A Cascaded Multimodal Framework for Automatic Social Communication Severity Assessment in Children with Autism Spectrum Disorder

Jihyun Mun[1], *Sunhee Kim*[2], *Minhwa Chung*[1]

[1]Department of Linguistics, Seoul National University, Republic of Korea
[2]Department of French Language Education, Seoul National University, Republic of Korea
{jhhh_1202, sunhkim, mchung}@snu.ac.kr

Abstract

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterized by deficits in social communication, affecting both language use and speech patterns. Since assessment relies on behavioral observations rather than standardized medical tests, developing an objective evaluation method is essential. Recognizing that ASD impacts both language and speech production, this study proposes a cascaded multimodal framework for ASD severity assessment. The framework processes raw audio, generates transcriptions via automatic speech recognition, and extracts linguistic and acoustic features using speech-language foundation models. Given the atypical suprasegmental and segmental speech characteristics in ASD, two speech foundation models are employed. A co-attention mechanism then integrates these representations to estimate severity. Achieving a Spearman's correlation of 0.5629 with human ratings, the proposed approach offers a scalable, fully automated ASD assessment tool.

Index Terms: autism spectrum disorder, multimodal, cascaded framework, automatic speech recognition, automatic assessment

1. Introduction

Autism Spectrum Disorder (ASD) is a group of neurodevelopmental conditions characterized by impairments in social communication and interaction, alongside restricted, repetitive patterns of behavior [1]. A defining characteristic of ASD is persistent deficits in social communication, which manifest as atypical language use and distinct speech production patterns [2]. ASD affects multiple linguistic domains, including vocabulary, syntax, morphology, and pragmatics, all of which are essential for effective social interaction. In terms of speech production, children with ASD often exhibit atypical patterns at both the suprasegmental and segmental levels [3, 4, 5]. In suprasegmental aspects, they may demonstrate unconventional prosody, such as monotone or exaggerated pitch contours, pitch variation, and speech energy distribution [3, 6]. At the segmental level, articulatory and phonological development is often delayed, as reflected in atypical vowel formant patterns and pronunciation errors that are both developmentally inappropriate and deviate from typical patterns [3, 7]. Such speech and language challenges hinder social engagement and have long-term implications for social integration.

Timely intervention is crucial for mitigating ASD symptoms and improving quality of life [8]. Although typical developmental processes promote natural acquisition of core social skills, children with ASD often require structured interventions to develop joint attention, social referencing, and verbal communication skills [9]. Currently, no standardized medical test exists for ASD assessment, and clinicians commonly rely on observational instruments such as the Autism Diagnostic Observation Schedule (ADOS). These methods are vulnerable to subjectivity, examiner bias, and variability in caregiver reports [10]. Delays in assessment can impede early and proper therapeutic interventions, emphasizing the urgent need for objective, automated assessment tools.

Given the prominent role of speech and language in ASD-related deficits, a number of studies have explored computational approaches leveraging speech or language features for ASD diagnosis [11, 12, 13, 14, 15]. However, only a limited number of studies have addressed the quantification of ASD severity using speech or language features [16, 17, 18], and even fewer have proposed integrated frameworks that jointly leverage both acoustic–prosodic and lexical cues [16, 19, 20, 21]. Most existing approaches still rely on handcrafted acoustic features or manual transcripts, requiring substantial domain expertise and potentially introducing subjectivity. Although recent advances in deep learning enable automatic extraction of high-level representations from raw data [22], and deep learning-based methods have shown promise in ASD detection and severity assessment [16, 17, 23], there remains a lack of fully automated systems that capture both acoustic and linguistic factors without relying on human-generated transcripts.

