



Developing an End-to-End Framework for Predicting the Social Communication Severity Scores of Children with Autism Spectrum Disorder

Jihyun Mun¹, Sunhee Kim², Minhwa Chung¹

¹Department of Linguistics, Seoul National University, Republic of Korea

²Department of French Language Education, Seoul National University, Republic of Korea

jhhh.1202@snu.ac.kr, sunhkim@snu.ac.kr, mchung@snu.ac.kr

Abstract

Autism Spectrum Disorder (ASD) is a lifelong condition that significantly influencing an individual's communication abilities and their social interactions. Early diagnosis and intervention are critical due to the profound impact of ASD's characteristic behaviors on foundational developmental stages. However, limitations of standardized diagnostic tools necessitate the development of objective and precise diagnostic methodologies. This paper proposes an end-to-end framework for automatically predicting the social communication severity of children with ASD from raw speech data. This framework incorporates an automatic speech recognition model, fine-tuned with speech data from children with ASD, followed by the application of fine-tuned pre-trained language models to generate a final prediction score. Achieving a Pearson Correlation Coefficient of 0.6566 with human-rated scores, the proposed method showcases its potential as an accessible and objective tool for the assessment of ASD.

Index Terms: autism spectrum disorder, speech recognition, language model, prompt tuning, end-to-end framework, automatic assessment

1. Introduction

Autism Spectrum Disorder (ASD) is defined as a lifelong condition that significantly affects an individual's communication abilities and their interaction within society [1]. Children with ASD experience social deficits, communication difficulties, and atypical behavior patterns, including impaired socio-communicative interactions and a limited range of interests and activities [1, 2].

Early diagnosis and intervention are critical due to the profound impact of ASD's symptomatic behaviors on foundational developmental processes. Early intervention is particularly pivotal for social development, as initial social capabilities and deficits inform intervention outcomes and treatment strategies [3, 4]. In clinical environments, standardized diagnostic tools like the Autism Diagnostic Observation Schedule, 2nd edition (ADOS-2), are employed [5]. However, the use of standardized tools for evaluating children presents numerous challenges, including expertise scarcity leading to delayed or overlooked diagnoses [6], potential bias from subjective interpretations by caregivers or evaluators [7], and the extended duration of the evaluation process, which can burden both children and their caregivers and may reduce the children's concentration. Consequently, there is a pressing need for developing objective and precise methodologies which diagnose and predict severity for early diagnosis and intervention of ASD [8, 9].

Recent advancements in automated methods for predicting ASD severity incorporate a range of technologies, including

MRI [10, 11, 12], fMRI [13], EEG signals [9, 14], and genetic and environmental factors [15]. Despite their efficacy, these methods often require specialized equipment and expertise, presenting barriers to widespread adoption [6]. In contrast, speech data offers a more accessible and less intrusive alternative [16], providing a viable option for diagnosing and assessing the severity of ASD. Studies have concentrated on the pragmatic aspects of language, including the appropriate use of language across various social contexts, particularly in children with ASD in comparison to their typically developing (TD) peers [17, 18, 19]. They underscored that children with ASD frequently exhibit atypical language behaviors in social contexts, thereby emphasizing the complex relationship between linguistic and social challenges. The utilization of speech data not only circumvents the limitations associated with other diagnostic materials but also leverages the unique linguistic characteristics of children with ASD. This underscores the potential of linguistic materials for the automated diagnosis and severity prediction of ASD [20, 21, 22], offering a promising direction for enhancing accessibility and reducing the reliance on extensive resources and specialized knowledge.

Machine learning techniques have been applied to identify ASD based on linguistic indicators [20, 21], with traditional methods requiring meticulous feature selection, a process that is time-intensive and highly specialized [23]. Deep learning approaches offer an alternative by deriving more abstract representations [24], such as using lexical embeddings from a fine-tuned BERT model for ASD diagnosis [22]. However, deep learning models necessitate large datasets, which poses a challenge for ASD research due to the typically small available datasets. Pre-trained language models (PLMs), fine-tuned on specific tasks, leverage extensive pre-training corpora to mitigate this issue [25].

A notable concern when applying PLMs to classification tasks is the potential misalignment between the objectives during pre-training and fine-tuning [26]. The integration of natural language prompts in fine-tuning PLMs, a technique known as prompt tuning, aligns the model's objectives with those of the pre-training phase, thereby enhancing performance on specific tasks in the context of limited data [26, 27].

