TOWARDS PRECISION CHARACTERIZATION OF COMMUNICATION DISORDERS USING MODELS OF PERCEIVED PRAGMATIC SIMILARITY

Nigel G. Ward[1], *Andres Segura*[1], *Georgina Bugarini*[1],
Heike Lehnert-LeHouillier[2], *Dancheng Liu*[3], *Jinjun Xiong*[3], *Olac Fuentes*[1]

[1]University of Texas at El Paso, [2] New Mexico State University, [3] University at Buffalo
United States

ABSTRACT

The diagnosis and treatment of individuals with communication disorders offers many opportunities for the application of speech technology, but research so far has not adequately considered: the diversity of conditions, the challenges of limited data, and the role of pragmatic deficits. This paper explores how a general-purpose model of perceived pragmatic similarity may overcome these limitations. It shows that a simple model can capture utterance aspects that are relevant to diagnosis of autism and of specific language impairment, outlines how it might support several use cases for clinicians and clients, and analyzes its performance and limitations.

Index Terms— precision medicine, dialog behaviors, social interaction, pretraining, low-resource tasks, autism, atypical behavior

1. MOTIVATIONS

Communication disorders affect many people's lives, and the potential for speech technology to help has inspired much research. Most successes in this area, however, follow a pattern [1], in being limited to categorical diagnoses, limited to cases where ample and convenient labeled data is available [2], and limited in terms of helping clients understand their conditions. It is also poorly aligned with the modern concept of "precision medicine" [3], and the aim of deeper modeling of conditions with complex etiologies (such as psychiatric disorders [4, 5, 6], where an important goal is discover latent factors, and eventually ties to genetic, neural, developmental, environmental, or other root causes). While research in this pattern has led to models that perform well on many curated data sets, this pattern is not suitable for many applications, and translation to clinical practice has been slow [7]. We see three general needs.

Need 1: Precise Diagnosis and Categorization

Many individuals with communication disorders have non-stereotypical conjunctions of symptoms, each perhaps present to varying degrees [7]. For example, someone diagnosed with autistism may have some characteristics mildly or not at all, and may also have some elements of anxiety, etc. There are also many people with subclinical traits, some of whom may still wish to better understand their condition and learn to communicate better. Such people can potentially benefit from information beyond being tagged with one of a few possible labels. While previous speech technology work on communicative disorders has mostly targeted simplistic diagnosis, we aim to support more nuanced characterizations. These could be useful for clinicians and also for clients seeking to self-monitor, especially for individualized treatments. Incidentally, while there are many qualms about automated rather than human-provided diagnoses, AI to support clients in better understanding their conditions is less likely to be concerning.

Need 2: Small-Data Modeling

Obtaining enough data to build models of communications disorders is always a challenge, often due to various privacy concerns. In the usual research pattern, an assumption of heterogeneity within each disorder is taken to justify grouping data across many individuals. This conveniently boosts the usable data size, but at the cost of imprecise categorization, as noted above. Further, many factors affect language skills, so the reference data should ideally be limited to speech from people with the same age, gender, demographic, family environment, dialect, and so on [8], also reducing the amount of relevant data. Moreover, the increasing desire to evaluate communicative effectiveness in naturalistic conditions — especially in dialog — rather than via read speech, may further reduce the available amount of comparable and good-quality data. These considerations impel us to seek ways to get more out of limited data.

Need 3: Consideration of Pragmatic Aspects

For some communication disorders, especially the neurogenic ones, the greatest issue is difficulty in correctly conveying or perceiving *pragmatic* intents, as the abilities to use speech functionally and in interaction can be the most important aspects for communicative effectiveness and acceptance by peers [9]. Typical pragmatic functions include, for example, the intent to praise, to mark contrast or novelty, to change the topic, to yield the turn or take it, and so on. Phonological, lexical and syntactic deficits are easier to measure, and have been the focus of most modeling to date, but pragmatic behaviors and perceptions also matter; social success depends not only on the intent of the speaker but also on interlocutors' perceptions.

This paper explores how these needs may be addressed with a general-purpose model of perceived pragmatic similarity: one that approximates human judgments of "how pragmatically similar are [any two utterances] in terms of the overall feeling, tone, and intent" [10]. Incidentally, unlike *semantic* similarity [11, 12, 13, 14, 15], modeling *pragmatic* similarity is something new [10, 16].

2. CLASSIFYING SPEAKERS BY CONDITION

Our first set of experiments focused on Need 3. We hypothesized that a model of pragmatic similarity would be adequate for capturing utterance aspects relevant to communicative disorders, and tested this by building a classifier and evaluating it on two diagnostic tasks.