To address these limitations, this study proposes a novel, fully automated cascaded multimodal framework for assessing the social communication severity of children with ASD. The framework processes raw audio waveforms through a cascaded pipeline that integrates automatic speech recognition (ASR) and speech-language foundation models. The ASR module generates transcriptions, which are then analyzed alongside the raw audio input using speech-language foundation models. Given that children with ASD exhibit atypical suprasegmental and segmental speech characteristics, we leverage two speech foundation models for speech representation. To enhance modality interaction, the framework incorporates a co-attention mechanism that generates speech-conditioned text embeddings and text-conditioned speech embeddings, which are then fused to generate a final representation for prediction. By eliminating the need for manually engineered features and human transcription, this approach offers a scalable and objective solution, potentially enabling earlier and more consistent clinical assessment and intervention.

The remainder of this paper is structured as follows: Section 2 details the proposed methodologies, Section 3 describes the experimental setup and results, Section 4 discusses the findings, and Section 5 concludes the paper.

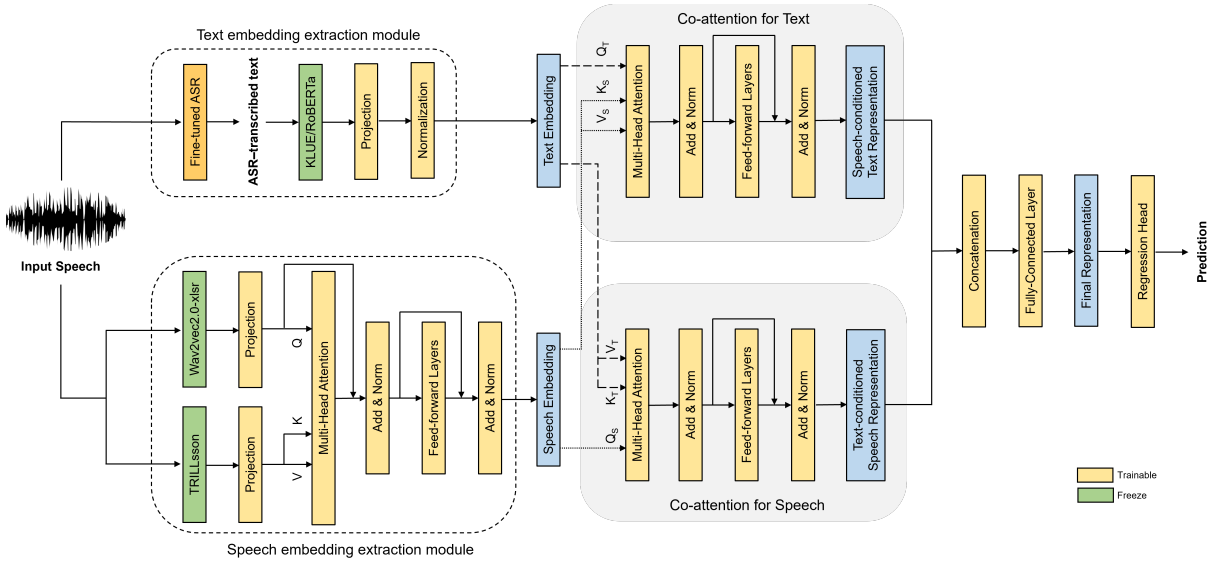

Figure 1: *Overview of the proposed ASD severity prediction framework*

2. Methods

Figure 1 illustrates the overall architecture of the proposed framework, which consists of three main components: (i) Speech and text embedding extraction, (ii) Co-attention modules for multimodal fusion, and (iii) Feature fusion followed by a final regression head for ASD severity prediction.

2.1. Speech and Text Embedding Extraction

To leverage both speech and language-specific information, we extract embeddings from each modality using pre-trained foundation models. All foundation model weights are frozen, without additional training, to fully utilize their inherent knowledge.

2.1.1. Speech Embedding Extraction

We employ two complementary speech foundation models: TRILLsson [24] for suprasegmental information and wav2vec2.0-xlsr [25] for segmental features.

