Fusion of data augmentation for improved dysarthria severity classification

Paban Sapkota[1], Bhanuja Karumuru[1†], Hemant Kumar Kathania[1*†], Sudarsana Reddy Kadiri[2†]

[1]Department of Electronics and Communication Engineering, National Institute of Technology Sikkim, Ravangla, India.
[2]Speech Analysis and Interpretation Laboratory (SAIL), University of Southern California, Los Angeles, USA.

*Corresponding author(s). E-mail(s): hemant.ece@nitsikkim.ac.in;
Contributing authors: phec230006@nitsikkim.ac.in;
b200087@nitsikkim.ac.in; skadiri@usc.edu;
†These authors contributed equally to this work.

Abstract

Currently, there are very few corpora available online for dysarthric speech. Collecting large amounts of dysarthric speech data poses significant challenges, including locating individuals with dysarthria and recording their speech in a suitable setting. To address the issue of data scarcity, this paper proposes fusion of different data augmentation techniques: speaking-rate modification, pitch modification, formant modification, and vocal-tract length perturbation (VTLP). Our baseline setup employs an artificial neural network (ANN) model with mel-frequency cepstral coefficients (MFCC) features. A suitable feature set was identified through feature-based analysis in a speaker-dependent manner to enhance the baseline. Subsequently, each of the four augmentation methods was explored by varying their respective modification factors to find the optimal modification factor for each method individually. Following this, combinations of augmentation techniques were examined to determine the best augmentation strategy. Later, the study was extended for the speaker-independent settings as well. The findings indicate that augmentation can significantly aid in the severity classification of dysarthria, especially addressing the data scarcity issue. Furthermore, the classification performance improves with the proposed fusion of multiple different augmentation techniques. With the proposed data augmentation fusion methods, relative improvements of 42.86% and 12.22% were observed in the TORGO database, and 18.78% and 5.29% in the UAspeech database, for

speaker-dependent and speaker-independent settings, respectively, in the task of classifying dysarthria severity levels.

Keywords: Dysarthria, data-scarcity, severity, data augmentation, speaker-independent

1 Introduction

Dysarthria, a motor speech disorder, profoundly impacts the ability to produce coherent and fluent speech [1]. Gathering data on dysarthric speech is challenging due to its scarcity. Accurate classification of dysarthric speech severity is essential for guiding clinical interventions [2] and designing personalized treatment plans [3] tailored to individual severity levels. Adjusting the clinical approaches based on dysarthria severity enables continuous monitoring and adaptation to changing speech patterns, ensuring effective treatment [4]. Moreover, robust classification systems for dysarthria severity can assist in customizing assistive speech devices to meet diverse speech challenges [5].

The literature extensively examined various machine learning (ML) techniques designed to automate tasks related to severity classification [6–8]. For instance, leveraging deepspeech logits as features has proven effective in achieving high classification accuracy [9]. Another study has shown notable improvements through the integration of prosody and spectral acoustic features for severity classification [7]. In [10], perceptually enhanced Fourier transform spectrograms and Constant-Q transform spectrograms were studied, demonstrating superior performance in the dysarthric severity classification task. Additionally, [11] analyzed the suitability of phonation, articulation, prosody, and glottal features, highlighting that prosody and articulation features are particularly effective for dysarthria severity estimation. Furthermore, there is a clear shift towards employing neural network architectures instead of traditional ML-based classifiers to improve severity classification models [5]. The data augmentation techniques have demonstrated significant utility in the field of dysarthric speech recognition. For example, works like [8, 12] employed temporal and speed perturbation-based data augmentation using healthy speech showing improved performance. Additionally, studies in [13, 14] utilized speaker-dependent data augmentation approaches to model spectro-temporal differences between dysarthric and healthy speakers, contrarily, another study [15] emphasized that the speed perturbation based data augmentation method outperformed vocal tract length perturbation in dysarthric speech recognition tasks. However, there is limited work that includes the implementation and analysis of augmentation techniques for dysarthric speech severity classification.

The existing research highlights a notable gap in addressing data scarcity issues and the exploration of data augmentation techniques in the task of dysarthria severity classification. Therefore, motivated by this, we investigated four prominent data augmentation techniques in our work: speaking rate modification [16], pitch modification [17, 18], formant modification [19], and vocal tract length perturbation [20, 21] to mitigate issues related to data scarcity and introduce more variability into the speech

database. First we have enhanced the baseline classification performance by incorporating additional acoustic and prosodic features alongside the baseline MFCC features [7] for TORGO [22] and UAspeech [23] dysarthria speech databases. Subsequently, we implemented a combination of these data augmentation methods for the severity level classification task. This comprehensive strategy aims to significantly improve the accuracy and reliability of dysarthria severity classification, addressing an area that warrants further exploration.

The key highlights of this paper are outlined below:

- We have explored combinations of acoustic and prosodic features to enhance the baseline system for classifying severity levels (very low, low, medium, and high) using the TORGO and UAspeech dysarthria speech databases.
- We investigate four prominent augmentation techniques in speaker-dependent scenario—speaking rate modification, pitch modification, formant modification, and vocal tract length perturbation (VTLP) to address data scarcity and enhance the performance of the severity classifier.
- We proposed fusion of augmentation methods, resulting in a robust classification system tailored for the TORGO and UAspeech dysarthria speech databases.
- Further, proposed fusion of augmentation methods was also explored in a speaker-independent manner, aiming to enhance the robustness of the classifier against variability in speech characteristics across individuals for both the databases.

The paper's subsequent sections are arranged as follows: Section 3.2 provides a comprehensive description of the database used in this study, emphasizing its relevance to the task. Section 2 describes the data augmentation techniques applied, including modifications of speaking rate, pitch, and formants, as well as the use of vocal tract length perturbation (VTLP). In Section 3, we outline the experimental setup, and Section 4 presents an in-depth analysis of acoustic and prosodic feature utilization to improve the baseline classifier. This section also examines the impact of various data augmentation techniques and introduces our proposed combination strategy to enhance the robustness of the severity classification system. Finally, Section 5 summarizes the key findings and discusses their potential implications for future research and applications.

2 Data Augmentation Strategies

This section describes the four data augmentation methods investigated in this study: speaking rate modification, pitch modification, formant modification, and vocal tract length perturbation. A simplified block diagram of the overall process is presented in Figure 1. In the following subsections, we detail the four types of speech modification techniques utilized in the data augmentation process of this study.

