Speech Corpus for Korean Children with Autism Spectrum Disorder: Towards Automatic Assessment Systems

Seonwoo Lee[1], Jihyun Mun[1], Sunhee Kim[2], Minhwa Chung[1, 3]
[1]Department of Linguistics, Seoul National University,
[2]Department of French Language Education, Seoul National University
[3]Interdisciplinary Program in Artificial Intelligence, Seoul National University
Gwanak-ro 1, Gwanak-gu, Seoul, Republic of Korea
{lsw5220, jhhh_1202, sunhkim, mchung}@snu.ac.kr

Abstract

Despite the growing demand for digital therapeutics for children with Autism Spectrum Disorder (ASD), there is currently no speech corpus available for Korean children with ASD. This paper introduces a speech corpus specifically designed for Korean children with ASD, aiming to advance speech technologies such as pronunciation and severity evaluation. Speech recordings from speech and language evaluation sessions were transcribed, and annotated for articulatory and linguistic characteristics. Three speech and language pathologists rated these recordings for social communication severity (SCS) and pronunciation proficiency (PP) using a 3-point Likert scale. The total number of participants will be 300 for children with ASD and 50 for typically developing (TD) children. The paper also analyzes acoustic and linguistic features extracted from speech data collected and completed for annotation from 73 children with ASD and 9 TD children to investigate the characteristics of children with ASD and identify significant features that correlate with clinical scores. The results reveal some speech and linguistic characteristics in children with ASD that differ from those in TD children or another subgroup of ASD categorized by clinical scores, demonstrating the potential for developing automatic assessment systems for SCS and PP.

Keywords: autism spectrum disorder (ASD), speech corpus, social communication severity, pronunciation proficiency, acoustic and linguistic features

1. Introduction

Children with ASD (Autism spectrum disorder) exhibit deficits related to social communication skills (Qualls and Corbett, 2017). In Wetherby (2006), several examples highlight challenges in initiating interactions with others, responding to social overtures, using appropriate vocal tones and gestures, understanding another person's perspective, and engaging in reciprocal conversations. These deficits lead to fewer social interactions than typically developing (TD) children because children with ASD encounter greater difficulties when attempting to interact with others and display limited awareness of their social challenges.

Speech production is one of the difficulties they have, which has received comparatively less attention (Wolk and Brennan, 2013) even though 80% of children with ASD are able to produce at least one word for communication (Turner et al., 2006; Charman et al., 2005). They exhibit atypical prosody , including the higher fundamental frequency and its greater variability (Diehl and Paul, 2012; Fusaroli et al., 2017; Lyakso et al., 2017; Bonneh et al., 2011; McCann and Peppé, 2003); slower speech rate (Bone et al., 2012), longer duration of utterances, frequent and prolonged pauses (Diehl and Paul, 2012; Fusaroli et al., 2017); and poor voice quality (Bone et al., 2012). Articulatory and phonological skills are also delayed in that some children with ASD score below the lower limit of the normal range on standardized articulation tests (Rapin et al., 2009; Cleland et al., 2010; Shriberg et al., 2011). A substantial 41% of school-aged children with ASD produce pronunciation errors that are not only developmentally delayed but also deviate from typical patterns (Cleland et al., 2010; Wolk and Brennan, 2013), which can significantly hinder effective communication.

The diagnosis of ASD reflects these aspects. According to American Psychiatric Association (2013), ASD is diagnosed based on difficulties in communication and interaction, as well as restricted interests and repetitive behaviors, which affect daily functioning. In the clinical setting, standardized diagnostic tests, such as the Autism Diagnostic Observation Schedule, 2nd edition (ADOS-2) (Lord et al., 2012), are conducted. It not only measures the difficulties in social communication (Qualls and Corbett, 2017), but also contains an item for prosody evaluation. However, evaluating children using standardized tests is challenging. The shortage of expertise has led to delayed or missed diagnoses (Li et al., 2022). Moreover, results of the standardized tests could be biased by subjectivity from caregivers or evaluators (Frigaux et al., 2019). The evaluation process lasting longer than an hour also increases children's and their

caregivers' burden as well as decreases children's concentration.

Therefore, digital therapeutics (DTx) for children with ASD have been actively developed in the screening, diagnosis, and treatment of ASD (Wu et al., 2023). Recent studies for DTx systems for children with ASD have utilized a machine learning approach using video data. Kojovic et al. (2021) introduced a machine learning approach that differentiated between children with ASD and TD by recognizing actions in video data. In Cilia et al. (2021), visual data from eye-tracking was pivotal for the diagnosis of ASD. However, collecting video data costs a lot given that it necessitates specialized equipment like eye trackers or cameras, which can be also distracting for children (Albo-Canals et al., 2013). In contrast, speech data offers distinct advantages over video recordings. Audio recordings are more cost-effective and straightforward than video recordings, as well as less intrusive (Clemente, 2008).

In spite of the advantages of speech, there is an absence of a DTx system evaluating children with ASD based on speech data. It can be attributed to the scarcity of speech corpora for machine learning. There are several speech corpora for children with ASD. Dutch ASymmetries Corpus in the ASDBank (Kuijper et al., 2015) is composed of audio recordings of 46 children with ASD. Each child was on a structured storytelling task. The basic demographics such as age and gender are available on the website. The USC CARE corpus (Black et al., 2011) contains speech from 46 children with ASD and 14 children without ASD. The speech recordings were from entire ADOS sessions, which enabled the corpus to include ADOS code scores and final ADOS diagnosis. The CSLU Autism speech corpus (Gale et al., 2019) is utilized for improving the performance of speech recognition. It contains 30 autistic children and 13 children with specific language impairments. The speech was recorded during a recalling sentence task, a sub-task of a standardized language assessment. As the task is short, the total duration of the speech is one and a half hours.

