pISSN 2005-8063
eISSN 2586-5854
2023. 6. 30.
Vol.15 No.2
pp. 53-59

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2023.15.2.053

Knowledge-driven speech features for detection of Korean-speaking children with autism spectrum disorder*

Seonwoo Lee[1] · Eun Jung Yeo[1] · Sunhee Kim[2] · Minhwa Chung[1,**]

[1]*Department of Linguistics, Seoul National University, Seoul, Korea*
[2]*Department of French Language Education, Seoul National University, Seoul, Korea*

Abstract

Detection of children with autism spectrum disorder (ASD) based on speech has relied on predefined feature sets due to their ease of use and the capabilities of speech analysis. However, clinical impressions may not be adequately captured due to the broad range and the large number of features included. This paper demonstrates that the knowledge-driven speech features (KDSFs) specifically tailored to the speech traits of ASD are more effective and efficient for detecting speech of ASD children from that of children with typical development (TD) than a predefined feature set, extended Geneva Minimalistic Acoustic Standard Parameter Set (eGeMAPS). The KDSFs encompass various speech characteristics related to frequency, voice quality, speech rate, and spectral features, that have been identified as corresponding to certain of their distinctive attributes of them. The speech dataset used for the experiments consists of 63 ASD children and 9 TD children. To alleviate the imbalance in the number of training utterances, a data augmentation technique was applied to TD children's utterances. The support vector machine (SVM) classifier trained with the KDSFs achieved an accuracy of 91.25%, surpassing the 88.08% obtained using the predefined set. This result underscores the importance of incorporating domain knowledge in the development of speech technologies for individuals with disorders.

Keywords: autism spectrum disorder (ASD), disorder detection, speech features, prosodic features

1. Introduction

Autism spectrum disorder (ASD) is diagnosed by deficits in social interaction as well as restricted and repetitive behaviors or interests, according to the Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5) (American Psychiatric Association, 2013). Though diagnosis of ASD is mostly based on observing a child's behavior during a clinical evaluation process, some standardized diagnostic tests such as Autism Diagnostic Observation Schedule, 2nd edition (ADOS-2) contain specific items to evaluate speech and language of a child, particularly, with respect to prosody (Lord et al., 2012).

* This work was supported by the 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).

** mchung@snu.ac.kr, Corresponding author

Received 23 May 2023; Revised 24 June 2023; Accepted 24 June 2023

Speech of children with ASD is typically demonstrated as monotonous or exaggerated because of its atypical prosody. When it comes to pitch, the fundamental frequency (F0) of ASD children is higher than that of typicaly developing (TD) children (Diehl & Paul, 2012; Fusaroli et al., 2017; Lyakso et al., 2017), though it is a controversial topic. The range of the F0 varies greatly, from very narrow to very wide (Bonneh et al., 2011; Fusaroli et al., 2017; McCann & Peppé, 2003). The speech rate of ASD children is slower than their TD peers (Bone et al., 2012). It could be due to longer utterances (Diehl & Paul, 2012; Fusaroli et al., 2017) and syllable duration (Fusaroli et al., 2017). Additionally, the pauses are longer and more frequent in the speech of ASD children (Fusaroli et al., 2017). Voice quality is also a contributing factor to their abnormal speech. The degree of atypical speech in ASD children has a positive correlation with jitter and jitter variability and a negative correlation with harmonic-to-noise ratio (HNR) (Bone et al., 2012).

These abnormal speech patterns of ASD children have motivated researchers to focus on speech signals to detect ASD (Asgari et al., 2021; Beccaria et al., 2022; Cho et al., 2019; Lee et al., 2023; MacFarlane et al., 2022; Mohanta et al., 2020). It would reduce the time and effort spent for diagnosis of ASD and could be easily utilized for screening of the disorder. Automated ASD detection models have been developed based on hand-crafted features (Asgari et al., 2021; Cho et al., 2019; MacFarlane et al., 2022; Mohanta et al., 2020), or, recently, deep learning (Lee et al., 2023). Hand-crafted features were to reflect the characteristics specific to the dataset used, which resulted in difficulty in generalizing to other datasets. On the other hand, deep learning approaches suffered from the black box issue, making it challenging to interpret the results.

