Interspeech 2024
1-5 September 2024, Kos, Greece

Automatic Children Speech Sound Disorder Detection with Age and Speaker Bias Mitigation

Gahye Kim[1*], *Yunjung Eom*[1*], *Selina S. Sung*[1,4*], *Seunghee Ha*[2], *Tae-Jin Yoon*[3], *Jungmin So*[1†]

[1]Sogang Univ., [2]Hallym Univ., [3]Sungshin Women's Univ., Republic of Korea,
[4]University of Wisconsin-Madison, WI, USA

{gahye007, yun2972}@sogang.ac.kr, ssung23@wisc.edu,
shha@hallym.ac.kr, tyoon@sungshin.ac.kr, jso1@sogang.ac.kr

Abstract

Addressing speech sound disorders (SSD) in early childhood is pivotal for mitigating cognitive and communicative impediments. Previous works on automatic SSD detection rely on audio features without considering the age and speaker bias which results in degraded performance. In this paper, we propose an SSD detection system in which debiasing techniques are applied to mitigate the biases. For the age bias, we use a multi-head model where the feature extractor is shared across different age groups but the final decision is made using the age-dependent classifier. For the speaker bias, we augment the dataset by mixing the audios of the multiple speakers in the same age group. When evaluated with our Korean SSD dataset, the proposed method showed significant improvements over previous approaches.

Index Terms: speech sound disorder, child speech, audio classification

1. Introduction

As children grow older, their ability to maneuver language and speech becomes more crucial, as speaking is a key human skill that allows the pursuit of fundamental aspects of life such as education, social interaction, and communication [1]. Speech sound disorders (SSDs) hinder children's ability to articulate their words and thoughts [2, 3]. It can profoundly impact their emotional well-being and social interactions, which leads to challenges such as feelings of isolation, anxiety, and attention issues [4]. More and more children are suffering from speech disorders. During the 2022 pandemic, the number of SSD diagnoses of children from age 2 to 12 increased significantly. [5]. When a child with SSD is left untreated until the age of six, approximately 15% of the patients tend to also experience language developmental disability [6]. As an intervention, such as speech therapy, can significantly improve the child's condition when carried out early, detection of SSDs is pivotal in allowing children with SSDs to smoothly transition into adulthood [7–10].

However, relying solely on professional speech-language pathologists (SLPs) is unrealistic due to the limited number of SLPs available. This invokes the need for a more efficient clinical approach. Accordingly, utilizing deep learning models as tools for systematic automated detection of SSD is becoming increasingly favorable for its greater accessibility and efficiency [9]. If automatic SSD detection becomes viable with such an approach, it will allow for an initial assessment of a child's speech within the home setting.

*equal contribution. †corresponding author. This work was supported by the National Research Foundation of Korea (NRF-2021S1A5A2A03064795).

Research in the field of neural networks has attempted to construct an effective classifier based on training neural networks such as CNN and RNN [11, 12]. A Transformer-based audio classification model has also been shown effective for the automated detection of disorder speech [13, 14]. Speaker embedding methods such as x-vectors and i-vectors have demonstrated promising results in previous studies [15–17]. They are speaker-level feature representations obtained from MFCCs or posterior features from DNN. When integrated into the SVM classifier, they improve the model's ability to detect abnormal speech patterns in SSD speech data [9]. Recently, formant and duration analysis have been proposed as valuable features for detecting disordered speech [7].

One key guideline used for child SSD diagnosis is examining their pronunciation of certain words in comparison to that of healthy speakers in the subject's age group [9]. In other words, the age of the patient determines the boundary that serves as the baseline for diagnosing speech disorders. This indicates that when building a systematic classifier for SSD detection, the subject's age must be taken into consideration. Particularly in younger subjects, distinguishing between typically developing (TD) and SSD presents significant challenges due to the natural articulation errors inherent in early childhood development. If these features are not considered during the training process of SSD detection models, it can result in a bias leading to the misclassification of young TD subjects as having SSDs. In this study, we aim to mitigate such bias induced by age.