2.1 Speaking-rate Modification

The implementation of speaking rate modification employed Time Scale Modification (TSM) using the Real-Time Iterative Spectrogram Inversion Look-Ahead (RTISI-LA) algorithm [16, 24, 25]. A scaling factor $\mathtt{s}$, where $0.5 \leq s \leq 2$, was systematically varied

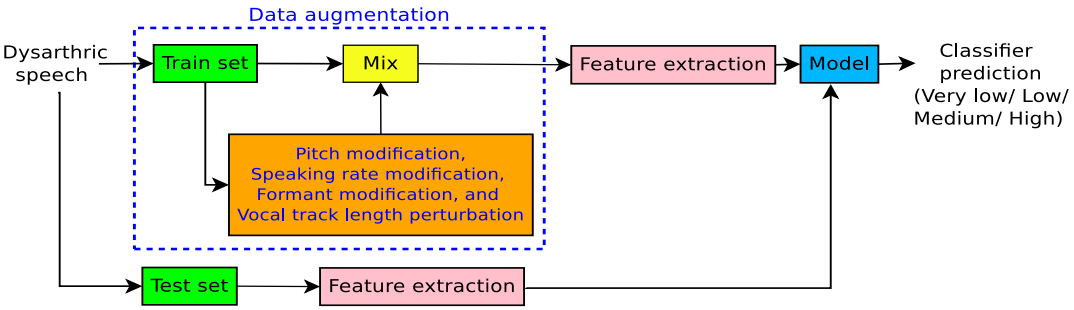

Fig. 1 A general block diagram illustrating the implementation process of data augmentation techniques.

during the experimentation in increments of 0.1. The range of s was selected based on perceptual constraints and prior studies, where values below 0.5 introduce severe temporal distortion and intelligibility loss, while values above 2 result in unnatural speech stretching and artifacts. This algorithm processes the audio signal through Short-Time Fourier Transform Modification (STFTM) [26]. During processing, each frame of the input signal is iteratively reconstructed using a window function [27]. The modified frame length, $L = 256 \cdot \mathtt{s}$, where 256 denotes the baseline frame length (in samples) used in the STFT analysis, chosen as a trade-off between time and frequency resolution, and the corresponding hop size influences the STFT process, thereby affecting the overall modification of the signal. The resulting modified audio is subsequently utilized for further augmentation. A critical aspect of this process is the introduction of phase perturbation [27], which effectively mitigates resonance effects during reconstruction.

2.2 Pitch Modification

The pitch modification technique employed in this study utilized the Real-Time Iterative Spectrogram Inversion Look-Ahead (RTISI-LA) algorithm, as described in [18]. This algorithm allows for pitch adjustments based on a scaling factor τ. The parameter τ was varied within the range of 0.5 to 2 with a step size of 0.1. This method employs Short-Time Fourier Transform Modification (STFTM) for processing audio signals. Initially, frame length and hop size are set, and the semitone value is converted into a scaling factor with their mathematical relation, $\tau = 2^{\left(\frac{n}{12}\right)}$, where '$n$' represents the specified semitone value. The algorithm iterates through the signal frames, applying a modified window function and reconstructing the signal iteratively to achieve the desired pitch modification. The window function is adjusted based on the modified frame length, which is influenced by the semitone value, thereby contributing to the resampling process. These adjustments are essential for tailoring the tonal characteristics of the output audio in accordance with the specified semitone value. Notably, the choice of semitone value significantly affects the iterative signal reconstruction process [25, 27]. The resulting modified audio is subsequently utilized for augmentation purposes.

2.3 Formant Modification

Formant modification is a critical technique in speech signal processing, utilized to manipulate the spectral characteristics of an audio signal by adjusting its formants [19]. Formants correspond to the resonant frequencies within the speech spectrum, significantly influencing the quality and timbre of the produced sound. This technique employs Linear Prediction (LP) models to analyze short-time segments of the speech signal.

LP analysis is a well-established method in speech processing that models the spectral envelope of a signal using an all-pole filter [28]. The coefficients of this filter, referred to as LP coefficients, capture the resonant characteristics of the signal. These LP coefficients are estimated using the autocorrelation method, which solves the normal equations via the Levinson–Durbin recursion [29, 30]. In the context of formant modification, these LP coefficients are crucial as they are adjusted to achieve the desired alterations in the spectral envelope.

In the specific context of children's speech modification, previous studies have demonstrated the efficacy of formant modification techniques using LP analysis [19]. This involves applying LP analysis to short-time segments of the signal and manipulating the poles of the LP model through a warping function parameterized by α ($-1 \leq \alpha \leq 1$ with a step size of 0.05). The step size of 0.05 is empirically chosen to ensure gradual and controlled modification of formant frequencies while maintaining computational efficiency. The optimal value of α is selected by evaluating performance across this range to avoid unstable spectral distortions. This adjustment allows systematic shifting of the formant frequencies, thereby altering the spectral characteristics of the speech signal. The modified LP coefficients resulting from this process are then used to synthesize the modified signal.

Key steps in the formant modification algorithm include:

- Performing LP analysis on short-time segments of the audio signal, using the formula, given in [19] as:

$$\hat{S}(z) = \left(\sum_{k=1}^{P} a_k z^{-k} \right) S(z), \tag{1}$$

  where $\hat{S}(z)$ and $S(z)$ denote the Z-transforms of the prediction and original speech signal, respectively, and a_k are the LP coefficients. Here, P denotes the order of the LP model, i.e., the number of past samples used for prediction.

- Modifying LP coefficients through pole warping using a warping factor α. The unit delay filters are replaced by an all-pass filter $D(z)$ to warp the frequency scale [31], using a first-order filter,

$$D(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \tag{2}$$

  where, α is the warping factor, with a value in the range of $-1 < \alpha < 1$.
- Synthesis of the modified speech signal using the adjusted LP coefficients.

2.4 Vocal Tract Length Perturbation

The configuration of the vocal tract varies significantly among different speakers, resulting in considerable inter-speaker variability in speech characteristics. To address this variability, Vocal Tract Length Normalization (VTLN) was initially introduced as a technique to standardize vocal tract configurations [32]. In contrast, Vocal Tract Length Perturbation (VTLP) is employed to introduce controlled variability into speech data by simulating different vocal tract lengths.

The VTLP process begins with a time-domain audio segment $x(t)$, which is transformed into the frequency domain as $X(f)$ through its Fourier transform. VTLP then applies a perturbation factor β along the frequency axis of $X(f)$, altering the spectral envelope to produce an output $Y(f) = X(\beta f)$. Here, β is chosen from a discrete set where values such as $0 \leq \beta < 1$ simulate shorter vocal tracts, while $\beta > 1$ simulates longer vocal tracts. This technique preserves the original audio duration while modifying the spectral characteristics of the audio segment.

In our experiments, we explored perturbation factors ranging from 0.98 to 1.08, using a step size of 0.02, following the strategy adopted in prior studies [33, 34]. This range was empirically selected to introduce realistic and perceptually acceptable variations in vocal tract length. Values close to 1 ensure that the modified speech remains natural, while small deviations simulate mild inter-speaker anatomical differences. Larger deviations were avoided as they lead to excessive spectral distortion and reduced intelligibility. This approach enables VTLP to effectively introduce controlled variations in speech signals, specifically targeting the spectral characteristics associated with different vocal tract lengths.

3 Experimental Setup and Classifier Configuration

In the speaker-dependent studies, the dataset was divided into 80% for training the classifier and 20% for testing. This split was conducted in a way that maintained the 80:20 ratio across all severity levels. The randomized stratification method was employed to ensure that stratification was based directly on the severity level, preserving both the proportional representation and the randomness in the selection of samples from each class in both the training and testing sets.