These corpora exhibit limitations in developing an automatic assessment system for the following reasons: First of all, the amount of speech from children with ASD is limited. While there are more than 40 children with ASD (Kuijper et al., 2015; Black et al., 2011), the speech duration might be limited to improve the performance of machine learning models. For instance, the USC CARE corpus has a total duration of 50 hours, but this also includes speech from clinicians. Another concern is the unavailability of clinical scores which is crucial for the purpose of machine learning, especially supervised learning. While some assessments have been implemented during the construction of the corpora, the clinical scores are not provided, except for the USC CARE corpus. However, the ADOS codes in this corpus were assessed by one of three psychologists, which may introduce some degree of subjectivity. Furthermore, there is an absence of clinical scores from speech and language pathologists, who possess specialized expertise in speech and language areas. No study analyzed the speech or language features of children with ASD in correlation with the clinical scores, accordingly.

As for the Korean language, any specialized speech corpus for children with ASD is not available yet, while Korean speech corpora for various disorders have been recently constructed including dysarthria (Choi et al., 2012), cleft palate (Lee et al., 2012), and multiple speech disorders with various etiology, which is released by AI-Hub[1].

To overcome the absence of a suitable speech corpus for automatic assessment systems tailored for children with ASD, this paper presents a speech corpus composed of recordings of Korean children with ASD and clinical scores regarding social communication severity (SCS) and pronunciation proficiency (PP). The speech within the corpus is then examined in terms of acoustic and linguistic attributes in relation to the clinical scores to figure out significant features for automatic assessment systems for children with ASD. This corpus is the first speech corpus of children with ASD that includes perceptually evaluated SCS and PP by 3 speech and language pathologists (SLPs). Moreover, it is also the first speech corpus of Korean children with ASD.

2. Speech Corpus of Children with ASD

2.1. Data Collection

Aiming to capture speech and language traits of children with ASD, the speech recordings are obtained during speech and language evaluation sessions conducted by certificated SLPs at speech and language therapy centers in Korea. For all speakers, written content was obtained from the speakers or their guardians. Participants' autonomy or right was respected by allowing them to discontinue recording at any time. Each session includes standardized tests for speech and language, which are widely used in Korea, measuring articulation, receptive and expressive language skills, and vocabulary understanding and production, as well as natural conversation in a clinical setting. According to the child's chronological age

[1]Website: https://www.aihub.or.kr/

and language development, different tests are conducted. Specific standardized tests are as follows:

- Articulation: Assessment of Phonology and Articulation for Children (APAC) (Kim et al., 2007)
- Language skills: Sequenced Language Scale for Infants (SELSI) (Kim et al., 2003a), Preschool Receptive-Expressive Language Scale (PRES) (Kim et al., 2003b), Language Scale for School-aged Children (LSSC) (Lee et al., 2015)
- Vocabulary: Korean version of Macarthur-Bates Communicative Development Inventories (K M-B CID) (Pae and Kwak, 2011), Receptive & expressive vocabulary test (REVT) (Kim et al., 2009)

The articulation evaluation is an essential part of each session to identify the distinctive pronunciation characteristics of children with ASD. During the process, a child names pictures or repeats utterances provided by an SLP. Even if a child is unable to complete the articulation test, a portion of the test is still recorded. A session generally lasts around 1 hour, and the duration of a child's utterances during a session is expected to be approximately 5 minutes. However, the length of both the session and the child's speech can vary based on the child's chronological age and attention span. Throughout the session, both children and SLPs have their speech recorded. A Logitech Blue Yeti microphone is located at the ceiling's center, preventing a child from being distracted. Starting in June 2022, the corpus will comprise 300 children with ASD and 50 TD children by the end of 2024, encompassing at least 25 hours of speech from children with ASD. Children with ASD are determined based on either DSM-4 or DSM-5 criteria. The corpus incorporates meta information including chronological age, gender, and language evaluation outcomes.

2.2. Transcription and Annotation

All the audio recordings of the sessions are transcribed and annotated by trained transcribers. Given the challenges of deciphering a child's intended speech, the transcription is done phonemically in the Korean alphabet. However, this approach poses limitations for analyzing linguistic aspects and pronunciation-related attributes of children, impeding the identification of inappropriate use of language or pronunciation errors. To overcome the limitation, a transition to orthographic transcription is being planned. This shift will facilitate a more comprehensive analysis, enabling the extraction of linguistic traits for automatic SCS evaluation and automatic mispronunciation detection.

The transcriptions are enriched with annotations. Annotations for pre-processing the audio data include overlap, noise, low volume, as well as non-linguistic sounds like coughs or laughter. In order to probe linguistic attributes, immediate echolalia, off-topic utterances or delayed echolalia, exclamation, and long pauses are identified and labeled. To identify speech sound errors that arise during the articulation evaluation, a target word and mispronounced phonemes are presented together, along with a special symbol. For anonymization, audio segments containing identifiable information and their corresponding transcriptions are masked with silence and a specific symbol, respectively.