Moreover, it is also difficult to obtain sufficient data for TD/SSD detection which calls for a need for data augmentation. If the model lacks sufficient relevant features, the model may inadvertently learn irrelevant features, leading to a decline in its intended performance. In an SSD detection system, the learning of features associated with the speaker, such as voice tone or volume, rather than those indicative of SSD, could result in a speaker bias. Consequently, this may yield a model reliant on speaker-specific features rather than utilizing discriminative features for SSD detection. To combat speaker bias in the model, suitable data augmentation strategies are required. While conventional audio augmentation methods are widely used in audio classification [18, 19], there exists only a few research done on augmentation methods that reflect the characteristics of disordered speech [20]. To ensure proper learning from SSD speech data, effective augmentation techniques are essential for the model to perform optimally across all pathological speakers [18, 20, 21].

In this paper, we propose debiasing techniques to mitigate age and speaker bias in the SSD dataset. As the basic pipeline, we follow the widely used method of finetuning a pre-trained model for TD/SSD classification. To address the speaker bias, we apply data augmentation on the training dataset by randomly

10.21437/Interspeech.2024-1799

mixing audio samples from multiple speakers of the same age group. Additionally, to mitigate the speaker bias, we implement a model with multiple classifiers. Our proposed multi-head classification model is designed to account for the fundamental differences in articulatory abilities based on the speaker's age. It employs a shared feature learning mechanism but adopts distinct classifiers tailored to each age group. Our results show that both the augmentation method and the multi-head classification model improve our baseline TD/SSD detection accuracy. Especially, many children aged 2 to 4, who were misclassified as SSD before, were correctly classified.

Our contributions can be summarized as follows. Firstly, we collected a previously non-existent SSD dataset of Korean children. Secondly, we propose a methodology to mitigate age and speaker bias through model architecture and dataset augmentation. Finally, we evaluated our proposed method against existing methods, demonstrating superior performance in terms of Unweighted Average Recall (UAR), Macro F1, and accuracy.

2. Korean Child Speech Database

2.1. Data Collection

We collected a 4.4-hour child speech dataset that has been carefully annotated for our study. Participants were recruited as part of a large-scale project aimed at establishing speech development profiles of Korean-speaking children. This study included 709 Korean-speaking children aged 2 to 10 years, encompassing both children with normal speech and those with SSD. The age distribution is shown in Table 1.

Table 1: *Participant age distribution*

Age	2	3	4	5	6	7	8	9	10	Total
TD	32	66	66	74	56	54	51	19	5	423
SSD	1	74	73	59	45	21	7	6	0	286
Count	33	140	139	133	101	75	58	25	5	709

The participants include a total of 423 children with normal speech. These children had no reported speech or language challenges and showed no signs of hearing, physical, or cognitive development problems. Additionally, the study included 286 Korean-speaking children with SSD, who exhibited speech problems due to cleft palate, congenital anomalies, neurological impairments, intellectual deficits, or sensorineural hearing loss, as reported by their caregivers.

Graduate students majoring in speech-language pathology were trained to deliver consistent and uniform instructions to participants in the data collection. During the assessment, each examiner and child sat facing each other. Target words from the Assessment of Phonology and Articulation for Children (APAC) and the Korean Articulation and Phonology Profile (K-APP) were presented through illustrated pictures. These tests are standardized articulation and phonology tests specifically designed for Korean-speaking children. The target word list contains 80 words ranging from 1 to 5 syllables. The examiner presented the picture and asked what the word was to prompt a voluntary response from the child. If the child did not respond spontaneously, the examiner modeled the word and then elicited the child's response through delayed imitation. Table 2 shows examples of target words in our dataset.

Table 2: *Examples of target words*

syallables	words	
1-syllable	뱀 (baem)	빗 (bit)
	학 (hak)	쌀 (ssal)
2-syllables	나무 (namu)	딸기 (ttalgi)
	단추 (danchu)	그네 (geune)
3-syllables	색종이 (saekjongi)	눈사람 (nunsaram)
	호랑이 (horangi)	옥수수 (oksusu)
4-syllables	크레파스 (keurepaseu)	파인애플 (painaepeul)
	할아버지 (harabeoji)	미끄럼틀 (mikkeureomteul)
5-syllables	아이스크림 (aiseukeurim)	엘리베이터 (ellibeiteo)

2.2. Data Cleaning

There were instances in which children would utter words that were not the target words or overlapped with the instructions from the speech therapist. After recording, Praat [22] was used for data cleaning. The audio intervals including the target words were tagged with phonetic transcription. Utterances with excessive noise or overlapping speech were excluded. Finally, word level audio files containing the target words were saved with their transcriptions in our database.

As a result, our dataset contains on average about 30 ± 19 target words per speaker. The average duration of each sample is 0.8 ± 0.5 seconds. Samples that are shorter than 0.3 seconds are omitted due to the lack of sufficient vocal content. The word-level audio sample distribution is shown in Table 3.

