



Automatic Detection of Speech Sound Disorder in Child Speech Using Posterior-based Speaker Representations

Si-Ioi Ng¹, Cymie Wing-Yee Ng², Jiarui Wang¹, Tan Lee¹

¹Department of Electronic Engineering, The Chinese University of Hong Kong

²Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University

siioing@link.cuhk.edu.hk, tanlee@ee.cuhk.edu.hk

Abstract

This paper presents a macroscopic approach to automatic detection of speech sound disorder (SSD) in child speech. Typically, SSD is manifested by persistent articulation and phonological errors on specific phonemes in the language. The disorder can be detected by focally analyzing the phonemes or the words elicited by the child subject. In the present study, instead of attempting to detect individual phone- and word-level errors, we propose to extract a subject-level representation from a long utterance that is constructed by concatenating multiple test words. The speaker verification approach, and posterior features generated by deep neural network models, are applied to derive various types of holistic representations. A linear classifier is trained to differentiate disordered speech in normal one. On the task of detecting SSD in Cantonese-speaking children, experimental results show that the proposed approach achieves improved detection performance over previous method that requires fusing phone-level detection results. Using articulatory posterior features to derive i-vectors from multiple-word utterances achieves an unweighted average recall of 78.2% and a macro F1 score of 78.0%.

Index Terms: child speech, speech sound disorder, speaker representation, articulatory feature, speech attributes.

1. Introduction

Speech sound disorder (SSD) is a common type of communication disorders affecting 2.5% - 13.8% of children aged below 8 [1–4]. When growing up, children are expected to master the language's speech sounds in stages and be able to self-correct the pronunciation mistakes. A significant portion of children may encounter persistent difficulties in correctly producing certain speech sounds after the expected stage of acquisition. The symptom is one diagnostic attribute of SSD in clinical speech therapy. Untreated children with SSD are prone to unsatisfactory social and educational outcomes [5, 6]. Early diagnosis and intervention are necessary for effective treatment and rehabilitation. At present, clinical diagnosis of SSD is carried out by qualified speech and language pathologists (SLP). The evaluation of the speech sound inventories of children reveals the presence of SSD.

To timely identify children with SSD and refer to SLP for intervention, automated detection of speech disorder is considered a highly desirable approach for assessment and/or screening a large population of children. Detection of SSD is commonly formulated as a task of distinguishing disordered speech sounds from typical ones at phoneme-level, based on acoustic speech signals. Siamese neural networks were adopted to contrast hypothetically disordered consonant segments with typical ones [7, 8]. Posterior features were derived from automatic

speech recognition systems to facilitate mispronunciation detection in disordered child speech [9, 10]. Considering that consonant error could alter the acoustical characteristics of its neighbouring vowel, our recent work proposed to detect consonant errors from consonant-vowel speech segments [11]. The phoneme-level error detection requires automatic time alignment of the target phoneme segments. Time alignment could be inaccurate for disordered speech recorded in naturalistic communication scenarios. Inaccurate segments would affect the efficacy of model training and the detection performance.

Taking a different perspective, detection of SSD can also be formulated as a problem of measuring the overall goodness of the child's speech production, without going into specific parts of a long utterance. The reliability concern of segment boundaries is thus bypassed. Paralinguistic features defined in the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) represent the overall acoustic characteristics of a speech utterance using statistical functionals [12]. They were successfully applied to detect atypical speech, including disordered speech with phonological and articulation problems [13]. The approach of speaker verification (SV) was also investigated in classifying disordered speech [14–17]. In [18, 19], SSD subjects were identified by a text-dependent SV system trained on single-word utterances. The word-level detection results were fused to give a subject-level score as an overall assessment of word articulation.

Spectral features, e.g. Mel-frequency cepstrum coefficients (MFCCs), have been routinely used to extract speaker representations. Posterior features derived from deep neural network (DNN) classifiers are closely related to the linguistic properties of speech [20, 21]. They particularly characterize the articulation aspect of speech sounds. Compared to conventional spectral features, posteriors are believed to convey more pertinent information that can help detect atypical speech sounds [7, 9, 10]. We expect that using speaker representations extracted from posterior features would effectively differentiate disordered children from typically developing (TD) ones.

