Applied Acoustics 210 (2023) 109412

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Automatic speaker verification system for dysarthric speakers using prosodic features and out-of-domain data augmentation

Shinimol Salim [a,*], Syed Shahnawazuddin [b], Waquar Ahmad [a,*]

[a] Electronics and Communication Department, National Institute of Technology, Calicut 673601, India
[b] Electronics and Communication Department, National Institute of Technology, Patna 800005, India

ARTICLE INFO

Article history:
Received 20 December 2022
Received in revised form 11 March 2023
Accepted 29 April 2023
Available online 29 May 2023

Keywords:
Automatic speaker verification system
Dysarthria
Duration modification based data augmentation
MFCC
Prosody
i-vector
x-vector

ABSTRACT

A communication disorder is an impairment of a person's ability to talk or communicate appropriately. Dysarthria is a common neuro-motor speech communication disorder that can be caused by neurological damage. Dysarthria may affect the articulation, phonation, and prosody of a speaker. Dysarthria patients have poor neuromotor coordination and other physical impairments, making it difficult to utilize an interactive keyboard or other user interfaces. The ASV system can make biometric applications more accessible to dysarthric speakers by eliminating the need for them to remember cumbersome and unique authentication numbers and passwords. In this paper, we presented a study on developing an automatic speaker verification (ASV) system for dysarthria patients with varying speech intelligibility to assist them in remote access control and voice-based biometric applications. In the initial part of our proposed approach, we included a duration modification-based data augmentation module in the front end of the ASV system. Since prosody deficits are one of the early indicators of dysarthria, we investigated the role of prosodic variables in combination with the traditional Mel-frequency cepstral coefficients (MFCC). The prosodic variables explored in this study include pitch, loudness, and voicing probability. Separate *i*-vector and *x*-vector models are trained and compared using individual MFCC, prosodic variables, and their combinations. The experimental results showed that the proposed approach based on combining MFCC and prosody features along with duration-modification-based data augmentation produced promising results.

1. Introduction

Due to rising demand for security applications and the invention of new technologies over the last few decades, the biometric sector has become one of the primary growth drivers. Because of differences in vocal-tract shape, larynx size, and other voice production organs, no two people sound alike. Aside from physical differences, each person has his own speaking style, pronunciation pattern, vocabulary choice, and so on. Because of all of these factors, speech can be used as a voice biometric. Voice authentication is more versatile, accurate, and non-intrusive than conventional biometric technologies. Speaker Verification (SV) is the biometric assignment of confirming a claimed identity by analyzing a spoken sample of the claimant's voice. Therefore it is considered as one of the foremost helpful biometric characteristics for human–machine interaction. The major application of speaker verification system is in an on-site access control or in a network-based access control. When it comes to handling security concerns with remote telephone access to a variety of telephone-based applications, such as voice banking, credit card authorization, payment and transaction authentication, password reset system, telephone trading, access to confidential information and so forth, an ASV system is an appropriate solution. With the use of an ASV system, users can access any privileged data or services from any location in the world [1–3].

Several speaker modeling techniques have been used in speaker verification task. A simple factor analysis method was proposed in [4] to describe a low-dimensional speaker and channel-dependent space. Here speech utterance was represented by a vector called *i*-vectors. To compare *i*-vectors and to enable same-or-different speaker decisions, a probabilistic linear discriminant analysis (PLDA) classifier was utilised. Recently, deep neural networks (DNN) were used to map variable-length utterance to fixed dimensional embeddings called *x*-vectors. The *x*-vectors are then centered and projected using Linear Discriminant Analysis (LDA) and modeled by PLDA[4–6]. The *i*-vector and the *x*-vector structure,

* Corresponding authors.
E-mail addresses: shinimolsalim@gmail.com (S. Salim), s.syed@nitp.ac.in (S. Shahnawazuddin), waquar@nitc.ac.in (W. Ahmad).

https://doi.org/10.1016/j.apacoust.2023.109412
0003-682X/

extracted using deep neural network architectures, has recently become the state-of-the-art method for speaker verification.

Speech production mechanism necessitates rapid and synchronized muscle control. This include muscles in our face, tongue, lips, throat, soft palate as well as pharynx. Neurological damage to nerves controlling these muscles can cause speech communication disorders. Dysarthria is a type of neurological speech disorder in which a person's tongue, larynx and surrounding muscles are unable to govern each other which, in turn, affects speech intelligibility, pitch, articulation, phonation and prosody. Cancer, Parkinson's disease, stroke, head injury, cerebral palsy, and muscular dystrophy are some of the causes of dysarthria. Dysarthria can be characterised by flat monotonous speech with poor articulation, sound hoarse, strained or presence of excessive nasalization. Disordered speech prosody, poor pronunciation, less intelligibility, interword delays, and disfluencies are some of the earliest cues of dysarthria. It can cause disturbance in excitation, vocal tract setup, larynx and speed with which articulators change position etc. The quality of phonation, pitch, and loudness of speech are affected by laryngeal problems. The speaker's breathing may be shallow, and he or she may have difficulties matching exhalation with vocalisation. In dysarthric speech, the involvement of the soft palate frequently results in the impression of excessive nasal sounds. The degree of dysarthria can range from mild to severe. As the condition worsens, speech becomes nearly unintelligible [7–9].

Since dysarthria patients have poor neuro motor coordination and other physical impairments that make it challenging for them to use interactive keyboard or touch screen applications. The ASV system, which eliminates the need for unique identification numbers and passwords, can make biometric applications more accessible to dysarthric speakers. Speaker recognition can be utilized for adjusting both speech recognition and an automatic assessment system for dysarthric speakers [10]. Automatic speaker verification technology can significantly improve the quality of life of people with dysarthria through applications in personalized user interfaces, forensics, surveillance, authentication, and security control for confidential information areas [41]. Therefore, it should be given adequate consideration. Efforts in this area are required to construct an effective and robust ASV system that caters to the needs of those suffering from dysarthria.