Table 3: *The word-level data distribution*

Age	2	3	4	5	6	7	8	9	10	Total
TD	593	1508	1811	1945	1558	1612	1584	627	172	11410
SSD	39	2925	2453	1838	1391	654	188	126	0	9614

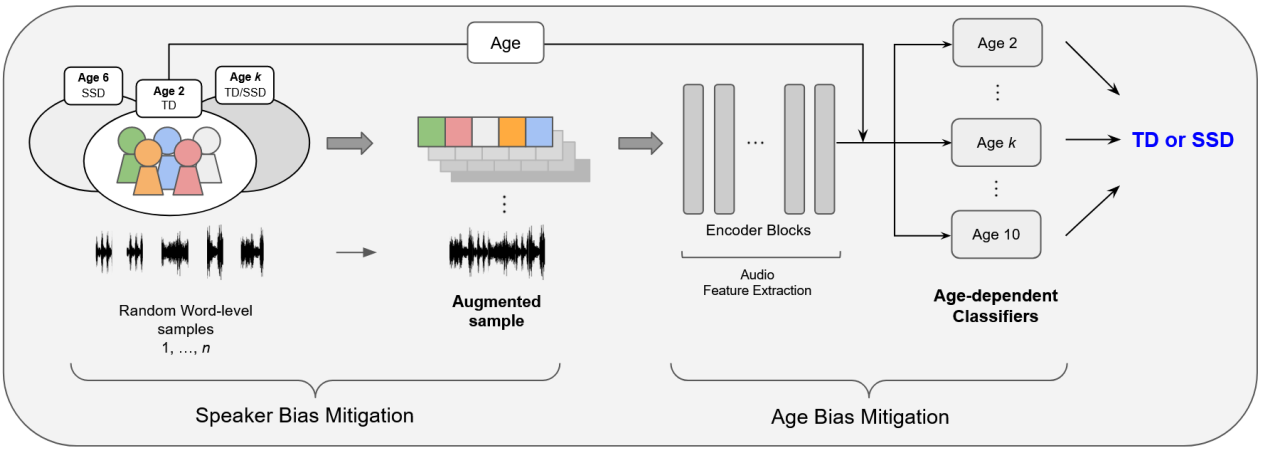
3. Proposed System Design

Our overall system design is described in Figure 1. We use Whisper [23], a widely used pre-trained automatic speech recognition (ASR) model, for the downstream task of audio classification. The Whisper model has demonstrated strong performance in audio-related tasks such as speech recognition, translation, and language identification. However, the model presents two significant challenges for our task. Firstly, the model must differentiate between TD and SSD while considering the age to avoid performance degradation due to age bias. Secondly, due to insufficient data, the model may learn spurious features that do not contribute to accurately classifying TD and SSD, instead relying on factors like tone and voice, leading to a risk of speaker bias. To address these challenges, we implement age-based multi-classifiers in conjunction with data augmentation. An alternative approach could have involved creating separate models for each age group and training them independently. However, as we will demonstrate, our findings indicate that our multi-head model performs better.

3.1. Age-dependent Multi-Head Model

Our goal is to have the model learn general audio features that decide pronunciation proficiency while determining TD/SSD

Figure 1: *Our overall system design*



based on age. Accordingly, we use a common backbone for audio feature extraction, and then attach multiple classification heads to reach a final classification decision. The Whisper model processes the input audio data sampled at 16,000 Hz, which is then padded to be 30 seconds long. The sample is then converted into a spectrogram. It passes through two feature extraction layers and 24 encoder blocks in the Whisper encoder. The projection layer adjusts the feature dimension and average pooling further compresses the feature. The feature is then fed into one of the age-dependent classifiers. Only the weights of the classifier corresponding to the ages present in each batch are updated. We design the following loss function for our Multi-head model:

$$L_{\text{total}} = \frac{1}{B}\sum_{i=1}^{M} a_i L_i \quad \text{where} \quad \sum_{i=1}^{M} a_i = B \tag{1}$$

where B is the batch size, M is the number of heads, and L_i is the binary cross entropy. a_i is the number of occurrences of a specific age within a batch. The sum of a_i is the batch size so that only the weight for each age can be updated in the batch.

3.2. Data Augmentation

We utilize data augmentation in our proposed system for the following reasons: (1) to prevent the model from overfitting and learning with speaker bias, and (2) to highlight the TD/SSD features unique for each age group.

