Automatic Speech Recognition and Assessment Systems Incorporated into Digital Therapeutics for Children with Autism Spectrum Disorder

Seonwoo Lee[1], Jihyun Mun[1], Sunhee Kim[1], HyunJu Park[2], Suvin Yang[2], HyunDon Kim[3], SeungJae Noh[3], WonBin Kim[3], and Minhwa Chung[1]

[1] Seoul National University, Gwanak-ro 1, 08826 Seoul, Republic of Korea
{lsw5220,jhhh_1202,sunhkim,mchung}@snu.ac.kr

[2] Gachon University, Seongnam-daero 1342, 13120 Seongnam, Republic of Korea
{phj8747,tnqls3931}@gachon.ac.kr

[3] PlaytoCure, Wiryeseoil-ro 1-gil 21-7, 13647 Seongnam, Republic of Korea
{everlastkim,sj418,teyrunia}@playtocure.kr

Abstract. Children with autism spectrum disorder (ASD) frequently encounter challenges in social communication and interaction, which necessitates continuous, comprehensive interventions to enhance their communication skills. Despite increasing interest in digital therapeutics (DTx), research on speech-utilizing interventions for children with ASD remains limited. This study introduced speech-based technologies integrated into DTx software designed to support the development of communicative skills in children with ASD. We compiled a large speech corpus from both children with ASD and typically developing children, which included clinical scores on social communication severity and speech production, rated by certified speech and language pathologists. Then three speech-based technologies were developed: automatic speech recognition for verbal interaction within the DTx, an automatic assessment model for social communication severity to monitor progress, and an automatic speech production assessment model to facilitate speech production skills. The results were promising, demonstrating a syllable error rate of 12.36% in automatic speech recognition for target keywords, a correlation coefficient of 0.71 for assessing social communication severity, and a correlation coefficient of 0.75 for speech production assessment. These technologies are expected to improve the accessibility of interventions for children with ASD, overcoming barriers related to location, time, and human resources.

Keywords: Automatic speech recognition · Automatic assessments · Children with autism spectrum disorder(ASD) · Digital Therapeutics

1 Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition, with difficulties in social communication being one of the primary symptoms for its diagnosis according to the Diagnostic and Statistical Manual of Mental Disorders

(5th ed.; DSM–5) [1]. It is recommended that diagnosis and, consequently, interventions start as early as possible because the earlier the intervention, the better the prognosis tends to be. To enhance the communicative skills, interventions for comprehensive aspects of communication are required, which encompass the formation, transmission, reception, understanding of messages, and exchanging feedback during communication breakdowns [2]. However, continuous interventions from a young age can lead to physical and financial burdens for both children and their guardians.

Recent advancements and the widespread use of mobile device technologies have spurred research into digital therapeutics (DTx) for intervention [3]. DTx, utilizing software, can be mass-produced at a lower cost, thus alleviating constraints related to time, space, and manpower in treatment. However, in the context of ASD, there are only a few DTx products worldwide, such as Canvas Dx which assists in diagnosis [4]. This indicates research into DTx related to autism, especially those focused on intervention, is globally underdeveloped. Speech or voice-based DTx is rapidly growing, given that speech can be easily input into digital devices and continuously monitored. Consequently, speech has been used as a biomarker in disease diagnosis [5] and in the DTx for the rehabilitation of communication disorders[6, 7]. Considering that over 80% of children with ASD are capable of verbal communication[8, 9], speech-based DTx could play a crucial role in facilitating the improvement of their communication skills.

This paper introduced speech-based technologies applied to DTx specifically aimed at improving the communicative skills of children with ASD. For this purpose, a speech corpus of Korean children with ASD, with clinical scores rated by experts, was constructed. The speech-based technologies included automatic speech recognition (ASR) for keywords, automatic social communication severity assessment, and automatic speech production assessment. ASR is an essential component for enabling children with ASD to verbally interact with DTx. The ASR system was tailored to the speech characteristics of ASD on the fine-tuning framework. The automatic social communication severity assessment system was applied to monitor children's progress, while the automatic speech production assessment was applied to improve delayed speech production skills. The two assessment models were trained on various acoustic features related to the speech characteristics of children with ASD. These acoustic features could be a basis for the assessment results, which would enhance interpretability and reliability in clinical settings.