2.3. Clinical Scores

2.3.1. Social Communication Severity Level

The social communication severity of each child with ASD is evaluated by three SLPs holding a national certificate in speech and language pathology and possessing over three years of clinical experience, who did not participate in the data recording process. These evaluators classify the children's social communication severity based on their audio recordings and transcriptions. The evaluation criteria are adopted from the social communication component of the diagnosis criterion of DSM-5 criteria (American Psychiatric Association, 2013). The severity is categorized into three levels, considering the extent of required support. The levels are as follows:

- Level 1 (REQUIRING SUPPORT): Have difficulty initiating social interactions, may exhibit unusual or unsuccessful responses to social advances made by others, may seem to have decreased interest in social interactions.
- Level 2 (REQUIRING SUBSTANTIAL SUPPORT): Exhibit marked delays in verbal and non-verbal communication, have limited interest or ability to initiate social interactions, and have difficulty forming social relationships with others, even with support in place.
- Level 3 (REQUIRING VERY SUBSTANTIAL SUPPORT): Have very limited initiation of social interaction and minimal response to social overtures by others and may be extremely limited in verbal communication abilities.

The final severity levels are determined through a voting process, wherein a level is selected if it obtains agreement from more than 2 SLPs.

2.3.2. Pronunciation Proficiency Score

The three SLPs who conduct the SCS evaluation also assess pronunciation proficiency in children with ASD. The same audio recordings and transcriptions for SCS evaluation are used for PP evaluation. The evaluation criteria are adopted from the item within ADOS-2 (Lord et al., 2012) on voice and intonation. The criteria are outlined as follows:

- Score 1: The intonation is not peculiar or strange, displaying typical and appropriate modulation
- Score 2: Less variation in pitch and tone modulation. Monotonic, exaggerated, or occasionally peculiar intonations are observed
- Score 3: Unusual intonation or inappropriate pitch and stress result in markedly monotonous or mechanical sounds lacking inflection. The child emits peculiar cries or sounds.
- Excluded: Limited voice production for the evaluation. While exhibiting normal crying, other vocalizations are scarce.

The final pronunciation scores are calculated by averaging all scores obtained by the three SLPs and subsequently rounding it.

3. Initial Analysis of Acoustic and Linguistic Features

3.1. Current version of the corpus

Starting from 2022 until now, 113 children with ASD (93 boys, 18 girls, and 2 not reported) and 9 children with TD (6 boys and 3 girls) have participated in audio recording. The annotation and evaluation for clinical scores are complete for 73 children with ASD (59 boys, 12 girls, and 2 not reported) and all TD children. Among children with ASD, 2 children's chronological age has not been reported yet. Except for the 2 children, the mean chronological age of 73 children with ASD is 7;8 years. For the 9 TD children, the mean chronological age is 7;10 years. For children with ASD, the total number of utterances both from children with ASD and SLPs is 57,856 (58 hours and 22 minutes), and the number of utterances only from children with ASD is 20,841 (14 hours and 17 minutes). For children with TD, the number of utterances both from TD children and SLPs is 7,032 (6 hours 42 minutes) and that of utterances only from TD children is 2,231 (1 hour and 16 minutes).

The number of children in the TD group and the three ASD subgroups is summarized in Table 2. For SCS, 73 children with ASD are categorized into three subgroups based on their SCS levels: level 1, level 2, and level 3. A lower level indicates better social communication skills. During the analyses of PP, audio recordings of 21 children with ASD are excluded, who did not contain sufficient speech for evaluation. As a result, features from 52 children with ASR are used in the subsequent analysis. These children with ASD are further categorized into three subgroups based on their PP scores: score 1, score 2, and score 3. A lower score indicates better pronunciation.

Group	*N*	mean CA (year;month)	Dur.-all (*N* of utt.)	Dur.-ch. (*N* of utt.)
ASD	73	7;8	58h 22m (57,865)	14h 17m (20,841)
TD	9	7;10	6h 42m (7,032)	1h 16m (2,231)

N for the number; CA for chronological age
Dur-all for the entire corpus; Dur-ch. for children's recordings

Table 1: Basic information of the current version of the corpus

	TD	ASD				
		1	2	3	excl.	Total
SCS	9	25	25	23	-	**73**
PP		15	23	14	21	**73**

excl. denotes exclusion due to limited amount of speech production

Table 2: The number of children in each group divided by social communication severity (SCS) levels and pronunciation proficiency (PP) scores

The inter-rater reliability of each clinical score is assessed using the Intraclass Correlation Coefficient. For SCS evaluation, the inter-rater reliability is 0.939 with a 95% confidence interval of 0.910 to 0.960 ($p<0.001$). Following the PP evaluation, the calculated Intraclass Correlation Coefficient value for 52 children with ASD is 0.941 with a 95% confidence interval of 0.912 to 0.961 ($p<0.001$). One child in pronunciation score 2 has been excluded from the following analysis due to the limited number of utterances.

3.2. Acoustic Analysis

Acoustic features Various low-level acoustic features are extracted to analyze the speech of children with ASD in terms of SCS and PP. The PP score would be directly related to the pronunciation features such as the percentage of correctly pronounced phonemes or vowel space-related features. However, features pertaining to segmental errors cannot currently be extracted since the speech is not orthographically transcribed. As a preliminary study for automatic assessment models, low-level acoustic features are selected.

