INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A Study on Using Duration and Formant Features in Automatic Detection of Speech Sound Disorder in Children

Si-Ioi Ng[1], *Cymie Wing-Yee Ng*[2], *Tan Lee*[1]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong
[2]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University

siioing@link.cuhk.edu.hk, cymie.ng@polyu.edu.hk, tanlee@ee.cuhk.edu.hk

Abstract

Speech sound disorder (SSD) in children is manifested by persistent articulation and phonological errors on specific phonemes of a language. Automatic SSD detection can be done using features extracted from deep neural network models. Interpretability of such learned features is a major concern. Motivated by clinical knowledge, the use of duration and formant features for SSD detection is investigated in this research. Acoustical analysis is performed to identify the acoustic features that differentiate between the speech of typical and disordered children. On the task of SSD detection in Cantonese-speaking children, the duration features are found to outperform the formant feature, and surpass previous methods that use paralinguistic feature set and speaker embeddings. Specifically, the duration features achieve a mean unweighted average recall of 71.0%. The results enhance the understanding of SSD, and motivate further use of temporal information of child speech in SSD detection.

Index Terms: child speech, speech sound disorder, detection of pathological speech, vowel duration, formant frequency.

1. Introduction

Speech sound disorder (SSD) refers to one of the most common developmental disorders in which children encounter persistent difficulties in correctly pronouncing certain speech sounds after the expected age of acquisition. The most common symptom of SSD is manifested as substitution, omission, distortion, and insertion of speech sound(s) in a word. Population studies report that the prevalence of SSD ranges from 2% to 25% in children aged below 7 [1–3]. Timely diagnosis of SSD is crucial to effective treatment and rehabilitation. Clinical assessment of SSD is carried out by qualified speech-language pathologists (SLP) [4]. The assessment aims to analyze the child's phonetic inventory and patterns of speech sound errors that are likely related to SSD. Given the scant resource of speech-language pathology services in public [5, 6], automated detection tools for SSD are desired to alleviate the burden on SLPs and benefit a large population of children at risk.

Automatic detection of SSD is the task of distinguishing abnormal speech sound production from typical one based on acoustic speech signals. Acoustical characteristics of disordered speech are routinely captured by fixed-dimensional embeddings derived from deep neural network (DNN) models. In [7–9], Siamese neural network was trained to derive phone embeddings from monophone/diphone segments to detect speech sound errors in disordered child speech. DNN trained for speaker verification [10], or pre-trained by self-supervised learning [11], was applied to derive the speaker embeddings that reflect the overall goodness of speech sound production. However, interpreting the acoustic information encapsulated in DNN embeddings is a difficult task. This limitation hinders the practical use of these systems in clinical applications.

Features designed with domain knowledge that describe the speech characteristics of disordered speech are alternative options to DNN embeddings. The hand-crafted features are interpretable, of which the extraction does not require DNN, and is computationally efficient. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), as described in [12], was shown to be efficacious in various tasks related to speech disorders [13, 14]. To further improve detection performance and enhance understanding of the disorder, attempts have been made to design novel features that describe specific acoustic characteristics of disordered speech. In [15], the use of features concerning the segmental distortion, loudness, and hyper-nasality in the speech was shown to outperform eGeMAPS in distinguishing between Apraxia and Dysarthria. In [16], features related to utterance duration, speaking rate, pitch, and intensity of speech were used to determine the empathy level in conversational speech.

Speech sound acquisition involves different abilities, including phonological knowledge, speech perception, motor learning, cognitive- and meta-awareness, etc. Difficulties in any of these abilities can lead to multiple effects on the produced speech, e.g. atypical distribution of speech segment duration and formant frequencies as compared to typically-developing (TD) speech [17, 18]. The present study investigates the use of duration feature and formant feature in automatic detection of SSD in children. The duration features include the duration of words and duration of long and short vowels. The formant feature include the first, second, and third formant frequencies of the long vowels. Acoustical analysis is carried out with a set of words selected for articulation test. The duration and formant features are analyzed to reveal the developmental differences in speech characteristics in TD and disordered speakers. Statistical tests, i.e. *z-test*, are performed to determine the parameters that reflect the discrepancy between the TD and disordered speech. In SSD detection, speaker-level feature vectors are constructed from duration and formant features derived from word utterances. We will compare the feature vectors with the paralinguistic feature and the speaker embeddings as applied in existing detection approaches.