As an evaluation metric, we have chosen the classification error rate (CER). Along with the overall class CER, we have incorporated the performances of each of the severity class individually. For example, k number of samples are in the test set in total, where vl, l, m, and h are number of samples from very low (VL), low (L), medium (M), and high (H) severity respectively,

$$CER_{(\text{Overall})} = \frac{k - (\text{no. of correct predictions})}{k} \times 100 \tag{3}$$

Equation (3) gives the overall CER. The individual CER for Very Low severity was calculated as

$$CER_{(\text{VL})} = \frac{vl \text{ - (no. of correct predictions from } vl)}{vl} \times 100. \tag{4}$$

Similarly, the CER for the other classes were computed in the same manner. This approach allows us to assess both the overall performance and the performance for each severity level.

3.1 Classification Model Details

Baseline experiments were conducted using four different classifiers: support vector machine (SVM), K-nearest neighbors (KNN), random forest (RF), and artificial neural networks (ANN). A 13-dimensional set of Mel-frequency cepstral coefficients (MFCC) features was selected for the initial analysis.

Prior to training, all feature sets were standardized using z-score normalization to ensure zero mean and unit variance. The normalization parameters were estimated on the training set and consistently applied to validation and test sets to prevent data leakage and ensure stable convergence across models.

Classical Machine Learning Classifiers: From classical models, three classifiers were employed to provide diverse decision-making perspectives. A linear Support Vector Machine (SVM) [35] was adopted to balance classification performance and computational efficiency, with the regularization parameter optimized via 5-fold cross-validation, yielding an optimal value of $C = 2$. The K-Nearest Neighbors (KNN) [36] classifier was utilized as a non-parametric baseline to capture local decision boundaries, where the number of neighbors was tuned in the range of 10–100 using 5-fold cross-validation; feature standardization ensured consistent distance computation across dimensions. Furthermore, a Random Forest (RF) [37] classifier was implemented as an ensemble of bootstrap-aggregated decision trees, with the number of trees optimized between 10 and 200. This ensemble approach improves generalization by reducing variance and enhances robustness against overfitting in heterogeneous speech feature representations.

Artificial Neural Network (ANN) Classifier [38]: The ANN serves as the primary classifier for the augmentation-based study and is designed to model non-linear relationships in speech features. The network consists of a fully connected feedforward architecture with an input layer of dimension 13, followed by five hidden layers comprising 13, 128, 64, 32, and 16 neurons, respectively. Each hidden layer employs the Rectified Linear Unit (ReLU) activation function to introduce non-linearity and facilitate hierarchical feature learning. The output layer is configured with neurons corresponding to the number of severity classes (three for TORGO and four for UASpeech), and utilizes a softmax-based probabilistic formulation implemented via sparse categorical cross-entropy loss. Model optimization is performed using the Adam optimizer, enabling adaptive learning rates and stable convergence during training. The network is trained for 100 epochs with validation monitoring to assess generalization performance. During inference, class probabilities are obtained from the output layer, and final predictions are generated using the argmax operation. In addition to overall classification accuracy, label-wise accuracies are computed and averaged to provide a balanced evaluation across severity classes.

The chosen architecture reflects a trade-off between representational capacity and computational efficiency, making it suitable for extensive augmentation-based experimentation while maintaining stable training behavior.

The output layer consisted of three nodes for the TORGO dataset [22] and four nodes for the UAspeech dataset [23], corresponding to the number of severity classes in their respective databases. The description of the databases is presented in the following subsection.

3.2 Database Description

In this study, we utilized two of the most widely used and well-established dysarthric speech databases in research: the TORGO [22] and UASpeech [23] database.

3.2.1 TORGO Dysarthric Database

The TORGO database was developed through a collaboration between the Computer Science and Speech-Language Pathology departments at the University of Toronto and the Holland-Bloorview Kids Rehabilitation Hospital [22]. It contains recordings from eight individuals with speech impairments (three women and five men) and seven without impairments (three women and four men), with participants aged between 16 and 50 years. Recordings were captured using two microphone types: an array microphone at a 44.1 kHz sampling rate and a head-mounted electret microphone at 16 kHz. For consistency, all acoustic signals were down-sampled to 16 kHz across the dataset [39].

The TORGO database utilizes the standardized Frenchay Dysarthria Assessment (FDA) to assess dysarthria severity [22]. This assessment examines 28 perceptual dimensions of speech, including reflexes, respiration, facial and oral movements, and intelligibility. Experienced speech-language pathologists administer the FDA, providing each dysarthric speaker with an overall score [40]. Based on these scores, severity levels—Very Low, Low, Medium, and High—are assigned, offering a standardized measure of dysarthria severity suitable for both research and clinical use [22].

3.2.2 UAspeech Dysarthric Database

The UAspeech database is one of the most comprehensive public resources for dysarthric speech, containing recordings from individuals with dysarthria and matched control participants. This dataset includes recordings from 28 speakers: 15 individuals diagnosed with cerebral palsy and 13 age-matched healthy controls.

The dataset features both read and spontaneous speech samples, offering a realistic representation of everyday communication. Each participant recorded 765 isolated words, organized into three sets (A, B, and C), with 155 words common to all sets. These include 19 computer commands, 26 radio-alphabet letters, 10 digits, and the 100 most frequent words from the Brown corpus of written English. An additional 100 words per set were selected from Project Gutenberg novels [23].

The TORGO corpus includes approximately 15 hours of recordings, with around 6 hours from dysarthric speakers and 9 hours from healthy speakers. In comparison, the UAspeech corpus contains roughly 100 hours of recordings, comprising about 66 hours from dysarthric speakers and 34 hours from control speakers, all sampled at 16 kHz. These corpora provided valuable resources for our research on dysarthric speech.

Severity level distributions for dysarthric speakers in both the TORGO and UAspeech databases are presented in Table 1.

Table 1 Description of severity levels for dysarthric speakers in the TORGO [22] and UAspeech [23] corpora. The letters "M" and "F" represent male and female speakers, respectively.

Severity	**Speaker-ID**	
Level	**TORGO**	**UAspeech**
Very Low	F03, F04, M03	F05, M08, M09, M10, M14
Low	F01, M05	F04, M05, M11
Medium	M01, M02, M04	F02, M07, M16
High	-	F03, M01, M04, M12

4 Results and Discussion

This section presents the results of the data augmentation experiments, with separate analyses for the speaker-dependent and speaker-independent cases in Sections 4.1 and 4.2, respectively. The speaker-dependent studies were conducted in detail to identify the optimal modification factors for each of the four augmentation techniques. These optimal factors were subsequently applied in the speaker-independent experiments.

4.1 Speaker-Dependent Results

The speaker-dependent study commenced with the careful selection of the baseline classifier system (Section 4.1.1). This was followed by the implementation of individual augmentation techniques (Section 4.1.2). Subsequently, combinations of augmentation methods were explored (Section 4.1.3). The results of the speaker-dependent experiments utilizing both the TORGO and UAspeech databases are presented concurrently.