2 Related Work

This section provides an overview of the existing research on speech corpora of children with ASD and speech-based technologies for children with ASD, including areas such as ASR, automatic assessment for severity in relation to social communication deficit, and automatic speech production assessment. The most extensive speech corpus for children with ASD to date is the USC CARE corpus [10], which contains speech from 46 children with ASD and 14 non-ASD children, totaling 50 hours, including speech from clinicians. It also incorporates clinical scores from ADOS (Autism Diagnostic Observation Schedule) [11] and final diagnoses. The Dutch Asymmetries Corpus [12] in ASDBank and the CSLU Autism speech corpus [13] contain a smaller amount of speech from children with ASD and a restricted range of utterances, which can be attributed to the challenges associated with recruiting children with ASD and recording their speech.

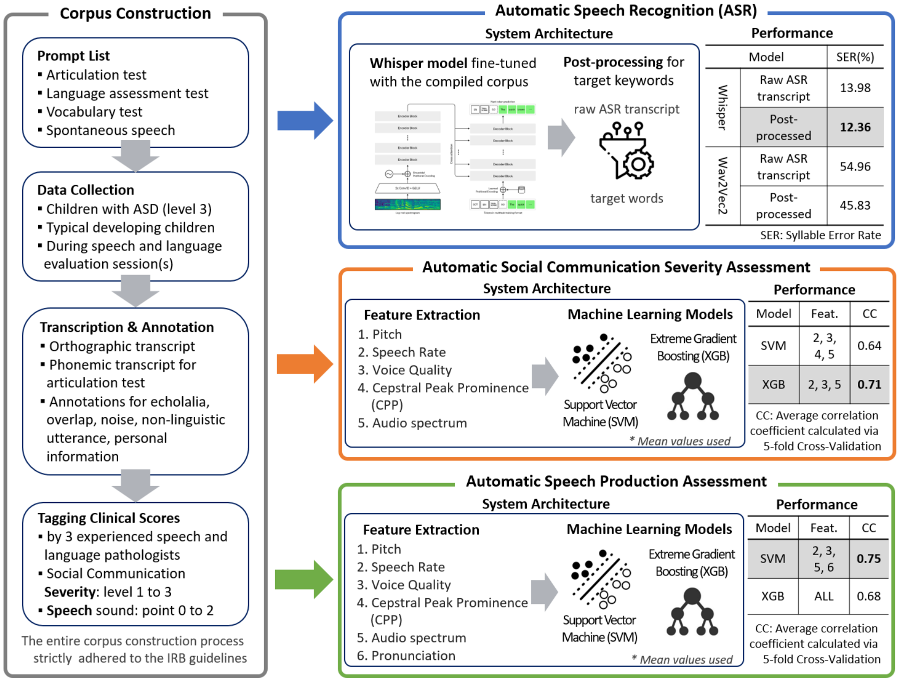
Research on ASR for children with ASD has been limited. A deep learning and transfer learning-based ASR model achieved a word-level speech recognition accuracy of 74% [13]. A study utilizing a large-scale acoustic model and a commercial ASR model showed word recognition rates over 80% in children with ASD with higher IQs and around 50% in those with IQs below 70 [14]. [13] employed 1.5 hours of small-scale speech data. [14] did not leverage the unique speech characteristics of children with ASD because they did not train the models using any speech data of them.