TRILLsson is specifically designed for paralinguistic tasks such as emotion recognition, speaker identification, and dysarthria classification [24], and it has been applied in diagnosing and assessing various neurocognitive and speech-related disorders [26, 27, 28]. We use a TRILLsson variant based on an Audio Spectrogram Transformer architecture. To retain rich suprasegmental information, we extract the penultimate layer embeddings, preserving sequence-level temporal variations before dimensionality reduction. Wav2vec2.0 embeddings are known to capture phonological and syllabic information effectively [29, 30]. For segmental speech characteristics, we extract the last hidden state from the wav2vec2-large-xlsr-53 model. We adopt this model to leverage its multilingual training data, which includes Korean.

Since TRILLsson and wav2vec2.0 operate at different convolutional strides, their extracted embeddings differ in sequence length. We apply linear interpolation to align TRILLsson's output with wav2vec2.0's sequence length and then combine the two set of embeddings using a transformer-like processing block. Specifically, we treat wav2vec2.0 embeddings as the query and TRILLsson embeddings as the key and value in a multi-head attention mechanism. This process enriches segmental tokens with suprasegmental information. The output is then passed through a series of residual connection and feed-forward layers to obtain the final speech embedding representation.

2.1.2. Text Embedding Extraction

Due to the difficulty of obtaining human transcriptions for ASD speech, we incorporate an ASR model to generate transcriptions. We fine-tune the Whisper-large-v2 [31] ASR model on ASD and TD (typically developing) children's speech data from our training set, reducing the syllable error rate (SER) from 93.05% to 30.86%.

Text embeddings are generated from an ASR-transcribed or manually transcribed text using a pre-trained Korean language model, klue/RoBERTa-base [32]. We use the final hidden state to capture semantic and contextual information. The extracted embeddings pass through a projection layer to align with the speech embeddings. Then they undergo a normalization step before being used in the subsequent fusion process.

2.2. Co-attention for Multi-modal Fusion

We employ co-attention transformer blocks [33] to fuse speech and text representations. Two parallel attention processes are performed:

- Speech-conditioned text representation: text embeddings serve as the query, and speech embeddings serve as the key and value, yielding speech-informed linguistic features.
- Text-conditioned speech representation: speech embeddings serve as the query, and text embeddings serve as the key and value, capturing text-informed speech features.

This bidirectional co-attention mechanism enhances modality-specific features by modeling cross-modal dependencies.

2.3. Feature Fusion and Prediction

We concatenate the speech-conditioned and text-conditioned outputs, then feed the result through a fully connected layer,

producing a single multimodal representation. Finally, this representation is passed to a regression head that predicts the severity level of ASD. Unlike classification-based approaches, we adopt a regression framework to capture the ordinal nature of ASD severity levels.

3. Experiments

3.1. Data

We use the Korean ASD corpus introduced by [7], which provides detailed information on severity measurements and participant demographics. In this corpus, each child's social communication severity is independently evaluated by three certified speech-language pathologists, and the average of these three ratings serves as the child's overall severity score. Higher severity levels indicate greater social communication impairment. The dataset is split into training, validation, and test sets at an 8:1:1 ratio in a speaker-independent manner, maintaining stratified severity distribution. Table 1 summarizes the dataset.

Table 1: *Data distribution*

ASD Severity	# of Speakers	# of Utterances (Duration)
Level 1	95	31357 (14h 44m 55s)
Level 2	61	11881 (6h 8m 55s)
Level 3	56	4181 (3h 17m 22s)

3.2. Experimental Setup

We implement all experiments in PyTorch following the framework described in Section 2.

3.2.1. Hyperparameters

We set the projection dimension to 512 for all linear layers and multi-head attention modules. Specifically, both speech and text embeddings are projected to a 512-dimensional space before entering the co-attention blocks. Key hyperparameters include:

- Model initialization: He initialization
- Optimizer: AdamW with learning rate of 3e-5 and weight decay of 1e-3
- Loss function: Mean Squared Error (MSE)
- Batch size: 64
- Number of heads for multi-head attention: 4
- Dropout: 0.1 in all attention and feed-forward layers
- Number of epochs: Up to 200
- Early stopping: Training stops if validation loss does not improve for 20 consecutive epochs
- Gradient clipping: Global norm clipped to 1.0

3.2.2. Seed Ensemble

To mitigate the variability introduced by random initialization, we train five models using different random seeds (0-4) and average their predictions during inference. This approach enhances the stability and robustness of the model's predictions.