2.1. Classification Method

Our method is based on the assumption that the utterances of people with a condition will be mostly similar to those of other people

	model prediction	
	ASD	NT
speakers with autism diagnoses	10	4
neurotypical speakers	1	13

Table 1. Autistic vs Neurotypical discrimination

	model prediction	
	SLI	TD
SLI speakers	36	31
typically developing speakers	3	64

Table 2. SLI vs TD discrimination

with the same condition, and conversely for those without the condition. This assumption is probably not unrealistic, unlike the stronger assumption, often made for many conditions, that there is a core underlying cause or marker. For autism, for example, it is known that the prosodic behaviors are diverse, as seen by conflicting results in the literature [17, 18, 19, 20], referencing both inappropriately loud speech and too soft voice, and both monotone pitch and overly wide pitch range. Use of a similarity-based method can be robust to such diversity. Further, a model that provides a similarity metric can support nearest-neighbor classifiers, which are often the best choice for tasks with only limited data available.

The specific pragmatic similarity model we used was Segura's [10]. In brief, this computes, for each utterance, features using a pretrained HuBert model, and estimates similarity using the cosine of the two feature vectors. Our classification algorithm leverages this. It has three steps. First, it classifies each utterance of an unknown speaker as likely representing a condition or not, based on the majority of the 7 nearest neighbors (kNN) in the reference data. It does this 24 times, using all 24 layers of HuBert features in turn. Second, it makes an overall classification of each clip based on the majority vote of the per-layer classifications. Finally it classifies the speaker as having a condition or not based on the label assigned to majority of his or her utterances. We tested this algorithm leave-one-speaker-out style on two datasets.

2.2. Evaluation for Autism

We first used the NMSU dataset of age-matched autistic (ASD) and neurotypical (NT) adolescents engaged in a find-the-difference task with a confederate [21]. For this data set there were 28 speakers, with typically 5-10 minutes of data for each. As seen in Table 1, the accuracy was 82%, with 1 neurotypical and 4 autistic speakers misclassified. Lacking publicly-available data for evaluation, there is no way to benchmark this method, but these results do not seem inferior to those reported by the most comparable previous work [22, 23]. To clarify, we do not claim this to be an advance in autism detection, but rather a demonstration that a very simple model, whose parameters are fixed and require no tuning on autism-specific data, still has substantial discriminative power. On further examination, we found that the misclassified speakers were of three categories: having inadequate data (few utterances), being among the youngest neurotypicals, or having autism diagnoses but relatively low ADOS scores. Thus the mispredictions do not seem to be attributable to failures of the similarity model or the kNN approach.

	model prediction	
	ASD	NT
speakers with autism diagnoses	13	8
neurotypical speakers	1	6

Table 3. ASD vs NT discrimination without ASD data

2.3. Evaluation for Specific Language Impairment

We next used the Edmonton Narrative Norms Instrument [24], a corpus of children ages 4 to 10 retelling stories, served at Talkbank, including children with Specific Language Impairment (SLI) [25], and typically developing (TD) children. SLI is not known to involve a pragmatics-related deficit, but we chose this data because of availability and our interest in younger children's speech. The audio was segmented into utterances. As there was more data in the typically-developing category, we downsampled it to 67 speakers, while matching the age distribution of the SLI children. There were an average of about 35 utterances per child. The results are fair, as seen in Table 2. Examining dependencies on age revealed the discriminations to be at chance for the oldest group, the 10-year-olds, but better for other ages. For comparison, we built a baseline using the average length of utterance, in seconds, which classified children whose average utterance length was less than 70% of the TD average as SLI. While its performance was above chance, our similarity-based method was far more accurate. Overall, this suggests that our model is capturing aspects of similarity beyond the purely pragmatic. Depending on the purpose, this can be a good thing.

2.4. Performance without Condition-Specific Data

Regarding Need 2, a common low-resource scenario is one in which we have only data from the typical/normal population, with no condition-specific data. If we assume that communicative disorders are departures from the norm, it should be possible to identify an individual as impaired or not, even without any disorder-specific data. If so, then we can use a similarity metric to detect atypicality: if many of a speaker's utterances are not similar to anything produced by the reference/normal population, then they may have some condition. Of course, such a vague characterization would not be useful for diagnosis, but could support screening to refer children for professional evaluation.