Studies on automatic severity assessment for children with ASD, especially utilizing speech, are scarce, while research on automatic detection or screening for ASD has been more active[15–19]. [20] predicted clinical scores related to socio-communication using embeddings extracted from speech, containing acoustic and lexical features. Although this study incorporated lexical and acoustic features, the use of deep neural network structures made the interpretation of results difficult. Research focusing on the automatic evaluation of speech production skills in children with ASD is almost nonexistent. This might be attributable to a relative lack of interest compared to other areas, such as interaction difficulties [21]. [22] introduced a model evaluating prosody by recognizing intonation curves according to sentence structure. However, it was limited to intonation for assessing a single grammatical aspect.

3 Speech Corpus of Korean Children with ASD

A speech corpus of Korean children with ASD, which is currently under construction, was utilized. The process of construction is summarized in Figure 1. The participants included children diagnosed with ASD based on DSM-5 [1] criteria, along with a control group of typically developing children. Prospective participants suspected of having ASD were either referred to an affiliated hospital for free assessments and medical consultations or underwent an ASD screening test, the Childhood Autism Rating Scale (2nd ed; CARS2) [23], conducted by experts at the Center for Integrative Development and Psychology (CIDP) of Gachon University. If they met the diagnosis criteria, support for participation was provided. The participants were recruited through partnerships with various institutions that have a Memorandum of Understanding with the CIDP, including users of the CIDP. Audio recordings were made during speech and language evaluations conducted by speech and language pathologists.

The recordings were segmented and transcribed by utterance. Phonetic transcription was conducted alongside orthographic transcription, especially for ut-



Corpus Construction

Prompt List
- Articulation test
- Language assessment test
- Vocabulary test
- Spontaneous speech

Data Collection
- Children with ASD (level 3)
- Typical developing children
- During speech and language evaluation session(s)

Transcription & Annotation
- Orthographic transcript
- Phonemic transcript for articulation test
- Annotations for echolalia, overlap, noise, non-linguistic utterance, personal information

Tagging Clinical Scores
- by 3 experienced speech and language pathologists
- Social Communication **Severity**: level 1 to 3
- **Speech** sound: point 0 to 2

The entire corpus construction process strictly adhered to the IRB guidelines

Automatic Speech Recognition (ASR)

System Architecture

Whisper model fine-tuned with the compiled corpus

Post-processing for target keywords

raw ASR transcript

target words

Performance

	Model	SER(%)
Whisper	Raw ASR transcript	13.98
	Post-processed	**12.36**
Wav2Vec2	Raw ASR transcript	54.96
	Post-processed	45.83

SER: Syllable Error Rate

Automatic Social Communication Severity Assessment

System Architecture

Feature Extraction
1. Pitch
2. Speech Rate
3. Voice Quality
4. Cepstral Peak Prominence (CPP)
5. Audio spectrum

Machine Learning Models

Extreme Gradient Boosting (XGB)

Support Vector Machine (SVM)

* Mean values used

Performance

Model	Feat.	CC
SVM	2, 3, 4, 5	0.64
XGB	2, 3, 5	**0.71**

CC: Average correlation coefficient calculated via 5-fold Cross-Validation

Automatic Speech Production Assessment

System Architecture

Feature Extraction
1. Pitch
2. Speech Rate
3. Voice Quality
4. Cepstral Peak Prominence (CPP)
5. Audio spectrum
6. Pronunciation

Machine Learning Models

Extreme Gradient Boosting (XGB)

Support Vector Machine (SVM)

* Mean values used

Performance

Model	Feat.	CC
SVM	2, 3, 5, 6	**0.75**
XGB	ALL	0.68

CC: Average correlation coefficient calculated via 5-fold Cross-Validation

Fig. 1. Overview of the speech technologies applied to a digital therapeutics: corpus construction process; system architecture and performance for automatic speech recognition, automatic social communication severity assessment, and automatic speech production assessment system

terances during the articulation assessment test. Specialized symbols were used to identify specific traits indicative of ASD, including immediate echolalia, context-independent utterances interpreted as delayed echolalia, soft speech, and extended pauses. To protect privacy, personal information such as names and addresses was anonymized by silencing in the audio recordings and masking the corresponding transcriptions with a special symbol. To reduce transcription and annotation inaccuracies, a cross-verification procedure was implemented, involving two independent transcribers.