The feature set employed in the previous study

(Lee et al., 2023), along with features related to cepstral prominence peak (CPP), are extracted. The feature set proved effective in capturing the speech characteristics of children with ASD, distinguishing them from TD children. The feature set encompasses five categories of acoustic features: pitch, audio spectrum, speech rate, voice quality, and CPP. The features within each category are as follows:

- Pitch: the mean, standard deviation, maximum, minimum, median, 25th percentile, and 75th percentile values of fundamental frequencies
- Audio spectrum: a log energy and the 12-dimensional MFCCs, which represent the shape of the vocal track
- Speech rate: total duration, pause duration, speaking duration, speaking rate, articulation rate, average syllable duration, the number of pauses, and the ratio of speech duration and total duration
- Voice quality: jitter, shimmer, HNR (harmonics-to-noise ratio), the number of voice breaks, and the percentage of voice breaks
- CPP (cepstral peak prominence): CPP without voice detection (raw audio), CPP with voice detection

Mel-frequency cepstral coefficients (MFCCs) provide an alternative feature set for capturing features related to pronunciation. The low-order coefficients represent the vocal tract configuration, which is associated with articulation. Therefore, log energy and 12-dimensional MFCCs are extracted for the analysis. CPP features are a more reliable measure of voice quality and can be applied to running speech samples (Heman-Ackah et al., 2003). Spectrum features are extracted using the Librosa library (McFee et al., 2022) and pitch, voice quality, and speech rate features are extracted using the Parselmouth library (Jadoul et al., 2018) in Python. CPP features are extracted through Praat (Boersma and Weenink, 2023).

Analysis methods The acoustic features are extracted from the speech data from the current version of the corpus. The feature values are averaged for each child because the SCS levels and PP scores are rated at the speaker level.

A comparative analysis is carried out among the 4 groups: the three ASD groups and the TD group. In cases where both the assumptions of normality and homogeneous variance are met, a one-way ANOVA is performed, followed by Bonferroni post hoc tests. Alternatively, if either normality assumption or homogeneity of variance assumption is violated, a Kruskal-Wallis test is conducted, followed by Dunn's post hoc test. These statistical analyses are conducted using SciPy library (Virtanen et al., 2020) and Scikit-posthocs (Terpilowski, 2019) in Python. Then the same feature values are normalized to have a zero mean and unit variance to be examined if each of them has a significant correlation with SCS levels and PP scores. Spearman's rank correlation coefficients are computed in Python using SciPy library (Virtanen et al., 2020), as the feature values are continuous whereas scores are categorical. The Spearman's rank correlation coefficient ρ between pronunciation and each feature is calculated using SciPy library (Virtanen et al., 2020) in Python.

Results The results of multiple group comparisons and post hoc tests are detailed in Table 3. This table also presents Spearman's rank correlation coefficients between each feature and the SCS levels or PP scores, accompanied by their respective p-values. The comparative analysis shows that certain acoustic features display differences within ASD subgroups categorized by SCS level, though no acoustic feature distinguishes TD from any ASD subgroups. In case of MFCCs, the 5th, 7th, and 12th MFCCs of severity level 1 are different from those of severity level 2, and all of them except for the 12th MFCC also differ from severity level 3. Regarding voice quality features, jitter and the percentage of voice breaks exhibit significant differences within ASD subgroups. Pitch-related and CPP features show significant variations within these subgroups.

In contrast, distinct patterns emerge for PP scores. Some acoustic features show differences between TD children and ASD subgroups categorized by PP score. The 2nd, 3rd, 8th, and 10th MFCCs of TD are different from those of pronunciation score 1. All of them except for the 2nd MFCC also differ from pronunciation score 2. The 8th MFCC is different from that in pronunciation score 3. Among the 3 subgroups of ASD, features in voice quality and CPP differed. It implies that voice quality-related features would play a major role in the perceptual pronunciation proficiency evaluation of children with ASD.

Correlation analyses between acoustic features and SCS levels indicate that voice quality-related features, especially CPPs, have a moderate relationship with SCS levels. This suggests that children with ASD exhibiting higher SCS levels might demonstrate poorer vocal control. The correlation analysis based on PP scores presents that voice quality-related features, such as jitter and CPPs, are moderately related to the scores. The positive coefficient ρ of jitter corresponds to its general interpretation, as higher jitter is associated