Building on recent methodological advancements and leveraging the distinctive benefits of prompt tuning in contexts with limited data, this paper proposes an end-to-end (E2E) framework that incorporates a prompt tuning methodology for predicting the severity of social communication in children with ASD. The deployment of prompt tuning methodologies necessitates the transcription of audio recordings. However, manual transcription presents several challenges, including high costs, limited availability, and issues with scalability. To overcome these challenges, we integrate an Automatic Speech Recogni-

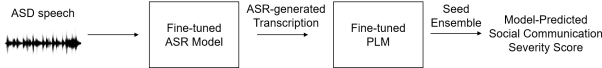


Figure 1: *Proposed E2E framework for automatically predicting the social communication severity scores of children with ASD*

tion (ASR) model into our framework, enabling the derivation of final prediction scores directly from raw speech data. The comprehensive framework utilizes an ASR model, specifically fine-tuned with speech data from children with ASD, followed by the application of fine-tuned PLMs and an ensemble method to generate a final prediction score.

The remainder of the paper is organized as follows: Section 2 details the methodologies employed, Section 3 outlines the experimental setup, Section 4 presents the results, Section 5 discusses the findings, and Section 6 concludes the study.

2. Methods

This study introduces an E2E framework that incorporates fine-tuned ASR models, fine-tuned PLMs, and a seed ensemble method for predicting the social communication severity scores in children with ASD, as depicted in Figure 1.

2.1. Automatic Speech Recognition Model

We selected two pre-trained multilingual ASR models for this purpose: wav2vec2-xls-r-300m [28] and whisper-large-v2 [29]. To tailor these models to the nuances of speech from TD children and children with ASD, we fine-tune each model using speech data specific to these groups.

2.2. Fine-tuning Pre-trained Language Models

The study further involves fine-tuning three PLMs—KR-BERT [30], KLUE/roberta-base [31], and KR-ELECTRA-Discriminator [32]—employing three distinct approaches: traditional fine-tuning, manual prompting, and p-tuning. These models were chosen for prompt-based fine-tuning due to their demonstrated effectiveness in text classification tasks, as evidenced by prior research [26, 33, 34]. Training incorporates ten different initialization seeds to increase robustness and mitigate the effects of random initialization.

2.2.1. Traditional Fine-tuning

Fine-tuning adapts a model pre-trained on a vast dataset to a smaller, task-specific dataset [35], effectively leveraging the extensive knowledge acquired during pre-training [25] for specific downstream tasks. In this process, a regression head is attached to the model. The [CLS] token, representing the input sequence comprehensively, facilitates the prediction of a continuous severity score.

2.2.2. Manual Prompting

Manual prompting involves crafting specific input prompts to direct the behavior of transformer models towards generating desired outputs. By designing appropriate prompts, it’s possible to utilize the extensive knowledge embedded in these models for performing specific tasks without additional task-specific training. In this approach, a regression head initializes the models, and a template guides the model to focus on predicting the

Table 1: *ASR and PLM fine-tuning speakers*

	Train		Test	
	ASD	TD	ASD	TD
ASR tuning data	152 (11h 15m 10s)	32 (2h 54m 51s)	16 (2h 14m 53s)	8 (35m 48s)
PLM tuning data	87 (73 M, 13 F, 1 U) (mean 9;1, std 5;8)	32 (13 M, 11 F, 8 U) (mean 5;11, std 2;10)	16 (13 M, 3 F) (mean 12;11, std 7;3)	8 (1 M, 3 F, 4 U) (mean 4;1, std 2)

M: male (boy), F: female (girl), U: unreported
mean and std refer to the mean and standard deviation of chronological age of speakers

social communication severity score from the input text.

2.2.3. Automated Prompting: P-Tuning

P-tuning, introduced by [36], advances beyond manual prompting by parameterizing prompts and optimizing them alongside the model’s parameters during fine-tuning, allowing the model to autonomously identify the most effective prompts for a task. For this study, the p-tuning approach is implemented using the PEFT library [37]. Models are initialized with a regression head, and virtual tokens are incorporated and tuned specifically for the task, optimizing the models’ predictions.

2.3. Seed Ensemble for Robust Prediction

To mitigate the variability introduced by the randomness in model initialization and to improve the overall performance, a seed ensemble technique is employed. For each PLM, we aggregate the predictions from the ten individually fine-tuned models (one per seed) to formulate a singular and more accurate prediction.

