In-Domain Data Augmentation to Enhance Severity Level Classification of Dysarthria from Speech

Bhanuja Karumuru, Paban Sapkota, and Hemant Kathania
Department of Electronics and Communication Engineering, National Institute of Technology Sikkim, Sikkim, India
b200087@nitsikkim.ac.in, phec230006@nitsikkim.ac.in, and hemant.ece@nitsikkim.ac.in

***Abstract*—In this paper, we present our endeavor to construct an automatic dysarthria severity level classification system tailored for low-resource dysarthric speech datasets. The scarcity of available speech data from dysarthric speakers poses a significant challenge to training an effective classification system. Addressing this challenge, we devised a robust baseline system by blending a distinctive set of features, encompassing temporal, prosodic, and spectral information, with the traditional MFCC features. This amalgamation aptly captures the nuanced characteristics of dysarthric speech, facilitating efficient model training. To tackle the constraints of low-resource conditions, we explored four prominent augmentation techniques: Speaking Rate, Pitch, Formant, and Vocal Tract Length Perturbation (VTLP) modification based data augmentation for the task of severity classification. The explored data augmentation gives a significant reduction in the classification error rate (CER), and VTLP-based data augmentation is superior among others. Further, we also investigated combinations of explored data augmentation methods, fortifying the reliability of our dysarthria severity classification system. The combined novel augmentation gives a noteworthy relative improvement of 42.86% over the baseline on the dysarthric severity classification.**

***Index Terms*—Dysarthria, severity classification, speech modification, speech augmentation**

I. Introduction

Dysarthria, recognized as a motor speech disorder, significantly impacts an individual's ability to produce intelligible and fluent speech [1]. Acquiring dysarthric speech data is challenging due to its inherently low-resource nature. Characterized by impaired muscle control over the speech articulators, resulting distinct speech patterns requires specialized data collection methods. This complicates the development of effective models and algorithms for dysarthric speech analysis. The classification of dysarthric speech severity is crucial for guiding clinical interventions [2] and developing personalized treatment strategies [3] based on prior knowledge of the patient's severity level. Additionally, for continuous long-term monitoring, adjusting treatment plans according to severity enables dynamic adaptation to evolving speech patterns [4]. Another potential application of a robust dysarthria speech severity classification system is in customizing designs for assistive speech systems and devices to address various speech difficulties [5].

In the literature, there are extensive exploration of a plethora of techniques equipped with Machine Learning (ML), demonstrating the potential to objectify tasks related to severity classification [6]–[8]. For instance, the output of deepspeech logits as features have been explored in [9], yielding high accuracy in classification. Another study showcasing a combination of prosodic and spectral acoustic features have demonstrated a noteworthy enhancement in the accurate classification of severity [10]. Meanwhile, others have focused on enhancing the models used as classifiers, emphasizing a shift in trends from traditional ML-based classifiers to promising neural network architectures for severity classification objectives [5].

Our research endeavors focus on advancing dysarthria severity classification through the application of multiple machine learning techniques to identify a suitable base model. Then we have boosted the baseline classification performance by introducing additional acoustic and prosodic features to the base MFCC features [7]. We harnessed the TORGO database of dysarthric speech, renowned for its comprehensive dysarthric speech recordings, as the cornerstone of our investigation [11]. We have proposed a novel approach of combination of data augmentation methods for the severity level classification task. For that we have inquired four prominent data augmentation techniques - Speaking Rate modification [12], Pitch Modification [13]–[15], Formant Modification [16], [17], and Vocal Tract Length Perturbation (VTLP) [18]. The augmentations have been applied only to the system with model and feature-set that gave remarkable baseline outcomes. This comprehensive strategy, integrating multiple data augmentation methods, aims to enhance the accuracy and reliability of dysarthria severity classification, an area that has not been thoroughly explored. The key contributions of this paper are outlined below:

- We have investigated acoustics and prosody features and their combinations to boost the baseline system for the classification of the severity level (very low, low, medium, and high) of dysarthria from speech.
- Explored four prominent augmentation techniques Speaking Rate Modification, Pitch Modification, Formant Modification, and Vocal Tract Length Perturbation (VTLP) to overcome the data scarcity and improve the performance of the severity classifier.
- Further down the line, combinations of augmentation methods have been thoroughly examined, giving us a robust classification system.