2. Background

2.1. About Cantonese

The present study concerns SSD in children who speak a specific language, namely Cantonese as spoken in Hong Kong. Cantonese is a major Chinese dialect widely spoken by mil-

10.21437/Interspeech.2023-937

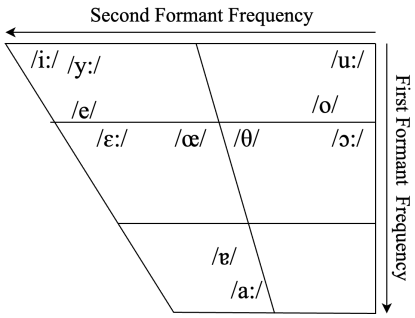

Figure 1: *Vowels in Cantonese [19].*

Table 1: *Partition of Speakers in experiments.*

Type	Partitions	Age			
		3	4	5	6
TD	Train	20	141	120	47
	Test	27	30	37	6
	Total	**47**	**171**	**157**	**53**
SSD	Train	6	25	11	8
	Test	23	37	31	9
	Total	**29**	**62**	**42**	**17**

lions of people. It is a monosyllabic and tonal language. Each Chinese character is pronounced as a single syllable carrying a lexical tone. Adopting the Cantonese inventory proposed by Bauer and Benedict, there are 19 initial consonants, 11 vowels and 6 lexical tones in Cantonese [19]. Vowel segments are involved in acoustical analysis and SSD detection in this work. The vowel inventory of Cantonese is illustrated in Figure 1.

2.2. Speech Database: CUCHILD

Experiments on speaker-level SSD detection are carried out with a large-scale child speech database named CUCHILD [20]. The database contains speech data from $1,986$ children aged 3 - 6 in Hong Kong local kindergartens. All speakers use Cantonese as their first language for daily communication. The speech materials consist of 130 Cantonese words of 1 to 4 syllables in length. These words cover Cantonese consonants and vowels of Cantonese. About 230 children in the database were found to have SSD by comparing their phonetic inventory, scores obtained from the Hong Kong Cantonese Articulation Test (HKCAT), and the patterns of speech sound errors [21]. Speech sound errors made by the SSD speakers were carefully annotated by four SLP trainees.

In this study, we use the speech data from 428 typically developing (TD) children and 150 disordered. The speech from 328 TD speakers is used to train the acoustic model of child speech, which is used for obtaining time alignments of word/vowel segments. Speech from 50 disordered speakers are involved in the training of SSD detection system. The remaining 100 TD and 100 SSD speakers are used for acoustical analysis, as well as performance evaluation. The age distributions of speakers in the training and test data are given in Table 1.

3. Proposed System and Feature Design

In standardized SSD assessment of young children, a set of designated test words is used for all subjects. The test words are selected purposely based on linguistic and clinical knowledge [21, 22]. The proposed system for SSD detection is depicted in Figure 2. For each speaker, multiple acoustical parameters are derived from the test words. These parameters include the duration of words of different syllable counts, the duration of the three short vowels /ɐ ɐ ɔ/, as well as the duration and the formant frequencies of the five long vowels /aː iː ɛː ɔː uː/. Production of words requires various degrees of coordination between the major articulators, while the production of vowels reveals different oral cavity configurations in vowel production. The above feature parameters are extracted to characterize the maturity of speech production in the child subject. The means and standard deviations of all acoustic parameters, and the age of the speaker, are used to construct a speaker-level feature vector. A support vector machine (SVM) is trained on the feature vectors to determine if the speaker is TD or disordered.

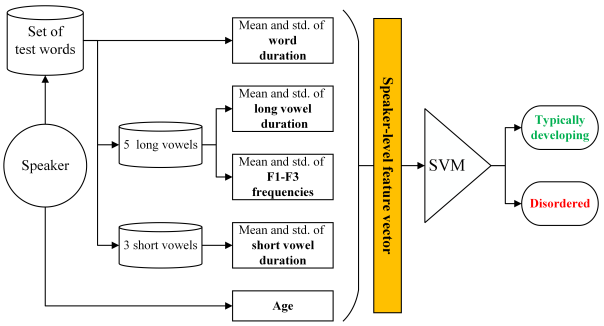

Figure 2: *Speaker-level SSD detection system.*

4. Acoustical Analysis

In this section, we analyze the acoustic parameters, i.e. duration and formants in TD and disordered children. For each parameter, a statistical test (*z-test*) is performed between TD and disordered speakers at the same age. The p-value is adjusted by the Benjamini-Hochberg procedure across age [23]. Disordered children typically delay motor skill development. They would be less skillful in executing the required steps of speech sound production. Therefore, we hypothesize that vowel and word segments in disordered speech are longer than in TD speech, and that the mean values of F1 to F3 in each type of vowel have discrepancies between TD and disordered speech. A significance threshold of 0.05 is used. The effect size is measured by Cohen's d.