3. Experiments

3.1. Data Preparation and Dataset Description

The speech samples were collected during linguistic assessment sessions conducted by certified speech-language pathologists (SLPs). The specifics of the data collection, transcription, and evaluation processes have been detailed in [38]. This study utilized speech data from 168 children diagnosed with ASD and 40 TD children. These participants were integral for fine-tuning the ASR models. Specifically, the ASD cohort included 103 children whose social communication severity was evaluated by three certified SLPs. The average of the three SLPs’ evaluations served as the severity score for the ASD children, while TD children were assigned a baseline score of zero. The datasets for evaluated ASD and TD children were employed for fine-tuning the PLMs. The overall dataset is described in Table 1. To ensure no overlap and maintain the integrity of the evaluation process, children included in the test set for PLM fine-tuning were excluded from the training dataset of the ASR model.

3.2. Fine-Tuning ASR Models

The ASR models, specifically wav2vec2 and whisper, are fine-tuned using Fairseq and Hugging Face’s Transformers, respectively. The Adam optimizer is utilized in both cases, with initial learning rates set to 3e-4 for wav2vec2 and 1e-5 for whisper. Given that Korean is a syllable-timed language, the performance of the fine-tuned models is evaluated using the syllable error rate (SER), achieving rates of 26.21% and 19.57%, respectively, after fine-tuning.

3.3. Fine-Tuning PLMs

For each tuning method, training spanned 40 epochs, utilizing a learning rate of $1e-5$, a batch size of 8, and the AdamW optimizer. The mean squared error is employed as the objective loss function. In manual prompting, the template "[text] the social communication severity score of the speaker is [MASK]" is used, with "[text]" replaced by actual dataset text and "[MASK]" serving as a placeholder. In p-tuning, experiments are conducted with 5, 10, 15, and 20 virtual tokens, setting the encoder's hidden size to 128. Differential learning rates are applied: $1e-5$ for both the base models and the prompt encoder, and $1e-3$ for the regression head.

3.4. Evaluation Metrics

The evaluation strategy includes two settings:

1. *Full-set setting*, where all available training data is used, reserving 20% for validation.
2. *Low-resource setting*, where only 20% of the full training data is accessible, following the methodology outlined by [34].

The evaluation metric employed is the Pearson Correlation Coefficient (PCC), which measures the relationship between the model's predicted output and the scores labeled by humans. To mitigate the effects of random initialization, each system's evaluation is executed ten times, each with a different random seed from PyTorch's random initialization setting. The final prediction is determined using the seed ensemble method.

4. Results

The study evaluates the effectiveness of the proposed framework, which integrates various ASR models, transcription types, PLMs, and tuning methods in predicting social communication severity in children with ASD across full-set and low-resource settings. The comprehensive results of our experiments are shown in Table 2.

As expected, human transcriptions consistently outperform ASR transcriptions. However, certain combinations of PLMs and tuning methods, specifically klue/roberta-base with p-tuning, reveal instances where ASR transcriptions surpass human transcriptions. In low-resource settings, the performance gap between human and ASR transcriptions diminishes, highlighting the potential of ASR transcriptions in scenarios of limited data availability. Remarkably, wav2vec2 transcription outperforms human transcription in specific cases when klue/roberta-base model is p-tuned, indicating a strong correlation with human-labeled scores (e.g., PCC of 0.6566 compared to 0.6216 with 20 virtual tokens). When comparing two ASR models, wav2vec2 transcriptions generally exhibit better performance than those from the whisper model, despite a higher syllable error rate.

The results demonstrate that the choice of PLM and the tuning method significantly affects the performance in predicting the severity score of social communication. In scenarios involving both ASR and human transcriptions within the full-set setting, fine-tuning and manual prompting tend to outperform p-tuning for the KR-BERT and KR-ELECTRA-Discriminator models. However, p-tuning shows superior performance with the klue/roberta-base model. This trend continues in the low-resource setting, where p-tuning enhances performance with human transcriptions for the KR-ELECTRA-Discriminator model.

Additionally, performance varies significantly based on the number of virtual tokens utilized in p-tuning. For example, with the KR-BERT model using ASR transcriptions, the PCC values range from negative to positive, indicating a shift from a negative to a moderate correlation with human-labeled scores. Similarly, with the KR-BERT model using human transcriptions in a low-resource setting, the correlation varies significantly from weak to moderate.