The paper's subsequent sections are arranged as follows: Section II details the explored database and its significance. Moving to Section III, we elaborate on employed augmentation methods, including Speaking Rate modification, Pitch

Modification, Formant Modification, and VTLP. Section IV outlines the experimental setup, highlights our investigation of acoustic/prosodic features to enhance the baseline classifier, and analyses in detail the effect of data augmentation techniques along with our proposed combination method for designing a robust severity classifier. Finally, Section V summarizes key findings and discusses possible implications.

II. Database

A. Torgo Database

The Torgo database was developed through a collaborative effort between the Computer Science and Speech-Language Pathology departments at the University of Toronto and the Holland–Bloorview Kids Rehabilitation Hospital [11]. The database includes recordings of eight people with speech difficulties (three women and five men) and seven without speech difficulties (three women and four men), covering ages 16 to 50 years. All the recordings in the TORGO database used two different types of microphones: an array microphone and a head-mounted electric microphone. The array microphone recorded at a sampling rate of 44.1 kHz, while the head-mounted electric microphone recorded at a sampling rate of 22.1 kHz. Down-sampling at 16 kHz was performed on the acoustic signals [19].

In the Torgo database, dysarthria severity is determined through a comprehensive evaluation using the standardized Frenchay Dysarthria Assessment (FDA) [11]. This assessment considers 28 perceptual dimensions of speech, covering reflexes, respiration, facial and oral movements, and intelligibility. Experienced speech-language pathologists conduct the assessment, assigning each dysarthric speaker an overall FDA score [20]. The severity levels (e.g., VERY LOW, LOW, MEDIUM, HIGH) is then assigned based on these scores, providing a standardized measure of dysarthria severity for research and clinical purposes [11]. Severity level for each speaker is given in Table I.

TABLE I
TORGO CORPUS: SEVERITY LEVEL OF THE DYSARTHRIC SPEAKERS

Severity Level	Speaker ID
Very Low	F03, F04, M03
Low	F01, M05
Medium	M01, M02, M04

III. Explored Augmentation Methods

Data augmentation is instrumental in addressing the challenges posed by data scarcity. Given the limited availability of speech data from dysarthric speakers, augmentation serves as a viable solution to enhance the corpus size. Our exploration encompasses four speech modification-based augmentation methods: Speaking Rate, Pitch, Formant, and VTLP modifications, with all classes being subject to augmentation.

A. Speaking Rate Modification

The speaking rate modification was facilitated through Time Scale Modification (TSM) based on Real-Time Iterative Spectrogram Inversion with Look-Ahead (RTISI-LA) algorithm [12]. A scaling factor s, where $0.5 \leq \mathrm{s} \leq 2$, was varied during experimentation with a step size of 0.1. The algorithm utilizes Short-Time Fourier Transform Modification (STFTM) to process the audio signal [21]. The primary processing loop iterates through the frames of the input signal, applying a window function and reconstructing the signal through an iterative reconstruction approach [22]. The modified frame length ($L = 256 \cdot \mathrm{s}$) and the modified hop size ($S = \frac{L}{4}$) influence the STFT process, impacting the overall modification of the signal. The resulting modified audio is then utilized further in augmentation. A crucial component of the process is that it introduces phase perturbation [22], effectively mitigating resonance effects during reconstruction.

B. Pitch Modification

The implementation of pitch modification also employed the RTISI-LA algorithm as detailed in [14], which allows adjusting its pitch based on a specified semitone value represented by a scale factor τ, where τ was varied with a step size of 0.1 ($0.5 \leq \tau \leq 2$). This approach too utilizes a STFTM in the processing of the audio signal. The frame length and hop size are initially set, and the semitone value is converted to a scaling factor. The processing loop iterates through the frames of the input signal, applying the modified window function and reconstructing the signal through an iterative reconstruction process. The window used in the resampling process is adapted based on the modified frame length influenced by the semitone value. These adjustments play a crucial role in achieving the desired pitch modification, offering flexibility in tailoring the output audio's tonal characteristics based on the specified semitone value. It is to be noted that the window function is adapted for pitch modification, and the semitone value significantly influences the resampling process in the iterative reconstruction of the signal [23]. The resulting modified audio is then utilized for the augmentation.