4.1. Data preparation

To locate the word and vowel segments, forced alignment is applied with a Gaussian Mixture Model - Hidden Markov Model (GMM-HMM) based acoustic model. The acoustic model is trained with Mel-frequency cepstral coefficients (MFCC) extracted every 10 ms with a 25 ms Hamming window. Formant frequencies are extracted using the Praat software by linear predictive analysis with Burg's algorithm [24–27]. Children typically have higher formant frequencies than adults due to the short vocal tract [28]. Depending on the height and frontness of the vowels, the ceiling values of formant frequencies are empirically set to 8000 Hz for vowel /iː/, 7500 Hz for vowels /ɛː/, 7000 Hz for vowels /aː/, and 4500 Hz for /ɔː uː/. Five formant values are measured at each time frame. The formant contours are smoothed using a 3-point median filter. The medians of smoothed frequency contours are computed [29]. To exclude possible outliers from acoustical analysis, we retain the vowel duration, word duration, and formant frequencies that fall within the range of 5th-95th percentile across all measurements.

Table 2: *Duration of the five Cantonese long vowels, measured in milliseconds (* p<0.05, ** p<0.01, *** p<0.001).*

Age	/aː/			/iː/			/uː/			/ɛː/			/ɔː/		
	TD	SSD	d	TD	SSD	d	TD	SSD	d	TD	SSD	d	TD	SSD	d
3	$312_{\pm 10}$	$341_{\pm 10}$	0.297^{***}	$170_{\pm 8}$	$192_{\pm 9}$	0.255^{***}	$132_{\pm 6}$	$147_{\pm 7}$	0.245^{***}	$274_{\pm 7}$	$320_{\pm 8}$	0.627^{***}	$252_{\pm 10}$	$264_{\pm 10}$	0.131
4	$317_{\pm 9}$	$334_{\pm 10}$	0.173^{*}	$181_{\pm 8}$	$190_{\pm 9}$	0.118^{**}	$152_{\pm 6}$	$152_{\pm 6}$	0.002	$291_{\pm 8}$	$315_{\pm 8}$	0.307^{**}	$237_{\pm 9}$	$266_{\pm 10}$	0.301^{*}
5	$303_{\pm 10}$	$340_{\pm 10}$	0.383^{***}	$170_{\pm 8}$	$184_{\pm 8}$	0.164^{***}	$141_{\pm 6}$	$143_{\pm 6}$	0.042	$283_{\pm 8}$	$308_{\pm 8}$	0.301^{**}	$233_{\pm 9}$	$267_{\pm 10}$	0.368^{***}
6	$292_{\pm 8}$	$299_{\pm 9}$	0.085	$166_{\pm 8}$	$163_{\pm 8}$	0.045	137 ± 5	$138_{\pm 6}$	0.019	$273_{\pm 6}$	$270_{\pm 7}$	0.036	$226_{\pm 7}$	$233_{\pm 7}$	0.146

Table 3: *Duration of the three Cantonese short vowels, measured in milliseconds.*

Age	Short vowel								
	/ɐ/			/e/			/o/		
	TD	SSD	d	TD	SSD	d	TD	SSD	d
3	$126_{\pm 5}$	$145_{\pm 6}$	0.351^{***}	$197_{\pm 8}$	$210_{\pm 8}$	0.162	176.+-8	$195_{\pm 9}$	0.231^{*}
4	$126_{\pm 5}$	$136_{\pm 5}$	0.211^{***}	$195_{\pm 7}$	$202_{\pm 7}$	0.089	$175_{\pm 7}$	$176_{\pm 7}$	0.017
5	$120_{\pm 5}$	$137_{\pm 5}$	0.362^{***}	$186_{\pm 7}$	$204_{\pm 8}$	0.242^{**}	$173_{\pm 8}$	$192_{\pm 8}$	0.249^{**}
6	$109_{\pm 4}$	$118_{\pm 4}$	0.216^{**}	$173_{\pm 6}$	$179_{\pm 6}$	0.103	$159_{\pm 7}$	$151_{\pm 5}$	0.146