Features		Social Communication Severity			Pronunciation Proficiency		
		Inter-group Difference		Correlation	Inter-group Difference		Correlation
		Statistics†	Post-hoc‡	PCC	Statistics†	Post-hoc‡	PCC
Audio Spectrum	log energy	F=0.197	-	−0.035	F=0.676	-	0.209
	1st MFCC	F=**2.920***	-	**−0.252***	H=7.374	-	0.129
	2nd MFCC	F=0.169	-	0.048	F=**4.234****	TD>ASD1	0.039
	3rd MFCC	F=0.608	-	0.094	H=**17.394****	TD<ASD1,ASD2	**0.289***
	4th MFCC	F=0.771	-	−0.091	F=2.140	-	0.062
	5th MFCC	F=**5.109****	ASD1<ASD2,ASD3	**0.344****	H=5.978	-	0.233
	6th MFCC	F=1.193	-	0.046	H=7.354	-	−0.032
	7th MFCC	F=**6.241****	ASD1<ASD2,ASD3	**0.310****	H=5.669	-	**0.296***
	8th MFCC	F=0.952	-	0.145	F=**9.241*****	TD<ASD1,ASD2,ASD3	−0.039
	9th MFCC	F=0.960	-	0.156	H=5.840	-	0.203
	10th MFCC	F=1.056	-	**0.258***	F=**3.547***	TD<ASD1,ASD2	−0.197
	11th MFCC	H=1.677	-	0.103	H=6.336	-	−0.006
	12th MFCC	F=**4.209****	ASD1<ASD2	**0.248***	H=**7.870***	-	**0.313***
Speech Rate	total duration	H=5.830	-	0.023	H=1.564	-	−0.082
	speech duration	F=1.540	-	0.099	F=2.040	-	−0.084
	speaking rate	F=1.709	-	0.059	F=**5.762****	ASD1<ASD2	**0.328***
	articulation rate	H=3.706	-	0.138	H=**14.203****	ASD1<ASD2	0.235
	number of pauses	H=0.299	-	0.138	F=0.401	-	−0.126
	average syllable duration	F=0.869	-	−0.143	H=**10.142****	ASD1>ASD2	−0.202
	phone ratio	H=5.940	-	-0.075	H=7.261	-	**0.308***
	pause duration	H=4.329	-	0.146	H=3.003	-	−0.120
Voice Quality	jitter	H=**17.763*****	ASD1<ASD2,ASD3	**0.334****	H=**12.914****	ASD1<ASD3	**0.493*****
	shimmer	H=3.557	-	0.124	H=**8.147***	ASD1<ASD2	**0.275***
	HNR (harmonic-to-noise ratio)	H=2.955	-	-0.077	H=**8.777***	ASD1>ASD2	**−0.331***
	number of voice breaks	F=**3.206***	-	**0.259***	F=1.905	-	0.100
	percentage of voice breaks	H=**16.916****	ASD1<ASD3	**0.357*****	H=5.392	-	0.259
Pitch	median	F=**5.289****	ASD1,ASD2<ASD3	**0.315****	F=**3.031***	-	0.020
	mean	F=**5.236****	ASD1,ASD2<ASD3	**0.320****	F=2.758	-	0.044
	standard deviation	F=**5.870****	ASD1,ASD2<ASD3	**0.323****	F=**4.092***	TD<ASD3	0.254
	minimum	H=1.267	-	0.099	H=1.593	-	−0.025
	maximum	F=**5.937****	ASD1<ASD3	**0.374*****	F=**4.030***	-	0.200
CPP	CPP w/o voice detection	H=**25.120*****	ASD1,ASD2>ASD3	**−0.446*****	H=**11.555****	ASD1,ASD2>ASD3	**−0.399****
	CPP w/ voice detection	F=**8.484*****	ASD1>ASD2,ASD3	**−0.417*****	H=**12.810****	ASD1,ASD2>ASD3	**−0.485*****

† F in statistics resulted from one-way ANOVA test and H from Kruskal Wallis test.
‡ The number following ASD represents a severity level or pronunciation score.
PCC denotes Pearson Correlation Coefficient between the feature values and the severity levels or pronunciation proficiency scores.
$^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$

Table 3: Analysis results of acoustic features for inter-group differences among TD and three ASD subgroups and correlation with clinical scores

with greater variance in the voice, leading to poorer speech quality. In contrast, the correlation is negative for CPP features. It is because a lower CPP value is linked with poorer vocal control (Heman-Ackah et al., 2003). Notably, they are the features which show significant differences within the ASD groups. It further supports the role of voice quality-related features in pronunciation evaluation.

3.3. Linguistic Analysis

Linguistic features Since SCS levels are closely related to children's social interactions, we aim to capture their characteristics not only from acoustic features but also from linguistic features.

To extract linguistic features, The Linguistic Feature Toolkit (LFTK) (Lee and Lee, 2023) is employed. The LFTK is a comprehensive toolkit that comprises over 200 handcrafted features collected and categorized from previous research in areas such as text readability assessment, automated essay scoring, fake news detection, and paraphrase detection. We carefully select a linguistic feature set that takes into consideration the data collection process and the specific characteristics of the Korean language.

The linguistic feature set encompasses six categories of features, each designed to capture various aspects of the text data:

- Wordsent: This category includes basic word and sentence counts, such as the total number of words, total number of stop words, total number of syllables, total number of unique words, and total number of sentences.

- Partofspeech: These features pertain to part-of-speech properties and cover aspects like the total number of adjectives, adpositions, adverbs, nouns, and verbs. It also includes

Features		Social Communication Severity		
		Inter-group Difference		Correlation
		Statistics†	Post-hoc‡	PCC
Wordsent	total # of words	H=**19.772*****	TD>ASD3	**−0.410*****
	total # of stop words	H=**14.625****	TD>ASD3	**−0.341****
	total # of syllables	F=1.033	-	**−0.326****
	total # of unique words	H=**26.131*****	ASD1>ASD2,ASD3	**−0.458*****
	total # of sentence	H=**30.757*****	ASD1>ASD2,ASD3	**−0.501*****
Part of Speech	total # of adjective	H=**22.939*****	ASD1>ASD2,ASD3	**−0.427*****
	total # of adposition	F=2.068	-	−0.167
	total # of adverb	F=**3.901****	ASD1>ASD3	**−0.309****
	total # of nouns	F=**4.392****	ASD1>ASD3	**−0.413*****
	total # of verbs	H=**29.492*****	ASD1>ASD2,ASD3	**−0.473*****
	total # of unique adjective	H=**26.697*****	ASD1>ASD2,ASD3	**−0.466*****
	total # of unique adposition	H=**12.771****	ASD1>ASD3	**−0.314****
	total # of unique adverb	H=**22.681*****	ASD1>ASD2,ASD3	**−0.423*****
	total # of unique nouns	H=**21.996*****	ASD1>ASD3	**−0.429*****
	total # of unique verbs	H=**31.176*****	ASD1>ASD2,ASD3	**−0.491*****
Average Wordsent	average # of words per sentence	F=2.480	-	**0.356*****
Average Part of Speech	average # of adjective per word	F=1.185	-	−0.187
	average # of adposition per word	F=**2.779***	ASD1<ASD3	**0.235***
	average # of adverb per word	H=2.015	-	0.014
	average # of nouns per word	F=1.288	-	−0.070
	average # of verbs per word	H=**25.859*****	ASD1>ASD2,ASD3	**−0.468*****
	average # of adjective per sentence	F=1.877	-	0.133
	average # of adposition per sentence	F=1.894	-	**0.329****
	average # of adverb per sentence	H=6.998	-	**0.247***
	average # of nouns per sentence	F=2.118	-	**0.246***
	average # of verbs per sentence	F=1.189	-	−0.039
Lexical Variation	simple adjective variation	F=0.511	-	0.048
	simple adposition variation	F=1.628	-	**−0.229***
	simple adverb variation	H=0.972	-	−0.039
	simple noun variation	F=0.550	-	0.084
	simple verb variation	F=0.015	-	0.111
Type Token Ratio	simple TTR	F=0.104	-	−0.082