C. Formant Modification

Formant modification [16] is a crucial technique in speech signal processing, employed to alter the spectral characteristics of an audio signal by adjusting its formants. Formants are resonant frequencies in the speech spectrum that play a significant role in determining the quality and timbre of the produced sound. The implementation of formant modification highlights the process of fitting Linear Predictive Coding (LPC) models to short-time segments of the audio signal.

LPC analysis is a powerful method in speech processing that models the spectral envelope of a signal [17]. The key idea is to represent the speech signal as the output of an all-pole filter, with the coefficients of this filter providing insights into the resonant characteristics of the signal. For formant modification, LPC coefficients are analyzed and modified to achieve the desired changes in the spectral envelope.

In the context of modifying children's speech, the study in [24] boasts the effectiveness of formant modification through LPC analysis. The technique applies LPC analysis to short-time segments of the signal, and warps the poles of the LPC model to modify the formant. The modified LPC coefficients are subsequently utilized to synthesize the altered signal.

The primary steps we followed in the formant modification algorithm are outlined as follows:

- We performed LPC analysis on short-time segments of the audio signal.
- Modified the LPC coefficients with pole warping, denoted by factor α, $(-1 \leq \alpha \leq 1)$ with step size of 0.05 to acquire the modified formant.
- Synthesized the modified signal with the warped LPC coefficients.

The formant modification technique provides a means to manipulate the spectral characteristics of speech signals by adjusting the LPC coefficients. Implementation in this paper showcases the practical application of formant modification in the task of dysarthric speech classification.

D. VTLP

The configuration of the vocal tract exhibits variability among different speakers. To address such inter-speaker variations, Vocal Tract Length Normalization (VTLN) was introduced as a technique [25]. VTLN is then applied in a reverse manner, leading to Vocal Tract Length Perturbation (VTLP), which introduces variability to speech data by simulating different vocal tract lengths. The process involves taking a time-domain audio segment $x(t)$ and representing it in the frequency domain as $X(f)$, which is the Fourier transform of $x(t)$. VTLP then applies a perturbation factor β, selected from a discrete set (e.g., $0 \leq \beta < 1$ simulates the shorter vocal tract while $\beta > 1$ is responsible for longer vocal tract), along the frequency axis of $X(f)$, resulting in the output $Y(f)$ = X(βf). This perturbation technique modifies the spectral envelope of the audio segment while maintaining the original audio duration. VTLP is a method to introduce controlled variability in speech signals, specifically targeting the spectral characteristics associated with different vocal tract lengths. This mechanism allows VTLP to introduce variations in the spectral envelope of the audio segment while maintaining the original duration of the audio. We have experimented with the perturbation factors ($0.98 \leq \beta \leq 1.08$) with a step size of 0.02.

IV. Results and Analysis

A. Experimental setup and Baseline Result

The baseline experiments were performed with four different types of classifiers - K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Networks (ANN). The 13-dimensional set of MFCC features representing a concise depiction of the audio spectrum, were selected for the initial set of study. We chose the Classification Error Rate (CER) as a suitable metric for the performance evaluation. The SVM, KNN, and RF models went through 5-fold cross-validation to get the best parameters. For the SVM classifier, kernel was set to linear and regularization parameter had a value of 2. The KNN was tuned to 13 number of neighbors. While the Random Forest (RF) classifier had 120 number of trees. The ANN had an input layer matching the dimension of input features. It included a total of four hidden layers of 128, 64, 32, and 16 dimensions respectively. Rectified Linear Unit (ReLU) were selected as activation function for each of the hidden layers and Softmax for the output layer. The output layer consist of three nodes in alignment to the number of severity classes present in the database. The training, testing, and validation splits were distributed as 70%, 15%, and 15%, respectively. Table II sketches the baseline results in terms

TABLE II
THE BASELINE CLASSIFICATION RESULTS, PRESENTED AS CER IN %, FOR THE EXPLORED CLASSIFIERS. THE PERFORMANCE ASSESSMENT IS BASED ON MFCC FEATURES.