5. Discussion

The results highlight a complex relationship between transcription types, PLM selection, tuning methods, and data availability in the automated assessment of ASD severity.

The diminishing performance disparity between human and ASR transcriptions in low-resource settings underscores the proposed method's potential in enhancing the accessibility and scalability of ASD severity assessment. This trend suggests that ASR technology may serve as a feasible alternative to human transcription in situations where resources are limited.

The generally better performance of the wav2vec2 model over the whisper model, despite the latter's lower error rate, indicate that there are aspects of speech relevant to ASD severity that are captured by wav2vec2 but ignored by whisper due to its disfluency removal. It is known that children with ASD display various types of speech disfluencies, such as sound and syllable repetitions, interjections, within-word breaks, and final sound prolongations [39]. The whisper model's tendency to eliminate speech disfluencies, including filler words, hesitations, and repetitions [40], contrasts with the wav2vec2 model's capability to detect disfluencies or stuttering. Therefore, accurately capturing the characteristics of ASD speech, including speech disfluencies, necessitates the selection of an appropriate ASR model that retains these critical speech features. This consideration is pivotal in developing effective diagnostic tools and interventions for ASD, highlighting the importance of choosing an ASR model that aligns with the nuanced requirements of ASD speech.

The varied performance across PLMs under different tuning methods highlights the necessity of meticulous consideration for each PLM-tuning combination. The klue/roberta-base model's effective response to p-tuning, across both transcription types, suggests its potential as a powerful tool in optimizing PLMs, particularly in data-constrained environments. Additionally, the number of virtual token significantly influences performance differences. Although the number of prompt tokens greatly impacts few-shot performance, a larger number of prompt tokens is not always better; it depends on the amount of training data [36]. In practice, we should determine the optimal number of prompt tokens through model selection, highlighting the need for careful consideration of tuning settings.

6. Conclusion

This study proposes an E2E framework, incorporating fine-tuned ASR models and PLMs, for automatically predicting social communication severity in children with ASD. Demonstrating a PCC of 0.6566, the experimental results affirm the framework's utility, especially in data-limited situations.

Key contributions of this paper include the introduction of an automated method for predicting the social communication severity score in children with ASD from raw speech data, the development of an E2E framework that eliminates the need for human transcription, and the validation of this frame-

Table 2: Pearson correlation coefficient with human-labeled scores

		Full-set setting			Low-resource setting			
		ASR transcription		Human transcription	ASR transcription		Human transcription	
		Wav2vec2	Whisper		Wav2vec2	Whisper		
KR-BERT	Fine-tuning	0.2791	0.1984	0.5516**	0.4471*	0.2253	0.5817**	
	Manual	0.3637	0.1624	0.4701*	0.4204*	0.0869	0.5032**	
	P-tuning	5	-0.1992	-0.1576	0.4483*	0.1808	-0.0346	0.1119
		10	0.3861	0.2129	0.4511*	0.3815	0.2841	0.5595**
		15	0.3367	0.1077	0.3139	0.1491	-0.0602	0.3409
		20	-0.0047	-0.0410	0.4663*	-0.2881	-0.1187	0.0050
klue/roBERTa-base	Fine-tuning	0.3880	0.2070	0.4322*	0.3806	0.2271	0.3972	
	Manual	0.0761	0.0846	0.2859	0.3515	0.1445	0.4367*	
	P-tuning	5	0.5587**	0.4980*	0.4207*	0.6117**	0.5633**	0.5731**
		10	0.5431**	0.5109*	0.5181**	0.6333***	0.6183**	0.6217**
		15	0.5343**	0.4331*	0.4812*	0.6163**	0.6166**	0.6230**
		20	0.5852**	0.5330**	0.5472**	0.6566**	0.5854**	0.6216**
KR-ELECTRA-Discriminator	Fine-tuning	0.4649*	0.4315*	0.9019***	0.3425	0.1735	0.6454***	
	Manual	0.4452*	0.2109	0.7645***	0.4207*	0.1556	0.6925***	
	P-tuning	5	0.0750	0.0605	0.6546***	0.1652	0.0564	0.7138***
		10	-0.1221	0.0485	0.6509***	0.2095	0.1134	0.7117***
		15	0.0335	0.0491	0.7164***	0.1552	0.0948	0.7273***
		20	-0.2002	-0.0258	0.5335**	0.3324	0.2717	0.7654***

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

work’s effectiveness in data-restricted settings. These achievements indicate the practical applicability of the framework in real-world ASD severity assessments, a field where acquiring large datasets is often challenging.