Recently, in order to mitigate the limitations of conventional approaches, some studies have proposed the detection of ASD speech using predefined acoustic feature sets such as Geneva Minimalistic Acoustic Standard Parameter Set (GeMAPS), extended GeMAPS (eGeMAPS) (Eyben et al., 2016), and COMputationsl PARalinguistics ChallengE (ComParE) which are extracted by openSMILE toolkit (Eyben et al., 2010) in Python. They are proposed as a standard acoustic feature set for voice analysis to facilitate easy comparison and merger across studies on speech signals (Eyben et al., 2016). The GeMAPS set contains 18 low-level descriptors (LLD) and their functionals (arithmetic mean and coefficient of variation). The 18 LLDs are related to frequency, amplitude, spectrum, and time. The eGeMAPS set has additional 7 LLDs in the cepstral and frequency domains, some of which are dynamic. The ComParE set includes 44 LLDs whose extra domain includes voicing probability and extra delta features are added. The total number of features in GeMAPS, eGeMAPS, and ComParE is 62, 88, and 6,373, respectively. The easiness to extract features and the general coverage for acoustics have attracted researchers to utilize predefined feature sets to detect a disorder, including Alzheimer's disease (Haider et al., 2020), voice disorders (Barche et al., 2020) as well as ASD (Beccaria et al., 2022; Cho et al., 2019).

One of the limitations is that most features included in the predefined set do not hold significant clinical relevance, because they were proposed for general acoustic research purposes. For the detection of a disease, a subset of features was selected and even compressed to reduce the number of features. This process may not consider the actual speech characteristics of people with ASD, which does not ensure that the features irrelevant to the properties of ASD are excluded. An additional concern arises that the results are hard to be interpreted when the features are compressed by feature extraction techniques. It is not easy to accurately explain what each compressed feature denotes and how it contributes to the result. Therefore, the compressed features could lose their ability to provide evidence for diagnosis. Despite the dramatic development of technology, technologies for disorders should be utilized with considerable care.

This paper aims to investigate the efficacy of knowledge-driven speech features (KDSFs) representing characteristics of children with ASD for detecting children with ASD. Then it will be directly compared with a predefined set, eGeMAPS, for the first time. The higher performance of KDSFs would indicate the importance of injecting disorder-specific knowledge into the development of technologies for disordered speech. The rest of this paper is organized as follows. Section 2 introduces related works of detection of ASD with predefined feature sets. The next section describes methods, including a speech dataset and features used. Then experimental settings and results are followed in section 4. Section 5 discusses the experimental results, followed by a conclusion in section 6.

2. Related Work

Two studies made direct use of the predefined feature sets for the detection of autistic speech. Beccaria et al. (2022) utilized the eGeMAPS feature set to analyze Italian-speaking ASD children's speech and classify children with ASD from their TD peers. They selected to use the set, considering the number of features and its dynamic features. The 16 features significantly different from age- and gender-matched TD children were selected to train various classifiers to avoid the situation where features unrelated to the symptoms lead to lower performance. Selected features included pitch falling slope, F2 frequency, F2 bandwidth, jitter, shimmer, HNR, loudness, loudness of rising slope, loudness of 20th percentile, spectral flux of voiced segments, slope between 0–500 Hz of unvoiced segments, and slope of 500–1,500 Hz of unvoiced segments. Among various classifiers, support vector machine (SVM) showed stable performances across metrics, achieving 83% of accuracy and 80% of F1-score on the test set.

Cho et al. (2019) extracted speech features using the ComParE feature set. Extra text features were also utilized to reflect linguistic characteristics such as lexicons, turn-taking, and part-of-speech. All features were extracted from both children with ASD and TD. Among more than 6,000 ComParE features, 44 LLDs with 4 functionals of mean, median, standard deviation (SD), and interquartile range (IQR), which is calculated by subtracting 25th percentile value from 75th percentile value, were used. By feature selection, more than 17 acoustic features of AD children were selected, including voicing probability, median of HNR-delta, IQR of root mean square-delta, median of 6th and 7th Mel-frequency cepstral coefficient (MFCC)-delta, IQR of HNR median or 3rd MFCC-delta, mean and median of 6th MFCC, median zero crossing rate, and median and IQR of F0-delta. Due to the large number of features from speech and text, principal component analysis was implemented to reduce the feature dimension into 10. The gradient boosting classifier achieved an accuracy of 75.71% (65.71% for ASD and 85.71% for TD children).