4.1.1 Baseline: MFCC versus Feature-Set

Baseline experiments were conducted using four different classification models: SVM, KNN, RF, and ANN, with 13-dimensional MFCC features as the initial base features. For both the TORGO and UAspeech databases, the ANN outperformed the SVM, KNN, and RF classifiers, achieving CERs of 6.32% and 6.98%, respectively.

Building on previous studies [5, 41], we enhanced baseline performance by integrating various acoustic and prosodic features alongside MFCC features. The optimal combination comprised the mean values of Zero Crossing Rate (ZCR), Spectral Centroid, Spectral Entropy, Spectral Crest, Spectral Flatness, Spectral Rolloff, Pitch, Root Mean Square Energy (RMSE), and Log Energy, in addition to MFCC features. A summary of the considered features and their respective types is provided in Table 2. This resulted in a comprehensive 22-dimensional feature set (13 MFCC + 9 additional features) that significantly improved the representation for the severity classification

Table 2 List of features combined with MFCCs to enhance baseline classification performance. The selected features were chosen based on their proven effectiveness in prior research studies [5] and [41].

Type	Temporal	Spectral	Prosodic
Features	ZCR, RMSE, Log Energy	Spectral Centroid, Spectral Entropy, Spectral Crest, Spectral Flatness, Spectral Rolloff	Pitch

task. The explored feature combinations culminated in a distinct set of features, which led to enhanced baseline performances across all four classifier models. The classification performance, assessed in terms of Classification Error Rate (CER), for both the baseline and the improved systems, is detailed in Table 3.

Table 3 Classification results (CER %) for the explored classifiers, comparing baseline CERs of SVM, KNN, RF, and ANN models (lower CERs indicate better classification performance). The evaluation is based on MFCC features and a combined feature set for both TORGO and UAspeech datasets. "Clx." refers to the classifier used, and "Ovrl." represents the overall classification CER.

Clx.	TORGO								UAspeech									
	Baseline				Improved				Baseline					Improved				
	Ovrl.	VL	L	M	Ovrl.	VL	L	M	Ovrl.	VL	L	M	H	Ovrl.	VL	L	M	H
SVM	21.16	15.16	11.89	15.27	17.23	11.67	9.49	13.30	25.48	14.02	35.77	30.46	29.89	21.28	11.38	26.71	27.45	26.55
KNN	7.09	5.56	4.14	4.47	7.85	6.32	3.60	5.78	7.37	4.99	3.53	2.75	3.45	3.76	2.20	1.84	1.53	1.95
RF	7.31	5.56	3.93	5.13	6.22	4.47	3.05	4.91	9.97	6.74	4.80	3.88	4.53	4.96	3.04	2.32	2.10	2.51
ANN	**6.32**	3.92	3.92	4.68	**3.71**	2.73	4.98	3.65	**6.98**	6.82	8.67	6.10	6.48	**2.29**	1.82	2.79	1.83	3.02

From Table 3, we can observe that there are notable improvements in classification performance when compared to using MFCC features alone. This indicates that MFCC features may not sufficiently capture the severity level information from speech samples of dysarthric speakers. Incorporating features from both spectral and prosodic domains significantly enhances severity classification performance. Among the classifiers, the Artificial Neural Network (ANN) consistently outperformed the others, even after the baseline was enhanced with prosodic and spectral features. Specifically, the ANN achieved the lowest CERs of 3.71% for the TORGO dataset and 2.29% for the UAspeech dataset. Building on these findings, the study then proceeded with data augmentation, using these results as the baseline.

4.1.2 Individual Augmentation Results

The individual augmentation experiments involved varying the modification factors within their specified ranges, as detailed in Section 2. The speaking rate (SR) scaling factor and pitch modification (PM) factor (s and τ) were adjusted between $0.7 \leq s \leq$

1.2, while the formant modification (FM) warping factor (α) was incremented in steps of 0.05 within the range from -0.01 to 0.01. For the VTLP experiment, the vocal tract length perturbation factor (β) was explored within the range $0.98 \leq \beta \leq 1.08$. The resulting CERs were plotted against these modification factors for both the TORGO and UAspeech datasets, as depicted in Figure 2. We can observe from the Figure

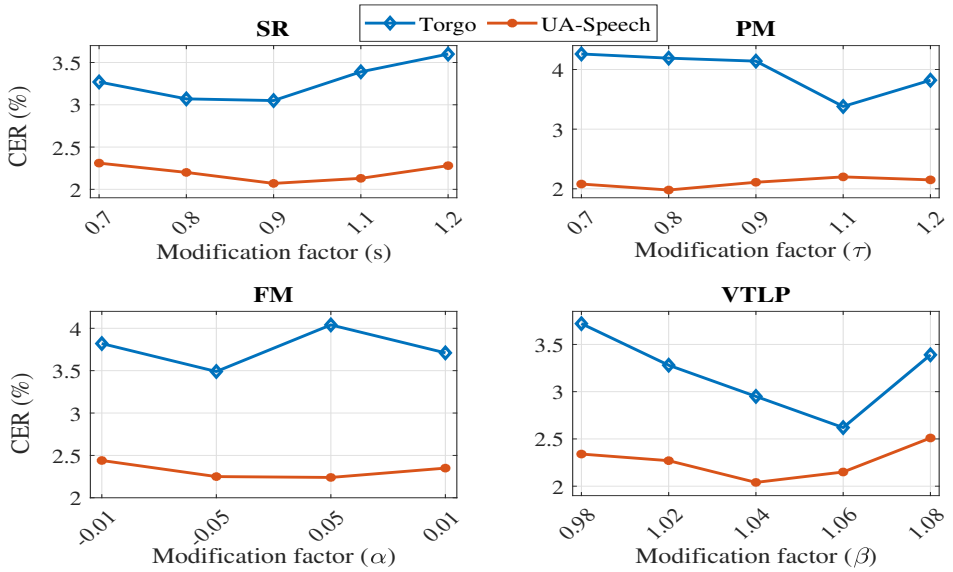

Fig. 2 Performance of data augmentation methods with variations in modification factor on TORGO and UAspeech databases. Here, the modification factors s, τ, α, and β are the scaling factor, pitch modification factor, warping factor, and vocal tract length perturbation factor, respectively.

2, for the TORGO database, the modification factors s=0.9, τ =1.1, α =-0.05, and β =1.06 were found to be the best factors. In case of the UAspeech database, the best modification factors for data augmentations were found to be s=0.9, τ =0.8, α =-0.05, and β =1.04. The detailed results, for these modification factors, in terms of CERs is shown in Table 4.

Performances of the four methods in the Table 4 indicate that all data augmentation techniques were effective, outperforming the baseline CER. Among the methods, Vocal Tract Length Perturbation (VTLP) achieved the best results for the TORGO dataset, with a CER of 2.62%, reflecting a relative improvement of 29.38% over the baseline. In contrast, Pitch Modification (PM) excelled with the UAspeech dataset, achieving a CER of 1.98%, which corresponds to a relative improvement of 13.54% compared to the baseline. Utilizing these optimal modification factors from each augmentation method, we proceeded to conduct combination experiments within this study.