3.2.3. Evaluation Metrics

We employ Spearman's correlation coefficient (SCC) to measure the correlation between the predicted ASD severity scores and the ground truth labels. SCC is suitable for ordinal data and quantifies correlation in rank ordering.

3.3. Results

We conducted a series of experiments to evaluate the effectiveness of the proposed multimodal framework illustrated in Figure 1. Key components-speech encoder, text encoder, and text input type (manual vs. ASR)-were systematically varied to assess their individual contributions.

In the unimodal settings (Configurations 1–5), we separately evaluated speech-only and text-only models. For speech, we tested embeddings derived from two pretrained models and their combination. For text, we compared embeddings generated from manual transcripts and ASR-generated transcripts, both processed by the same text encoder. In the multimodal settings (Configurations 6–11), we explored various combinations of speech and text inputs. Configuration 10 isolated the effect of ASR input by substituting manual transcripts, while Configuration 11 represents the full model, incorporating both speech encoders and ASR transcripts.

Table 2 summarizes the SCCs for all configurations in the ASD severity estimation task.

Table 2: *Spearman's correlation coefficients (SCCs) of uni- and multi-modal configurations for ASD severity assessment*

Type	Experimental Configuration	SCC
Unimodal	(1) Wav2Vec2.0 Embedding	0.4971
	(2) TRILLsson Embedding	0.1235
	(3) Speech Embedding: (1) + (2)	0.4346
	(4) Manual Text Embedding	0.1452
	(5) ASR Text Embedding	0.1831
Multimodal	(6) Wav2Vec2.0 + Manual Text	0.5385
	(7) Wav2Vec2.0 + ASR Text	0.5308
	(8) TRILLsson + Manual Text	0.5099
	(9) TRILLsson + ASR Text	0.4863
	(10) Proposed w/o ASR: (3) + (4)	0.5615
Proposed	**(11) Proposed (Full Model): (3) + (5)**	**0.5629**

As shown in Table 2, our fully automated framework (11) that fuses segmental (wav2vec2.0) and suprasegmental (TRILLsson) speech embeddings with ASR-derived text embedding via co-attention achieves the highest SCC of 0.5629.

Among speech-only configurations, wav2vec2.0 alone (1) achieves an SCC of 0.4971. In contrast, TRILLsson alone (2) results in a lower SCC of 0.1235. Merging these two speech embeddings without text (3) results in an SCC of 0.4346, which remains lower than that of wav2vec2.0 alone. Text-only configurations (4)-(5) show SCCs of 0.1452 for manual transcriptions and 0.1831 for ASR text. Integrating text with a single speech embedding enhances performance. Specifically, wav2vec2.0 combined with manual text (6) achieves 0.5385, while wav2vec2.0 with ASR text (7) obtains 0.5308; TRILLsson with manual text (8) yields 0.5099, and with ASR text (9) 0.4863. The best performance is observed when both segmental and suprasegmental speech embeddings are incorporated with text embeddings (10)-(11), with the fully automated framework (11) using ASR text achieving the highest SCC of 0.5629.

4. Discussion

In this paper, we propose a cascaded multimodal framework for estimating ASD severity by integrating segmental (wav2vec2.0) and suprasegmental (TRILLsson) speech embeddings with text embeddings derived from ASR transcriptions. Experimental results consistently indicate that a multimodal approach, combining segmental and suprasegmental speech representations with textual features, provides a more comprehensive representation of ASD-related communication impairments compared to unimodal methods.