Classifier	Overall	VL	L	M
SVM	21.16	15.16	11.89	15.27
KNN	7.09	5.56	4.14	4.47
RF	7.31	5.56	3.93	5.13
ANN	**6.32**	3.92	3.92	4.68

of CER (%), exhibiting MFCC features, for SVM, KNN, RF and ANN classifiers. Along with the overall class CER, we have incorporated the performances of each of the severity class individually as well. For example, k number of samples are in the test set in total, where vl, l, and m are number of samples from Very Low (VL), Low (L), and High (H) severity respectively, then $\frac{k-(\text{no. of correct predictions})}{k} \times 100$ gives the overall CER. Meanwhile, the individual CER for Very Low severity was calculated as $\frac{\text{vl - (no. of correct predictions from vl)}}{\text{vl}} \times 100$. Similarly, the CER for the other two classes were computed in the same way. The ANN model outperformed the other three classifiers, demonstrating the lowest error in classification with a 6.32% CER. This underscores the effectiveness of the ANN model, portraying superior performance in severity classification task. It can be attibuted to the influence of hyperparameter tuning on model's performance, emphasizing the critical role of selecting the appropriate model parameters for accurate severity classification.

B. Boosting the Baseline Performance

To boost the performance of the baseline we have explored different temporal, spectral and prosodic features [26] and their combinations. The mean value for Zero Crossing Rate (ZCR), Spectral Centroid, Spectral Entropy, Spectral Crest, Spectral Flatness, Spectral Rolloff, Pitch, Root Mean Square Energy (RMSE), and Log Energy in combination with MFCC features were found as the best suited feature combination for our severity classification task. The aforementioned features followed the extraction illustrated in [5] and [27].

The unique combination of 22-dimensional (13+9) features

TABLE III
IMPROVED BASELINE RESULTS (IN CER (%)) FOR THE CLASSIFIERS WE EXPLORED. EVALUATION BASED ON THE COMBINED FEATURE SET WE'VE EMPLOYED.

Classifier	Overall	VL	L	M
SVM	17.23	11.67	9.49	13.30
KNN	7.85	6.32	3.60	5.78
RF	6.22	4.47	3.05	4.91
ANN	**3.71**	2.73	4.98	3.65

led the baseline to outperform the case when the models utilized only MFCC features. This superiority stems from the diverse feature combination, capturing various aspects of the speech signal and enhancing the representation of the classification task. The significant improvements in classification can be observed in Table III. SVM, RF, and ANN

classifiers showed tremendous reduction in CER. The ANN model still outclassed other classifiers with a CER of 3.71%. The detailed results in Table III provide valuable insights into the effectiveness of different classifiers with this expanded feature set.

C. Effect of data augmentation

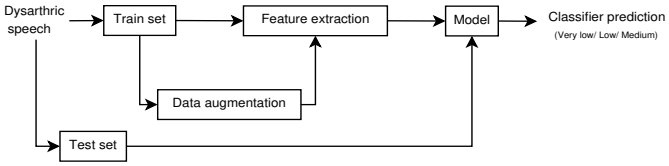

Fig. 1. Block diagram of proposed system

To tackle data scarcity and capture more acoustic and speaker variability, we have explored four prominent augmentation techniques: Speaking Rate Modification, Pitch Modification, Formant Modification, and Vocal Tract Length Perturbation (VTLP). Block diagram of proposed system is given in Figure 1. In Figure 1, data augmentation experiments were conducted utilizing each technique separately. This was achieved by modifying only the training set and then combining it with the original training set to train the classifier for severity classification. For Speaking rate, the modification factor s was varied through 0.7 to 1.2 with a step size of 0.1. In the case of pitch modified augmentation, the experiments were conducted for factor $0.8 \leq \tau \leq 1.2$ with stride of 0.1. Formant modification was carried out for four different values of factor α as shown in the Figure 2. And the VTLP was conducted in range of $0.98 \leq \beta \leq 1.08$ with a 0.02 interval. The results are vividly depicted in Figure 2.