The compiled corpus underwent evaluation by three nationally certified speech and language pathologists, each possessing over five years of experience in the field of ASD intervention. These professionals assigned social communication severity (hereinafter referred to as severity) levels and speech production (hereinafter referred to as speech) scores to each child with ASD.

To present, the recruitment process has involved 281 children diagnosed with ASD and 50 typically developing children. From this cohort, audio recordings have been obtained for 180 children with ASD and 40 children exhibiting typical development. As of this writing, clinical score tagging has been finalized for 113 children with ASD.

The process of developing the speech corpus was approved by the Institutional Review Board (IRB) of Gachon University (IRB No: 1044396-202207-HR-136-01). All participants were involved through their or their guardians' own initiative. Written consent was obtained from each speaker or their guardians. Participants' autonomy was respected by allowing them to discontinue recording at any time. For more detailed information about the speech corpus, you can refer to [24].

4 ASR and Automatic Assessment Systems

4.1 System Overview

The structures of the ASR, automatic severity and speech assessment systems are illustrated in Figure 1. For ASR, two multilingual ASR models, Whisper [25] and Wav2Vec2.0-XLSR [26], were fine-tuned using speech data from children with ASD. The fine-tuning process aimed to adapt each ASR model to better capture the speech characteristics unique to ASD. The Whisper model utilizes supervised learning techniques, significantly enhancing its performance across a wide range of datasets by scaling up weakly supervised speech recognition training, which leads to high transferability. In contrast, the Wav2Vec2.0-XLSR model focuses on learning cross-lingual speech representations by unsupervised pre-training on raw speech waveforms from multiple languages, making it adept at handling multilingual data. To enhance the model's performance on specific target keywords essential for ASD intervention, a post-processing step was added to extract these keywords.

The automatic severity and speech assessment systems followed a similar framework, with the speech assessment incorporating an additional pronunciation feature. The specific features are as follows:

- Pitch: Mean, standard deviation, minimum, maximum, and range of fundamental frequency
- Speech Rate: Speaking rate, articulation rate, number and duration of pauses, phone ratio, and average syllable duration
- Voice Quality: Jitter, shimmer, Harmonic-to-Noise Ratio, number of voice breaks, and percentage of voice breaks
- Cepstral Peak Prominence (CPP): Calculated before and after voice detection
- Audio spectrum: Log energy and 12 mel-frequency cepstral coefficients, reflecting the shape of the vocal tract
- Pronunciation: Phoneme error rate of the hypothesized phoneme sequence which was obtained by an acoustic model, in comparison with canonical phoneme sequence

Feature values for each utterance were averaged per speaker. Combinations of each feature category were then input into two regression-based machine learning models: Support Vector Machine (SVM) and Extreme Gradient Boosting (XGB). The model's performance was measured by the Spearman correlation coefficient between experts-assessed clinical scores and predicted scores, averaged over 5-fold cross-validation.

4.2 Performance of Each System

The performance of each system is summarized in Figure 1. The Whisper model showed significantly better performance than the Wav2Vec2.0-XLSR model, with the fine-tuned Whisper model yielding a 13.98% of syllable error rate (SER). Post-processing further improved the SER to 12.36%, marking an 11.6% relative improvement. This achievement is comparable to previous studies on ASR for children with ASD [13, 14].