In the present study, we investigate the possibility of transforming the word articulation test for SSD detection into a holistic assessment of multiple word utterances. Prescribed target words are grouped together, and the subject-level judgement is given based on the overall discrepancy between disordered speech and typical speech. The detection does not require precisely locating, analysing and classifying individual phones/words. Towards capturing the coarse-grained speech characteristics of long utterances, speaker representations such as i-vectors and x-vectors have been investigated [22, 23]. Using word-level utterances ensures that every single error in the pronunciation can be covered, and speech duration would not be too short for feature extraction [24]. In view of the limited amount of child speech data, we need to determine the

Table 1: *Speech attributes of Cantonese.*

Category	Feature	Phoneme
Manner	Plosive	p p ^h t t ^h k k ^h k ^w k ^{wh}
	Nasal	m n ŋ
	Affricate	ts ts ^h
	Fricative	s f h
	Glide	j w
Aspiration	Liquid	l
	Aspirated	p ^h t ^h k ^h k ^{wh} ts ^h
Place	Unaspirated	p t k k ^w ts
	Alveolar	t t ^h ts ts ^h s j
	Lateral	l
	Labial	p p ^h w m
	Velar	k k ^h ŋ
	Labio-Velar	k ^w k ^{wh}
	Labio-dental	f
	Vocal	h
Vowel/Semi-vowel		a: i: e: e: e: e: o: u: y: ɐ ɪ ɵ ʊ

type of speaker representation that can perform well in the low-resource scenario. We also investigate the use of knowledge-driven posteriors in the extraction of speaker representations. Grouping the phonemes based on speech attributes is adopted in the training of posterior extractors, as speech attributes were shown to provide diagnostic information about SSD symptoms [7, 10]. We will compare the proposed speaker representations for subject-level SSD detection with phone-level and word-level approaches.

2. Child Speech Database

This research is focused on SSD in Cantonese-speaking preschool children. Cantonese is a major Chinese dialect widely spoken by millions of people in Hong Kong, Macau, Guangdong Provinces of Mainland China, and many overseas Chinese communities. It is a monosyllabic and tonal language. Each Chinese character is pronounced as a single syllable carrying a lexical tone. There are 19 initial consonants, 11 vowels, 10 diphthongs, 6 final consonants and 6 lexical tones [25, 26].

Experiments on subject-level SSD detection are carried out with a large-scale child speech database named CUCHILD [27]. The database contains speech data from 1,986 kindergarten children aged 3 - 6 in Hong Kong. All speakers use Cantonese as their first language. The speech materials consist of 130 Cantonese words of 1 to 4 syllables in length. These words cover all Cantonese consonants and vowels. Each subject in CUCHILD was formally assessed with the Hong Kong Cantonese Articulation Test (HKCAT) [28]. About 230 children in the database were found to have SSD. Most of these children committed speech sound errors more frequently in the initials consonants than vowels.

We select speech data from 415 subjects in CUCHILD, including 265 typically developing (TD) children and 150 disordered. 48% of the TD and 61% of the disordered children are aged 3-4. Speech sound errors made by the disordered speakers are carefully annotated by four student clinicians from speech therapy programmes. Given the heterogeneous acoustic environment of speech recording, for some subjects, a few spoken word items could not be located or were contaminated by background noise. 125 ± 14 word utterances are available for each speaker. For each disordered speakers, 45 ± 25 word utterances are annotated as having speech sound errors. 5-fold cross-validation experiments are carried out. 80% of the TD and disordered speakers are used in model training, i.e. about 5.5 hours of training data in each fold. Speech data from the training speakers does not appear in the test set.