While these studies demonstrate the effectiveness of predefined feature sets in detecting the speech of ASD children, the importance

of proper feature selection is also clearly revealed. A large number of features were extracted initially, but only a subset of the features was actually used. This could be explained by the fact that the feature sets may not always reflect clinical intuition and significance due to their large number. As a result, researchers are prompted to select or even compress features for the purpose of improving performance as well as avoiding overfitting. However, as mentioned earlier, the reduced features are difficult to interpret, thereby diminishing the practicality of using a classification model in clinical settings.

3. Method

3.1. Speech Database

The speech database used for the experiment is an initial version of a speech database of Korean-speaking children with ASD, which is under construction for developing a digital therapeutic program. In this version of the database, there are 64 ASD children and 9 TD children as controls. Every child with ASD is diagnosed by either DSM-4 or DSM-5 criteria. Speech from a child with ASD is excluded for the experiment because of the small number of utterances. All the utterances in the database are in Korean. Unfortunately, the database would not be publicly open because of its privacy issue.

The speech is recorded during speech and language evaluation sessions at various speech and language centers in Korea. The utterances are recorded through a microphone (Logitech Blue Yeti) which is attached to the center of the ceiling. Initially, the microphone was placed on a desk between a child and a speech and language pathologist (SLP). However, it was later moved to the current location because the presence of the microphone and its twinkling light distracted some children.

Each session consists of a standardized Korean articulation assessment using the Assessment of Phonology and Articulation for Children (APAC) (Kim et al., 2007), and a conversation with an SLP. Articulation assessment is conducted to capture the articulatory characteristics of ASD children, which can be utilized to develop an automatic speech recognition model and pronunciation evaluation model for the digital therapeutic program. APAC is selected because every phoneme in the Korean language is included within the target words. During each session of the assessment, a child is asked to name objects or actions described on colored pictures. If the child makes any mispronunciations, the SLP asks the child to name it again. When a child does not respond spontaneously, the SLP would name it and request the child to repeat it. Conversations are designed to elicit words that would be used for the digital therapeutic program. The topics of the conversations are related to children's everyday routines and play. The audio recording of each session is manually segmented into utterances and then transcribed and annotated for overlap, noise, immediate echolalia, off-topic utterances, exclamation, non-linguistic sounds (such as laugh or cough), long pauses, and low-volume sounds. In addition, utterances spoken during the articulation assessment are phonemically transcribed with a special symbol when there are any mispronunciations. Any audio segments associated with the child's identification are masked. The current version of the database does not include metadata about the children's chronological age, gender, diagnosis-related scores, and language and speech performance. This information will be added to the final version.

The process of collecting and pre-processing the database is approved by the Institutional Review Board (IRB) of Gacheon University (IRB No: 1044396-202207-HR-136-01) and written informed consent is obtained from each speaker or their parents or caregivers.

3.2. Knowledge-Driven Speech Features (KDSFs)

In order to reflect the speech characteristics of ASD children, prosodic features that are identified in previous studies as different from their TD peers are extracted as KDSFs, including pitch, voice quality, and speech rate. Pitch-related features are to capture the higher F0 and variations in the range of F0 of children with ASD (Bonneh et al., 2011; Diehl & Paul, 2012; Fusaroli et al., 2017; Lyakso et al., 2017; McCann & Peppe´, 2003). Voice quality-related features are extracted to reflect the fact that voice quality is correlated with the abnormality of speech in children with ASD (Bone et al., 2012). Features associated with speech rate are selected to take into account slower speech rate, which could be influenced by longer utterances, syllable duration, pause duration, and frequent pauses (Bone et al., 2012; Diehl & Paul, 2012; Fusaroli et al., 2017).