Table 4 The CERs (%) for the best modification factor for all four augmentation methods (speaking rate, pitch, formant, and VTLP modification) are tabulated here. 'T' and 'UA' stands for the TORGO and the UAspeech databases, respectively. The parameters 's', 'τ', 'α', and 'β' represent the modification factors for speaking rate, pitch, formant, and VTLP, respectively. "Ovrl." represents the overall classification CER.

Aug.	Mod. Factor		TORGO				UAspeech				
	T	UA	Ovrl.	VL	L	M	Ovrl.	VL	L	M	H
Baseline	–	–	3.71	2.73	4.98	3.65	2.29	1.82	2.79	1.83	3.02
Ori.+SR (s)	0.9	0.9	3.05	2.28	6.57	2.64	2.07	1.34	3.01	2.27	2.21
Ori.+PM (τ)	1.1	0.8	3.38	1.82	7.29	3.81	**1.98**	1.72	2.7	1.48	2.21
Ori.+FM (α)	-0.05	-0.05	3.49	2.28	6.57	3.81	2.25	2.14	2.14	2.35	2.45
Or.+VTLP (β)	1.06	1.04	**2.62**	1.59	5.11	2.93	2.04	1.37	2.96	1.83	2.45

4.1.3 Augmentation Combination Results

Initially, experiments were conducted with combinations of two augmentation methods, followed by investigations involving combinations of three methods. Finally, we explored the combination of all four methods to determine the most effective augmentation strategy. The modification factors selected for these experiments were based on those identified as optimal for each of the four methods, as detailed in Table 4. The results of the augmentation combination experiments are presented in Table 5.

Table 5 Performance comparison of various augmentation integration methods is presented in terms of CER(%), with emphasis on the relative improvements (R. I.) from the baseline for both TORGO and UA Speech datasets. The parameters 's', 'τ', 'α', and 'β' here represent the best modification factors (from the results in Section 4.1.2) for speaking rate, pitch, formant, and VTLP, respectively.

Strategic Combination of	TORGO					UAspeech					
Augmentation Methods	Overall	VL	L	M	R. I.	Overall	VL	L	M	H	R. I.
Baseline	3.71	2.73	4.98	3.65	–	2.29	1.82	2.79	1.83	3.02	–
Orig. + τ + s	3.49	1.59	5.84	4.99	8.57	2.11	1.44	2.31	2.22	2.9	7.86
Orig. + τ + α	3.27	2.96	8.76	1.47	14.29	2.18	1.07	3.31	3.01	2.17	4.80
Orig. + s + α	**2.18**	1.59	5.11	1.76	**42.86**	2.06	1.72	2.79	2.00	2.01	10.04
Orig. + τ + β	2.83	2.05	8.03	1.76	25.71	**1.86**	1.07	2.48	2.31	2.13	**18.78**
Orig. + s + β	2.94	1.37	6.57	3.52	22.86	1.96	1.62	2.66	1.48	2.29	14.41
Orig. + α + β	3.05	2.51	6.57	2.35	20.00	2.07	1.79	3.09	1.26	2.33	9.61
Orig. + τ + s + β	3.49	3.19	8.76	1.76	8.57	2.03	1.52	2.79	2.00	2.17	11.35
Orig. + τ + α + β	2.62	1.59	6.57	2.35	31.43	1.92	1.05	3.27	1.26	2.7	16.16
Orig. + α + s + β	2.40	1.82	5.11	2.05	37.14	2.14	1.69	3.79	1.83	1.61	6.55
Orig. + τ + s + α	3.60	3.19	7.30	2.64	5.71	2.15	1.99	2.83	1.66	2.25	6.11
Orig. + τ + s + α + β	2.73	1.59	7.30	2.35	26.47	1.86	1.39	2.7	1.26	2.37	18.78

Following the strategic combination of augmentation methods, we identified the optimal combination for minimizing the CER on the TORGO dataset as the pairing of speaking rate modification (s=0.9) and formant modification (α=-0.05). This combined augmentation achieved a CER of 2.18%, reflecting a substantial relative improvement of 42.86% from the baseline. The results from the TORGO dataset indicate that combining methods that perform well individually does not necessarily lead to improved outcomes. Although VTLP augmentation demonstrated superior performance on its own, it did not surpass the effectiveness of the combination of speaking

rate modification and formant modification when integrated with other methods. Notably, this trend did not apply to the UAspeech dataset.

The augmentation combination experiments on the UAspeech dataset yielded optimal results with the combination of pitch modification (τ=0.8) and Vocal Tract Length Perturbation (VTLP) (β=1.04) methods. This combination achieved the lowest CER of 1.86%, representing an 18.78% relative improvement over the baseline. The strong individual performances of these augmentation methods in separate experiments further emphasize the effectiveness of their combination for enhanced overall performance.

4.2 Speaker-Independent Results and Analysis

To evaluate the robustness of the modification factors determined from the speaker-dependent experiments, we strategically applied them in speaker-independent scenarios. The dataset was divided into training and testing sets, ensuring that one speaker from each severity level was included in the test set, while the remaining speakers were utilized for training.

Table 6 The distribution of speakers into training and testing sets into distinct data splits (D.S.) for the speaker-independent experimentation on the TORGO and UAspeech databases. 'VL', 'L', 'M', and 'H' denote severity levels Very Low, Low, Medium, and High, respectively. Three unique combinations were necessary to cover all eight speakers in the TORGO database, whereas a minimum of five combinations were required to encompass all fifteen speakers in the UAspeech database test set.

Database	Setup	Training Data				Testing Data			
		VL	L	M	H	VL	L	M	H
TORGO	D.S.-I	F04, M03	M05	M02, M04	-	F03	F01	M01	-
	D.S.-II	F03, M03	F01	M01, M04	-	F04	M05	M02	-
	D.S.-III	F03, F04	F01	M01, M02	-	M03	M05	M04	-
UAspeech	D.S.-I	M09, M10, M14, F05	F04, M11	F02, M16	F03, M04, M12	M08	M05	M07	M01
	D.S.-II	M08, M10, M14, F05	M05, M11	F02, M07	F03, M01, M12	M09	F04	M16	M04
	D.S.-III	M08, M09, M14, M10	F04, M05	M07, M16	F03, M04, M12	F05	M11	F02	F03
	D.S.-IV	M08, M09, M14, F05	M11, M05	M07, M16	M01, M04, M12	M10	M11	M07	M12
	D.S.-V	M08, M09, M10, F05	M11, M05	M07, F02	M01, M12, F03	M14	F04	M16	M04

For the TORGO dataset, a minimum of three splits was required to guarantee representation of all dysarthric speakers in the test sets. In contrast, the UAspeech dataset necessitated at least five splits for comprehensive coverage. We selected the combined feature set based on prior studies and opted for the artificial neural network

(ANN) model as our classifier for the baseline. The distribution of speakers across these splits is outlined in Table 6.

To ensure coverage of all dysarthric speakers, we conducted speaker-independent analyses on both the TORGO and UAspeech databases.