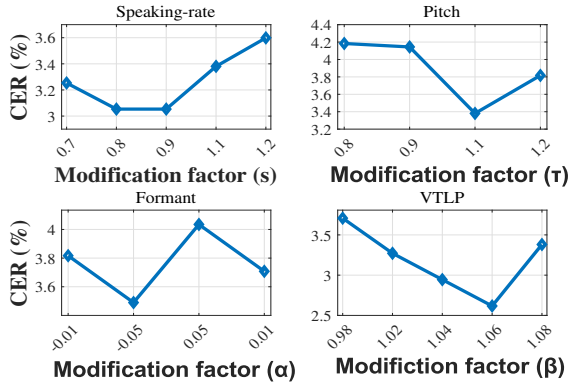

Fig. 2. Performance of data augmentation with varying modification factor.

It is evident from the Figure 2 that for s = 0.9, τ = 1.1, α = -0.05, and β = 1.06, we achieved the best results for overall severity classification in dysarthria speech. Detailed performance severity wise (very low/ low/ medium) with these best factors are tabulated in Table IV. Among the explored techniques, the augmentation with the VTLP modification method outperformed others, with the lowest CER of 2.62% for overall severity classification, and for severity-wise classification, CER is 1.59%, 5.11%, and 2.93% for very low, low, and medium severity in dysarthria speech, respectively.

D. Combination of Augmentation Techniques

Further investigating the impact of augmentations, the subsequent series of experiments explored various unique combinations of augmentation techniques for severity classification. As depicted in Table III, our boosted ANN base classifier served as the benchmark. Table V then highlights the combinations that exhibited notably superior performance compared to this baseline.

TABLE IV
THE CERS (%) FOR THE BEST MODIFICATION FACTOR FOR ALL FOUR AUGMENTATION METHODS (SPEAKING RATE, PITCH, FORMANT AND VTLP MODIFICATION) ARE TABULATED HERE. ALONG WITH OVERALL CER, CERS FOR EACH SEVERITY IS ALSO PRESENTED FOR THE BEST FACTOR.

Augmentation method	Modification factor	Overall CER(%)	Severity CER (%)		
			VL	L	M
Baseline	–	3.71	2.73	4.98	3.65
Orig. + Speaking Rate	s = 0.9	3.05	2.28	6.57	2.64
Orig. + Pitch	τ = 1.1	3.38	1.82	7.29	3.81
Orig. + Formant	α = -0.05	3.49	2.28	6.57	3.81
Orig. + VTLP	β = 1.06	**2.62**	1.59	5.11	2.93

TABLE V
PERFORMANCE COMPARISON OF VARIOUS AUGMENTATION INTEGRATION METHODS IS DELINEATED IN TERMS OF CER(%), WITH EMPHASIS ON THE RELATIVE IMPROVEMENTS (R. I.) FROM THE BASELINE.

Strategy for Augmentaion Combinations	Overall CER (%)	Severity CER (%)			R. I. (in %)
		VL	L	M	
Baseline	3.71	2.73	4.98	3.65	–
Orig. + Pitch + Speaking rate	3.49	1.59	5.84	4.99	8.57
Orig. + Pitch + Formant	3.27	2.96	8.76	1.47	14.29
Orig. + Speaking rate + Formant	**2.18**	1.59	5.11	1.76	**42.86**
Orig. + Pitch + VTLP	2.83	2.05	8.03	1.76	25.71
Orig. + Speaking rate + VTLP	2.94	1.37	6.57	3.52	22.86
Orig. + Formant + VTLP	3.05	2.51	6.57	2.35	20.00
Orig. + Pitch + Speaking rate + VTLP	3.49	3.19	8.76	1.76	8.57
Orig. + Pitch + Formant + VTLP	2.62	1.59	6.57	2.35	31.43
Orig. + Formant + Speaking rate + VTLP	2.40	1.82	5.11	2.05	37.14
Orig. + Pitch + Speaking rate + Formant	3.60	3.19	7.30	2.64	5.71