MFCCs as cepstral features are additionally extracted given that they encode general acoustic characteristics associated with spectral shape, timbre, and linguistic content of the audio signal. They are widely utilized for the detection of autistic speech (Beccaria et al., 2022; Mohanta & Mittal, 2022) as well as other disorders such as dysarthria (Yeo et al., 2021), Parkinson's disease (Benba et al., 2015), and voice disorders (Barche et al., 2020). Specific features are described as follows. The number in parentheses denotes the total number of features in the category:

1) Pitch (7): A voice report is first generated, and F0-related values are extracted from it. Pitch features include mean, SD, maximum, minimum, median, 25th percentile, and 75th percentile values of F0s.
2) Voice quality (5): Voice quality-related features include jitter, shimmer, HNR, the number of voice breaks, and the percentage of voice breaks. Jitter refers to the cycle-to-cycle frequency variation, while shimmer refers to the cycle-to-cycle amplitude variation. HNR measures the ratio between the harmonic components and the non-harmonic components (noise) in the voice signal. The number of voice breaks is the count of abrupt interruptions in the voice signal, and the percentage of voice breaks denotes the proportion of time in the voice signal containing abrupt interruptions.
3) Speech rate (8): Speech rate-related features include 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. Speaking duration is calculated by subtracting the pause duration from the total duration. Speaking rate is calculated as the total duration divided by the number of syllables, and articulation rate is calculated as the speaking duration divided by the number of syllables. Average syllable duration is a reciprocal of articulation rate, calculated as the number of syllables divided by the speaking duration.
4) MFCCs (13): Cepstral features include the first 13 MFCCs. From a speech signal, 12-dimensional MFCCs and a log

Table 1. Classification results (%)

Feature set		Accuracy	ASD			TD		
			Precision	Recall	F1-score	Precision	Recall	F1-score
eGeMAPS	All (88)	88.08	88.12	**99.90**	93.64	**78.57**	2.66	5.14
	RFE (70)	88.29	88.62	99.43	93.72	65.31	7.73	13.82
KDSFs	All (29)	**91.25**	92.23	98.33	**95.18**	76.85	40.10	52.70
	RFE (19)	90.78	**94.55**	94.99	94.77	62.50	**60.39**	**61.43**
Subset of KDSFs	Pitch+MFCC (18)	91.08	93.05	**97.09**	95.03	**69.37**	47.58	56.45
	VQ+MFCC (18)	91.90	**95.45**	95.32	95.39	66.51	**67.15**	66.83
	SR+MFCC (21)	**92.13**	95.40	95.66	**95.53**	67.98	66.67	**67.32**

The best performances are in bold. The number in parenthesis denotes the number of features used for training the classifiers.
ASD, autism spectrum disorder; TD, typical development; eGeMAPS, extended Geneva Minimalistic Acoustic Standard Parameter Set; RFE, recursive feature elimination; KDSFs, knowledge-driven speech features; VQ, voice quality; MFCC, Mel-frequency cepstral coefficient; SR, speech rate.

energy set are extracted. Then each of the 13-dimensional MFCCs are time-averaged.

Pitch, voice quality, and speech rate-related features are extracted with Parselmouth (Jadoul et al., 2018) in Python and cepstral features using the Librosa library (McFee et al., 2022) in Python.

4. Experiments

4.1. Experimental Settings

The dataset contains 14,967 utterances (10 hours and 21 minutes) of ASD children and 2,086 (1 hour and 18 minutes) of TD children. The dataset is first divided into a training set and a test set in a ratio of 8:2 based on the speakers. During training, the small number of utterances from TD children would cause an issue related to a data imbalance. To handle the issue, three augmentation techniques of SpecAugment (Park et al., 2019) of time warp, frequency masking, and time masking are applied to only TD children's utterances in the training set, resulting in 1,588, 1,588, and 1,643 extra utterances, respectively. Consequently, the total number of training utterances of ASD and TD children is 11,974 and 6,473, respectively. The total number of test utterances of ASD and TD children is 2,993 and 414, respectively, without any augmented utterances to avoid any potential biases that may arise from learning characteristics specific to the augmentation. There are no overlapping speakers in the training and test datasets.

The KDSFs features are extracted from each utterance. In order to be compared with the effectiveness of the KDSFs, the eGeMAPS features (version 2.1) (Eyben et al., 2016) are extracted as the baseline feature set. SVMs with the radial basis function (RBF) kernel are trained for classification of children with ASD and TD children, using the Scikit-learn toolkit (Pedregosa et al., 2011) in Python. The optimal parameters are obtained by implementing grid search with stratified 5-fold cross-validation. For both C and γ, the search space is [0.001, 0.01, 0.1, 1, 10, 25, 50, 100]. An SVM classifier is trained with the KDSF set and its subsets with spectral features to compare the effectiveness of each prosodic feature. A baseline classifier is trained with the entire eGeMAPS features. To identify useful features as well as reduce the number of features, a feature selection algorithm, recursive feature elimination (RFE) is additionally implemented to train both feature sets. An SVM with the linear kernel is used as an estimator during the feature selection process. Ten CPUs (Intel(R) Xeon(R) Gold 5220R CPU @ 2.20 GHz) are utilized for training each classifier.