The modification factors for data augmentation used in these combination experiments are s=0.9, τ=1.1, α=-0.05, and β=1.06, for the TORGO database, whereas s=0.9, τ=0.8, α=-0.05, and β=1.04 were utilized for the UAspeech database. These augmentation factors were selected based on the results of individual augmentation studies conducted for the speaker-dependent cases, as detailed in Section 4.1.2. The results of the augmentation combination experiments with TORGO and UAspeech databases are presented in Table 7.

Table 7 Comparison of overall CER under different augmentation combinations across three data splits (TORGO) and five data splits (UASpeech) datasets. Relative improvement over baseline: 11.06% (TORGO) and 1.55% (UASpeech). τ: pitch modification factor, s: speaking rate scale factor, α: formant warping factor, β: VTLP factor.

Combination	TORGO				UAspeech					
	D.S.-I	D.S.-II	D.S.-III	Avg.	D.S.-I	D.S.-II	D.S.-III	D.S.-IV	D.S.-V	Avg.
Baseline	55.82	51.16	47.67	51.55	66.17	62.35	59.57	6.06	76.71	54.17
Orig. + τ + s	59.27	47.03	46.37	50.89	65.92	64.11	59.83	7.09	81.21	55.63
Orig. + τ + α	56.14	47.50	**39.22**	47.62	66.73	63.86	**54.70**	6.06	79.86	54.24
Orig. + s + α	66.65	47.87	45.19	53.24	63.47	67.79	58.17	7.12	84.21	56.15
Orig. + τ + β	59.99	50.88	43.56	51.48	67.39	59.81	60.25	6.72	79.02	54.64
Orig. + s + β	59.08	48.42	45.07	50.86	66.78	64.86	57.35	7.79	82.58	55.87
Orig. + α + β	**55.01**	45.32	46.65	48.99	63.18	65.61	57.45	6.46	**76.21**	53.78
Orig. + τ + s + β	57.32	49.44	48.15	51.64	67.25	66.21	55.75	5.90	78.59	54.74
Orig. + τ + α + β	56.77	**41.52**	39.26	**45.85**	**60.83**	67.15	56.76	**5.22**	83.66	54.72
Orig. + α + s + β	61.98	52.22	42.87	52.36	63.26	**59.57**	54.81	5.37	83.65	**53.33**
Orig. + τ + s + α	61.35	49.72	44.74	51.94	65.43	62.79	59.35	7.32	83.86	55.75
Orig. + τ + s + β + α	64.3	48.7	42.02	51.67	63.93	62.62	57.35	6.72	80.82	54.29

Performance on the TORGO database yielded an average Classification Error Rate (CER) of 45.25%, reflecting a relative improvement of 12.22% over the baseline average CER of 51.55%. As illustrated in Table 7, various data splits favored different augmentation combinations. For instance, in Data Split I (D.S.-I), the combination of formant modification and Vocal Tract Length Perturbation (VTLP) proved most effective. In Data Split II (D.S.-II), the best results were achieved with a combination of pitch modification, formant modification, and VTLP. In Data Split III (D.S.-III), the combined augmentation of pitch modification and VTLP led to a significant reduction in CER compared to the baseline.

Similar trends were observed in the UAspeech database. In D.S.-I, the combination of pitch, formant, and VTLP modifications yielded the lowest CER. For D.S.-II, the most effective combination was formant modification, speaking rate modification, and VTLP. D.S.-III demonstrated the best results with the combination of pitch and formant modifications. Likewise, Data Splits IV and V yielded optimal results with specific combinations: pitch, formant, and VTLP for D.S.-IV, and formant and VTLP for D.S.-V. Averaging the individual CERs across the five splits resulted in a CER of 51.31%, representing a 5.29% relative improvement over the baseline.

These results indicate that the proposed augmentation techniques not only perform well in classifying the severity levels of dysarthria in speaker-dependent scenarios but also significantly enhance dysarthria severity classification in speaker-independent cases.

4.3 Computational Cost Analysis

In addition to classification performance, we analyze the computational efficiency of the proposed approach in terms of training time, testing time, and peak memory consumption. These metrics are evaluated for all baseline classifiers as well as the proposed augmentation fusion framework on both TORGO and UASpeech datasets.

Table 8 Computational cost analysis in terms of peak memory usage (MB) and execution time (s) for TORGO and UASpeech datasets.

Method	**TORGO**				**UASpeech**			
	Memory (MB)		**Time (s)**		**Memory (MB)**		**Time (s)**	
	Train	Test	Train	Test	Train	Test	Train	Test
SVM	1.0055	0.1564	53.1062	0.0625	12.3076	1.1888	7366.0925	6.9573
KNN	3.0405	0.3755	124.1323	0.2852	33.2567	3.6535	3017.9379	5.1291
RF	1.2743	0.1265	420.7610	0.0340	13.6099	1.3269	1690.7999	0.4654
ANN (Baseline)	4.2968	0.3732	301.3499	0.6047	8.1988	1.2139	493.5310	2.1342
Augmentation Fusion + ANN	**5.3402**	**0.3959**	**453.6887**	**0.6365**	**27.8700**	**1.9759**	**5268.3036**	**2.5850**

As shown in Table 8, classical machine learning models such as SVM and KNN exhibit lower memory consumption but incur significantly higher training times, particularly for larger datasets. For instance, SVM requires 7366.09 s for training on UASpeech, which limits its scalability. Similarly, KNN demonstrates increased runtime and memory usage due to its instance-based nature.

The ANN baseline provides a balanced trade-off between computational cost and performance, with moderate training time and memory usage across both datasets. In comparison, the proposed augmentation fusion framework introduces a moderate increase in computational cost. For TORGO, the training time increases from 301.35 s to 453.69 s and memory usage from 4.30 MB to 5.34 MB. For UASpeech, the training time increases from 493.53 s to 5268.30 s and memory usage from 8.20 MB to 27.87 MB. This increase is primarily attributed to the expanded training data resulting from augmentation (approximately threefold increase), particularly in the best-performing combinations such as Orig. $+ s + \alpha$ for TORGO and Orig. $+ \tau + \beta$ for UASpeech under speaker-dependent settings, for which this cost analysis was conducted. Despite this increase, the testing time remains comparable across all methods, indicating that the proposed approach does not introduce significant inference overhead, with increased performance gains (relative improvements of 42.86% and 18.78%). This is particularly important for practical deployment scenarios where real-time or near real-time inference is required.

Therefore, the proposed framework achieves a favorable balance between classification performance and computational efficiency. Although the augmentation fusion

strategy incurs additional training cost due to the enlarged and more diverse training data, the inference cost remains comparable to the baseline models. This ensures that the proposed approach does not introduce additional overhead during deployment. Furthermore, it maintains practical memory requirements and efficient inference, making it well-suited for real-world and resource-constrained applications.

4.4 Comparison with Prior Works

Table 9 presents prior work on the classification of dysarthric speech severity using the UAspeech and TORGO databases, providing a clear comparison with our proposed method.