To test this idea, we again used the NMSU and ENNI data sets, with two measures of typicality: an utterance-by-utterance model and a speaker-centroid model. In the former, we rated speaker typicality by how similar their utterances were, on average, to the closest 3 of those of all the typical speakers (while leaving out the speaker himself, if in the typical set). In the second model, we rated speaker typicality by the distance between the centroid of all their utterances and the centroid of all typically-developing speakers. For both models, we varied the thresholds post hoc, however performance was smooth, so this likely only slightly overstates the actual utility. For the ENNI data there was no benefit with either model, but for the NMSU data both models did moderately well, with the centroid model doing slightly better. In particular, a cosine threshold of 0.97 gave the performance shown in Table 3.

Unsurprisingly, the performance is weaker than when exploiting ASD-specific data, but there is still some value. We plan to investigate whether this can be improved by conditioning on the context and/or by excluding utterances that are atypical for the speaker.

3. SUPPORT FOR NEW USE CASES

Having established that a pragmatic similarity model can capture utterance aspects relevant to communication disorders, this section considers the posibility for Need 1. We propose five use cases:

Use Case A: Detecting Atypical Speakers. Given a collection of recordings of people, clinicians might like a tool to automatically screen for atypical speakers, identified as those whose utterances are low in similarity to those of other speakers, especially when diagnostic norms do not exist.

Use Case B: Finding Similar Speakers. Clients or clinicians may like a tool able to identify a very similar speaker in a dataset. Enabling clients to hear speech from similar others, that is, people whose behaviors are perceived similarly, might help them understand how they themselves are perceived by other people. This could help them more vividly understand their prospects, or help them match with appropriate support groups. Further, while some clinicians may have enough experience to recall how previous clients, similar in some way to the current one, responded to various interventions, novice clinicians may lack such experience. Thus such a tool could be useful for clinician training or to provide support for "looking up" similar clients via their recordings in a dataset.

Use Case C: Finding Typical Utterances. Clinicians may want the ability to quickly find one or more typical or representative utterances by the speaker, as a way to focus their attention for detailed analysis [26] or to choose sample utterances to include in reports.

Use Case D: Finding Comparable Utterances. Clinicians may want to be able to use a specific utterance from a child of interest as a "query" and have the system retrieve similar utterances from other speakers in the dataset.

Use Case E: Identifying Atypical Utterances. Clinicians may also like to have automatic support for finding *atypical* utterances, to be able to hear, for a single speaker, utterances that they have produced that are saliently non-typical (low in similarity to anything in the reference data), and thus likely to be possibly perceived negatively by peers. Further, detection of atypical utterances could be useful in a preprocessing step to classification or to Use Cases A and B, for the sake of excluding outlier utterances that may reflect noise or some unusual transient speaker state, and are thus likely uninformative. Alternatively, some atypical utterances might be informative as children's "leading-edge" behaviors [27].

Thus, A and B involve identifying individuals, and C, D, and E identifying utterances of interest. Today, the assessment and diagnosis of dialog abilities and behavior patterns is labor-intensive. Discussion with speech and language pathologists (SLPs), followed by a survey of 18 practitioners, indicated that they see the most potential in Use Cases C, A, and D, in that order. We suspect the value may be even greater as behavior sampling evolves from data collected in structured ways to data collected in more naturalistic conditions, which may be voluminous. In short, there are many ways in which a model of pragmatic similarity could support clinicians.

4. MODEL VALIDATION

Previous evaluation of Segura's model [10] considered mostly overall correlations, but these are not ideally informative. For example, a slightly improved version of Segura's model, in which we use only features from HuBert's 24th layer, and downselect these using a custom feature-selection process to maximize performance, was shown, in our own previous work [28], to correlate 0.74 with human judgments, which is not far below human inter-annotator agreement. However, these correlations were computed over a wide span of similarities, and thus were heavily swayed by the basic ability to separate rather similar pairs from completely dissimilar pairs.

This ability is not adequate for several of our use cases (B, C, and D), which need a model that is also able to identify the pragmatically *most* similar utterances. This ability is also the critical one for use in nearest-neighbor classifiers, as illustrated earlier.

We accordingly set out to measure this ability. We hypothesized that the model would be able to discriminate the most similar utterances from those that were not similar or less similar. We tested this using a task inspired by Use Case D, finding similar utterances from a different speaker, which seems technically to be the most challenging of the use cases. The rest of the section describes the experiment.