Within the array of combinations, the distinctive blend of Speaking Rate and Formant modifications, employing the optimal factors detailed in Section IV-C, showcased the most promising results. Augmenting the Original (Orig.) training set yielded the minimal error of 2.18% (CER) in accurately classifying severity levels, with a Relative Improvement (R.I) of 42.86% as compared to boosted baseline system. For severity-wise classification (very low/low/medium), this combination outperformed all others, achieving CERs of 1.59%, 5.11%, and 1.76%, respectively. One noticeable point is that the individual classification of the low-severity category showed no significant improvement. This may be attributed to two factors: firstly, the categorization of the severity of speaker M05, which is categorized as low but is actually low-medium; and secondly, the limited amount of data available in the test set for the low-severity class. The test set contains data in the ratios of 4:1:3 for very low, low, and medium severity, respectively. The modified Speaking rate and Formant were able to optimize the classifier by incorporating additional native temporal and spectral variations in the speech while addressing data scarcity. We also experimented on combination of three techniques in augmentation for the classification. One of such combination did show good R.I. of 37.14%, however the combination of Formant and Speaking Rate remained superior overall.

V. Conclusions

In this study, we strived to advance the development of robust severity classification systems for dysarthric speech. An in-depth exploration of machine learning techniques, acoustic /prosodic features and data augmentation methods are being presented in this work. Leveraging the TORGO dysarthric database, our study not only establishes a robust baseline using a unique feature set but also demonstrates the efficacy of augmenting the training data with techniques: Speaking Rate Modification, Pitch Modification, Formant Modification, and VTLP. Our proposed method of combining augmentation techniques for the task of severity classification boasts its effectiveness.

The analysis of different features showed that introduction of additional acoustic and prosodic features as input to classification model can be helpful in accurate classification of dysarthria severity levels. Evidence of it was when a boost in the system base performance was seen. On comprehensive analysis, we found that a combination of Speaking Rate and Formant modifications yields exceptional results, achieving the least overall error rate of 2.18% (CER) in severity classification, for their best modification factors which were obtained during individual augmentation experiments. In comparison to the baseline, our method achieved a 42.86% improvement in classier's performance. This improvement can be attributed to the modification method's contribution by introducing additional inter-speaker temporal and spectral variations leading classifier towards robustness. Further experiments showed that combining more than two techniques also has potential in improving the system performance, particularly Pitch + Speaking Rate + VTLP, exhibits balanced performance across severity levels. However, the combination of Formant and Speaking Rate remains best suited for the task in hand. This work contributes greatly to the field, providing a robust method of classifying severity with the proposed combination of augmentation techniques, addressing the challenges posed by the low-resource nature of dysarthric speech data.

References

[1] Daniel Kempler and Diana Van Lancker. Effect of speech task on intelligibility in dysarthria: A case study of parkinson's disease. *Brain and language*, 80(3):449–464, 2002.

[2] Catherine Mackenzie. Dysarthria in stroke: a narrative review of its description and the outcome of intervention. *International journal of speech-language pathology*, 13(2):125–136, 2011.

[3] Amlu Anna Joshy and Rajeev Rajan. Automated dysarthria severity classification using deep learning frameworks. In *28th European Signal Processing Conference (EUSIPCO)*, pages 116–120. IEEE, 2021.

[4] Kathryn M Yorkston, Edythe A Strand, and Mary RT Kennedy. Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology*, 5(1):55–66, 1996.

[5] Chitralekha Bhat, Bhavik Vachhani, and Sunil Kumar Kopparapu. Automatic assessment of dysarthria severity level using audio descriptors. In *ICASSP*, pages 5070–5074. IEEE, 2017.

[6] Amlu Anna Joshy and Rajeev Rajan. Automated dysarthria severity classification: A study on acoustic features and deep learning techniques. *Transactions on Neural Systems and Rehabilitation Engineering*, 30:1147–1157, 2022.

[7] Bassam Ali Al-Qatab and Mumtaz Begum Mustafa. Classification of dysarthric speech according to the severity of impairment: an analysis of acoustic features. *IEEE Access*, 9:18183–18194, 2021.

[8] Bhavik Vachhani, Chitralekha Bhat, and Sunil Kumar Kopparapu. Data augmentation using healthy speech for dysarthric speech recognition. In *Interspeech*, pages 471–475, 2018.