4.2. Metrics

The metric for the performance is accuracy which considers both results for ASD and TD. True positive (TP) is the case a true ASD utterance is classified as ASD, and true negative (TN) is when a true TD utterance is classified as TD. In contrast, a false positive (FP) is a true TD utterance that is classified as ASD, and a false negative (FN) is a true ASD utterance that is classified as TD. Precision, recall, and F1-score from both groups of children are also calculated in addition to accuracy.

Accuracy, precision, and recall are calculated as in equations (1), (2), (3), respectively. F1-score is calculated as the harmonic mean of precision and recall. The precision, recall, and F1-score from the perspective of TD children can be calculated by switching P and N in the equations.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$\mathrm{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

4.3. Results

The classifier trained with all the KDSFs (C=10, γ=0.001) outperformed the baseline classifier trained with the entire eGeMAPS feature set (C=10, γ=0.001) in accuracy on the test set. KDSFs achieve 91.25% of accuracy, while the entire eGeMAPS feature set achieves 88.08% of accuracy. The proposed feature set also achieves better F1-score (95.18%) and precision (92.23%) for ASD, and better F1-score (52.70%) and recall (40.10%) for TD than the eGeMAPS set. The results are described in Table 1. It takes 1 hour and 17 minutes and 3 hours and 36 minutes to train each classifier.

When RFE is implemented for feature selection, KDSFs also shows better performance than eGeMAPS on the test set in

accuracy, precision and F1-score for ASD, and recall and F1-score for TD, which is the same pattern as the full feature set. For KDSFs, 19 out of 29 features are selected after RFE, and the accuracy reaches 90.78% on the test set (C=10, γ=0.001). The F1-score slightly declines for children with ASD (94.77%), but improves for TD children (61.43%), compared to the full KDSF set. The 19 selected features are related to voice quality, speech rate, and cepstral values. Specific features are listed below:

1) Voice quality (3): jitter, shimmer, HNR
2) Speech rate (8): total duration, pause duration, speaking duration, speaking rate, articulation rate, average syllable duration, the number of pauses, and ratio of speech duration and total duration
3) MFCCs (7): the 1st to 11th MFCCs except for 5th, 7th, and 8th

Out of 88 eGeMAPS features, 70 are selected for training the classifier, which shows 88.29% of accuracy on the test set (C=25, γ =0.001). F1-score for ASD children is similar to the entire feature set (93.72%), while F1-score for TD children improves to 13.82%. Specific features are as below: (μ for mean; SDN for normalized SD; V for voiced segments; UV for unvoiced segments):

1) Frequency (17): F0 (μ, SDN), percentile range of 0–2, values of 20th/50th/80th percentiles, frequency of the first formant (F1), the second formant (F2), and the third formant (F3) (μ, SDN), bandwidth of F1/F3 (μ, SDN), bandwidth of F2 (μ)
2) Amplitude (10): loudness (μ, SDN), percentile range 0–2, values of 20th/50th/80th percentiles, falling and rising slope (μ, SD)
3) Voice quality (6): jitter (μ, SDN), shimmer (μ, SDN), HNR (μ, SDN)
4) Speech rate (6): length of UV (μ, SD), length of V (μ, SD), V per second (i.e. pseudo syllable rate), the number of loudness peaks per second
5) Spectrum (30): the ratio of energy of the spectral harmonic peak at F1/F2/F3 center frequency to the energy of the spectral peak at F0 (μ, SDN), alpha ratio of UV (μ), alpha ratio of V (μ, SDN), Hammarberg Index of UV (μ), Hammarberg Index of V (SDN), harmonic difference between H1 and H2/H1 and H3 (μ), 1st/2nd/3rd/4th MFCC (μ), 1st/2nd/3rd/4th MFCC of V (μ), spectral slope of 0–500 Hz/500–1,500 Hz of V/UV (μ), spectral flux of UV (μ), spectral flux of V/all segments (μ, SDN)
6) Equivalent sound level

The training time spent for KDSFs and eGeMAPS features including feature selection procedure is 16 hours and 54 minutes, and 27 hours and 41 minutes, respectively.