The majority of research in the automatic speech processing area related to dysarthric speech is based on dysarthric speech severity assessment, speech intelligibility enhancement and automatic speech recognition [11,12]. Few studies have attempted to address the issue of automatic speaker recognition in order to provide biometric applications for dysarthria patients. Automatic speaker identification and intelligibility assessment system was proposed in [8] combining auditory cues with MFCC features, and a comparison of GMM and SVM modeling approaches was performed. A dysarthric speaker identification system that uses a deep belief network for feature extraction and a multi-layer neural network for classification was reported in [7], and comparisons with MFCC features were made. A Speaker recognition system using three sets of feature representations such as *i*-vectors, bottle-neck-neural-network-based features, covariance-based features, and a multi-class SVM classifier is implemented in [13]. It is worth mentioning here that the specific recognition task in a commercial system is more concerned with verification tasks than identification task. However, automatic speaker verification received no attention in the context of dysarthric speakers. Therefore we present our efforts in that direction through the study detailed in this paper.

The main contributions of this paper is as follows:

- To the best of our knowledge, this is the first work dealing in detail the nuances of developing an ASV system for dysarthria patients
- Investigated prosody based measures for ASV system in the context of dysrthria speakers
- We have also investigated the role of duration modification based data augmentation to deal with the paucity of domain-specific speech data in order to develop effective ASV system
- We have studied the effects of combining MFCC and prosodic features on an ASV system developed for dysarthric speakers. Separate set of ASV systems are trained using both individual MFCC features and prosodic features as well as their combination. When prosodic features were combined with MFCC features, verification performance is significantly improved. This technique was found not only to be useful for dysarthric speakers but also for normal speakers
- A study was conducted to demonstrate the differences in F0 contour dynamics and time varying loudness of control speakers and dysarthria speakers
- Investigated average phoneme duration to evaluate the speaking rate and compare the phone duration of dysarthria and control speech and observed a significant difference in average phoneme duration between healthy speakers and dysarthric speakers
- Experimental comparison of *x*-vector and *i*-vector-based speaker embeddings is also performed in the context of dysarthric speaker ASV
- Severity wise analysis of system performance was done to understand the effect of data augmentation along with MFCC and prosody features with variation in the disease severity of the dysarthria speakers
- The paper also compares the duration modification based data augmentation technique with other related methods and analyzed the performance of dysarthric speakers with different severity

The remainder of the paper is organized as follows: An analysis of prosody features for speech from dysarthria speakers is explained in Section 3. Proposed speaker verification system is presented in Section 4, Experimental evaluations and results are discussed in Section 5. Finally, Section 6 concludes this paper.

2. An analysis of prosody features for dysarthric speech

2.1. Motivation

Speaker characteristics are exhibited in the speech signals from each speaker as a result of the anatomical differences in the speech-producing organs as well as due to the differences in the learned or acquired speaking habits. Prosody conveys a person's habitual speaking traits. It is a branch of linguistics that deals with the representation and interpretation of speaking patterns of a speaker. Prosodic characteristics include intonation, stress, and rhythm, each of which is a complex perceptual quantity conveyed mainly through three acoustic parameters namely pitch, duration, and energy. Prosodic deficits are one of the hallmark characteristics of dysarthria. Dysarthria causes dysregulation of the respiratory, laryngeal, and supralaryngeal systems, which lead to prosodic disturbances. Neurological damage causes disturbance in the timing and precision of articulatory muscle movement, which is required for normal prosody. Speakers with dysarthria often exhibit exaggerated, redundant and atypical prosodic features when compared to healthy speakers. Nature of prosodic abnormality may vary with the type and severity of dysarthria [14,15]. According to Monrad-Krohn [16], prosodic variation in dysarthric speech might be Hyperprosody (excessive or exagger-

ated prosody), Dysprosody (distorted prosody), or Aprosody (attenuation or lack of normal prosody)

Dysarthria is characterised by a lack of control over vocal-tract contractions, limiting the range and speed of laryngeal movement. Prosodic abnormalities can be caused by a lack of control over one's vocal folds. Prosodic impairment is one of the prevalent symptoms of dysarthria. Neurological damage causes disturbance in the timing and precision of articulatory muscle movement, which is required for normal prosody. Previous studies have demonstrated that the prosodic patterns of dysarthria speakers differ from those of healthy speakers. The degree of prosodic impairment varies by type of dysarthria and severity of the speech disorder within that category. Prosodic abnormalities can appear as irregular variations in pitch, loudness, duration, and rhythm [17,18]. Studies conducted in [19] indicate that combining prosody with machine learning algorithms could effectively differentiate between healthy and dysarthric speech. While several researchers have investigated prosody features for dysarthria severity assessment methods, no studies have used prosody-based measures for automatic speaker verification systems in the context of dysarthria speakers. The contribution reported in this paper demonstrates that integrating prosodic variables with MFCC features in the front-end processing of an ASV system can significantly improve the accuracy of the system for dysarthria speakers. The following subsections of this section go through the 3 prosody features employed in this study.