Despite these promising results, the framework faces interpretability challenges. In domains such as ASD diagnosis and assessment, the models’ decision-making processes must be transparent to ensure trust and reliability in their practical application [41]. Interpretability becomes even more crucial in extremely data-limited situations, where variability in results can be substantial. However, the highest-performing model employs P-tuning of the PLM, which utilizes virtual tokens as learnable parameters. These parameters are inherently non-interpretable and untrackable, obscuring the model’s decision logic and further complicating the issue of interpretability.

Future research will explore instruction tuning methodologies that could provide “chain-of-thought” reasoning [42], potentially enhancing the interpretability of model predictions. The goal is to align high predictive accuracy with clear, understandable outputs, ensuring that the models not only predict with high precision but also provide interpretable and actionable insights for clinical use.

7. Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant

funded by the Korea government(MSIT) [No.2022-0-00223, Development of digital therapeutics to improve communication ability of autism spectrum disorder patients].

8. References

- [1] S. Al Shirian and H. Al Dera, “Descriptive characteristics of children with autism at autism treatment center, ksa,” *Physiology & behavior*, vol. 151, pp. 604–608, 2015.
- [2] J. S. Nicholas, J. M. Charles, L. A. Carpenter, L. B. King, W. Jenner, and E. G. Spratt, “Prevalence and characteristics of children with autism-spectrum disorders,” *Annals of epidemiology*, vol. 18, no. 2, pp. 130–136, 2008.
- [3] C. Lord and R. Luyster, “Early diagnosis of children with autism spectrum disorders,” *Clinical Neuroscience Research*, vol. 6, no. 3-4, pp. 189–194, 2006.
- [4] D. A. Zachor and E. B. Itzhak, “Treatment approach, autism severity and intervention outcomes in young children,” *Research in Autism Spectrum Disorders*, vol. 4, no. 3, pp. 425–432, 2010.
- [5] C. Lord, M. Rutter, R. J. Luyster, and K. Gotham, *Autism diagnostic observation schedule-2nd edition (ADOS-2)*, 2nd ed. Western Psychological Corporation, 2012.
- [6] J. Li, Z. Chen, G. Li, G. Ouyang, and X. Li, “Automatic classification of asd children using appearance-based features from videos,” *Neurocomputing*, vol. 470, pp. 40–50, 2022.
- [7] A. Frigaux, R. Evrard, and J. Lighezzolo-Alnot, “Adi-r and ados and the differential diagnosis of autism spectrum disorders: Interests, limits and openings,” *L’encephale*, vol. 45, no. 5, pp. 441–448, 2019.