Table 4: *Duration of test words of different syllable counts, measured in milliseconds.*

			Age			
			3	4	5	6
Syllable Count	1	TD	$502_{\pm 140}$	$528_{\pm 139}$	$507_{\pm 136}$	$475_{\pm 118}$
		SSD	$532_{\pm 151}$	$519_{\pm 138}$	$539_{\pm 150}$	$510_{\pm 170}$
		d	0.206^{***}	0.066	0.221^{***}	0.234
	2	TD	$749_{\pm 167}$	$760_{\pm 164}$	$724_{\pm 166}$	$685_{\pm 127}$
		SSD	$801_{\pm 187}$	$782_{\pm 172}$	$790_{\pm 177}$	$746_{\pm 196}$
		d	0.296^{*}	0.129^{*}	0.384^{***}	0.355^{*}
	3	TD	$1029_{\pm 205}$	$1019_{\pm 202}$	$974_{\pm 203}$	$928_{\pm 157}$
		SSD	$1046_{\pm 225}$	$1072_{\pm 210}$	$1054_{\pm 219}$	$995_{\pm 185}$
		d	0.076	0.257	0.377^{***}	0.386
	4	TD	$1399_{\pm 286}$	$1444_{\pm 296}$	$1342_{\pm 369}$	$1189_{\pm 180}$
		SSD	$1400_{\pm 307}$	$1457_{\pm 319}$	$1450_{\pm 274}$	$1274_{\pm 195}$
		d	0.002	0.040	0.323^{**}	0.444^{*}

4.2. Results of duration analysis

Table 2 and Table 3 report the statistics of vowel duration in speakers aged from 3 to 6. Vowel duration decreases in long vowels /aː ɔː/ and all short vowels as age increases. In long vowels /iː uː ɛː/, the decrease in vowel duration is observed in disordered speakers. This suggests a developmental difference in vowel production between TD and disordered speakers. From age 3 to 5, disordered speakers tend to have longer vowel duration than TD speakers, as evidenced by the statistical significance and the effect size. In speakers aged 6, the difference in vowel duration between TD and disordered speakers vanishes, except for the short vowel /ɐ/.

Table 4 reports the statistics of word duration. For all speakers, the word duration decreases as age increases. In tri-syllable and quad-syllable words, there is no statistical significance in word duration between TD and disordered speakers aged 3 and 4, whilst statistical significance is found in speakers aged 5 and 6. Between ages 3 and 6, the duration of mono-syllable and di-syllable words decreases by about 20 ms and 60 ms in TD and disordered speakers respectively. In tri-syllable words, the decrease in duration is about 100 ms in TD speakers and 60 ms in the disordered. In quad-syllable words, the decrease is about 200 ms in TD speakers and 130 ms in the disordered. The results suggest that disordered children could delay in mastering the motor skills to produce words of longer length.

4.3. Results of formant analysis

Tables 5, Table 6, and Table 7 show the statistics of the formant frequencies of the five long vowels. As age increases, F1 shows a decreasing trend in all vowels. In particular, the F1 values in TD and disordered speech at age 5 are significantly different in /aː ɛː ɔː/. A decrease in F2 is observed in /aː ɛː/, while F2 is steady in /iː uː ɔː/ as the age increases. The F2 frequency in TD and disordered children aged 3 is significantly different in /iː ɛː ɔː/. The differences are also observed in /iː uː ɛː ɔː/ in speakers aged 4, and in /aː iː ɔː/ in speakers aged 5. As age increases, the F3 frequency increases in back vowels /uː ɔː/, and decreases in /aː iː ɛː/. The F3 frequency in TD and disordered children aged 3 is significantly different in /ɛː ʊː ɔː/, and in /aː ɛː/ in speakers aged 6. The formant analysis suggests F1-F3 frequencies in different vowels could reflect the discrepancy between TD and disordered speakers.

5. Experiments on SSD Detection

To compose the speaker-level feature vector for SSD detection, vowel duration, word duration, and F1-F3 frequencies produce 16, 6, and 31 statistical measures respectively. Three types of feature vectors are experimented with to evaluate the efficacy of duration and formant frequencies for SSD detection. The first type contains the mean and standard deviation of F1-F3, i.e., only formant information. The second type contains the mean and standard deviation of word and vowel duration. Duration of quad-syllable words is not used in the experiment, as speech data of quad-syllable words in some of the speakers were contaminated by background noise, or could not be located during data post-processing. The third type of feature vector is the combination of duration and formant parameters. The speaker's age is appended to all three types of speaker vectors. As certain acoustic parameters could be less efficacious in reflecting the differences between TD and disordered speakers, recursive feature elimination (RFE) is applied to eliminate redundant parameters [30]. The number of parameters to be selected, k, is set empirically. The top-k most important parameters are determined based on a cross-validation classification experiment using speech data from the training set in Table 1.