4.1. Stimuli, Instruments, Subjects, and Procedure

We started with the DRAL corpus, which contains 2893 pragmatically-varied English utterances from conversations among 129 college students. Ideally, for a direct test of the model, we would have humans first identify, for each of these utterances, the most similar utterances in the corpus. We would then have been able to directly test the ability of the model, given any utterance, to find the most similar. However, having subjects listen to thousands of utterance pairs is not realistic. Instead, we downsampled.

Specifically, for each of 31 randomly selected reference utterances, we paired it with each of 10 candidates for the most similar other utterance, and presented these pairs to human subjects for judgments. The 10 candidates were chosen using the improved Segura model: Three were the top-three most similar, according to the model, and, as distractors, the seven whose similarity levels were at percentiles 99, 97, 95, 90, 80, 60, and 30, according to the model. In this way we ensured that very similar utterances were over-represented. All clips were 2-7 seconds long. To avoid speaker-identify confounds, each candidate utterance was constrained to be by a speaker different from the speaker of the corresponding reference utterance.

For each utterance pair, judges rated "how pragmatically similar are the two clips, in terms of the overall feeling, tone, and intent" on a scale from 1 to 5. They also recorded their top 3 rankings, which, although not really providing additional information, likely encouraged extra care with the judgments at the high end of the scale.

We recruited nine judges, mostly undergraduates, who we knew to be highly sensitive to the nuances of language, drawing mostly on the pool of those involved in our recent past studies. These subjects came in on a Saturday for a 4-hour session. Their compensation was 70 dollars plus lunch. We first provided an overview of the study and informal training, including listening to a variety of clip pairs and clip-pair sets, making judgments, and discussing the bases for judgments. (There was often variation, and from the discussions it was clear that this was often due to noticing different aspects or weighting them differently. For example, for some clips some judges reported focusing more on similarity in terms of the "feelings" conveyed, while others focused more on similarity in terms of the "sound".)

For each of the 31 sets, judges listened to the ten pairs, each comprised of the reference and the candidate separated by a beep. They rated each pair using QuestionPro sliders. Any judge could request repetitions of any pairs until they felt confident in their ratings. Typically over half of the pairs were re-listened to in this way, often multiple times. The judges were all in one room, so all heard the same stimuli the same number of times.

	Random Baseline	Our Method	Human Judges
Ratings Correlation	0.00	0.18	0.28
Recall@1 (= Precision@1)	10%	15%	20%
Recall@3	30%	43%	48%
Top 3 Intersection	0.90	1.12	1.24

Table 4. Ability to Identify the Most-Similar Pairs

4.2. Results

A good model is one for which the pairs that are predicted to be most similar are actually rated highly similar by the human judges. Table 4 shows the results with the improved model. The first row shows the correlations with human ratings, across all 310 pairs and all 9 judges. Clearly this is a hard task even for human judges. Further, we found that the model's ability to predict the averages of the human judgments, which we can take as an approximate gold standard, was much better, with a 0.30 correlation. We also examined whether the difference between the model's ratings of two pairs could serve as a measure of its confidence, and indeed there was a positive correlation between that difference and the number of times the system correctly predicted which of two pairs would be judged more similar. Incidentally, while we did not systematically explore the modeling space, we did obtain a better correlation by modifying Segura's model to normalize each feature to have 0 mean, across the entire training data, before taking the cosine, but this modification hurt performance on the more important metrics.

These metrics are seen in the next three rows of Table 4. First, we see that exact agreement on the #1 most similar clip was low, both for the system and the humans. This was not surprising, because all of the top 3 were designed to be very close in similarity. Second, we see that the human-judged #1 most similar clip was hard to identify even approximately, as it appeared in the system's top 3 choices only 43% of the time, and humans did only modestly better. The last measure is the most relevant: the number of clips in the top-3 human rankings that appeared in the model's top-3 rankings, averaged over all 9 judges across all 31 sets. As seen in the table, the system's performance was above expectation, and this was significant ($p < .005$, t-test, 31 samples), but below human performance.

To better understand the performance of the model, we did failure analysis over two overlapping subsets of the data. The first was the pairs for which the system's predicted similarity diverged most from the average human judgments. The second was the pairs that the model rated in the top 3, but whose average similarity ranking by the human judges was low, representing the failures that would be most harmful for the use cases. We found three common causes. The first common cause was pairs for which one or both utterances were acoustically unlike most of the training data. For example, these included an utterance with an ingressive exclamation, and several cases which were very quiet and reduced, apparently spoken between friends who understood each other well. In most use cases, such unusual utterances will likely be rare, so this may not be problematic in practice. The second common cause was that the model was apparently often sensitive to pragmatic differences that were less important to our judges, such as different turn-taking intentions (hold versus yield), and the different evidentiality (statements based on first-person experience versus not). The third common cause was, conversely, that the model was sometimes apparently sensitive to dimensions of similarity that the human judges overlooked, or which may have been overshadowed by salient category distinctions. For example, this was the case for two clips, one in the form of a question and the other in the form of a statement, which were both seeking a confirmation of an amount or quantity.