In the comparison of four sets of features, namely KDSFs with or without feature selection, and eGeMAPS with or without feature, the model trained with all KDSFs demonstrates the highest accuracy. Furthermore, this model exhibits the highest F1-score for children with ASD. The classifier trained with selected KDSFs by RFE achieves the best F1-score regarding TD children. In terms of ASD detection, the highest precision is obtained with the selected KDSFs by RFE, while the highest recall is from all the eGeMAPS features. In contrast, for TD children, the highest precision is attained with the entire eGeMAPS features, while the best recall is obtained with the selected KDSFs by RFE.

Performances of classifiers trained with subsets of KDSFs are also depicted in Table 1, showing the classifier with speech rate features with MFCCs achieves the best in accuracy. Regarding recall, pitch features with MFCCs are better than others for detecting ASD children, while voice quality features with MFCCs are best for detection of TD children. The time for training the classifiers with each subset is between 1.5 and 2 hours.

5. Discussion

The classifiers trained with KDSFs outperform those trained with the predefined feature set in accuracy and F1-score for both ASD and TD, regardless of applying feature selection. Furthermore, KDSFs exhibit higher efficiency as training time for KDSFs is halved compared to that of eGeMAPS features. When using all features as input, the classifier with all KDSFs shows the highest accuracy, while a classifier with eGeMAPS features exhibits the lowest. Moreover, the best F1-score from the perspective of ASD and TD is observed from all KDSFs and selected KDSFs, respectively. Better performance of KDSFs is not surprising as they are designed to represent speech characteristics of ASD children's speech. The lowest performance of the entire eGeMAPS feature set could be explained by the large number of features. Too many features could deteriorate the performance by making the model overfit given data. The issue would become more challenging with a small and imbalanced dataset, which is often the case for disordered speech. Furthermore, some features would be unrelated to the traits of ASD children's speech. Therefore, the results imply the importance of reflecting domain knowledge into feature engineering.

Recall is one of the important metrics for the detection task. FNs should be minimized as missing FNs could lead to wrong diagnoses, which potentially impacts the outcomes of children with ASD. Regarding the significance of recall, the results would mislead that eGeMAPS is more effective than the proposed KDSFs as classifiers trained with eGeMAPS features exhibit higher recall for detecting children with ASD. However, in contrast to the high recall for ASD, recall for TD is observed extremely low. It is suspected that the classifiers with eGeMAPS features are overfitted for ASD utterances. It is also supported by the feature selection analysis. After reducing the number of features, the recall for TD utterances is improved.

The aforementioned analysis of recall underscores the importance of addressing the gap of performances in recall as well as F1-score between ASD and TD children. A great discrepancy denotes a biased classification result. When a dataset is limited in size and imbalanced, the classification result is likely to lean towards a class with more data points. Even though data augmentation of TD utterances is implemented to mitigate the problem, there remains still a tendency for bias. The results of KDSFs exhibit some differences between ASD and TD in recall and F1-score. However, when using the eGeMAPS features, a substantial number of TD utterances are incorrectly classified as ASD, resulting in much larger disparity in recall and F1-score between ASD and TD. It would lead to misguided conclusions without careful analysis. It is speculated that the classifier could have been negatively influenced during the training process by the relatively larger feature set compared to the dataset size.

In general, feature selection or feature extraction is conducted to reduce the feature dimension to address the problem caused by a

number of features. The experimental results indicate that the eGeMAPS feature set slightly benefits from feature selection in accuracy and all F1-scores, while KDSFs benefit in TD children's recall and F1-score. The reduced eGeMAPS set includes features from every domain (i.e. frequency, amplitude, voice quality, time, spectrum, and equivalent sound level). However, the reduced set from KDSFs includes voice quality, speech rate, and spectral features, excluding pitch-related features. Both reduced sets share jitter, shimmer, HNR, speech rate-related features, and MFCC 1st to 4th in common.

