	58.00	53.67	53.67	54.67
Acoustic	63.67	64.33	55.33	58.33	55.00
Lexical +Acoustic	63.67	61.33	61.00	**65.00**	58.00

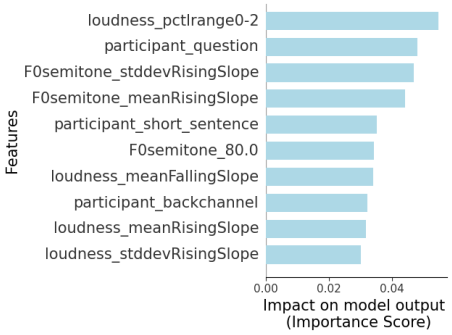
Table 7: *Autism prediction using hand-crafted features. We experimented with lexical features, acoustic features, or both. Acc denotes the overall prediction accuracy.*

Class	Precision	Recall	F1-score
TD	74.29	70.27	72.22
ASD	70.27	74.29	72.22
Acc (%)		72.22	

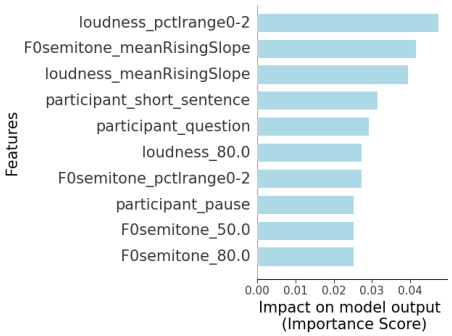
Table 8: *Autism prediction using both audio features extracted by the AST transformer, and with lexical features. TD denotes typically developing or neurotypical participants, and ASD denotes children with autism. Acc denotes prediction accuracy.*

5. Discussion and Conclusion

The close relationship between social communication skills and autism suggests that predicting conversation scores can facilitate clinical pre-screening, thus lowering critical barriers to early identification and support. Here we propose a framework for predicting perceived communication quality using short, natural conversations between a child and a non-expert adult interlocutor. We showed that a machine learning approach can achieve reasonable prediction performance. This approach can be used as an automatic and objective tool to monitor changes in conversational competence, for children with and without autism. Additionally, we utilized pre-trained transformers to extract expressive acoustic representations to improve the performance of conversation quality predictions. Our work also applied interpretability approaches for understanding model prediction with designed features. The selected features may be used to assist experts as they design and implement social communication support tools for autistic children. Our hand-crafted feature-based and deep-learning-based approaches can potentially be generalized to analyze the speech and conversation patterns of children with other types of language difficulties. As the aim of predicting conversation quality is to aid in the diagnosis and tracking of intervention progress in autistic children, the results of this classifier will not be shared directly with the children or their guardians and will only be evaluated by medical professionals alongside other assessments if applied in a clinical setting.



(a) *Feature importance for autistic children.*



(b) *Feature importance for TD children.*

Figure 4: *Important features identified by SHAP for different groups. The variables of the selected acoustic features are from eGeMAPSv02 configuration. The variables of selected lexical features are from Table 3.*

The limitations of our work include that we evaluated our approach on a relatively small dataset. In future work, we will analyze audio conversations between children and adults for various clinical purposes. We will also extend our work by integrating automated speech recognition and speaker diarization to create a complete pipeline for assessing the communication quality of children in real-world settings.

6. Acknowledgements

Research was supported by the Army Research Office (ARO) under Grant Number W911NF-20-1-0080, and by the US Defense Advanced Research Projects Agency (DARPA) under Contract FA8750-19-2-0201, and NSF-2125561. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense, the Army Research Office or the U.S. Government. This research was also supported partly by a gift from AWS AI for research in Trustworthy AI.

7. References

[1] D. American Psychiatric Association, A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders: DSM-5*. American psychiatric association Washington, DC, 2013, vol. 5.

[2] E. Fernell, M. A. Eriksson, and C. Gillberg, "Early diagnosis of autism and impact on prognosis: a narrative review," *Clinical epidemiology*, vol. 5, p. 33, 2013.

[3] S. Tang, N. Sun, D. L. Floris, X. Zhang, A. Di Martino, and B. T. Yeo, "Reconciling dimensional and categorical models of autism heterogeneity: A brain connectomics and behavioral study," *Biological Psychiatry*, vol. 87, no. 12, pp. 1071–1082, 2020, obsessive-Compulsive Disorder and Developmental Disorders. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0006322319318591

[4] C. G. McDonnell, C. C. Bradley, S. M. Kanne, C. Lajonchere, Z. Warren, and L. A. Carpenter, "When are we sure? predictors of clinician certainty in the diagnosis of autism spectrum disorder," *Journal of Autism and Developmental Disorders*, vol. 49, pp. 1391–1401, 4 2019.