From the 21,024 word-level audio samples, we concatenate the audio samples from the same speaker to carry out subject-level TD/SSD diagnosis. This creates a total of 709 subject-level samples. We augment the training dataset as follows: We randomly mix *n* word-level samples from a group of speakers with the same age and TD/SSD label, where *n* is the augmentation parameter. This way, each augmented data consists of mixed utterances from several speakers to represent the overall disorder characteristics at each age. The underlying assumption is that this will encourage the model to learn from the correlation of age and speech characteristics. In the augmentation, each word-level audio data was sampled without replacement. Thus, The rate of increase in dataset through augmentation is proportional to the inverse of *n*. The augmented dataset is created by merging both the subject-level and augmented samples.

4. Experiments

4.1. Experiment Setup

The model used in the experiment is "distil-whisper-medium.en" which is an optimized version of the Whipser ASR model [24]. We employ the encoder of the model as a feature extractor. All audio inputs are padded to be 30 seconds, using a padding value of 0. We use a batch size of 8 and train the model for 10 epochs on an NVIDIA GeForce RTX 3090 GPU. We divide our dataset into a 7:3 ratio for the training and test dataset. Speakers featured in the training set do not appear in the test set. The linear scheduler and AdamW optimizer are used to train the model with a learning rate of 3e-5 which was found optimal after empirical tuning.

For our experimentation, we use 3 different datasets: word-level, subject-level, and augmented dataset. The word-level dataset consists of audio samples which each containing a single word spoken by the subject. Subject-level dataset concatenates all the word-level audio spoken by each speaker. Lastly, the augmented dataset is the set containing subject-level data and augmented samples as described in 3.2. Using the the word-level dataset, we first reproduce a previous study [7] that utilized the syllable-level duration of the target words and the subject's age as features while training the SVM classifier. Then, we classify word-level data using Whisper, assigning labels based on the speaker's labels. Next, we experiment using the subject-level dataset. We train using three different DNN models: ResNet34, Wav2Vec2, and Whisper. Separate model, which is the method of using k-many different models for k many age groups and classifying them independently, is then experimented. We then train our proposed multi-head model where samples with the same age enter the corresponding age-specific classifier. Finally, we experiment on the augmented dataset, where each sample consists of multiple speakers from the same age group, with both Whisper and our proposed model. Each experiment was conducted using five distinct random seeds, and the results were calculated using a 95% confidence interval.

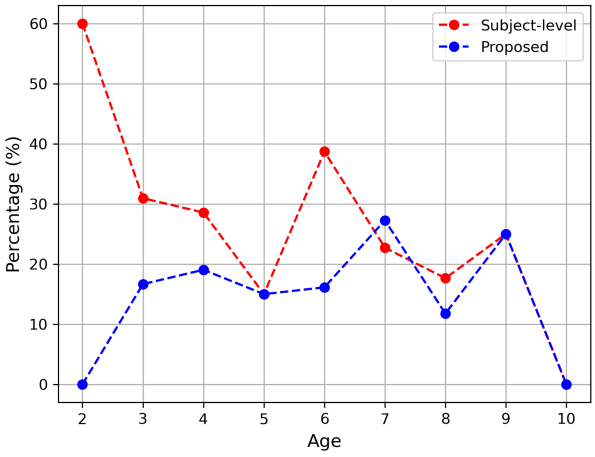
4.2. Results

Experiment results are reported in Table 4. We utilize 3 different evaluation metrics: UAR, Macro F1, and accuracy. In the table, the results are organized by dataset type, with each dataset trained and evaluated using several models. In the experiments,

Table 4: *TD/SSD classification results*

Method	Metric (Avg. ± CI, %)		
	UAR	Macro F1	Accuracy
Word-level dataset			
Duration + Age [7]	53.09	44.28	61.97
Whisper	69.47±0.69	69.58±0.69	70.37±0.66
Subject-level dataset			
ResNet34 [25]	69.70±2.80	70.16±2.98	72.77±2.54
Wav2Vec2 [26]	72.49±2.54	72.30±2.13	73.24±1.80
Whisper	72.76±1.43	73.29±1.46	75.12±1.26
Separate model	68.99±2.20	68.99±2.39	70.14±2.59
Whisper+Multi-head	75.17±2.11	75.69±2.30	77.27±2.53
Augmented dataset			
Whisper	74.36±2.50	74.99±2.48	76.90±2.08
Whisper+Multi-head *(Ours)*	**78.02±1.81**	**78.76±1.80**	**80.28±1.65**

Figure 2: *Misclassification rate by age*



we seek to answer the following research questions.