[9] Ayush Tripathi, Swapnil Bhosale, and Sunil Kumar Kopparapu. Improved speaker independent dysarthria intelligibility classification using deepspeech posteriors. In *ICASSP*, pages 6114–6118. IEEE, 2020.

[10] Bassam Ali Al-Qatab and Mumtaz Begum Mustafa. Classification of dysarthric speech according to the severity of impairment: an analysis of acoustic features. *IEEE Access*, 9:18183–18194, 2021.

[11] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46:523–541, 2012.

[12] Hemant K Kathania, S Shahnawazuddin, Waquar Ahmad, Nagraj Adiga, Sanjay Kumar Jana, and Arun B Samaddar. Improving children's speech recognition through time scale modification based speaking rate adaptation. In *International Conference on Signal Processing and Communications (SPCOM)*, pages 257–261. IEEE, 2018.

[13] Waquar Ahmad, Syed Shahnawazuddin, Hemant Kumar Kathania, Gayadhar Pradhan, and Arun B Samaddar. Improving children's speech recognition through explicit pitch scaling based on iterative spectrogram inversion. In *Interspeech*, pages 2391–2395, 2017.

[14] Hemant Kumar Kathania, Waquar Ahmad, Syed Shahnawazuddin, and Arun B Samaddar. Explicit pitch mapping for improved children's speech recognition. *Circuits, Systems, and Signal Processing*, 37:2021–2044, 2018.

[15] Xinglei Zhu, Gerald T. Beauregard, and Lonce Wyse. Real-time iterative spectrum inversion with look-ahead. In *2006 IEEE International Conference on Multimedia and Expo*, pages 229–232, 2006.

[16] Hemant Kumar Kathania, Sudarsana Reddy Kadiri, Paavo Alku, and Mikko Kurimo. A formant modification method for improved asr of children's speech. *Speech Communication*, 136:98–106, 2022.

[17] Alexander Johnson, Ruchao Fan, Robin Morris, and Abeer Alwan. Lpc augment: an lpc-based asr data augmentation algorithm for low and zero-resource children's dialects. In *IEEE ICASSP*, pages 8577–8581, 2022.

[18] Navdeep Jaitly and Geoffrey E Hinton. Vocal tract length perturbation (vtlp) improves speech recognition. In *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, volume 117, page 21, 2013.

[19] Kamil Lahcene Kadi, Sid Ahmed Selouani, Bachir Boudraa, and Malika Boudraa. Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge. *Biocybernetics and Biomedical Engineering*, 36(1):233–247, 2016.

[20] Kamil L Kadi and Sid Ahmed Selouani. 9 distinctive auditory-based cues and rhythm metrics to assess the severity level of dysarthria. *Signal and Acoustic Modeling for Speech and Communication Disorders*, 5:205, 2018.

[21] Siddharth Rathod, Monil Charola, and Hemant A Patil. Noise robust whisper features for dysarthric severity-level classification. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 708–715. Springer, 2023.

[22] Xinglei Zhu, Gerald T Beauregard, and Lonce L Wyse. Real-time signal estimation from modified short-time fourier transform magnitude spectra. *Transactions on Audio, Speech, and Language Processing*, 15(5):1645–1653, 2007.

[23] Yi Zheng and Romain Brette. On the relation between pitch and level. *Hearing research*, 348:63–69, 2017.

[24] Hemant Kumar Kathania, Sudarsana Reddy Kadiri, Paavo Alku, and Mikko Kurimo. Study of formant modification for children asr. In *IEEE ICASSP*, pages 7429–7433, 2020.

[25] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 346–348 vol. 1, 1996.

[26] Hemant Kumar Kathania, Viredner Kadyan, Sudarsana Reddy Kadiri, and Mikko Kurimo. Data augmentation using spectral warping for low resource children asr. *Journal of Signal Processing Systems*, 94(12):1507–1513, 2022.

[27] Stephanie Gillespie, Yash-Yee Logan, Elliot Moore, Jacqueline Laures-Gore, Scott Russell, and Rupal Patel. Cross-database models for the classification of dysarthria presence. In *Interspeech*, pages 3127–3131, 2017.