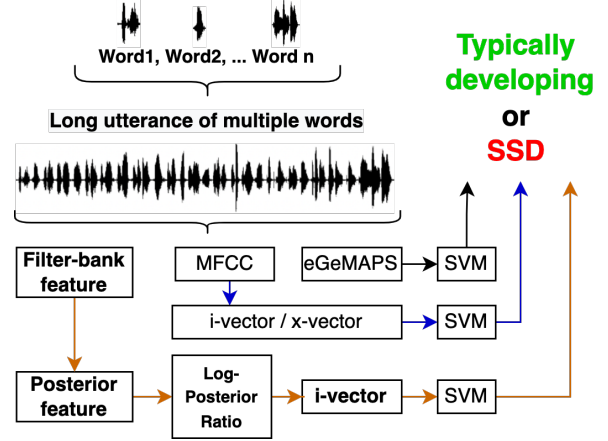


Figure 1: *Subject-level SSD detection system.*

3. Proposed System and Feature Design

3.1. System Architecture

In the standardized assessment of SSD for young children, a set of designated test words are used for all subjects. The test words are selected purposely based on linguistic and clinical knowledge to cover the target speech sounds (phonemes) to be evaluated [28, 29]. Our proposed system for subject-level SSD detection is shown as in Figure 1. For each child subject, speech segments that contain the test words are concatenated to construct a long utterance of multiple words. Front-end feature processing involves the computation of filter-bank features, MFCCs or eGeMAPS. I-vector or x-vector is derived from the MFCCs as a speaker-level feature representation. The i-vector can also be obtained from the posterior features generated by a DNN. A support vector machine (SVM) is trained on the i-vector, x-vector or eGeMAPS to determine if the subject is TD or disordered.

3.2. I-vector and x-vector

I-vector is a fixed-dimensional representation of variable-length utterance from a speaker. Consider an utterance i . It is modeled by a super-vector, defined as $\mu_i = m + Tw_i$, where m is the mean vector obtained from Gaussian Mixture Model - Universal Background Model (GMM-UBM). The GMM-UBM is the speaker-independent component trained on all speakers. T and w_i are the speaker-dependent components. T is the total variability matrix and w_i is the latent factor, known as the i-vector. Training of i-vector extractor is implemented by the expectation maximization (EM) algorithm.

x-vector is extracted from a DNN model [23]. A time-delay neural network (TDNN) architecture is adopted and trained with a cross-entropy objective function. Training of the TDNN requires splitting an utterance into segments. x-vectors of the segments are computed by aggregating the hidden representations of the TDNN via mean and variance pooling. They are averaged to derive the subject-level x-vectors.

I-vector and x-vector were applied to detecting speech-related diseases, e.g. dysarthria [14], Parkinson's disease [15], and oral cancer [16, 17]. It was common to train the extractors using out-of-domain corpus, e.g. from healthy adult speakers. In the present study, apart from using out-of-domain data, we attempt to use in-domain child speech data, which incorporates disordered speech in training.

3.3. Posterior feature for i-vector extraction

Consonant substitution is one of the major types of speech sound errors in child speech. The errors are usually described as specific categories of substitution patterns. For example, the Cantonese initial consonant /t/ being substituted by /k/ is known as the “backing” error. /ts^h/ substituted by /s/ is described as a “de-affrication” error. These errors correspond to the change of speech attributes, i.e., the manner and place of articulation, and/or the aspiration status. In the example of /t/→/k/, the place of articulation shifts from bilabial to alveolar, while the example of /ts^h/→/s/ is due to change in the manner of articulation from affricate to fricative. The grouping of phonemes based on speech attributes is listed in Table 1. The Cantonese phonemes are represented in the international phonetic alphabet (IPA) symbols.

Posterior features derived from a DNN based speech attribute/phone classifier provide information related to the contrast and similarity between phonemes. The classifier adopts the architecture of bi-directional gated recurrent unit (Bi-GRU). The inputs to the Bi-GRUs are filter-bank features. Posteriors are extracted from the Bi-GRU output at each time frame using a fully connected layer. The posteriors are transformed into log-posterior ratio (LPR), which is defined as

$$\text{LPR}_i = \log\left(\frac{p_i}{1 - p_i}\right), \quad (1)$$

where $i = 1, 2, \dots, N$, with N being the total number of phone/articulatory classes. p_i represents the posterior probability of the i -th class [30]. Subsequently, the LPRs are used for the i-vector modeling.