The eGeMAPS is used [12] as the baseline feature. Using the OpenSmile toolkit [31], 42 low-level descriptors (LLDs) are computed to characterize the spectral characteristics of the speech signal. A 88-dimensional feature vector is derived from the LLDs by statistical functionals. Besides, speaker embeddings, namely the i-vector and the x-vector as extracted from speaker verification (SV) systems, are adopted as another type of baseline features in this study [10]. The i-vector and x-vector are used to capture the holistic characteristic of child speech. The i-vector is extracted by a Gaussian Mixture Model - Universal Background Model [32], and the x-vector is extracted by a time-delayed neural network (TDNN) [33]. Both i-vector and

Table 5: *F1 frequencies in the five Cantonese long vowels, represented in semitones.*

Age	Long vowel /aː/			/iː/			/uː/			/ɛː/			/ɔː/		
	TD	SSD	d	TD	SSD	d	TD	SSD	d	TD	SSD	d	TD	SSD	d
3	$42.8_{\pm 2.1}$	$42.3_{\pm 2.3}$	0.199^{*}	$30.7_{\pm 2.5}$	$30.6_{\pm 2.6}$	0.055	$30.6_{\pm 2.1}$	$30.7_{\pm 2.3}$	0.035	$37.2_{\pm 2.1}$	$37.3_{\pm 2.3}$	0.047	$34.0_{\pm 2.1}$	$34.1_{\pm 2.0}$	0.050
4	$42.4_{\pm 2.1}$	$42.3_{\pm 2.1}$	0.041	$30.2_{\pm 2.5}$	$30.0_{\pm 2.4}$	0.066	$30.5_{\pm 2.2}$	$29.9_{\pm 2.1}$	0.251^{***}	$36.2_{\pm 2.1}$	$36.7_{\pm 2.4}$	0.221	$34.6_{\pm 2.3}$	$34.2_{\pm 2.3}$	0.176^{*}
5	$41.1_{\pm 2.1}$	$42.6_{\pm 2.0}$	0.686^{***}	$29.8_{\pm 2.6}$	$29.9_{\pm 2.8}$	0.023	$30.3_{\pm 2.3}$	$30.3_{\pm 2.3}$	0.016	$35.7_{\pm 2.2}$	$36.9_{\pm 2.3}$	0.541^{***}	$34.2_{\pm 1.9}$	$35.0_{\pm 2.0}$	0.407^{***}
6	$40.6_{\pm 2.0}$	$40.9_{\pm 2.0}$	0.151	$29.1_{\pm 2.3}$	$29.1_{\pm 2.3}$	0.029	$29.8_{\pm 2.3}$	$29.9_{\pm 2.2}$	0.042	$35.3_{\pm 1.8}$	$34.8_{\pm 1.8}$	0.294	$33.7_{\pm 1.6}$	$34.1_{\pm 1.6}$	0.260

Table 6: *F2 frequencies in the five Cantonese long vowels, represented in semitones.*