† F in statistics resulted from one-way ANOVA test and H from Kruskal Wallis test.
‡ The number following ASD represents a severity level.
PCC denotes Pearson Correlation Coefficient between the feature values and the severity levels.
*$p<0.05$, **$p<0.01$, ***$p<0.001$

Table 4: Analysis results of linguistic features for inter-group differences among TD and three ASD subgroups and correlation with social communication severity levels

counts of unique adjectives, adpositions, adverbs, nouns, and vowels.

- Avgwordsent: This category calculates the average number of words per sentence.
- Avgpartofspeech: It calculates averages related to part-of-speech features, including the average number of adjectives per word, adpositions per word, adverbs per word, nouns per word, verbs per word, adjectives per sentence, adpositions per sentence, adverbs per sentence, nouns per sentence, and verbs per sentence.
- Lexicalvariation: These features measure lexical variation and include variations in simple adjectives, adpositions, adverbs, nouns, and verbs.
- Typetokenratio: The type token ratio is used to capture the lexical richness of a text.

Each linguistic feature value is computed based on the transcription, which transcribes the entire recording session.

Analysis methods Linguistic analysis is conducted using the same methodology as acoustic analysis.

Results The results of both comparative and correlation analyses are provided in Table 4. Significant differences emerged across the four groups. Within the Wordsent features, key parameters like the total number of words, stop words, unique words, and sentences showed significant variations across groups. Significant differences were observed between severity level 1 and severity level 3 in terms of these parameters, emphasizing linguistic differences in children with varying social communication severity. Partofspeech features revealed significant variations, with post hoc analyses identifying specific group pairs with notable differences.

In the correlation analysis involving linguistic features and SCS levels, features reflecting a child's active participation, such as the total number of sentences and words, show moderate negative correlations with SCS levels. This suggests that children with higher SCS levels tend to participate less in sessions, resulting in limited verbal communication. Notably, more linguistic features show significant correlations with SCS levels compared to acoustic features, emphasizing the importance of linguistic analysis in examining children with ASD's speech. The relationships between linguistic features and social communication are complex, but these results offer insights into potential linguistic markers for further exploration in clinical contexts.

3.4. Discussion

The outcomes from statistical analyses point out distinct speech characteristics in children with ASD compared to TD children. It is notable that MFCCs are the major features differentiating TD and ASD subgroups divided by PP scores. Considering that MFCCs are more related to pronunciation than other features, it could reflect the poorer articulatory skills in children with ASD as reported in previous studies on their speech production (Rapin et al., 2009; Cleland et al., 2010; Shriberg et al., 2011; Wolk and Brennan, 2013). This finding also endorse the necessity for an analysis of pronunciation-related features in speech.

The features exhibiting differences between children with ASD and TD in this study do not overlap with the study on the differentiation of children with ASD from TD (Lee et al., 2023). It can be attributed to methodological differences, as this study computes group averages on a per-speaker basis, while Lee et al. (2023) computed group averages based on utterances.

Voice quality-related features show not only differences across subgroups of ASD but also correlation with SCS levels and PP scores. It is consistent with Bone et al. (2012) which revealed the relation between voice quality features and the degree of atypicality of speech in children with ASD. It is worth noting the relationship between SCS and voice quality. It could be explained by their difficulties in controlling vocal tones as reported in Wetherby (2006). Voice quality-related features could serve as indicators of the challenges these children face in effective communication and social engagement.

There were differences between results regarding SCS levels and those regarding PP scores. While MFCCs differed among the subgroups of ASD categorized by SCS levels, they differentiated ASD subgroups from TD from the perspective of pronunciation. Regarding the differences among the subgroups, pitch features exhibited variations in relation to SCS levels, whereas speech rate features differed among the subgroups based on PP scores. In addition to the acoustic features, linguistic features also demonstrated correlations with SCS levels, indicating the degree of participation. These contrasting results highlight the fact that assessing SCS and PP is based on distinct aspects of speech, emphasizing the multifaceted nature of speech evaluation in children with ASD.

Some features did not differ among the groups or showed a minor or negligible correlation with the clinical scores. Due to the substantial variability in speech characteristics among children with ASD, it is plausible that features could overlap within the subgroups. However, these features could still hold usefulness for developing an automatic evaluation model in that features utilized to train a classification model did not necessarily coincide with those showing significant differences between children with ASD and TD (Lee et al., 2023). The influence of each feature should be identified in further research.

4. Conclusion

A speech corpus of Korean children with ASD, which provides clinical scores on SCS and PP, is constructed for the first time and analyzed for automatic assessment systems for social communication and pronunciation. To reflect speech and linguistic aspects in children with ASD, interactions during speech and language evaluation sessions between children with ASD and SLPs were recorded. The speech corpus is to be composed of speech recordings collected from 300 children with ASD and 50 children with TD, with the current version containing speech from 73 children with ASD and 9 TD children.