Tripathi et al. [42] proposed a speaker-independent intelligibility assessment system utilizing novel features derived from DeepSpeech outputs, achieving improved classification in a four-class SVM framework on the Universal Access Speech database. H.M. et al. [47] investigated CNN-based classification of dysarthric speech intelligibility using perceptually enhanced Fourier and Constant-Q transform spectrograms, discovering that the Constant-Q transform outperformed other methods in word and sentence-level tasks. Additionally, Gurugubelli et al. [43] introduced perceptually

Table 9 Comparison with prior work on dysarthric speech severity classification in terms of CER (%).

Author	Method	SD/SI	CER (%)	Remarks
		UASpeech		
Tripathi et al. [42] [2020]	Deep posterior + SVM	SD	2.60	Train: E1, E3; Test: E2
		SI	46.10	LOSO setup
H.M. et al. [10] [2020]	Mel-spectrogram + CNN	SD	1.70	355 train / 100 test words
		SI	50.73	LOSO setup
Gurugubelli et al. [43] [2019]	PE-SFCC + i-vector (PLDA)	SI	39.22	LOSO; train/test word mismatch
Joshy et al. [44] [2022]	MFCC i-vector + DNN	SD	6.03	Common vs unseen words in test
		SI	50.78	LOSO setup
Joshy et al. [45] [2023]	Spectrogram + SE-CNN	SD	1.11	Unseen words used for testing
		SI	63.67	Gender-paired training per severity
Javanmardi et al. [46] [2024]	HuBERT + CNN	SI	51.99	LOSO per severity; balanced speakers
Proposed	Augmentation Fusion + ANN	SD	1.86	Multiple distinct splits evaluated
		SI	53.33	LOSO per severity
		TORGO		
Joshy et al. [44] [2022]	MFCC + CNN	SD	3.82	80%–20% split
Javanmardi et al. [46] [2024]	HuBERT + CNN	SI	50.17	LOSO per severity
Proposed	Augmentation fusion + ANN	SD	**2.18**	Stratified 80%–20% split
		SI	**45.85**	LOSO per severity

enhanced single frequency cepstral coefficients (PE-SFCC) for assessing dysarthric speech intelligibility. This approach leveraged human auditory perception and high-resolution single frequency filtering (SFF), achieving superior performance on the UAspeech database compared to state-of-the-art features. Joshy et al. explored various features and classifiers unique to the severity classification of dysarthric speech, such as i-vector based MFCCs utilized in DNN and CNN models, noting that i-vector MFCCs were particularly beneficial in speaker-independent cases [44, 48]. A recent study by Javanmardi et al. [46] employed pre-trained models for classification of dysarthric severity on the TORGO database, revealing that the HuBERT model's certain layers were more effective than others.

In contrast, our proposed method strategically combines data augmentation techniques, achieving either superior or comparable performance to recent studies in both severity-dependent and severity-independent classification tasks. Notably, our results with the TORGO database demonstrate improved classification performance in both speaker-dependent and speaker-independent scenarios, as illustrated in Table 9.

5 Conclusions

In this study, we aimed to advance the development of robust severity classification systems for dysarthric speech. We conducted a comprehensive exploration of machine learning techniques, acoustic and prosodic features, and data augmentation methods using the TORGO and UAspeech dysarthric speech databases. Our research establishes a solid baseline with a unique feature set while demonstrating the efficacy of augmenting the training data through techniques such as speaking rate modification, pitch modification, formant modification, and VTLP. The proposed fusion of these augmentation techniques highlights its effectiveness in enhancing severity classification, both in speaker-dependent and speaker-independent scenarios across both databases. The analysis of various features revealed that incorporating additional acoustic and prosodic features into the classification model significantly enhances the accuracy of dysarthria severity level classification for both databases. This improvement was evidenced by a marked boost in the system's overall performance.

Through a comprehensive analysis of the speaker-dependent experiments, we found that specific combinations of modification techniques yield exceptional results. On the TORGO database, the combination of speaking rate and formant modifications achieved the lowest overall error rate, with a CER of 2.18%. Similarly, on the UAspeech database, the combination of pitch and VTLP modifications outperformed all other combinations, achieving a CER of 1.86%. These results were attained using the optimal modification factors identified in the individual augmentation experiments. Compared to the baseline, our method improved the classifier's relative performance by 42.86% and 18.78%, respectively. These enhancements demonstrate that data augmentation effectively mitigates challenges such as data scarcity and temporal and spectral variability among speakers, thereby enhancing the classifier's robustness.

Further experiments conducted in speaker-independent settings validated the effectiveness of the proposed augmentation combination methods. It was observed that combining more than two techniques has the potential to improve system performance

significantly. Different combinations of augmentation approaches were found to be suitable for various data splits, demonstrating flexibility in application. When comparing average performances to the baseline, notable improvements were evident. The cost analysis indicates that the proposed augmentation fusion strategy achieves improved classification performance while maintaining inference cost comparable to the baseline. This work makes a significant contribution to the field by providing a robust method for classifying severity using the proposed combination of augmentation techniques. It effectively addresses the challenges posed by the low-resource nature of dysarthric speech data.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found here: https://www.cs.toronto.edu/ complingweb/data/TORGO/torgo.html and http://www.isle.illinois.edu/sst/data/UASpeech.

Declarations

- Funding : Not applicable
- Conflict of interest/Competing interests: Authors do not have any Conflict of interest/Competing interests.
- Ethics approval : Not applicable
- Consent to participate: Not applicable
- Consent for publication: Yes, all authors have read and agreed to the for publication.
- Code availability: Not applicable
- Authors' contributions: All authors contributed equally to this work.

References

[1] Kodrasi, I., Pernon, M., Laganaro, M., Bourlard, H.: Automatic and perceptual discrimination between dysarthria, apraxia of speech, and neurotypical speech. In: ICASSP, pp. 7308–7312 (2021). IEEE

[2] Al-Ali, A., Al-Maadeed, S., Saleh, M., Naidu, R.C., Alex, Z.C., Ramachandran, P., Khoodeeram, R., Kumar, R.: The detection of dysarthria severity levels using ai models: A review. IEEE Access (2024)

[3] Joshy, A.A., Rajan, R.: Automated dysarthria severity classification using deep learning frameworks. In: 28th European Signal Processing Conference, pp. 116–120 (2021). IEEE

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

[5] Bhat, C., Vachhani, B., Kopparapu, S.K.: Automatic assessment of dysarthria severity level using audio descriptors. In: ICASSP, pp. 5070–5074 (2017). IEEE

[6] Joshy, A.A., Rajan, R.: 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] Al-Qatab, B.A., Mustafa, M.B.: Classification of dysarthric speech according to the severity of impairment: an analysis of acoustic features. IEEE Access **9**, 18183–18194 (2021)

[8] Vachhani, B., Bhat, C., Kopparapu, S.K.: Data augmentation using healthy speech for dysarthric speech recognition. In: Interspeech, pp. 471–475 (2018)

[9] Tripathi, A., Bhosale, S., Kopparapu, S.K.: Improved speaker independent dysarthria intelligibility classification using deepspeech posteriors. In: ICASSP, pp. 6114–6118 (2020). IEEE

[10] H M, C., Karjigi, V., Sreedevi, N.: Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech. IEEE Transactions on Neural Systems and Rehabilitation Engineering **28**(12), 2880–2889 (2020)