2.2. Pitch

Pitch is a perceptual aspect of sound. The acoustic representation of pitch is the fundamental frequency (F0) of vibration of the vocal fold. F0 is the lowest frequency of periodic signal during the phonation of voiced segments. Because of changes in the physical structure of the vocal folds across speakers, F0 is speaker-specific. The first harmonic of the voice (F0) is controlled by varying either the subglottal pressure or the laryngeal tension, or a combination of both. Several factors influence the dynamics of the F0 contour, including the speaking style of a speaker, identity of the sound spoken, context, position of phrases/words, language, intonation rules, and the type of sentence. The speaker-specific information in the F0 contour can be used to model a speaker [20]. Pitch is a typical dysarthric indicator that exposes differences not only between healthy and dysarthric speakers but also between dysarthric speakers with varying degrees of severity levels. Speakers with low severity are monotonic, and high severity dysarthric speakers tend to have much higher values of F0.

In order to better understand the differences in F0 contour of normal and dysarthric speaker, an additional study was conducted to demonstrate the difference in F0 contour and is shown in Fig. 1. The F0 contours shown in Fig. 1 confirm the findings by Kent and Rosenbek reported in [21]. The F0 contour is a pattern that extends across multiple syllables, with each syllable represented by a line segment on the plot, and the duration of individual syllables varies. For a dysarthria speaker syllables can blend together in a way that makes it difficult to distinguish where one syllable ends and another begins. This can create a flattened or indistinct pattern of F0 contour. Additionally, some syllables may be pronounced with suprasyllabic nasalization, which can further contribute to this flattened pattern. This pattern is referred to as fused because the nasalization can spread across multiple consecutive syllables, affecting the overall shape of the syllable chain [21]. F0 contour of a normal speaker is interrupted momentarily for unvoiced units and obstruents. Dysarthric speaker shows a larger separation of syllables, and nasalization is spread over successive syllables. F0 contour of a dysarthric speaker has limited inter-syllabic dependency and coherency. They have difficulty changing between

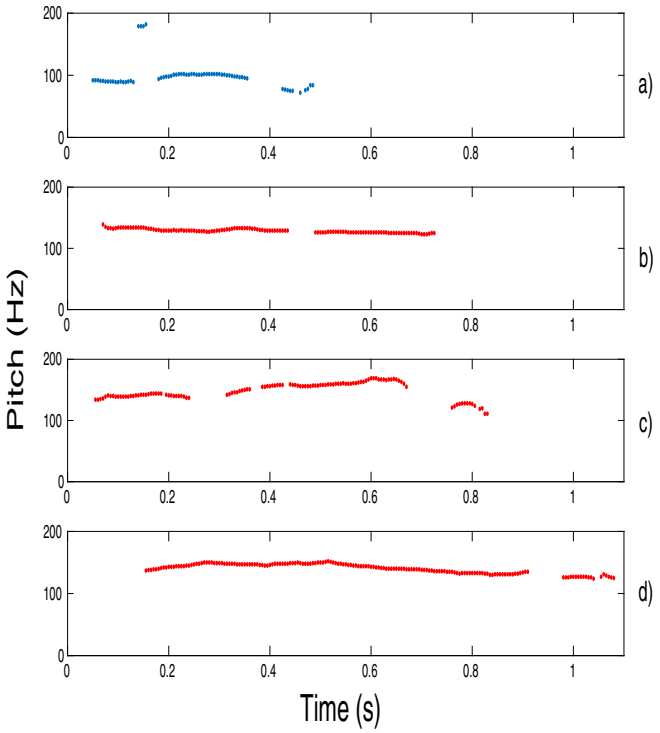

Fig. 1. Variation in F0 contour dynamics of (a) Control Speaker, (b), (c), and (d) Low severity, Medium Severity, and High Severity Dysarthric Speakers (Male), respectively. The considered examples are taken from the UA-Speech Database where each person speaks the same text *"Missouri"*.

voiced and unvoiced segments due to articulatory inaccuracy. An utterance spoken by a dysarthric speaker is longer in duration and monotone in pitch, the slower speaking-rate is due to nasalization and articulatory deficits. Interval between syllables increases as well, and a small variation in F0 can be observed across syllables. There is limited variation in syllable duration and syllable boundaries are indistinct because of constant articulation. Therefore, F0 pattern of a dysarthric speaker is constant in shape within syllables. Furthermore, the utterances from a dysarthric speaker are monopitch in nature. The most deviant or variable feature is that of voicing for the vowels. Vowels are lengthened for a dysarthric speaker in comparison to a normal speaker.

Fig. 2 demonstrates average of mean normalized F0 contour of 3 female control speakers as well as 3 female dysarthric speakers. Mean normalization helps to remove the effect of individual variations in F0 and facilitates comparison of F0 patterns between different speakers. It involves calculating the mean F0 of a speech signal and then dividing all F0 values in the signal by the mean value to obtain a normalized F0 contour. The graph clearly shows that dysarthric speakers have a higher F0 value than control speakers while producing vowels. Speakers with severe dysarthria often have significantly higher values of F0 than a speaker inflicted by mild dysarthria or normal healthy speakers. F0 variation is affected by the degree of severity level of dysarthria, as seen in Fig. 1. In English, approximating a sentence's prosody only by its F0 provides generally acceptable results because duration and intensity are closely related to F0 in this language. This study utilizes openSMILE feature extractor tool for estimating the prosody features. The algorithm is based on the Autocorrelation of Center-clipped Frames (ACF) in [22]. The signal is low-filtered at 900 Hz and divided into short-time frames of speech, and autocorrelation is applied to them. F0 is calculated from the resultant autocorrelated frames.