Ablation studies further highlight the role of each modality. Wav2vec2.0 alone serves as a strong predictor of ASD severity, suggesting that segmental features extracted by wav2vec2.0 offer robust predictive information. In contrast, using only TRILLsson embeddings significantly reduces performance. This outcome suggests that suprasegmental cues require additional segmental or lexical context for accurate severity estimation, as evidenced by the performance improvement when TRILLsson embeddings are fused with either wav2vec2.0 or text embeddings. However, merging wav2vec2.0 and TRILLsson without textual input results in lower performance than using wav2vec2.0 alone, likely due to overlapping acoustic features that introduce redundancy or interference. This finding aligns with prior research indicating that wav2vec2.0 encodes not only phonological but also suprasegmental aspects such as stress and pitch [30, 34]. Text-only configurations show weak predictive power, confirming that lexical information alone does not sufficiently capture the nuanced speech characteristics associated with ASD.

The effectiveness of the multimodal configuration is further demonstrated when text embeddings are integrated with speech embeddings. Adding textual features to a single speech embedding significantly enhances performance, underscoring the interaction between acoustic-prosodic and lexical markers in characterizing atypical language usage. Furthermore, the proposed multimodal approach, which integrates both segmental (wav2vec2.0) and suprasegmental (TRILLsson) speech embeddings along with text embeddings, consistently outperforms configurations that incorporate only a single speech embedding with text. The fusion of wav2vec2.0 and TRILLsson provides a richer and more complementary representation of ASD-related speech characteristics, capturing both fine-grained phonological details and broader prosodic patterns. This improvement is likely due to the complementary nature of segmental and suprasegmental cues, where wav2vec2.0 emphasizes phonological information while TRILLsson captures speech rhythm and intonation patterns. However, this integration is most effective when textual features are also included, reinforcing the crucial role of lexical context in resolving ambiguities and enhancing severity estimation. These findings highlight the importance of leveraging both speech representations simultaneously rather than relying on a single speech modality in multimodal settings.

Interestingly, the impact of ASR varies depending on the configuration. In text-only settings, ASR-generated text embeddings outperform manual transcriptions, likely because they retain phonetic and articulatory details that human transcribers tend to normalize, thereby preserving clinically significant deviations that may be lost in conventional orthography. However, when a single speech embedding is combined with text, ASR text embeddings underperform relative to manual transcriptions. This discrepancy may arise because ASR errors introduce noise that conflicts with acoustic features already extracted from the speech embedding. In this setting, where only one speech representation is available, transcription errors may overshadow the benefits of preserving articulatory deviations. Conversely, in the full multimodal model integrating both segmental and suprasegmental speech embeddings, ASR text slightly outperforms manual transcriptions. This suggests that a richer set of speech features allows the system to leverage clinically relevant phonological cues embedded in ASR text, mitigating the effects of transcription noise. A qualitative analysis of ASR outputs and human transcriptions reveals notable differences that likely reflect developmental delays and articulation challenges in children with ASD. For instance, while a human transcriber may normalize a child's utterance to standard orthography (e.g., [p^ho.ham]), the ASR system may generate a form closer to the actual pronunciation (e.g., [pu.aŋ]), capturing difficulties with aspirated sounds or syllable-initial consonants. Such discrepancies indicate that ASR-generated text retains valuable markers of speech production deficits, thereby contributing to more accurate ASD severity prediction.

Compared with existing approaches, the highest reported SCC of 0.5629 is slightly lower than the 0.567 achieved by [16], which relies on handcrafted acoustic features and manual transcriptions. Nonetheless, our fully automated system eliminates the need for extensive human annotation and feature engineering, offering greater scalability and applicability in real-world clinical and research settings. By enabling direct processing of raw speech through both speech and text models, the proposed framework efficiently captures ASD-related segmental and suprasegmental speech characteristics, as well as linguistic deficits, thereby validating the feasibility and effectiveness of an automated ASD severity assessment approach.