3.4. Paralinguistic features

The eGeMAPS is a minimal set of paralinguistic parameters measured on speech signals [12]. A collection of low-level descriptors (LLDs) are computed to characterize the speech signal’s frequency, energy, and spectral aspects at frame-level. The LLDs include acoustic measures such as pitch, formant frequency, MFCC, shimmer, jitter and spectral energy, etc. An 88-dimensional feature vector is derived by applying statistical functionals to the LLDs. Extraction of eGeMAPS features is implemented by the OpenSmile toolkit [31].

4. Fusing Phone/Word-level Results

Conventionally, SLPs observe the child’s speech sound errors in specific parts of the test words. The results are recapitulated for diagnosis of the child. Our proposed approach based on speaker-level representation is compared with methods in which word-level or phone-level error detection is done first on individual test words and subsequently combined to obtain the subject-level decision [11, 13, 18].

Phone-level detection is implemented based on our previous work in [11]. A DNN model is trained on the classification task to extract fixed-dimension embeddings of consonant-vowel segments. For detection, each test embedding \mathbf{x}_{test} is compared with the reference embeddings \mathbf{x}_{ref} from TD speakers in a pairwise manner. Pairs of embeddings from TD speech are generated to train a logistic regression (LR) classifier based on the similarities given by $\|\mathbf{x}_{test} - \mathbf{x}_{ref}\|_1$. The classification output tells whether the consonant-vowel segments in the test-reference pair are different. For each speaker, the accuracy score of detecting each initial consonant is stacked to derive

a speaker-level representation that describes the child’s performance on the articulation test. A SVM is trained on the representations to classify the test subject as TD or disordered.

For word-level detection, i-vector, x-vector, or eGeMAPS can be extracted from single-word utterances. These representations are similar to the speaker-level representations introduced in Section 3, except that one representation would be obtained from each test word. Training labels of word-level representations are inherited from the speaker-level diagnostic labels, i.e. TD or SSD. An LR classifier is trained on the word-level representations. The majority of word-level classification results determines the subject-level result.

5. Experimental Set-up

5.1. I-vector and x-vector extraction

The i-vector and x-vector systems are implemented with the Kaldi toolkit [32]. Four i-vector systems are evaluated in our experiments. Training of the first system follows the approach in [18] using MFCCs and single-word utterances. The other three systems are trained on long utterances of multiple words for extracting speaker-level representation. The input features are MFCCs, speech attribute LPR and phone LPR, respectively. The i-vector systems use 256 mixture components, and the dimension of i-vectors is 100.

Three different x-vector systems are trained and compared with the i-vector systems. The network architecture in [23] is used with that the size of hidden layers reduced by half. The generated x-vector has a dimension of 256. The first x-vector system is built using an out-of-domain corpus called CUSENT, which contains about 20 hours of Cantonese speech from 76 adults speakers [33]. The network is trained with speaker labels. The other two x-vector systems are trained with child speech, using long utterances of multiple words, and single-word utterances, respectively. When single-word utterances are used, discriminative training of x-vector extractors is carried out with the speaker labels. When utterances of multiple words are used, diagnostic labels (SSD or TD) are used to capture the difference between different TD and disordered speakers.

5.2. Posterior and embedding extractors

The phone-level or speech-attribute posterior extractors and the consonant-vowel embedding extractors all use the same neural network model with 3 layers of Bi-GRUs. The input acoustic features are 80-dimensional filter-bank features extracted every 0.01 second, and mean and variance normalised. Training speech data consist of consonant-vowels segments extracted obtained by forced alignment. The embedding and posterior extractors are trained as a softmax-based multi-class classification task with cross-entropy objective functions. The number of classification targets for phone and consonant-vowel units is 33 and 173, respectively. Multi-task training is carried out on 16 speech attributes for speech attribute posterior extractors. The Adam optimizer is applied with a learning rate of 0.0001 for the posterior extractors and 0.001 for the embedding extractors [34]. The posterior extractors are built using the Phonet toolkit [35] toolkit. The consonant-vowel embedding extractors are implemented by PyTorch [36].