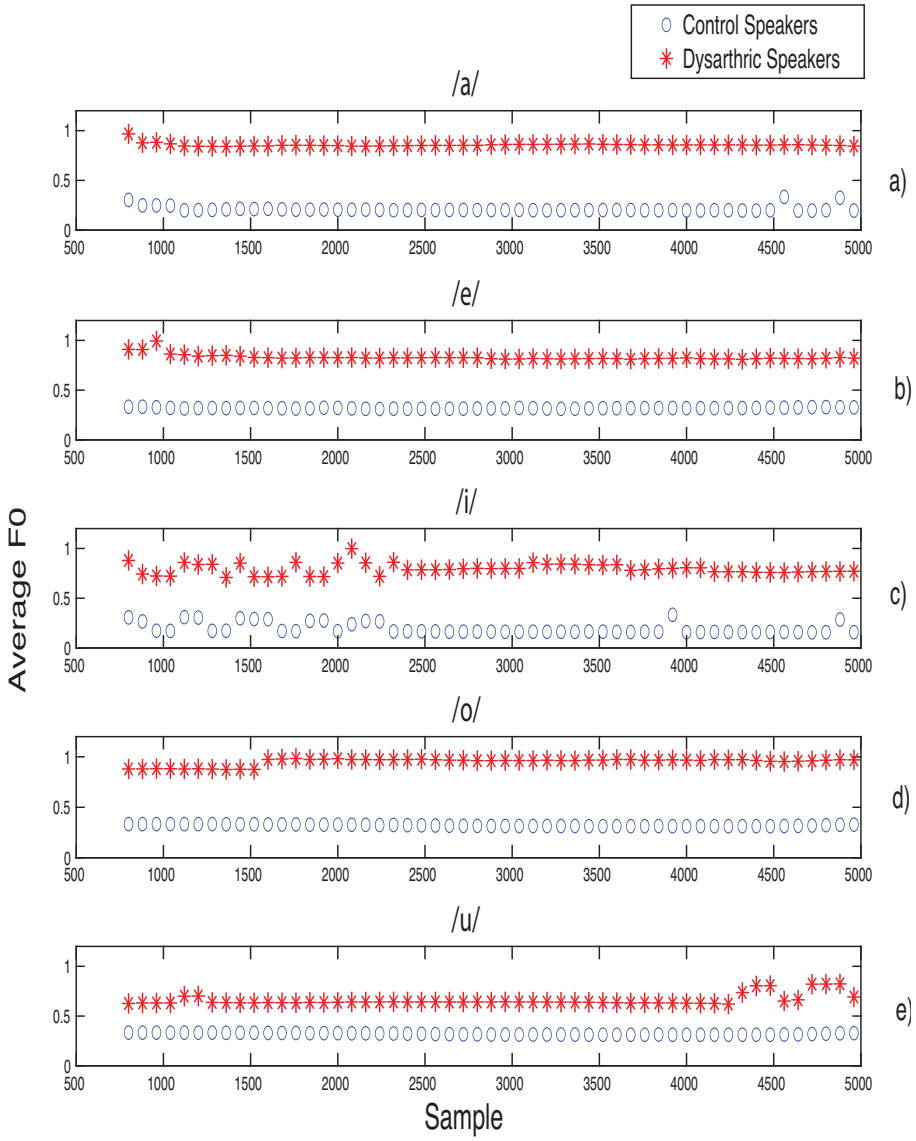

Fig. 2. Average of normalized F0 contour of 3 female Control Speakers and 3 female Dysarthric Speakers from Torgo database for the vowel segments (a)/a/, (b)/e/, (c)/i/, (d)/o/, and (e)/u/, respectively.

2.3. Loudness

Loudness is a perceptual attribute of sound in which the impulse-like nature of glottal excitation in speech production plays a significant role. The acoustic representation of loudness is the distribution of spectral energy or intensity of a sound. Because of the pressure of the vocal-tract system on the vocal source during production, the perceived loudness is dependent on the type of speech sound. A person's behavioral state also has a significant influence on the loudness perceived in a spoken signal. Therefore loudness is governed by both physiological as well as behavioural characteristics of a speaker. Because of differences in speaking style and accent, the dynamics of the energy contour vary among speakers. Dysarthria is characterized by a neurophysiological failure to organize or implement motor gestures. Laryngeal and phonation damage will affect the ability of a person to adjust the loudness of speech, resulting in monoloud speech.

The time-varying loudness of a control speaker and dysarthric speakers of three severity levels, namely low, medium and high, is depicted in Fig. 3. Dysarthric speakers tend to have less intensity variation and reduced control over loudness. Energy in dysarthric speech is more distributed. Small and gradual variation of intensity can be seen across syllables. The number of cycles made by the vocal folds in one second defines the fundamental frequency, which is a natural result of the length of the vocal folds. Vocal intensity is related to the subglottic pressure of the air column. This subglottic pressure is influenced by a variety of parameters, like amplitude of vibration and the tension of the vocal folds. Both pitch and loudness are somehow related. The intensity varies with frequency, since increasing the laryngeal tonus results in a stronger glottic pressure and, as a result, intensity increases. Therefore high voices appear to be more intense. Vowels carry the majority of the energy in comparison to other speech sounds. Fig. 4 shows the average of mean normalized acoustic loudness plot of 3 female control speakers as well as 3 female dysarthric speakers from Torgo database, where the loudness values are normalized to remove the speaker-dependent variations. The average vowel loudness for dysarthric speakers is higher than that of healthy con-