Age	Long vowel /aː/			/iː/			/uː/			/ɛː/			/ɔː/		
	TD	SSD	d	TD	SSD	d	TD	SSD	d	TD	SSD	d	TD	SSD	d
3	$51.8_{\pm 1.9}$	$51.6_{\pm 1.7}$	0.120	$57.5_{\pm 2.4}$	$58.1_{\pm 2.2}$	0.236^{***}	$42.4_{\pm 2.7}$	$42.3_{\pm 2.6}$	0.026	$57.5_{\pm 1.3}$	$57.0_{\pm 1.5}$	0.325*	$41.9_{\pm 1.9}$	$42.5_{\pm 1.9}$	0.360^{***}
4	$51.6_{\pm 1.8}$	$51.5_{\pm 1.7}$	0.037	$57.5_{\pm 2.3}$	$57.9_{\pm 2.3}$	0.176^{***}	$42.4_{\pm 2.5}$	$41.9_{\pm 2.6}$	0.220^{***}	$57.1_{\pm 1.1}$	$56.7_{\pm 1.3}$	0.302^{*}	$42.5_{\pm 1.9}$	$42.2_{\pm 1.8}$	0.171^{*}
5	$51.2_{\pm 1.8}$	$51.6_{\pm 1.7}$	0.249^{**}	$57.7_{\pm 2.2}$	$58.0_{\pm 2.1}$	0.136^{***}	$42.3_{\pm 2.5}$	$42.2_{\pm 2.4}$	0.043	$56.7_{\pm 1.0}$	$56.7_{\pm 1.1}$	0.036	$42.5_{\pm 1.8}$	$42.8_{\pm 1.7}$	0.173^{*}
6	$50.8_{\pm 2.0}$	$50.3_{\pm 1.9}$	0.280	$57.8_{\pm 2.0}$	$57.6_{\pm 1.8}$	0.151	$42.5_{\pm 2.5}$	$42.3_{\pm 2.4}$	0.091	$56.5_{\pm 1.0}$	$56.4_{\pm 1.2}$	0.076	$42.2_{\pm 1.7}$	$42.4_{\pm 1.5}$	0.145

Table 7: *F3 frequencies in the five Cantonese long vowels, represented in semitones.*

Age	Long vowel /aː/			/iː/			/uː/			/ɛː/			/ɔː/		
	TD	SSD	d	TD	SSD	d	TD	SSD	d	TD	SSD	d	TD	SSD	d
3	$62.7_{\pm 1.6}$	$62.6_{\pm 1.6}$	0.069	$63.8_{\pm 1.3}$	$64.1_{\pm 1.3}$	0.187^{***}	$52.9_{\pm 1.8}$	$53.2_{\pm 1.8}$	0.168^{**}	$64.3_{\pm 1.3}$	$64.3_{\pm 1.2}$	0.008	$49.9_{\pm 2.7}$	$50.6_{\pm 2.5}$	0.267^{*}
4	$62.2_{\pm 1.4}$	$62.1_{\pm 1.5}$	0.105	$63.6_{\pm 1.1}$	$63.7_{\pm 1.1}$	0.088^{*}	$53.0_{\pm 1.6}$	$53.0_{\pm 1.7}$	0.022	$64.1_{\pm 1.1}$	$63.8_{\pm 1.1}$	0.285	$50.8_{\pm 2.9}$	$50.3_{\pm 2.6}$	0.161
5	$61.7_{\pm 1.5}$	$61.7_{\pm 1.4}$	0.007	$63.4_{\pm 1.0}$	$63.5_{\pm 1.1}$	0.088^{*}	$53.2_{\pm 1.6}$	$53.3_{\pm 1.6}$	0.038	$63.4_{\pm 1.0}$	$63.3_{\pm 1.1}$	0.136	$51.5_{\pm 2.8}$	$51.3_{\pm 2.8}$	0.066
6	$61.2_{\pm 1.2}$	$60.6_{\pm 1.1}$	0.487^{*}	$63.3_{\pm 0.9}$	$63.1_{\pm 1.0}$	0.246^{**}	$53.8_{\pm 1.3}$	$53.6_{\pm 1.5}$	0.087	$63.0_{\pm 0.8}$	$63.0_{\pm 0.8}$	0.002	$52.0_{\pm 2.6}$	$52.3_{\pm 2.5}$	0.133

x-vector extractors are trained on the MFCCs extracted from speech data in the training set. The dimensions of the i-vector and x-vector are 100 and 512 respectively, which are further reduced to 30 using principal component analysis (PCA). Both i-vector and x-vector systems are implemented by Kaldi [34].

Using scikit-learn [35], the SVM for speaker-level SSD detection is trained on the proposed feature vectors, paralinguistic features, or speaker embeddings. A linear kernel and a regularization parameter of 1.0 are used.