5.3. Back-end classifier for SSD detection

SVM and LR classifiers for phone-, word- and subject-level detection are trained on speaker-level representations or paralin-

Table 2: Classification performance of embedding and posterior extractors.

Extractor	Metric	Performance
Consonant-vowel	Accuracy	88.0 (0.9)
Phone		70.8 (1.7)
Speech attribute	Max. UAR (Vocal)	95.4 (1.0)
	Min. UAR (Velar)	83.1 (0.6)

Table 3: Performance of subject-level SSD detection using the proposed approach.

Method	Dim. Reduction	Classifier	UAR	Macro F1
		Subject		
eGeMAPs	-	SVM	65.1 (5.3)	63.9 (5.1)
x-vector			65.7 (6.2)	65.2 (6.1)
x-vector _(CUSENT)			68.2 (5.9)	67.1 (7.0)
i-vector _(CUSENT)	-	SVM	69.9 (3.6)	68.6 (4.0)
	LDA		72.0 (5.8)	71.3 (5.3)
i-vector	-		71.7 (6.3)	71.7 (6.3)
	LDA		72.5 (7.3)	72.4 (7.7)
i-vector (phone LPR)	-		74.5 (7.9)	73.8 (6.5)
	LDA		76.2 (6.5)	76.0 (6.6)
i-vector (speech attribute LPR)	-		76.9 (5.3)	76.5 (5.6)
	LDA		78.2 (4.6)	78.0 (4.9)

guistic features using sklearn [37]. We assign positive label ‘1’ to the representations of disordered speakers and negative label ‘0’ to the TD ones. The SVM uses a linear kernel. The regularization parameter of SVM and LR is 1.0.

6. Results and Discussion

The classification performance of the embedding and posterior extractors is evaluated on the speech data from TD speakers in each cross-validation fold. Results are reported in Table 2 in terms of classification accuracy and unweighted average recall (UAR). Overall, the extractors produce a fairly satisfactory performance in classifying the consonant-vowels, phones and speech attributes. The embeddings or posterior features extracted from child speech utterances are deemed applicable and effective to subject-level SSD detection.

The 5-fold cross-validation results of subject-level SSD detection using the proposed approach are given in Table 3 in terms of UAR and the macro F1 score. With long utterances of multiple words, the i-vector approach surpasses x-vector and eGeMAPs. The i-vector approach remains competitive in the low-resource scenario. Training i-vector and x-vector extractors on out-of-domain adult speech data lead to comparable performance, where i-vector slightly outperforms x-vector. In-domain child speech is preferable in the training of the i-vector system. Using eGeMAPs is less competitive in the detection. Given SSD is mainly manifested in phonological errors, the paralinguistic information captured by eGeMAPs may be less informative about the contrast between TD and disordered speech.

Using speech attribute LPR in the i-vector extraction achieves the best detection performance, followed by phone LPR. The speech attribute LPR describes multiple changes in articulation and aspiration status. It contains richer diagnostic information than the phone LPR. Reducing the dimensionality of i-vector with linear discriminant analysis (LDA) further improves the detection and reduces the mis-classification of disordered speakers, giving a UAR of 78.2% and a macro F1 score of 78.0%.

The results of the best performed i-vector system are analysed using a two-tailed z-test. A significant threshold of 0.05 is assumed. The mis-classified subjects are found to be signif-

Table 4: Performance of subject-level SSD detection based on aggregating word-/phone-level detection results.

Method	Classifier		UAR	Macro F1
	Segment	Subject		
eGeMAPs	LR	Majority	59.4 (8.0)	58.5 (9.5)
x-vector			63.7 (5.6)	63.7 (6.0)
i-vector			65.6 (1.9)	65.9 (2.3)
Consonant-vowel embedding	LR	SVM	75.6 (4.3)	72.6 (3.9)

icantly younger than correctly classified subjects. Given word-level ground-truth annotations, the average number of errors in the false negative subjects is significantly less than in the true positive subjects. From the clinical perspective, TD children of younger age, i.e. about 3-4, are allowed to make more mistakes than older children in the articulation test. The development of their phonemic inventories and oral motor skills has just begun. The ambiguity caused by erroneous pronunciations in young TD and disordered subjects makes SSD detection more challenging.