To explore the acoustic and linguistic characteristics in relation to the clinical scores, comparative analyses, and correlation analyses are performed. The children with ASD are divided into three subgroups based on their SCS levels and

PP scores, respectively, and the differences in the acoustic features among the three ASD groups and TD children are examined. As a result, features within MFCCs, voice quality, and pitch are different among the groups divided by the SCS levels. Certain features within MFCCs, speech rate, voice quality, and pitch exhibit significant differences across the groups divided by PP scores. Voice quality-related features show a significant moderate correlation with SCS levels and PP scores. SCS levels are also in significant moderate correlations with linguistic features embracing various surface lexical features. These findings underscore the multifaceted nature of evaluation for children with ASD and are also invaluable for clinicians and researchers working to enhance our understanding of ASD and develop effective automatic assessment systems.

One limitation of these analyses is the absence of features directly associated with SCS and PP. At this moment, it is unable to identify the exact content of the utterances and mispronounced speech sounds as it is phonemically transcribed. Once orthographic transcriptions are provided, the analyses will be extended with the PP- and SCS-related features. Future work involves developing automatic evaluation models for SCS and PP, reflecting the analysis findings.

5. Acknowledgements

The process of developing the speech corpus is approved by the Institutional Review Board (IRB) of Gachon University (IRB No: 1044396-202207-HR-136-01) and written informed consent is obtained from each speaker or their caregivers. This work was partly 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] and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [No.2021-0-01343-004, Artificial Intelligence Graduate School Program (Seoul National University)].

6. Bibliographical References

Jordi Albo-Canals, Marcel Heerink, Marta Diaz, Vanesa Padillo, Marta Maristany, Alex Barco, Cecilio Angulo, Ariana Riccio, Lauren Brodsky, Simone Dufresne, et al. 2013. Comparing two lego robotics-based interventions for social skills training with children with asd. In *2013 IEEE RO-MAN*, pages 638–643. IEEE.

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders*, 5th edition edition. American Psychiatric Association.

Matthew P. Black, Daniel Bone, Marian E. Williams, Phillip Gorrindo, Pat Levitt, and Shrikanth Narayanan. 2011. The USC CARE corpus: child-psychologist interactions of children with autism spectrum disorders. In *Interspeech 2011*, pages 1497–1500. ISCA.

Paul Boersma and David Weenink. 2023. Praat: doing phonetics by computer (version 6.3.10).

Daniel Bone, Matthew P. Black, Chi-Chun Lee, Marian E. Williams, Pat Levitt, Sungbok Lee, and Shrikanth Narayanan. 2012. Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist. In *Interspeech 2012*, pages 1043–1046. ISCA.

Yoram S. Bonneh, Yoram Levanon, Omrit Dean-Pardo, Lan Lossos, and Yael Adini. 2011. Abnormal speech spectrum and increased pitch variability in young autistic children. *Frontiers in Human Neuroscience*, 4:237.

Tony Charman, Emma Taylor, Auriol Drew, Helen Cockerill, Jo-Anne Brown, and Gillian Baird. 2005. 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.

Dae-Lim Choi, Bong-Wan Kim, Yeon-Whoa Kim, Yong-Ju Lee, Yongnam Um, and Minhwa Chung. 2012. Dysarthric speech database for development of QoLT software technology. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3378–3381, Istanbul, Turkey. European Language Resources Association (ELRA).

Federica Cilia, Romuald Carette, Mahmoud Elbattah, Gilles Dequen, Jean-Luc Guérin, Jérôme Bosche, Luc Vandromme, Barbara Le Driant, et al. 2021. Computer-aided screening of autism spectrum disorder: Eye-tracking study using data visualization and deep learning. *JMIR human factors*, 8(4):e27706.

Joanne Cleland, Fiona E. Gibbon, Sue J. E. Peppé, Anne O'Hare, and Marion Rutherford. 2010. Phonetic and phonological errors in children with high functioning autism and asperger

syndrome. *International Journal of Speech-Language Pathology*, 12(1):69–76.

Ignasi Clemente. 2008. Recording audio and video. *The Blackwell guide to research methods in bilingualism and multilingualism*, pages 177–191.

Joshua John Diehl and Rhea Paul. 2012. Acoustic differences in the imitation of prosodic patterns in children with autism spectrum disorders. *Research in autism spectrum disorders*, 6(1):123–134.

A Frigaux, R Evrard, and J Lighezzolo-Alnot. 2019. Adi-r and ados and the differential diagnosis of autism spectrum disorders: Interests, limits and openings. *L'encephale*, 45(5):441–448.

Riccardo Fusaroli, Anna Lambrechts, Dan Bang, Dermot M. Bowler, and Sebastian B. Gaigg. 2017. Is voice a marker for autism spectrum disorder? a systematic review and meta-analysis. *Autism Research*, 10(3):384–407.

Robert Gale, Liu Chen, Jill Dolata, Jan Van Santen, and Meysam Asgari. 2019. Improving ASR systems for children with autism and language impairment using domain-focused DNN transfer techniques: 20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language, INTERSPEECH 2019. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019-September:11–15.

Yolanda D. Heman-Ackah, Reinhardt J. Heuer, Deirdre D. Michael, Rosemary Ostrowski, Michelle Horman, Margaret M. Baroody, James Hillenbrand, and Robert T. Sataloff. 2003. Cepstral peak prominence: a more reliable measure of dysphonia. *The Annals of Otology, Rhinology, and Laryngology*, 112(4):324–333.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15.