The shared features also correspond to speech characteristics of children with ASD in the dataset. For ASD children, significant higher values than TD children are observed in jitter (t=8.32, p<0.001), shimmer (t=5.21, p<0.001), total duration of utterance (t=2.23, p=0.03), the ratio of speech duration to total duration (t=2.57, p=0.01), speaking rate (t=5.54, p<0.001), and articulation rate (t=5.38, p<0.001), while lower values in HNR (t=–12.10, p<0.001) and average syllable duration (t=–4.12, p<0.001). Speech duration (t=1.87, p=0.06) and pause duration (t=1.95, p=0.05) exhibit no significant difference, but tend to be higher in ASD utterances. As in studies on speech characteristics of ASD children, the speech utterances in the dataset are also distinct from TD children in jitter and HNR (Bone et al., 2012), speech rate (Bone et al., 2012), and duration of syllable and utterance (Fusaroli et al., 2017).

There are commonly selected eGeMAPS features between this work and Beccaria et al. (2022). Features related to voice quality, such as jitter, shimmer, and HNR are common not only in the eGeMAPS feature set but also in the KDSF set. In the frequency domain, F2 frequency and bandwidth are common. Loudness features include mean loudness, the value of the 20th percentile, and the SD of the rising slope. Among spectral features, the mean spectral flux of entire segments and voiced segments, and the mean slope of unvoiced segments of 0–500 Hz and 500–1,500 Hz are selected to train a classifier. There is nothing in common with Cho et al. (2019) because selected features in Cho et al. (2019) are mostly delta values and functionals that are not calculated in the eGeMAPS set.

Among the classifiers trained with each prosodic feature domain in conjunction with cepstral features, speech rate-related feature set performs best in terms of accuracy and F1-score of ASD and TD children. Meanwhile, the best precision and recall are from the pitch or voice quality domain for both ASD and TD. This indicates that all of the feature domain attribute to the improvement of KDSFs.

6. Conclusion

For the detection of ASD speech, KDSFs are utilized to reflect the prosodic characteristics of ASD. SVM classifiers are trained with KDSFs and eGeMAPS features from a speech database of Korean-speaking ASD and TD children. The highest performance is achieved by the KDSFs. This result indicates the importance of domain-specific knowledge to develop a model for disorder detection.

One of the limitations is the impossibility of controlling the age of speakers due to the absence of metadata in the speech database. It is planned to be appended, so we would conduct the experiments considering children's chronological age in the short future. Future works should consider more features as well. For example, features related to long pauses and loudness variability during a session could be included for the pragmatic aspects. Text features extracted from the transcript and annotations would also be useful. With the success of the KDSFs, the experiment would be expanded to classification of severity of ASD.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Association.

Asgari, M., Chen, L., & Fombonne, E. (2021). Quantifying voice characteristics for detecting autism. *Frontiers in Psychology*, *12*, 665096.

Barche, P., Gurugubelli, K., & Vuppala, A. K. (2020, October). Towards automatic assessment of voice disorders: A clinical approach. *Proceedings of the INTERSPEECH 2020* (pp. 2537-2541). Shanghai, China.

Beccaria, F., Gagliardi, G., & Kokkinakis, D. (2022, June). Extraction and Classification of Acoustic Features from Italian Speaking Children with Autism Spectrum Disorders. *Proceedings of the 2022 RaPID Workshop: Resources and ProcessIng of Linguistic, Para-Linguistic and Extralinguistic Data from People with Various Forms of Cognitive/Psychiatric/Developmental Impairments-within the 13th Language Resources and Evaluation Conference* (pp. 22-30). Marseille, France.

Benba, A., Jilbab, A., Hammouch, A., & Sandabad, S. (2015, March). Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease. *Proceedings of the 2015 International Conference on Electrical and Information Technologies (ICEIT)* (pp. 300-304). Marrakech, Morocco.

Bone, D., Black, M. P., Lee, C. C., Williams, M. E., Levitt, P., Lee, S., & Narayanan, S. (2012, September). Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist. *Proceedings of the INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association* (pp. 1043-1046). Portland, OR.

Bonneh, Y. S., Levanon, Y., Dean-Pardo, O., Lossos, L., & Adini, Y. (2011). Abnormal speech spectrum and increased pitch variability in young autistic children. *Frontiers in Human Neuroscience*, *4*, 237.