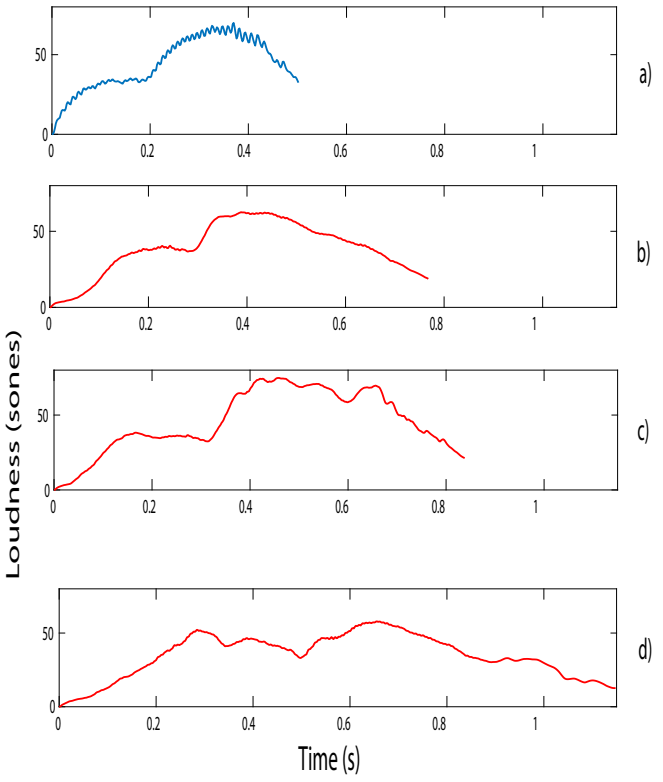

Fig. 3. Variation in loudness with time for (a) Control Speaker, (b), (c), and (d) Low severity, Medium Severity, and High Severity Dysarthric Speakers (Male), respectively. The considered examples are taken from the UA-Speech Database where each person speaks the same text *"Missouri"*.

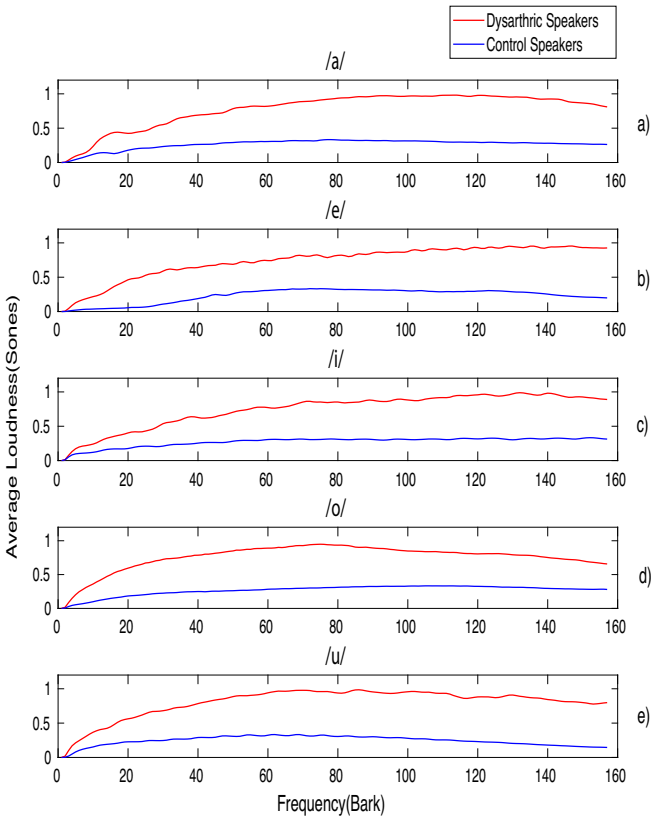

Fig. 4. Average of normalized loudness plot of 3 Control Speakers and 3 Dysarthric Speakers (Female) from Torgo database for the vowel segments (a)/a/, (b)/e/, (c)/i/, (d)/o/, and (e)/u/.

trol speakers. This is because of the higher frequency in producing vowels. However, utterances produced by a dysarthric speaker are monoloud in nature [23–25].

The openSMILE toolkit is used to extract the loudness characteristics. The loudness characteristics model the energy of each sound by calculating the sound amplitude at different intervals, simulating how it is perceived by the human ear. As a result, the extraction process is primarily dependent on two key criteria. First is, as the intensity of a stimulus increases, the hearing response also increases. Second, the sound perception is affected by both the spectral distribution and the length of the sound. Aside from that, loudness characteristics are extracted sequentially on a frame basis and combined into a loudness contour vector and it is expressed as as the normalised intensity raised to a power of 0.3 [26].

The average of mean normalized F0 contour and loudness contour of different control and dysarthria speaker sets provides a way to compare the overall F0 and loudness patterns of the two groups, taking into account individual variations within each group, and helped us to identify consistent differences in vowel sounds between the control and dysarthric groups in terms of their pitch and acoustic loudness. Table 1 presents the variance of the average of the normalized F0 contour and loudness for 3 female control and dysarthric speakers from Torgo database, specifically for vowel sounds/a/,/e/,/i/, and/u/.

To show a statistically significant difference between the average of the normalized F0 contour and loudness for both control and dysarthric speakers, we have conducted Welch's t-test. The Welch's t-test is a statistical test that can be used to determine if there is a significant difference between the means of two independent groups. To conduct Welch's t-test, first, the means and standard deviations of the two groups are calculated. Then, the t-value is calculated by dividing the difference between the two means by the standard error of the difference. Finally, the p-value is calculated from the t-value using the t-distribution with the appropriate degrees of freedom. If the p-value is less than the predetermined significance level of 0.05, then there is evidence of a statistically significant difference between the two groups being compared. Table 2 shows the p-value for Welch's t-test conducted with the average of the normalized F0 contour and loudness of control and dysarthric speakers for the vowel sounds. For all the vowel sounds p-value is below the significance level of 0.05; therefore, we can reject the null hypothesis and conclude that there is a statistically significant difference between the means of the control and dysarthria speakers.