Overall we see that the model has some ability to pick out the most-similar utterances, and no obvious show-stopper failure modes.

5. CONTRIBUTIONS AND OUTLOOK

Our first contribution is a discussion of ways in which a general pragmatic-similarity model could be a useful addition to the speech technology toolset for communication disorders, in terms both of unmet needs and of specific use cases.

Our second contribution is the finding that similarity can capture utterance aspects that are relevant to diagnoses (Tables 1 and 2). Since nothing in our approach was autism-specific or SLI-specific, we expect that this approach may be useful for many other conditions, or mixes of conditions.

Our third contribution is the finding that similarity can support discriminations without only small data and no task-specific training (Table 3). This indicates the potential value of shunting off the hard modeling work to a general model, in line with the increasingly popular strategy of pretraining on large, general data, and then using the model for specific tasks.

Our fourth contribution, relating to precise and richly-understandable characterization, is the finding that a pragmatic similarity model can support at least Use Case D: identifying most-similar utterances. Further, small informal studies of the other use cases are so far suggesting that the model can support them also.

Beyond applications, similarity models may also provide a new avenue for basic research on characterizing the space of communicative disorders.

We have obtained promising results despite using only a very simple similarity model, that uses only the speech signal and only considers single utterances. Future work should seek to improve and extend this: Beyond using the information in speech, cases where speakers are silent, contrary to expectation, are also highly informative. Beyond the speech signal, a model might consider multimodal information, such as gestures and activities. Beyond single utterances, a model might consider more context, such as the interlocutor's recent behavior, or the recent interaction style [29] as a proxy for the task and environment. Beyond acoustic-prosodic features, a model might also exploit the words. Beyond a black-box similarity model, one could use an explainable model built on perceptually-relevant prosodic features [30].

Even with our existing model, much work remains to be done. We need to test with more data and test actual utility through user studies.

Acknowledgments

We thank Arin Rahman for help with the analysis; Pamela Hadley, Carol Anne Miller, Hedda Meadan-Kaplansky and Tracy Preza for discussion of the SLP perspective; and Hyewon Suh and Julie Kientz for the survey of SLPs. This work was supported in part by the National Science Foundation and the Institute of Education Sciences, U.S. Department of Education, through Award # 2229873 – National AI Institute for Transforming Education for Children with Speech and Language Processing Challenges, and by NSF Award 2348085.

6. REFERENCES

[1] M. Milling, F. B. Pokorny, K. D. Bartl-Pokorny, and B. W. Schuller, "Is speech the new blood? Recent progress in AI-based disease detection from audio in a nutshell," *Frontiers in Digital Health*, vol. 4, 2022. paper id: 886615.

[2] T. Linzen, "How can we accelerate progress towards human-like linguistic generalization?," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5210–5217, 2020.

[3] K. B. Johnson, W.-Q. Wei, D. Weeraratne, M. E. Frisse, K. Misulis, K. Rhee, J. Zhao, and J. L. Snowdon, "Precision medicine, AI, and the future of personalized health care," *Clinical and translational science*, vol. 14, no. 1, pp. 86–93, 2021.

[4] T. R. Insel, "The NIMH research domain criteria (RDoC) project: Precision medicine for psychiatry," *American Journal of Psychiatry*, vol. 171, pp. 395–397, 2014.

[5] I. C. Passos, P. Ballester, F. D. Rabelo-da Ponte, and F. Kapczinski, "Precision psychiatry: The future is now," *The Canadian Journal of Psychiatry*, vol. 67, no. 1, pp. 21–25, 2022.

[6] P. W. Foltz, C. Chandler, C. Diaz-Asper, A. S. Cohen, Z. Rodriguez, T. B. Holmlund, and B. Elvevåg, "Reflections on the nature of measurement in language-based automated assessments of patients' mental state and cognitive function," *Schizophrenia Research*, vol. 259, pp. 127–139, 2023.

[7] V. P. Martin and J.-L. Rouas, "Why voice biomarkers of psychiatric disorders are not used in clinical practice? Deconstructing the myth of the need for objective diagnosis," in *International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pp. 17603–17613, 2024.