[11] Joshy, A.A., Parameswaran, P.N., Nair, S.R., Rajan, R.: Statistical analysis of speech disorder specific features to characterise dysarthria severity level. In: ICASSP, pp. 1–5 (2023)

[12] Xiong, F., Barker, J., Christensen, H.: Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. In: ICASSP, pp. 5836–5840 (2019). IEEE

[13] Jin, Z., Geng, M., Deng, J., Wang, T., Hu, S., Li, G., Liu, X.: Personalized adversarial data augmentation for dysarthric and elderly speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023)

[14] Wang, H., Jin, Z., Geng, M., Hu, S., Li, G., Wang, T., Xu, H., Liu, X.: Enhancing pre-trained asr system fine-tuning for dysarthric speech recognition using adversarial data augmentation. In: ICASSP, pp. 12311–12315 (2024). IEEE

[15] Geng, M., Xie, X., Liu, S., Yu, J., Hu, S., Liu, X., Meng, H.: Investigation of data augmentation techniques for disordered speech recognition. arXiv preprint arXiv:2201.05562 (2022)

[16] Roucos, S., Wilgus, A.: High quality time-scale modification for speech. In: ICASSP, vol. 10, pp. 493–496 (1985). IEEE

[17] Ahmad, W., Shahnawazuddin, S., Kathania, H.K., Pradhan, G., Samaddar, A.B.: Improving children's speech recognition through explicit pitch scaling based on

iterative spectrogram inversion. In: Interspeech, pp. 2391–2395 (2017)

[18] Moulines, E., Laroche, J.: Non-parametric techniques for pitch-scale and time-scale modification of speech. Speech communication **16**(2), 175–205 (1995)

[19] Kathania, H.K., Kadiri, S.R., Alku, P., Kurimo, M.: Study of formant modification for children asr. In: IEEE ICASSP, pp. 7429–7433 (2020)

[20] Jaitly, N., Hinton, G.E.: Vocal tract length perturbation (vtlp) improves speech recognition. In: Proc. ICML Workshop on Deep Learning for Audio, Speech and Language, vol. 117, p. 21 (2013)

[21] Claes, T., Dologlou, I., Bosch, L., Van Compernolle, D.: A novel feature transformation for vocal tract length normalization in automatic speech recognition. IEEE Transactions on Speech and Audio Processing **6**(6), 549–557 (1998)

[22] Rudzicz, F., Namasivayam, A.K., Wolff, T.: The torgo database of acoustic and articulatory speech from speakers with dysarthria. Language Resources and Evaluation **46**, 523–541 (2012)

[23] Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J.R., Huang, T.S., Watkin, K.L., Frame, S.: Dysarthric speech database for universal access research. In: Interspeech, vol. 2008, pp. 1741–1744 (2008)

[24] Zhu, X., Beauregard, G.T., Wyse, L.: Real-time iterative spectrum inversion with look-ahead. In: International Conference on Multimedia and Expo, pp. 229–232 (2006). IEEE

[25] Beauregard, G.T., Zhu, X., Wyse, L.: An efficient algorithm for real-time spectrogram inversion. In: Proceedings of the 8th International Conference on Digital Audio Effects, pp. 116–118 (2005)

[26] Rathod, S., Charola, M., Patil, H.A.: Noise robust whisper features for dysarthric severity-level classification. In: International Conference on Pattern Recognition and Machine Intelligence, pp. 708–715 (2023). Springer

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

[28] Johnson, A., Fan, R., Morris, R., Alwan, A.: Lpc augment: an lpc-based asr data augmentation algorithm for low and zero-resource children's dialects. In: IEEE ICASSP, pp. 8577–8581 (2022)

[29] Levinson, N.: The wiener rms error criterion in filter design and prediction, appendix b of wiener, n.(1949). Extrapolation, Interpolation, and Smoothing of Stationary Time Series (1949)

[30] Durbin, J.: The fitting of time-series models. Revue de l'Institut International de Statistique, 233–244 (1960)

[31] Laine, U.K., Karjalainen, M., Altosaar, T.: Warped linear prediction (wlp) in speech and audio processing. In: Proceedings of ICASSP, vol. 3, p. 349 (1994). IEEE

[32] Eide, E., Gish, H.: A parametric approach to vocal tract length normalization. In: IEEE ICASSP, vol. 1, pp. 346–3481 (1996)

[33] Kathania, H.K., Kadyan, V., Kadiri, S.R., Kurimo, M.: Data augmentation using spectral warping for low resource children asr. Journal of Signal Processing Systems **94**(12), 1507–1513 (2022)

[34] Sapkota, P., Kathania, H.K., Kadiri, S.R., Narayanan, S.: Improving end-to-end speech recognition for dysarthric speech through in-domain data augmentation. In: 2024 58th Asilomar Conference on Signals, Systems, and Computers, pp. 345–349 (2024). IEEE

[35] Vapnik, V.: Support-vector networks. Machine learning **20**, 273–297 (1995)

[36] Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE transactions on information theory **13**(1), 21–27 (1967)

[37] Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)

[38] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. nature **323**(6088), 533–536 (1986)

[39] Kadi, K.L., Selouani, S.A., Boudraa, B., Boudraa, M.: Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge. Biocybernetics and Biomedical Engineering **36**(1), 233–247 (2016)

[40] Kadi, K.L., Selouani, S.A.: 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)

[41] Gillespie, S., Logan, Y.-Y., Moore, E., Laures-Gore, J., Russell, S., Patel, R.: Cross-database models for the classification of dysarthria presence. In: Interspeech, pp. 3127–3131 (2017)

[42] Tripathi, A., Bhosale, S., Kopparapu, S.K.: Improved speaker independent dysarthria intelligibility classification using deepspeech posteriors. In: ICASSP, pp. 6114–6118 (2020)

[43] Gurugubelli, K., Vuppala, A.K.: Perceptually enhanced single frequency filtering for dysarthric speech detection and intelligibility assessment. In: ICASSP, pp. 6410–6414 (2019). IEEE

[44] Joshy, A.A., Rajan, R.: Automated dysarthria severity classification: A study on acoustic features and deep learning techniques. IEEE Transactions on Neural Systems and Rehabilitation Engineering **30**, 1147–1157 (2022)

[45] Joshy, A.A., Rajan, R.: Dysarthria severity assessment using squeeze-and-excitation networks. Biomedical Signal Processing and Control **82**, 104606 (2023)

[46] Javanmardi, F., Kadiri, S.R., Alku, P.: Pre-trained models for detection and severity level classification of dysarthria from speech. Speech Communication **158**, 103047 (2024)

[47] Chandrashekar, H., Karjigi, V., Sreedevi, N.: Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech. IEEE transactions on neural systems and rehabilitation engineering **28**(12), 2880–2889 (2020)

[48] Joshy, A.A., Parameswaran, P., Nair, S.R., Rajan, R.: Statistical analysis of speech disorder specific features to characterise dysarthria severity level. In: ICASSP, pp. 1–5 (2023). IEEE