RQ1: Does the proposed method improve the detection accuracy? First, the performance of a prior method [7], which utilized the duration and age features of audio, was assessed. This method yielded an accuracy of 61.97% on the word-level dataset. We then employed the Deep Neural Network(DNN) based method using Whisper on the same dataset and obtained the overall accuracy of 70.37%. We compared 3 DNN-based models on the subject-level dataset, and Whisper achieved the highest accuracy of 75.12%. We then applied our proposed multi-head method to Whisper, which led to a notable improvement of over 2% across all metrics, specifically from 75.12% to 77.27% in accuracy. Finally, to evaluate the effectiveness of the augmentation method, we compare Whisper trained with subject-level dataset to the Whisper with augmentation. Whisper with our augmentation method achieved the accuracy of 76.9%, which was a 1.8% increase. When both the multi-head model and the augmentation were employed, the accuracy rose to 80.28%, over 3.4% higher than the Whisper with augmentation. Overall improvement of 5.2% in the accuracy was observed solely due to our contribution. Further implications of this improvement is discussed in the following RQ2 and RQ3.

RQ2: Does the multi-head model reduce the age bias? The general feature extraction which occurs in the encoder blocks captures the differences in characteristics across TD samples and SSD samples. The samples then enter the age-dependent classifier in which the model further learns the age-specific articulation features and the varying TD/SSD boundary of each age. This improved the accuracy rate in nearly all of the age groups as shown in Figure 2. Furthermore, our method outperformed the separate model by over 7%, affirming the effectiveness of using a multi-head model. Most importantly, the misclassification rate in the age group of 2-4 decreased significantly. As shown in Figure 2, the proposed method successfully leveraged the model to adjust the different TD/SSD detection boundary for this age group, enabling it to correctly classify the age group of 2-4. Detecting SSDs in younger children is more effective because early treatment leads to greater effectiveness. Our proposed method has made progress in addressing the age bias to improve the overall accuracy.

Table 5: *Results varying augmentation parameters*

parameter	Metric (Avg. ± CI, %)		
	UAR	Macro F1	Accuracy
$n = 0$	72.76±1.43	73.29±1.46	75.12±1.26
$n = 5$	**78.02±1.81**	**78.76±1.81**	**80.28±1.65**
$n = 10$	77.23±1.80	77.87±1.90	79.34±1.84

RQ3: Does the augmentation method reduce the speaker bias? To investigate whether our data augmentation method was effective, we compare the metric results of the Whisper using subject-level dataset with and without the data augmentation method. The Whisper model of accuracy 75.12% increased to 76.90% when it was trained with the augmentation dataset. Our data augmentation reduced the features specific to each speaker by mixing samples from multiple speakers, but preserved the common features across each age. Therefore, the data augmentation, when combined with the proposed multi-head model, enhanced the accuracy due to the following reason: The speaker bias such as the tone, pitch, and duration were debiased by the mixing augmentation. This allowed the age feature to be emphasized, which aided the age-dependent classifier to correctly classify TD/SSD samples. As a result, when augmentation method was applied to the multi-head model, we achieved the highest accuracy. As shown in Table 5, we found that 5 words per sample effectively blended speaker-specific characteristics while ensuring the dataset remained sufficiently large.

5. Conclusion

In this paper, we presented the automatic speech sound detection for children with speaker and age bias mitigation. We have collected a dataset comprising TD/SSD subjects among Korean children aged 2 to 10. The age-dependent multi-head model was implemented to mitigate age bias and age-based mixing augmentation techniques were used to reduce speaker bias. When combined together, our proposed system showed high performance across all various evaluation metrics compared to previous methods. For future research, we plan to investigate noise reduction techniques and novel augmentation methods to address further biases that may exist in the dataset. In considering an approach akin to human methods for SSD detection, one might contemplate identifying incorrect pronunciations through speech recognition. It would be beneficial to leverage this aspect for further research.

6. References

[1] L. Sices, H. Taylor, L. Freebairn, A. Hansen, and B. Lewis, "Relationship between speech-sound disorders and early literacy skills in preschool-age children: Impact of comorbid language impairment," *Journal of Developmental and Behavioral Pediatrics*, vol. 28, no. 6, pp. 438–447, Dec. 2007.