[5] M. O. Mazurek, A. Curran, C. Burnette, and K. Sohl, "Echo autism stat: Accelerating early access to autism diagnosis," *Journal of Autism and Developmental Disorders*, vol. 49, pp. 127–137, 1 2019.

[6] S. M. Kanne and S. L. Bishop, "Editorial perspective: The autism waitlist crisis and remembering what families need," *Journal of Child Psychology and Psychiatry*, vol. 62, pp. 140–142, 2 2021.

[7] A. Garg, A. Parashar, D. Barman, S. Jain, D. Singhal, M. Masud, and M. Abouhawwash, "Autism spectrum disorder prediction by an explainable deep learning approach," *Computers, Materials and Continua*, vol. 71, pp. 1459–1471, 11 2021.

[8] S. Sadiq, M. Castellanos, J. Moffitt, M.-L. Shyu, L. Perry, and D. Messinger, "Deep learning based multimedia data mining for autism spectrum disorder (asd) diagnosis," in *2019 International Conference on Data Mining Workshops (ICDMW)*, 2019, pp. 847–854.

[9] J. Parish-Morris, M. Y. Liberman, C. Cieri, J. D. Herrington, B. E. Yerys, L. Bateman, J. Donaher, E. Ferguson, J. Pandey, and R. T. Schultz, "Linguistic camouflage in girls with autism spectrum disorder," *Molecular Autism*, vol. 8, p. 48, 12 2017.

[10] A. Sturrock, H. Chilton, K. Foy, J. Freed, and C. Adams, "In their own words: The impact of subtle language and communication difficulties as described by autistic girls and boys without intellectual disability," *Autism*, vol. 26, pp. 332–345, 2 2022.

[11] M. L. Cola, S. Plate, L. Yankowitz, V. Petrulla, L. Bateman, C. J. Zampella, A. D. Marchena, J. Pandey, R. T. Schultz, and J. Parish-Morris, "Sex differences in the first impressions made by girls and boys with autism," *Molecular Autism*, vol. 11, 6 2020.

[12] S. Cho, M. Liberman, N. Ryant, M. Cola, R. T. Schultz, and J. Parish-Morris, "Automatic detection of autism spectrum disorder in children using acoustic and text features from brief natural conversations," vol. 2019-September. International Speech Communication Association, 2019, pp. 2513–2517.

[13] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[14] S. Velupillai, H. Suominen, M. Liakata, A. Roberts, A. D. Shah, K. Morley, D. Osborn, J. Hayes, R. Stewart, J. Downs, W. Chapman, and R. Dutta, "Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances," *Journal of Biomedical Informatics*, vol. 88, pp. 11–19, 2018.

[15] N. A. Chi, P. Washington, A. Kline, A. Husic, C. Hou, C. He, K. Dunlap, and D. P. Wall, "Classifying autism from crowdsourced semistructured speech recordings: Machine learning model comparison study," *JMIR Pediatrics and Parenting*, vol. 5, 4 2022.

[16] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedule—generic: 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.

[17] A. B. Ratto, L. Turner-Brown, B. M. Rupp, G. B. Mesibov, and D. L. Penn, "Development of the contextual assessment of social skills (cass): A role play measure of social skill for individuals with high-functioning autism," *Journal of Autism and Developmental Disorders*, vol. 41, pp. 1277–1286, 9 2011.

[18] L. Gauder, L. Pepino, L. Ferrer, and P. Riera, "Alzheimer disease recognition using speech-based embeddings from pre-trained models," 2021.

[19] C. Howes, M. Purver, and R. McCabe, "Using conversation topics for predicting therapy outcomes in schizophrenia," *Biomedical informatics insights*, vol. 6, pp. BII–S11 661, 2013.

[20] D. Jurafsky, R. Ranganath, and D. McFarland, "Extracting social meaning: Identifying interactional style in spoken conversation," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. NAACL '09. USA: Association for Computational Linguistics, 2009, p. 638–646.

[21] 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.

[22] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *CoRR*, vol. abs/1201.0490, 2012. [Online]. Available: http://arxiv.org/abs/1201.0490

[24] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[25] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.