6. Results

5-fold cross-validation experiments on SSD detection are carried out with the test set as described in Table 1. The results of classification using the proposed speaker-level feature vectors, the paralinguistic features, and the speaker embeddings are given in Table 8, in terms of the unweighted average recall (UAR). Without applying RFE, the duration feature achieves $66.0 \pm 10.6\%$ UAR, and the formant feature achieves $57.5 \pm 7.9\%$ UAR. The duration feature outperforms eGeMAPS and x-vector. Combining formant feature and duration feature can help reduce the standard deviation. When RFE is applied, a performance gain is observed in both duration and formant features, and eGeMAPS. The duration feature, of which the dimension is reduced to 10 by RFE, achieves the best performance of $71.0 \pm 3.0\%$ UAR, whilst the eGeMAPS improves the performance to $69.5 \pm 6.6\%$ with the dimension being reduced from 88 to 70. The dimension-reduced duration feature also surpasses the i-vector approach, which has a UAR of $67.0 \pm 7.8\%$. Using RFE, the joint use of the formant and duration features achieves a 69.5 ± 8.6, which is still surpassed by the sole use of the duration feature. Overall, the vowel and word duration are shown to be promising features in the detection of SSD in children. For F1-F3 frequencies, despite significant differences in numerous vowels between TD and disordered speakers have been observed, the constructed formant feature is shown to be impotent to SSD detection.

Table 8: *SSD detection performance.*

Speaker-level Feature	Without RFE		With RFE	
	UAR	Dim.	UAR	Dim.
Duration+Age	$66.0_{\pm 10.6}$	23	$\mathbf{71.0_{\pm 3.0}}$	**10**
Formant+Age	$57.5_{\pm 7.9}$	31	$59.5_{\pm 2.9}$	20
Duration+Formant+Age	$66.0_{\pm 7.5}$	53	$69.5_{\pm 8.6}$	10
eGeMAPS	$61.0_{\pm 6.8}$	88	$69.5_{\pm 6.6}$	70
x-vector	$65.0_{\pm 8.5}$	256	-	-
i-vector	$67.0_{\pm 7.8}$	100	-	-

7. Conclusion

In this study, we demonstrate the use of small numbers of hand-crafted features in subject-level SSD detection. The design of hand-crafted features is inspired by clinical knowledge about child speech development. Acoustical analysis has shown that duration and formant frequencies of vowels, and duration of words, can reflect changes of speech characteristics in disordered speech. The cross-validation experiment of SSD detection has demonstrated that the use of duration feature is effective in SSD detection. It outperforms the formant feature, the conventional paralinguistic feature, and the speaker embeddings derived from state-of-the-art speaker verification systems. The present results serve to enhance our understanding of SSD, and motivate future designs of detection methods that utilize temporal information of child speech. Our future work will focus on erroneous consonant production that is difficult to be identified by SLPs, with the aim to further develop more novel and interpretable acoustic features for SSD detection.

8. Acknowledgement

This research is partially supported by a GRF project grant (Ref: CUHK 14208020) from Hong Kong Research Grants Council. It is also supported by a direct grant and a Research Sustainability Fund from the Research Committee of the Chinese University of Hong Kong, as well as a financial support by the Hear Talk Foundation under the project titled "Speech Analysis for Cantonese Speaking Children".

9. References

[1] J. Law, J. Boyle, F. Harris, A. Harkness, C. Nye *et al.*, "Prevalence and natural history of primary speech and language delay: findings from a systematic review of the literature," *International journal of language and communication disorders*, vol. 35, pp. 165–188, 2000.

[2] P. Eadie, A. Morgan, O. C. Ukoumunne, K. Ttofari Eecen, M. Wake, and S. Reilly, "Speech sound disorder at 4 years: Prevalence, comorbidities, and predictors in a community cohort of children," *Developmental Medicine & Child Neurology*, vol. 57, no. 6, pp. 578–584, 2015.

[3] Y. Wren, L. L. Miller, T. J. Peters, A. Emond, and S. Roulstone, "Prevalence and predictors of persistent speech sound disorder at eight years old: Findings from a population cohort study," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 4, pp. 647–673, 2016.

[4] S. Mcleod and E. Baker, "Speech-language pathologists' practices regarding assessment, analysis, target selection, intervention, and service delivery for children with speech sound disorders," *Clinical linguistics & phonetics*, vol. 28, no. 7-8, pp. 508–531, 2014.

[5] L. Ruggero, P. McCabe, K. J. Ballard, and N. Munro, "Paediatric speech-language pathology service delivery: An exploratory survey of australian parents," *International journal of speech-language pathology*, vol. 14, no. 4, pp. 338–350, 2012.

[6] N. McGill, S. McLeod, K. Crowe, C. Wang, and S. C. Hopf, "Waiting lists and prioritization of children for services: Speech-language pathologists' perspectives," *Journal of Communication Disorders*, vol. 91, p. 106099, 2021.

[7] J. Wang, Y. Qin, Z. Peng *et al.*, "Child Speech Disorder Detection with Siamese Recurrent Network Using Speech Attribute Features," in *Proc. Interspeech*, 2019, pp. 3885–3889.