As a comparison to the proposed detection approach, results of subject-level SSD detection based on aggregating phone-level and word-level detection results are given in Table 4. Aggregating phone-level detection results from consonant-vowel embeddings deliver competitive performance. The detection performance with a UAR of 75.6% and a macro F1 score of 72.6% is the best among detection methods based on word-level and phone-level results. The performance is also on par with the i-vector approach using MFCCs and utterances of multiple words. Information about context-dependency in consonant errors is effective for detection.

For SSD detection based on word-level results, the i-vector approach delivers a satisfactory performance with 65.6% in UAR and 65.9% in macro F1 score. $62 \pm 18\%$ of the word utterances are classified as correct pronunciations in the disordered speakers. $18.4 \pm 14.6\%$ of the word utterances are classified as erroneous in TD speakers. The results echo that disordered children can correctly produce a significant portion of the test words. TD children are also allowed to make age-appropriate errors. In the training of classifiers for word-level SSD detection, it would be sub-optimal to assume all word utterances from a speaker are errorless or related to SSD symptoms.

7. Conclusion

In this study, we demonstrate the use of knowledge-driven posterior-based features in subject-level SSD detection based on speaker representations. It has been shown that using fixed-dimensional speaker representation extracted from an aggregation of test words is a feasible and effective approach to detecting SSD in child speech. The proposed approach surpasses traditional methods that combine fine-grained classification results on individual target phones/words. In the low-resource scenario, the i-vector approach outperforms the x-vector one and paralinguistic features. For disorder symptoms that are mainly manifested in phonological errors, extraction of i-vector using posterior-based features can improve detection performance over conventional acoustic features. Our future work will focus on using self-supervised representation learning and neural confidence measure for subject-level SSD detection.