[8] A. S. Cohen, T. L. Renshaw, K. R. Mitchell, and Y. Kim, "A psychometric investigation of "macroscopic" speech measures for clinical and psychological science," *Behavior Research Methods*, vol. 48, pp. 475–486, 2016.

[9] K. Bottema-Beutel, "Glimpses into the blind spot: Social interaction and autism," *Journal of Communication Disorders*, vol. 68, pp. 24–34, 2017.

[10] N. G. Ward, A. Segura, A. Ceballos, and D. Marco, "Towards a general-purpose model of perceived pragmatic similarity," in *Interspeech*, 2024.

[11] M. Rouhizadeh, R. Sproat, and J. Van Santen, "Similarity measures for quantifying restrictive and repetitive behavior in conversations of autistic children," in *Association for Computational Linguistics. North American Chapter Meeting (NAACL)*, vol. 2015, pp. 117–123, 2015.

[12] J. Wieting, T. Berg-Kirkpatrick, K. Gimpel, and G. Neubig, "Beyond BLEU: Training neural machine translation with semantic similarity," in *ACL*, 2019.

[13] L. Besacier, S. Ribeiro, O. Galibert, and I. Calapodescu, "A textless metric for speech-to-speech comparison," *arXiv preprint arXiv:2210.11835*, 2022.

[14] S. Gehrmann, E. Clark, and T. Sellam, "Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text," *Journal of Artificial Intelligence Research*, vol. 77, pp. 103–166, 2023.

[15] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim, *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[16] K. Heffernan, A. Kozhevnikov, L. Barrault, A. Mourachko, and H. Schwenk, "Aligning speech segments beyond pure semantics," in *Findings of the Association for Computational Linguistics*, pp. 3626–3635, 2024.

[17] J. McCann and S. Peppe, "Prosody in autism spectrum disorders: A critical review," *International Journal of Language & Communication Disorders*, vol. 38, pp. 325–350, 2003.

[18] A. McAlpine, L. W. Plexico, A. M. Plumb, and J. Cleary, "Prosody in young verbal children with autism spectrum disorder," *Contemporary Issues in Communication Science & Disorders*, vol. 41, pp. 120–132, 2014.

[19] G. Kiss, *Prosody of Spontaneous Speech in Autism*. PhD thesis, Oregon Health & Science University, 2017.

[20] S. Wehrle, F. Cangemi, M. Kruger, and M. Grice, "Somewhere over the spectrum: Between singsongy and robotic intonation," in *Proceedings of 14th Associazione Italiana Scienze della Voce Conference*, pp. 1–19, 2019.

[21] H. Lehnert-LeHouillier, S. Terrazas, and S. Sandoval, "Prosodic entrainment in conversations of verbal children and teens on the autism spectrum," *Frontiers in Psychology*, vol. 11, p. 2718, 2020.

[22] 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 *Interspeech*, 2012.

[23] 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, pp. 1121–1136, 2015.

[24] P. Schneider, D. Hayward, and R. V. Dubé, "Storytelling from pictures using the Edmonton narrative norms instrument," *Journal of Speech Language Pathology and Audiology*, vol. 30, pp. 224–238, 2006.

[25] G. Sharma, X.-P. Zhang, K. Umapathy, and S. Krishnan, "Audio texture and age-wise analysis of disordered speech in children having specific language impairment," *Biomedical Signal Processing and Control*, vol. 66, 2021. paper id: 102471.

[26] I. S. Klatte, V. Van Heugten, R. Zwitserlood, and E. Gerrits, "Language sample analysis in clinical practice: Speech-language pathologists' barriers, facilitators, and needs," *Language, Speech, and Hearing Services in Schools*, vol. 53, no. 1, pp. 1–16, 2022.

[27] M. Rispoli and P. Hadley, "The leading-edge: The significance of sentence disruptions in the development of grammar.," *Journal of Speech, Language & Hearing Research*, vol. 44, no. 5, 2001.

[28] N. G. Ward and D. Marco, "A collection of pragmatic-similarity judgments over spoken dialog utterances," in *Linguistic Resources and Evaluation Conference*, 2024.

[29] N. G. Ward and J. E. Avlia, "A dimensional model of interaction style variation in spoken dialog," *Speech Communication*, vol. 149, pp. 47–62, 2023.

[30] N. G. Ward, D. Marco, and O. Fuentes, "Which prosodic features matter most for pragmatics?," in *IEEE ICASSP*, 2025.