2.4. Voice Quality

Speech characteristics related to the vocal fold are referred to as voice quality. These characteristics are directly associated with phonation and are based on the acoustical model of the vocal folds. Dysarthria patients have reduced control over their vocal folds, resulting in disordered voice quality. This results in a breathy, rough, and hoarse voice with harsh voice quality. One of the causes of voice quality disorder is hypernasality. The deterioration of voice quality intensifies as disease severity increases [18]. Dysarthric speakers have difficulty maintaining the phonation of voiced units. Our study utilized voicing probability for characterizing voice quality. Voicing probablity indicate the duration of voiced sound and it is the rate of voicing in a speech segment. Dysarthric speakers tend to have continuous voicing across syllables. The Autocorrelation Coefficient Function (ACF) can be used to determine the Harmonics to Noise Ratio (HNR), which represents the degree of acoustic periodicity and can be used to calculate voicing

Table 1
Table showing the variance of the average of the normalized F0 contour and loudness of both control and dysarthric speakers for the vowel sounds.

Vowel	F0 contour		Loudness	
	Control	Dysarthric	Control	Dysarthric
/a/	0.00216	0.00082	0.06202	0.00525
/e/	3.1721×10^{-5}	0.0015	0.05425	0.01031
/i/	0.00909	0.00643	0.05821	0.00480
/o/	7.9499×10^{-5}	0.00161	0.03454	0.00560
/u/	0.01439	0.00692	0.04325	0.00485

Table 2
p-value of the Welch's t-test conducted with average of the normalized F0 contour and loudness of control and dysarthric speakers for the vowel sounds.

Vowel	F0 contour	Loudness
/a/	2.5406×10^{-119}	4.5100×10^{-60}
/e/	8.0965×10^{-115}	3.0283×10^{-65}
/i/	7.6101×10^{-64}	6.4742×10^{-57}
/o/	1.0506×10^{-112}	1.1396×10^{-78}
/u/	3.2426×10^{-57}	2.6392×10^{-78}

probability. It is the ratio of the energy of the periodic part of speech to the energy of noise. It is computed as follows [26]:

$$HNR_n(dB) = 10log\left[\frac{ACF(T_0)}{ACF(0) - ACF(T_0)}\right] \quad (1)$$

$$ACF(m) = \frac{1}{N}\sum_{k=0}^{N-1-m} x(k).x(k+m) \quad (2)$$

where T_0 is the fundamental time period, N is the frame length and x(k) is the k^{th} sampling point in the n^{th} frame.

3. Duration modification based data augmentation

3.1. Necessity of data augmentation

The front-end of the ASV system comprises a duration-modification-based data augmentation module. The performance of DNN embedding appears to be extremely scalable as the amount of training data increases. As a result, these systems have been able to make use of large proprietary datasets with great success. To achieve a reasonable estimate of the model parameters, the TDNN architecture used in *x*-vector extraction needs the use of a large quantity of domain-specific data during training. Due to muscle weakness and exhaustion, a dysarthric speaker finds it difficult to speak for long periods of time [27]. As a result, collecting dysarthric speech data, particularly from individuals with severe dysarthria, is a difficult task. As a result, the dysarthric speech databases that are currently available only contain a limited amount of speech data from a small number of speakers. Training an *x*-vector-based system with a limited amount of dysarthric speech would result in under-fitting. Using a large amount of speech data from healthy speakers to train the TDNN, on the other hand, will bias the ASV system towards control participants. As a result, speakers who suffer from dysarthria will perform poorly. We employed data augmentation to address the problem of data scarcity and variability, as well as to improve the robustness of the model. Data augmentation is a technique that applies certain modifications to existing training data to generate new synthetic training samples, which are subsequently combined with the original dataset. We employ fourfold data augmentation, which combines the original data with three augmented copies. Our method makes use of data augmentation based on a duration modification strategy. The prime goal is to increase the amount of training data, increase the diversity of the acoustic characteristics of the training data, and to improve the generalization capability of the trained model.

3.2. Motivation for choice of duration modification as a technique for data augmentation

The rate at which individual speech units are pronounced is referred to as the speech rate. It varies according to the speech unit spoken, physiological characteristics, and emotional state of the speaker, hence it is speaker-specific. The speech rate is primarily determined by two factors: the type or context of speech production, as well as the speaking style. Dysarthria is a motor speech disorder that affects the muscles used for speech production, such as those controlling the lips, tongue, and vocal cords. As a result, dysarthric speakers may have difficulty producing clear and distinct speech sounds, making it hard for listeners to understand them. This can lead dysarthric speakers to compensate by speaking more slowly than control speakers. Dysarthria can cause atypical or variable speech rate correlated with longer pause time. Dysarthria speakers may speak more slowly and at a reduced tempo than healthy speakers, due to difficulty with tongue and lip movement. Speaking-rate refers to the number of syllables per second with pauses, whereas articulation rate refers to the number of syllables excluding pauses. Speakers with dysarthria have a lower rate in both cases because their pronunciations are generally prolonged [27].