5. Conclusion

This paper presents a fully automated multimodal framework that assesses the social communication severity of children with ASD by integrating segmental and suprasegmental speech embeddings with textual representations. Demonstrating an SCC of 0.5629 on a Korean ASD dataset, the proposed system demonstrates robust performance without manual feature extraction or transcription. The main contributions lie in designing a cascaded multimodal approach that unifies segmental and suprasegmental speech information with ASR-generated text, validating the effectiveness of this approach over unimodal configurations, and substantially reducing reliance on human intervention. These strengths underscore the framework's potential for clinical and research applications where time and specialized expertise are limited.

Future work may explore additional speech or language models and alternative fusion strategies to further enhance performance. Moreover, while this study focuses on severity assessment, upcoming research might unify both ASD diagnosis and severity estimation. Another promising avenue is a multitask learning approach to jointly evaluate social communication severity and other speech-related dimensions, such as pronunciation proficiency, leveraging shared linguistic and acoustic factors.

6. Acknowledgements

This work was supported by Mid-Career Bridging Program through Seoul National University.

7. References

[1] S. Faja and G. Dawson, "Autism spectrum disorder," *Child and Adolescent Psychopathology, Third Edition*, pp. 745–782, 2017.

[2] P. J. Prelock and N. W. Nelson, "Language and communication in autism: An integrated view," *Pediatric Clinics*, vol. 59, no. 1, pp. 129–145, 2012.

[3] E. Lyakso, O. Frolova, and A. Grigorev, "Perception and acoustic features of speech of children with autism spectrum disorders," in *Speech and Computer: 19th International Conference, SPECOM 2017, Hatfield, UK, September 12-16, 2017, Proceedings 19*. Springer, 2017, pp. 602–612.

[4] Y. Nakai, R. Takashima, T. Takiguchi, and S. Takada, "Speech intonation in children with autism spectrum disorder," *Brain and Development*, vol. 36, no. 6, pp. 516–522, 2014.

[5] F. Chen, L. Wang, G. Peng, N. Yan, and X. Pan, "Development and evaluation of a 3-d virtual pronunciation tutor for children with autism spectrum disorders," *PLoS One*, vol. 14, no. 1, p. e0210858, 2019.

[6] A. Skoufou, "Social interaction of preschool children with autism spectrum disorders (asd)-characteristics and educational approaches." *Online Submission*, vol. 6, no. 6, pp. 28–36, 2019.

[7] S. Lee, J. Mun, S. Kim, and M. Chung, "Speech corpus for korean children with autism spectrum disorder: Towards automatic assessment systems," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 15 160–15 170.

[8] S. Raj and S. Masood, "Analysis and detection of autism spectrum disorder using machine learning techniques," *Procedia Computer Science*, vol. 167, pp. 994–1004, 2020.

[9] T. Kodak and S. Bergmann, "Autism spectrum disorder: characteristics, associated behaviors, and early intervention," *Pediatric Clinics of North America*, 2020.

[10] P. McCarty and R. E. Frye, "Early detection and diagnosis of autism spectrum disorder: Why is it so difficult?" in *Seminars in Pediatric Neurology*, vol. 35. Elsevier, 2020, p. 100831.

[11] M. Asgari, A. Bayestehtashk, and I. Shafran, "Robust and accurate features for detecting and diagnosing autism spectrum disorders," in *Interspeech*, vol. 2013. NIH Public Access, 2013, p. 191.

[12] A. Mohanta, P. Mukherjee, and V. K. Mirtal, "Acoustic features characterization of autism speech for automated detection and classification," in *2020 National Conference on Communications (NCC)*. IEEE, 2020, pp. 1–6.

[13] F. Briend, C. David, S. Silleresi, J. Malvy, S. Ferré, and M. Latinus, "Voice acoustics allow classifying autism spectrum disorder with high accuracy," *Translational Psychiatry*, vol. 13, no. 1, p. 250, 2023.

[14] S. Lee, E. J. Yeo, S. Kim, and M. Chung, "Knowledge-driven speech features for detection of korean-speaking children with autism spectrum disorder," *Phonetics and Speech Sciences*, vol. 15, no. 2, pp. 53–59, 2023.