- [8] Y. Wang, J. Liu, Y. Xiang, J. Wang, Q. Chen, and J. Chong, "Mage: Automatic diagnosis of autism spectrum disorders using multi-atlas graph convolutional networks and ensemble learning," *Neurocomputing*, vol. 469, pp. 346–353, 2022.
- [9] H. Hadoush, M. Alafeef, and E. Abdulhay, "Automated identification for autism severity level: Eeg analysis using empirical mode decomposition and second order difference plot," *Behavioural brain research*, vol. 362, pp. 240–248, 2019.
- [10] E. Moradi, B. Khundrakpam, J. D. Lewis, A. C. Evans, and J. Tohka, "Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data," *Neuroimage*, vol. 144, pp. 128–141, 2017.
- [11] L. Q. Uddin, K. Supekar, C. J. Lynch, A. Khouzam, J. Phillips, C. Feinstein, S. Ryali, and V. Menon, "Salience network–based classification and prediction of symptom severity in children with autism," *JAMA psychiatry*, vol. 70, no. 8, pp. 869–879, 2013.
- [12] H. Kwon, J. I. Kim, S.-Y. Son, Y. H. Jang, B.-N. Kim, H. J. Lee, and J.-M. Lee, "Sparse hierarchical representation learning on functional brain networks for prediction of autism severity levels," *Frontiers in Neuroscience*, vol. 16, p. 935431, 2022.
- [13] S. Wang and N. C. Dvornek, "A metamodel structure for regression analysis: Application to prediction of autism spectrum disorder severity," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1338–1341.
- [14] Y. Zhang, S. Zhang, B. Chen, L. Jiang, Y. Li, L. Dong, R. Feng, D. Yao, F. Li, and P. Xu, "Predicting the symptom severity in autism spectrum disorder based on eeg metrics," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1898–1907, 2022.
- [15] M. Che, L. Wang, L. Huang, and Z. Jiang, "An approach for severity prediction of autism using machine learning," in *2019 IEEE international conference on Industrial Engineering and Engineering Management (IEEM)*. IEEE, 2019, pp. 701–705.
- [16] I. Clemente, "Recording audio and video," *The Blackwell guide to research methods in bilingualism and multilingualism*, pp. 177–191, 2008.
- [17] A. Philofsky, D. J. Fidler, and S. Hepburn, "Pragmatic language profiles of school-age children with autism spectrum disorders and williams syndrome," 2007.
- [18] J. Volden, J. Coolican, N. Garon, J. White, and S. Bryson, "Brief report: Pragmatic language in autism spectrum disorder: Relationships to measures of ability and disability," *Journal of autism and developmental disorders*, vol. 39, pp. 388–393, 2009.
- [19] I. Vogindroukas, M. Stankova, E.-N. Chelas, and A. Proedrou, "Language and speech characteristics in autism," *Neuropsychiatric Disease and Treatment*, pp. 2367–2377, 2022.
- [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] B. Ashwini, V. Narayan, and J. Shukla, "Spasht: Semantic and pragmatic speech features for automatic assessment of autism," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [22] 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.
- [23] T. Verdonck, B. Baesens, M. Óskarsdóttir, and S. vanden Broucke, "Special issue on feature engineering editorial," *Machine Learning*, pp. 1–12, 2021.
- [24] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [26] Y. Wang, J. Deng, T. Wang, B. Zheng, S. Hu, X. Liu, and H. Meng, "Exploiting prompt learning with pre-trained language models for alzheimer's disease detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [27] K.-W. Chang, Y.-K. Wang, H. Shen, I.-t. Kang, W.-C. Tseng, S.-W. Li, and H.-y. Lee, "Speechprompt v2: Prompt tuning for speech classification tasks," *arXiv preprint arXiv:2303.00733*, 2023.
- [28] A. Babu, C. Wang, A. Tjandra, K. Lakhota, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.
- [29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [30] S. Lee, H. Jang, Y. Baik, S. Park, and H. Shin, "Kr-bert: A small-scale korean-specific language model," *arXiv preprint arXiv:2008.03979*, 2020.
- [31] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh *et al.*, "Klue: Korean language understanding evaluation," *arXiv preprint arXiv:2105.09680*, 2021.
- [32] S. Lee and H. Shin, "Kr-electra: a korean-based electra model," <https://github.com/snunlp/KR-ELECTRA>, 2022.
- [33] Y. Wang, Y. Wang, Z. Peng, F. Zhang, L. Zhou, and F. Yang, "Medical text classification based on the discriminative pre-training model and prompt-tuning," *Digital Health*, vol. 9, p. 20552076231193213, 2023.
- [34] Y. Yao, B. Dong, A. Zhang, Z. Zhang, R. Xie, Z. Liu, L. Lin, M. Sun, and J. Wang, "Prompt tuning for discriminative pre-trained language models," *arXiv preprint arXiv:2205.11166*, 2022.
- [35] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.
- [36] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "Gpt understands, too," *AI Open*, 2023.
- [37] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, and S. Paul, "Pefit: State-of-the-art parameter-efficient fine-tuning methods," <https://github.com/huggingface/peft>, 2022.
- [38] S. Lee, J. Mun, S. Kim, and M. Chung, "Speech corpus for korean children with autism spectrum disorder: Towards automatic assessment systems," 2024.
- [39] L. W. Plexico, J. E. Cleary, A. McAlpine, and A. M. Plumb, "Disfluency characteristics observed in young children with autism spectrum disorders: A preliminary report," *Perspectives on Fluency and Fluency Disorders*, vol. 20, no. 2, pp. 42–50, 2010.
- [40] J. Louradour, "whisper-timestamped," <https://github.com/linto-ai/whisper-timestamped>, 2023.
- [41] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural computing and applications*, vol. 32, no. 24, pp. 18 069–18 083, 2020.
- [42] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.