To obtain a better understanding, we selected dysarthric and control speech utterances from the Torgo database and analysed it using Matlab and Praat software. As an example, Fig. 5 shows the waveform of the vowel sounds (/aa/,/ee/,/i/,/o/,/u/) from control and dysarthric speech utterances from speakers of Torgo database. The plot indicates that the vowel duration of dysarthric speech is longer than that of their control speech counterparts. As a result, the total duration for the identical set of sentences uttered by dysarthric and control speakers will be much longer in the case of dysarthric speakers. This is due to inter-word delays, frequent pauses, non-speech sounds, and phoneme elongation. Dysarthria can cause extension of syllable duration, vowel duration, word duration, and prolongation of the length of voiced segments. When assessing the speaking duration, utterance duration, or average phoneme duration, there is evidence of impairment in speech rate. We chose average phoneme duration to evaluate the speaking-rate since there is a correlation between phoneme duration and the total duration of utterance. To compare the phone durations of dysarthric and control speech, we used the Penn Forced Aligner to perform a forced alignment of the utterances at the phone level. After that, the alignment was manually checked and corrected in order to retrieve the phone duration. We observed a significant difference in average phoneme duration between healthy speakers and dysarthric speakers, and is shown in Fig. 6. Average phone duration of a dysarthric speaker is relatively longer than that of a control speaker. It is observed that nasal and vowel regions of dysarthric speech (Red) are of longer duration as compared to that of control speech (Blue). Table 3 shows the relative difference(%) in average

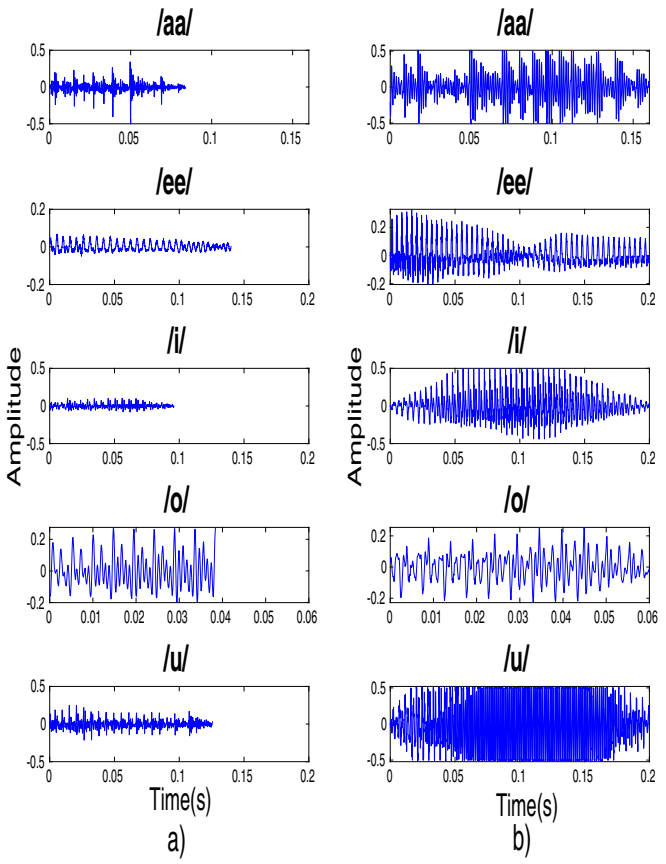

Fig. 5. Time domain waveforms for the vowel sounds /aa/, /ee/, /i/, /o/, and/u/ spoken by (a) a control speaker and (b) a dysarthric speaker from Torgo database.

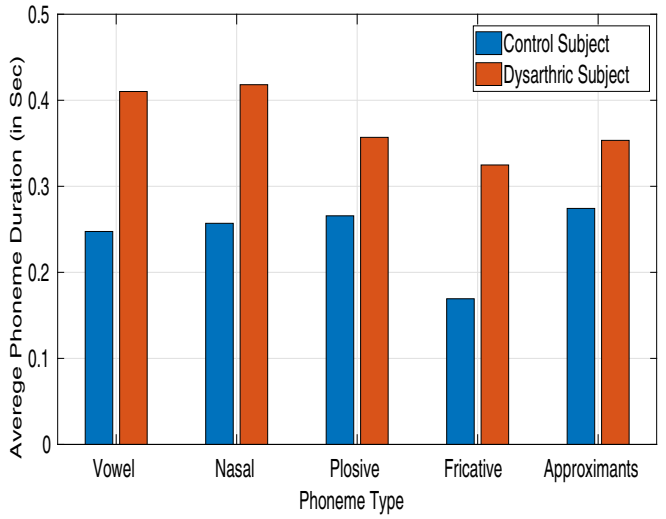

Fig. 6. Average phoneme duration (in sec) for Control subjects (Blue) and Dysarthric subjects (Red) from UA-Speech Corpus.

phone duration between dysarthric speaker and control speaker for four phoneme types namely vowel, nasal, plosive, fricative and approximants. It is evident from the table that vowel, sonorant, and fricative regions of dysarthric speakers are more lengthened than those of control speakers. Furthermore, the average phoneme duration is proportional to the severity of dysarthric speech; the average phoneme duration increases as the severity of the disease increases.

Table 3
The relative difference in average phoneme duration between control subjects and dysarthric subjects from UA-Speech Corpus.

Phoneme Type	Relative Difference(%)
Vowel	65.73
Nasal	62.65
Plosive	34.32
Fricative	91.79
Approximant	28.83

The main goal of data augmentation is to introduce the missing desired acoustic attribute into the training data. One of the missing attributes in the context of an ASV system for dysarthric speakers trained on control data is the increased average phoneme duration. Motivated by this, we propose to extend the duration of the training data from control speakers and then pool it into training. As a result, the ASV system will learn extended phoneme durations and, in time, will become more robust for dysarthria patients. This concept was adopted in this work, and we found that duration-modification-based data augmentation considerably improves performance over the baseline system and other types of data augmentation methods for dysarthria speakers.