For the assessment systems, the XGB model using features such as speech rate, voice quality, and audio spectrum achieved the highest correlation coefficient of 0.71 in the severity assessment. Meanwhile, the best performance for speech assessment, with a 0.75 correlation coefficient, was obtained by the SVM model trained with the pronunciation feature and the features employed in the XGB model for severity assessment. These common features in the best performing models indicate that children with ASD exhibit distinct characteristics in these areas. The features common to both the SVM and XGB models can be considered significant for the individual evaluation of severity and speech as well. For severity, these features included speech rate, voice quality, and audio spectrum; for speech, they were audio spectrum and pronunciation. However, further research is required for more detailed analysis, as performance may vary with different classifiers and features.

5 Conclusion and Future Works

The ASR and automatic assessment systems were developed as part of DTx software designed for children diagnosed with ASD, who possess minimal verbal communication skills. Although the size of the corpus is the largest of its kind to date, there still remain challenges in using a relatively small dataset for effective application in clinical settings. Nonetheless, the technologies have achieved performance comparable to that of existing ASR systems and automatic assessments for severity and speech production. As there are ongoing efforts to expand the speech corpus, it is anticipated that the speech-based technologies will be continually updated with new data. To further improve the performance, it is important to identify and leverage effective features for the automatic assessment models. Moreover, incorporating other features, such as linguistic features or standard deviation values to capture inter-individual variability, could enhance the performance. The advancement of these technologies and the development of DTx could significantly reduce physical and financial barriers, thus improving the accessibility of interventions.

Acknowledgments. 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].

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. American Psychiatric Association: Diagnostic and statistical manual of mental disorders: DSM-5, 5th edn. American Psychiatric Publishing, Washington, D.C. (2013)
2. Justice, L. M.: Communication sciences and disorder: An introduction, 1st edn. Upper Saddle River, NJ: Merrill/Prentice Hall (2006)
3. Washington, P., Park, N., Srivastava, P., Voss, C., Kline, A., Varma, M., ... Wall, D. P.: Data-driven diagnostics and the potential of mobile artificial intelligence for digital therapeutic phenotyping in computational psychiatry. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, **5**(8), 759–769 (2020)
4. Cognoa Homepage, https://cognoa.com/, last accessed 2024/03/25
5. Bowden, M., Beswick, E., Tam, J., Perry, D., Smith, A., Newton, J., Chandran, S., Watts, O., Pal, S.: A systematic review and narrative analysis of digital speech biomarkers in motor neuron disease. NPJ Digital Medicine **6**(1), 228 (2023)
6. Attwell, G. A., Bennin, K. E., Tekinerdogan, B: A systematic review of online speech therapy systems for intervention in childhood speech communication disorders. Sensors **22**(24), 9713 (2022)
7. Choi, M. J., Kim, H., Nah, H. W., Kang, D. W.: Digital therapeutics: Emerging new therapy for neurologic deficits after stroke. Journal of Stroke **21**(3), 242-–258 (2019)
8. Turner, L. M., Stone, W. L., Pozdol, S. L., Coonrod, E. E.: Follow-up of children with autism spectrum disorders from age 2 to age 9. Autism **10**(3), 243–265 (2006)
9. Charman, T., Taylor, E., Drew, A., Cockerill, H., Brown, J. A., Baird, G.: Outcome at 7 years of children diagnosed with autism at age 2: Predictive validity of assessments conducted at 2 and 3 years of age and pattern of symptom change over time. Journal of Child Psychology and Psychiatry **46**(5), 500–513 (2005)
10. Black, M. P., Bone, D., Williams, M. E., Gorrindo, P., Levitt, P., Narayanan, S: The usc care corpus: Child-psychologist interactions of children with autism spectrum disorders. In: Proceedings of Interspeech 2011, pp.1497–1500. ISCA, Florence, Italy (2011)
11. Lord C., Rutter M., DiLavore P. C., Risi S., Gotham K., Bishop S.: Autism diagnostic observation schedule: ADOS. 2nd edn. Torrance, CA: Western Psychological Services (2012)
12. Kuijper, S. J., Hartman, C. A., Hendriks, P.: Who is he? Children with ASD and ADHD take the listener into account in their production of ambiguous pronouns. PloS one **10**(7), e0132408 (2015)
13. Gale, R., Chen, L., Dolata, J., Van Santen, J., Asgari, M.: Improving asr systems for children with autism and language impairment using domain-focused dnn transfer techniques. In: Proceedings of Interspeech 2019, pp.11–15. ISCA, Graz, Austria (2019)