[2] S. McLeod, L. J. Harrison, L. McAllister, and J. McCormack, "Speech sound disorders in a community study of preschool children," *American Journal of Speech-Language Pathology*, vol. 22, no. 3, pp. 503–522, Aug. 2013.

[3] A. K. Namasivayam, D. Coleman, A. O'Dwyer, and P. van Lieshout, "Speech sound disorders in children: An articulatory phonology perspective," *Frontiers in Psychology*, vol. 2998, no. 2, Jan. 2020.

[4] G. 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, Feb. 2017.

[5] K. Health. (2023) Louder than words: Pediatric speech disorders skyrocket throughout pandemic.

[6] L. E. Ju, "The effects of speech sound disorder on vocabulary development," *Commun Sci Disord*, vol. 27, no. 4, pp. 868–878, 2022.

[7] T. L. S. Ng, S. W. Ng, "A study on using duration and formant features in automatic detection of speech sound disorder in children," in *Proc. INTERSPEECH 2023 – 24rd Annual Conference of the International Speech Communication Association*, Ireland, Dublin, Aug. 2023, pp. 4643–4647.

[8] M. Han and S. J. Kim, "Characteristics of functional speech sound disorders in korean children," *Annals of Child Neurology*, vol. 30, no. 1, pp. 8–16, Dec. 2021.

[9] S.-I. Ng, C. W.-Y. Ng, J. Wang, and T. Lee, "Automatic detection of speech sound disorder in child speech using posterior-based speaker representations," in *Proc. INTERSPEECH 2022 – 23th Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 2853–2857.

[10] J. Wang, Y. Qin, Z. Peng, and T. Lee, "Child speech disorder detection with siamese recurrent network using speech attribute features," in *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Sep. 2019, pp. 3885–3889.

[11] J. Liu, C. Ren, Y. Luan, S. Li, T. Xie, C. Seals, and M. Speights Atkins, "Speech disorders classification by cnn in phonetic e-learning system," in *Artificial Intelligence in HCI*, Cham, May. 2022, pp. 557–566.

[12] M. Krishnaveni, P. Subashini, and T. T. Dhivyaprabha, "Recurrent neural network model for the classification of tamil speech sound disorder signals," in *Proceedings of International Conference on Communication and Computational Technologies*, Singapore, Sep. 2023, pp. 745–759.

[13] 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 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 3618–3622.

[14] F. Javanmardi, S. Tirronen, M. Kodali, S. R. Kadiri, and P. Alku, "Wav2vec-based detection and severity level classification of dysarthria from speech," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.

[15] I. Laaridh, W. B. Kheder, C. Fredouille, and C. Meunier, "Automatic prediction of speech evaluation metrics for dysarthric speech," in *Proc. Interspeech 2017*, Aug. 2017, pp. 1834–1838.

[16] S. Quintas, J. Mauclair, V. Woisard, and J. Pinquier, "Automatic prediction of speech intelligibility based on x-vectors in the context of head and neck cancer," in *Interspeech 2020*, Shanghai (fully virtual conference), China, Oct. 2020, pp. 4976–4980.

[17] P. Kothalkar, J. Rudolph, C. Dollaghan, J. McGlothlin, T. Campbell, and J. H. Hansen, "Fusing text-dependent word-level i-vector models to screen 'at risk' child speech," in *Proc. INTERSPEECH 2018 – 19rd Annual Conference of the International Speech Communication Association*, Hyderabad, India, Aug. 2018, pp. 1681–1685.

[18] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 2340–2344.

[19] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, Sep. 2015, pp. 3586–3589.

[20] P. N. Sudro, R. K. Das, R. Sinha, and S. R. Mahadeva Prasanna, "Significance of data augmentation for improving cleft lip and palate speech recognition," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Tokyo, Japan, Dec. 2021, pp. 484–490.

[21] Y.-M. Kuo, S.-J. Ruan, Y.-C. Chen, and Y.-W. Tu, "Deep-learning-based automated classification of chinese speech sound disorders," *Children*, vol. 9, no. 7, Jul. 2022.

[22] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9, pp. 341–345, 01 2001.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, Hawaii, USA, Jul. 2023, pp. 28 492–28 518.

[24] S. Gandhi, P. von Platen, and A. M. Rush, "Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling," 2023.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Jun. 2016, pp. 770–778.

[26] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.