Cho, S., Liberman, M., Ryant, N., Cola, M., Schultz, R. T., & Parish-Morris, J. (2019, September). Automatic detection of autism spectrum disorder in children using acoustic and text features from brief natural conversations. *Proceedings of the INTERSPEECH 2019* (pp. 2513-2517). Graz, Austria.

Diehl, J. J., & Paul, R. (2012). Acoustic differences in the imitation of prosodic patterns in children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, *6*(1), 123-134.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y.., ... Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2), 190-202.

Eyben, F., Wöllmer, M., & Schuller, B. (2010, October). Opensmile: The Munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on Multimedia* (pp. 1459-1462). New York, NY.

Fusaroli, R., Lambrechts, A., Bang, D., Bowler, D. M., & Gaigg, S. B.

(2017). Is voice a marker for autism spectrum disorder? A systematic review and meta-analysis. *Autism Research*, *10*(3), 384-407.

Haider, F., de la Fuente, S., & Luz, S. (2020). An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, *14*(2), 272-281.

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1-15.

Kim, M. J., Pae, S., & Park, C. I. (2007). *Assessment of phonology and articulation for children (APAC)*. Incheon, Korea: Human Brain Research & Counseling.

Lee, J. H., Lee, G. W., Bong, G., Yoo, H. J., & Kim, H. K. (2023). End-to-end model-based detection of infants with autism spectrum disorder using a pretrained model. *Sensors*, *23*(1), 202.

Lord, C., Rutter, M., Luyster, R. J., & Gotham, K. (2012). *Autism diagnostic observation schedule, 2nd edition (ADOS-2)*. Los Angeles, CA: Western Psychological Corporation.

Lyakso, E., Frolova, O., & Grigorev, A. (2017, September). Perception and acoustic features of speech of children with autism spectrum disorders. *Proceedings of the 19th International Conference on Speech and Computer* (pp. 602-612). Hatfield, UK.

MacFarlane, H., Salem, A. C., Chen, L., Asgari, M., & Fombonne, E. (2022). Combining voice and language features improves automated autism detection. *Autism Research*, *15*(7), 1288-1300.

McCann, J., & Peppé, S. (2003). Prosody in autism spectrum disorders: A critical review. *International Journal of Language & Communication Disorders*, *38*(4), 325-350.

McFee, B., Metsai, A., McVicar, M., Balke, S., Thome´, C., Raffel, C., ... Kim, T. (2022). Librosa (Librosa version 0.9.2) [Python toolkit]. Retrieved from https://doi.org/10.5281/zenodo.6759641

Mohanta, A., & Mittal, V. K. (2022). Analysis and classification of speech sounds of children with autism spectrum disorder using acoustic features. *Computer Speech & Language*, *72*, 101287.

Mohanta, A., Mukherjee, P., & Mirtal, V. K. (2020, February). Acoustic features characterization of autism speech for automated detection and classification. *Proceedings of the 2020 National Conference on Communications (NCC)* (pp. 1-6). Kharagpur, India.

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019, September). SpecAugment: A simple data augmentation method for automatic speech recognition. *Proceedings of the INTERSPEECH 2019* (pp. 2613- 2617). Graz, Austria.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(2011), 2825-2830.

Yeo, E. J., Kim, S., & Chung, M. (2021, August). Automatic severity classification of Korean dysarthric speech using phoneme-level pronunciation features. *Proceedings of the INTERSPEECH 2021* (pp. 4838-4842). Brno, Czechia.

• **Seonwoo Lee**
Ph. D. Candidate, Dept. of Linguistics
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-9039
Email: lsw5220@snu.ac.kr
Fields of interest: Speech technology for people with disorders, Spoken language processing, Children with hearing loss

• **Eun Jung Yeo**
Ph. D. Candidate, Dept. of Linguistics
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-9039
Email: ej.yeo@snu.ac.kr
Fields of interest: Technology for disordered speech, Spoken language processing, Deep learning

• **Sunhee Kim**
Professor, Dept. of French Language Education
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-7693
Email: sunhkim@snu.ac.kr
Fields of interest: French phonetics, Spoken language processing

• **Minhwa Chung,** Corresponding author
Professor, Dept. of Linguistics
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea
Tel: +82-2-880-9195
Email: mchung@snu.ac.kr
Fields of interest: ASR, Spoken language processing, Computer-assisted language learning