8. References

- [1] S.-J. Kim, Y.-K. Ko, E.-Y. Seo, G.-A. Oh, S.-J. Kim, Y.-K. Ko, E.-Y. Seo, and G.-A. Oh, "Prevalence of speech sound disorders in 6-year-old children in korea," *Communication Sciences & Disorders*, vol. 22, no. 2, pp. 309–317, 2017.
- [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] S. A. Karbasi, R. Fallah, and M. Golestan, "The prevalence of speech disorder in primary school students in yazd-iran," *Acta Medica Iranica*, pp. 33–37, 2011.
- [4] 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.
- [5] G. R. Daniel and S. McLeod, "Children with speech sound disorders at school: Challenges for children, parents and teachers." *Australian Journal of Teacher Education*, vol. 42, no. 2, pp. 81–101, 2017.
- [6] E. R. Hitchcock, D. Harel, and T. M. Byun, "Social, emotional, and academic impact of residual speech errors in school-age children: A survey study," in *Seminars in speech and language*, vol. 36, no. 4. NIH Public Access, 2015, p. 283.
- [7] J. Wang, Y. Qin, Z. Peng, and T. Lee, "Child Speech Disorder Detection with Siamese Recurrent Network Using Speech Attribute Features," in *Proc. of 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. of Interspeech*, 2020, pp. 4476–4480.
- [9] A. Hair, G. Zhao, B. Ahmed, K. J. Ballard, and R. Gutierrez-Osuna, "Assessing posterior-based mispronunciation detection on field-collected recordings from child speech therapy sessions," *Proc. of Interspeech*, pp. 2936–2940, 2021.
- [10] M. Shahin and B. Ahmed, "Anomaly detection based pronunciation verification approach using speech attribute features," *Speech Communication*, vol. 111, pp. 29–43, 2019.
- [11] S.-I. Ng, C. W.-Y. Ng, J. Li, and T. Lee, "Detection of Consonant Errors in Disordered Speech Based on Consonant-Vowel Segment Embedding," in *Proc. of Interspeech*, 2021, pp. 2931–2935.
- [12] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *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] I. Laaridh, W. B. Kheder, C. Fredouille, and C. Meunier, "Automatic Prediction of Speech Evaluation Metrics for Dysarthric Speech," in *Proc. of Interspeech*, 2017, pp. 1834–1838.
- [15] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect parkinson's disease from speech," in *Proc. of ICASSP*, 2020, pp. 1155–1159.
- [16] I. Laaridh, C. Fredouille, A. Ghio, M. Lalain, and V. Woisard, "Automatic Evaluation of Speech Intelligibility Based on I-vectors in the Context of Head and Neck Cancers," in *Proc. of Interspeech*, 2018, pp. 2943–2947.
- [17] S. Quintas, J. Maclair, V. Woisard, and J. Pinquier, "Automatic Prediction of Speech Intelligibility Based on X-Vectors in the Context of Head and Neck Cancer," in *Proc. of Interspeech*, 2020, pp. 4976–4980.
- [18] P. V. Kothalkar, J. Rudolph, C. Dollaghan, J. McGlothlin, T. F. Campbell, and J. H. Hansen, "Automatic screening to detect 'at risk' child speech samples using a clinical group verification framework," in *Proc. of EMBC*, 2018, pp. 4909–4913.
- [19] P. Kothalkar, J. Rudolph, C. Dollaghan, J. McGlothlin, T. Campbell, and J. Hansen, "Fusing text-dependent word-level i-vector models to screen 'at risk' child speech," *Proc. of Interspeech*, pp. 1681–1685, 2018.
- [20] H. Behravan, V. Hautamäki, S. M. Siniscalchi, T. Kinnunen, and C.-H. Lee, "I-vector modeling of speech attributes for automatic foreign accent recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 29–41, 2015.
- [21] L. F. D'Haro, R. Cordoba, C. Salamea, and J. D. Echeverry, "Extended phone log-likelihood ratio features and acoustic-based i-vectors for language recognition," in *Proc. of ICASSP*, 2014, pp. 5342–5346.
- [22] 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.
- [23] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in *Proc. of Interspeech*, 2017, pp. 999–1003.
- [24] M. McLaren, R. Vogt, B. Baker, S. Sridharan, and S. Sridharan, "Experiments in svm-based speaker verification using short utterances." in *Proc. of Odyssey*, vol. 17, 2010.
- [25] E. Zee, "An acoustical analysis of the diphthongs in cantonese," in *Proc. of the International Congress of Phonetic Sciences*, vol. 2, 1999, pp. 1101–1104.
- [26] R. S. Bauer and P. K. Benedict, *Modern cantonese phonology*. Walter de Gruyter, 2011, vol. 102.
- [27] S.-I. Ng, C. W.-Y. Ng, J. Wang, T. Lee, K. Y.-S. Lee, and M. C.-F. Tong, "CUCHILD: A Large-Scale Cantonese Corpus of Child Speech for Phonology and Articulation Assessment," in *Proc. of Interspeech*, 2020, pp. 424–428.
- [28] 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.
- [29] R. Goldman and M. Fristoe, "Goldman-fristoe test of articulation: Third edition," *Pearson*, 2015.
- [30] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "On the projection of pllr for unbounded feature distributions in spoken language recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1073–1077, 2014.
- [31] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*, 2010, pp. 1459–1462.
- [32] D. Povey, A. Ghoshal, and B. et al., "The kaldi speech recognition toolkit," in *Proc. of ASRU*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.
- [33] T. Lee, W. K. Lo, P. C. Ching, and H. Meng, "Spoken language resources for cantonese speech processing," *Speech Communication*, vol. 36, no. 3-4, pp. 327–342, 2002.
- [34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of ICLR*, 2015.
- [35] J. C. Vázquez-Correa, P. Klumpp, J. R. Orozco-Arroyave, and E. Nöth, "Phonet: A tool based on gated recurrent neural networks to extract phonological posteriors from speech," in *Proc. of Interspeech*, 2019, pp. 549–553.
- [36] A. Paszke, S. Gross, and F. e. a. Massa, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.
- [37] F. Pedregosa, G. Varoquaux, and A. Gramfort *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.