14. O'Sullivan J, Bogaarts G, Kosek M, Ullmann R, Schoenenberger P, Chatham C, Nobbs D, Murtagh L, Lindemann M, Parish-Morris J, Liberman M, Aponte E, Dorn J, Lipsmeier F.: Automatic speech recognition for ASD using the open-source whisper model from openai. In: International Society for autism Research (INSAR) 2023 Annual Meeting. INSAR, Stockholm, Sweden (2023)
15. Cho, S., Liberman, M., Ryant, N., Cola, M., Schultz, R. T., Parish-Morris, J.: Automatic Detection of Autism Spectrum Disorder in Children Using Acoustic and Text Features from Brief Natural Conversations. In: Proceedings of Interspeech 2019, pp.2513–2517. ISCA, Graz, Austria (2019)
16. Ashwini, B., Narayan, V., Shukla, J.: SPASHT: Semantic and Pragmatic Speech Features for Automatic Assessment of Autism. In: Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.1–5. IEEE, Rhodes Island, Greece (2023)
17. Boughattas, N., Jabnoun, H.: Autism Spectrum Disorder (ASD) Detection Using Machine Learning Algorithms. In: Aloulou, H., Abdulrazak, B., de Marassé-Enouf, A., Mokhtari, M. (eds.) Participative Urban Health and Healthy Aging in the Age of AI. ICOST 2022. Lecture Notes in Computer Science, vol. 13287, pp. 225–233. Springer, Cham (2022). https://doi.org/https://doi.org/10.1007/978-3-031-09593-1_18
18. Farooq, M.S., Tehseen, R., Sabir, M., Atal, Z: Detection of autism spectrum disorder (ASD) in children and adults using machine learning. Scientific Reports **13**, 9605 (2023)
19. Zhao, Z., Tang, H., Zhang, X., Qu, X., Hu, X., Lu, J.: Classification of Children With Autism and Typical Development Using Eye-Tracking Data From Face-to-Face Conversations: Machine Learning Model Development and Performance Evaluation. Journal of medical Internet research **23**(8), e29328 (2021)
20. Chen, C. P., Gau, S. S. F., Lee, C. C.: Learning converse-level multimodal embedding to assess social deficit severity for autism spectrum disorder. In: Proceedings of 2020 IEEE International Conference on Multimedia and Expo, pp.1–6. London, U.K. (2020)
21. Wolk, L., Brennan, C.: Phonological investigation of speech sound errors in children with autism spectrum disorders. Speech, Language and Hearing **16**(4), 239-246 (2013)
22. Ringeval, F., Demouy, J., Szaszak, G., Chetouani, M., Robel, L., Xavier, J., ..., Plaza, M.: Automatic intonation recognition for the prosodic assessment of language-impaired children. IEEE Transactions on Audio, Speech, and Language Processing **19**(5), 1328–1342 (2010)
23. Schopler E., Van Bourgondien M.E., Wellman G.J., Love S.R.: Childhood Autism Rating Scale: CARS, 2nd edn. Western Psychological Services, Los Angeles (2010)
24. Lee S., Mun J., Kim S., Chung M.: Speech Corpus for Korean Children with Autism Spectrum Disorder: Towards Automatic Assessment Systems. arXiv preprint arXiv:2402.15539 (2024) [Accepted for LREC-COLING 2024]
25. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision, In: International Conference on Machine Learning, pp. 28492–28518, PMLR (2023)
26. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised cross-lingual representation learning for speech recognition. In: Proceedings of Interspeech 2021, pp.2426–2430. ISCA, Brno, Czech Republic (2021)