[8] S.-I. Ng and T. Lee, "Automatic Detection of Phonological Errors in Child Speech Using Siamese Recurrent Autoencoder," in *Proc. Interspeech*, 2020, pp. 4476–4480.

[9] W. Y. Ng, "The application of speech recognition technology in detecting speech sound errors in cantonese-speaking children," Ph.D. dissertation, The Chinese University of Hong Kong (Hong Kong), 2021.

[10] S.-I. Ng, C. W.-Y. Ng, J. Wang *et al.*, "Automatic detection of speech sound disorder in child speech using posterior-based speaker representations," in *Proc. Interspeech*, 2022, pp. 2931–2935.

[11] Y. Getman, R. Al-Ghezi, K. Voskoboinik, T. Grósz, M. Kurimo, G. Salvi, T. Svendsen, and S. Strömbergsson, "wav2vec2-based Speech Rating System for Children with Speech Sound Disorder," in *Proc. Interspeech 2022*, 2022, pp. 3618–3622.

[12] F. Eyben, K. R. Scherer, B. W. Schuller *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[13] M. Shahin, U. Zafar, and B. Ahmed, "The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 400–412, 2019.

[14] W. Xue, C. Cucchiarini, R. van Hout *et al.*, "Acoustic correlates of speech intelligibility: the usability of the eGeMAPS feature set for atypical speech," in *Proc. SLaTE*, 2019, pp. 48–52.

[15] I. Kodrasi, M. Pernon, M. Laganaro, and H. Bourlard, "Automatic discrimination of apraxia of speech and dysarthria using a minimalistic set of handcrafted features." in *Proc. Interspeech*, 2020, pp. 4991–4995.

[16] D. Tao, T. Lee, H. Chui, and S. Luk, "Characterizing Therapist's Speaking Style in Relation to Empathy in Psychotherapy," in *Proc. Interspeech*, 2022, pp. 2003–2007.

[17] T. Macrae, M. P. Robb, and G. T. Gillon, "Acoustic analysis of word and segment duration in children with speech sound disorder," *Asia Pacific Journal of Speech, Language and Hearing*, vol. 13, no. 2, pp. 77–86, 2010.

[18] W. Zhang, X. Gui, T. Wang, M. Ng, F. Yang, L. Wang, and N. Yan, "Acoustic Features Associated with Sustained Vowel and Continuous Speech Productions by Chinese Children with Functional Articulation Disorders," in *Proc. Interspeech 2018*, 2018, pp. 1696–1700.

[19] R. S. Bauer and P. K. Benedict, *Modern cantonese phonology*. Walter de Gruyter, 2011, vol. 102.

[20] S.-I. Ng, C. W.-Y. Ng, J. Wang *et al.*, "CUCHILD: A Large-Scale Cantonese Corpus of Child Speech for Phonology and Articulation Assessment," in *Proc. Interspeech*, 2020, pp. 424–428.

[21] P. Cheung, A. Ng, and C. K. S. To, "Hong kong cantonese articulation test. hong kong: Language information, sciences & research centre," *City University of Hong Kong*, 2006.

[22] R. Goldman and M. Fristoe, "Goldman-fristoe test of articulation: Third edition," *Pearson*, 2015.

[23] Y. Benjamini, D. Drai, G. Elmer *et al.*, "Controlling the false discovery rate in behavior genetics research," *Behavioural brain research*, vol. 125, no. 1-2, pp. 279–284, 2001.

[24] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, vol. 17, no. 1193. Amsterdam, 1993, pp. 97–110.

[25] N. Andersen, "On the calculation of filter coefficients for maximum entropy spectral analysis," *Geophysics*, vol. 39, no. 1, pp. 69–72, 1974.

[26] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [computer program]. version 6.0. 37," *Retrieved February 2018 from http://www.praat.org/*.

[27] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[28] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999.

[29] S. Yildirim, S. Narayanan, D. Byrd, and S. Khurana, "Acoustic analysis of preschool children's speech," in *Proc. 15th ICPhS*, 2003, pp. 949–952.

[30] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, pp. 389–422, 2002.

[31] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[32] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[33] D. Snyder, D. Garcia-Romero, G. Sell *et al.*, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[34] D. Povey, A. Ghoshal, G. Boulianne *et al.*, "The kaldi speech recognition toolkit," in *Proc. ASRU*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.