[15] J. Li, "Artificial intelligence-based detection of autism spectrum disorder using linguistic features," in *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*. IEEE, 2024, pp. 1–6.

[16] C.-P. Chen, S. S.-F. Gau, and C.-C. Lee, "Learning converse-level multimodal embedding to assess social deficit severity for autism spectrum disorder," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[17] J. Mun, S. Kim, and M. Chung, "Developing an end-to-end framework for predicting the social communication severity scores of children with autism spectrum disorder," in *Proc. Interspeech 2024*, 2024, pp. 1430–1434.

[18] M. Eni, I. Dinstein, M. Ilan, I. Menashe, G. Meiri, and Y. Zigel, "Estimating autism severity in young children from speech signals using a deep neural network," *IEEE Access*, vol. 8, pp. 139 489–139 500, 2020.

[19] H. Tanaka, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Linguistic and acoustic features for automatic identification of autism spectrum disorders in children's narrative," in *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 2014, pp. 88–96.

[20] 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." in *Interspeech*, 2019, pp. 2513–2517.

[21] H. MacFarlane, A. C. Salem, L. Chen, M. Asgari, and E. Fombonne, "Combining voice and language features improves automated autism detection," *Autism Research*, vol. 15, no. 7, pp. 1288–1300, 2022.

[22] M. Tang, P. Kumar, H. Chen, and A. Shrivastava, "Deep multimodal learning for the diagnosis of autism spectrum disorder," *Journal of Imaging*, vol. 6, no. 6, p. 47, 2020.

[23] 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)*. IEEE, 2019, pp. 847–854.

[24] J. Shor and S. Venugopalan, "Trillsson: Distilled universal paralinguistic speech representations," *arXiv preprint arXiv:2203.00236*, 2022.

[25] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[26] E. L. Campbell, J. Dineley, P. Conde, F. Matcham, K. M. White, C. Oetzmann, S. Simblett, S. Bruce, A. A. Folarin, T. Wykes *et al.*, "Classifying depression symptom severity: Assessment of speech representations in personalized and generalized machine learning models," in *INTERSPEECH 2023*, vol. 2023. ISCA, 2023, pp. 1738–1742.

[27] A. Favaro, Y.-T. Tsai, A. Butala, T. Thebaud, J. Villalba, N. Dehak, and L. Moro-Velázquez, "Interpretable speech features vs. dnn embeddings: What to use in the automatic assessment of parkinson's disease in multi-lingual scenarios," *Computers in Biology and Medicine*, vol. 166, p. 107559, 2023.

[28] S.-I. Ng, L. Xu, K. D. Mueller, J. Liss, and V. Berisha, "Segmental and suprasegmental speech foundation models for classifying cognitive risk factors: Evaluating out-of-the-box performance," in *Proc. Interspeech 2024*, 2024, pp. 917–921.

[29] P. C. English, J. D. Kelleher, and J. Carson-Berndsen, "Searching for structure: Appraising the organisation of speech features in wav2vec 2.0 embeddings," in *Proc. Interspeech 2024*, 2024, pp. 4613–4617.

[30] J. Yuan, N. Ryant, X. Cai, K. Church, and M. Liberman, "Automatic recognition of suprasegmentals in speech," *arXiv preprint arXiv:2108.01122*, 2021.

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[32] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh, J. Lee, J. Oh, S. Lyu, Y. Jeong, I. Lee, S. Seo, D. Lee, H. Kim, M. Lee, S. Jang, S. Do, S. Kim, K. Lim, J. Lee, K. Park, J. Shin, S. Kim, L. Park, A. Oh, J. Ha, and K. Cho, "Klue: Korean language understanding evaluation," 2021.

[33] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[34] A. de la Fuente and D. Jurafsky, "A layer-wise analysis of mandarin and english suprasegmentals in ssl speech models," *arXiv preprint arXiv:2408.13678*, 2024.