Min Jung Kim, Soyeong Pae, and Chang Il Park. 2007. *Assessment of Phonology and Articulation for Children (APAC)*. Human Brain Research & Counseling.

Young Tae Kim, Gyung Hun Hong, Kim KyungHee, Chang Hae-Seong, and Ju-Yeon Lee. 2009. *Receptive & expressive vocabulary test (REVT)*. Seoul Community Rehabilitation Center.

Young Tae Kim, KyungHee Kim, Hea Ryun Yoon, and Wha-soo Kim. 2003a. *Sequenced Language Scale for Infants (SELSI)*. Special Education Publishing.

Young Tae Kim, Tae-Je Seong, and YoonKyoung Lee. 2003b. *Preschool Receptive-Expressive Language Scale (PRES)*. Seoul Community Rehabilitation Center.

Nada Kojovic, Shreyasvi Natraj, Sharada Prasanna Mohanty, Thomas Maillart, and Marie Schaer. 2021. Using 2d video-based pose estimation for automated prediction of autism spectrum disorders in young children. *Scientific Reports*, 11(1):15069.

Sanne J. M. Kuijper, Catharina A. Hartman, and Petra Hendriks. 2015. Who Is He? Children with ASD and ADHD Take the Listener into Account in Their Production of Ambiguous Pronouns. *PloS One*, 10(7):e0132408.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Ji Eun Lee, Wook Eun Kim, Kwang Hyun Kim, Myung Whun Sung, and Tack Kyun Kwon. 2012. Research on construction of the korean speech corpus in patient with velopharyngeal insufficiency. *Korean J Otorhinolaryngol-Head Neck Surg*, 55(8):498–507.

Seonwoo Lee, Eunjung Yeo, Sunhee Kim, and Minhwa Chung. 2023. Knowledge-driven speech features for detection of korean-speaking children with autism spectrum disorder. *Phonetics and Speech Sciences*, 15:53–59.

YoonKyoung Lee, Hyunsook Heo, and Seungmin Jang. 2015. *Language Scale for School-aged Children (LSSC)*. Inpsyt.

Jing Li, Zejin Chen, Gongfa Li, Gaoxiang Ouyang, and Xiaoli Li. 2022. Automatic classification of asd children using appearance-based features from videos. *Neurocomputing*, 470:40–50.

Catherine Lord, Michael Rutter, Rhiannon J. Luyster, and Katherine Gotham. 2012. *Autism diagnostic observation schedule-2nd edition (ADOS-2)*, 2nd edition edition. Western Psychological Corporation.

Elena Lyakso, Olga Frolova, and Aleksey Grigorev. 2017. Perception and acoustic features of speech of children with autism spectrum disorders. In *Speech and Computer*, Lecture Notes in Computer Science, pages 602–612. Springer International Publishing.

Joanne McCann and Sue Peppé. 2003. Prosody in autism spectrum disorders: a critical review. *International Journal of Language & Communication Disorders*, 38(4):325–350.

Brian McFee, Alexandros Metsai, Matt McVicar, Stefan Balke, Carl Thomé, Colin Raffel, Frank Zalkow, Ayoub Malek, Dana, Kyungyun Lee, Oriol Nieto, Dan Ellis, Jack Mason, Eric Battenberg, Scott Seyfarth, Ryuichi Yamamoto, viktorandreevichmorozov, Keunwoo Choi, Josh Moore, Rachel Bittner, Shunsuke Hidaka, Ziyao Wei, nullmightybofo, Adam Weiss, Darío Hereñú, Fabian-Robert Stöter, Lorenz Nickel, Pius Friesch, Matt Vollrath, and Taewoon Kim. 2022. librosa/librosa: 0.9.2.

Soyeong Pae and Keum-Joo Kwak. 2011. *Korean MacArthur-Bates Communicative Development Inventories (K M-B CDI)*. Mindpress.

Lydia R Qualls and Blythe A Corbett. 2017. Examining the relationship between social communication on the ados and real-world reciprocal social communication in children with asd. *Research in autism spectrum disorders*, 33:1–9.

Isabelle Rapin, Michelle A. Dunn, Doris A. Allen, Michael C. Stevens, and Deborah Fein. 2009. Subtypes of language disorders in school-age children with autism. *Developmental Neuropsychology*, 34(1):66–84.

Lawrence D. Shriberg, Rhea Paul, Lois M. Black, and Jan P. van Santen. 2011. The hypothesis of apraxia of speech in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 41(4):405–426.

Maksim A. Terpilowski. 2019. scikit-posthocs: Pairwise multiple comparison tests in python. *Journal of Open Source Software*, 4(36):1169.

Lauren M. Turner, Wendy L. Stone, Stacie L. Pozdol, and Elaine E. Coonrod. 2006. Follow-up of children with autism spectrum disorders from age 2 to age 9. *Autism: The International Journal of Research and Practice*, 10(3):243–265.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261–272.

Amy M Wetherby. 2006. Understanding and measuring social communication in children with autism spectrum disorders. *Social and communication development in autism spectrum disorders: Early identification, diagnosis, and intervention*, 18(3):3–34.

Lesley Wolk and Christine Brennan. 2013. Phonological investigation of speech sound errors in children with autism spectrum disorders. *Speech, Language and Hearing*, 16(4):239–246.

Xuesen Wu, Haiyin Deng, Shiyun Jian, Huian Chen, Qing Li, Ruiyu Gong, and Jingsong Wu. 2023. Global trends and hotspots in the digital therapeutics of autism spectrum disorders: a bibliometric analysis from 2002 to 2022. *Frontiers in Psychiatry*, 14:1126404.