Data augmentation techniques such as background noise and room impulse response are typically used to make ASV systems less sensitive to environmental factors, the goal of duration-based augmentation is to introduce a missing attribute (i.e., increased phoneme duration) into the training data. However, it is important to note that the proposed duration-based augmentation approach is specifically targeted at training ASV systems for dysarthric speakers, who typically have longer phoneme durations than non-dysarthric speakers. Therefore, by introducing longer durations into the training data, the ASV system may actually become better equipped to recognize and verify dysarthric speakers, who have longer durations as a result of their condition. Furthermore, we have incorporated other acoustic cues of prosody, such as pitch, loudness, and voice quality into the ASV system to improve its overall performance. Ultimately, the effectiveness of the proposed duration-based data augmentation approach will depend on its ability to improve the accuracy and robustness of the ASV system for dysarthric speakers.

3.3. Duration modification Technique

Duration modification is the process of generating a new speech signal by including the intended modification into the duration of an utterance. An explicit signal processing technique based on identifying instants around glottal closure (GC) and glottal opening (GO) were used as described in [28], to modify the duration of the speech. This gives flexibility at the glottal cycle level. It consist of three main tasks: (i) compute the instants of glottal closure and glottal opening for the given speech data using Zero Frequency Filtering (ZFF) method, (ii) determining the modified GC and GO epoch sequence according to the desired duration, and (iii) synthesize a duration modified speech signal from the modified epoch sequence [29].

3.3.1. ZFF method for determining instants of Glottal Closure and Glottal Opening

The following steps are utilized in processing the voice signal to obtain the GC and GO instant and zero frequency filtered signal [30].

1. Difference the input speech signal s[n] to eliminate any time varying low frequency bias in the signal

$$x[n] = s[n] - s[n-1] \tag{3}$$

2. Pass the differenced speech signal x[n] twice to an ideal resonator that operates at zero frequency

$$r_1[n] = -\sum_{k=1}^{2} b_k r_1[n-k] + x[n] \tag{4}$$

$$r_2[n] = -\sum_{k=1}^{2} b_k r_2[n-k] + r_1[n] \tag{5}$$

where b_1, b_2= −2,-1 respectively
3. Remove the trend in filtered signal r_2[n] by a moving average filter

$$r[n] = r_2[n] - \frac{1}{2N+1}\sum_{m=-N}^{N} r_2[n+m] \tag{6}$$

where 2 N + 1 corresponds to window size of average pitch period computed over segment of speech
4. Trend removed signal r[n] is called the zero-frequency filtered (ZFF) signal, and the instants of significant excitation correspond to the positive zero crossings of the filtered signal.

3.3.2. Duration Modification using GC and GO epoch sequence

Duration modification is the task of changing the duration of a given utterance in a desired manner. To modify the duration of speech, a new speech signal is created by adjusting the duration of the original utterance. This involves generating a new sequence of epochs (distinct points in the signal) from the original signal, regardless of whether they come from voiced or unvoiced parts. The process for generating this new epoch sequence is similar to the method used for residual modification, as described in [31]. For generating the desired epoch sequence for duration modification, the original epoch sequence is resampled according to the desired modification factor α.

The next step after obtaining the modified epochs is to generate the speech signal. To do this, we need to identify the original epochs that are closest to the modified epochs. Then, we extract speech samples around the original epochs and place them starting from the corresponding new epoch. However, since the interval of the new epoch is different from that of the original epoch, we may need to delete some speech samples or add new ones to fill the new epoch interval. To delete the required number of speech samples, we remove them from the tail end of the selected speech samples. To insert the required number of speech samples, we resample about 10% of the tail end of the selected speech samples and add them to the end. So the duration modification consist of 3 main tasks:

1. Derive the GC and GO instant from the input speech signal
2. Derive a modified epoch sequence according to desired duration modification rate α
3. Derive a duration modified speech signal from the modified epoch sequences

α is the duration modification rate of the training data. If $\alpha < 1$, time stretching occurs, which means the duration of the reconstructed audio increases. If $\alpha > 1$, it can result in time compression, and the duration of the reconstructed audio decreases. α=1 indicates that no modification have been made.

4. ASV system architecture

An ASV system to automatically verify speakers with dysarthria was developed in this work, and the system architecture is depicted in Fig. 7. The system consists of three main parts: duration-modification-based data augmentation module, feature extraction, and classifier. Speech data from healthy speakers and its duration-modified version are pooled together during the training phase. Data augmentation contributes to increasing the diversity of acoustic conditions captured by training data and introducing missing desirable characteristics. The duration modification method extends the duration of all phonemes in the utterances uniformly. Detailed description of duration modification based data augmentation module is given in Section 3. To eliminate the non-speech sound units, the pooled training speech is passed through an energy-based voice activity detection (VAD) module. Following that, front-end feature extraction is performed, followed by Cepstral Mean and Variance Normalization (CMVN) [5]. The training phase of the system is shown in Fig. 8. In the feature extraction stage, both MFCC and prosodic features were computed. Separate classifiers based on *i*-vector and *x*-vector are trained using MFCC, prosody features and combination of these features. *x*-vector and *i*-vector extraction is based on the system proposed in [5]. After voice detection on the speech data from dysarthric speakers, feature extraction and feature normalisation are performed during the evaluation or testing phase. The TDNN-based extractor is used to obtain *x*-vectors, and the GMM-UBM recipe is utilised to create *i*-vectors. Finally, PLDA scoring on features from test utterances and models trained during the training phase will be used to make a verification decision.

4.1. Feature Extraction and PLDA based Scoring

A time-varying vocal-tract system generates speech [40]. When compared to control or healthy speakers, dysarthric speakers have substantially less control over the vocal-tract system, resulting in significant variations in the acoustic and overall dynamic features of the speech signal [32]. Features can be low-level features as well as high-level features. Low-level features are physical traits like source features (excitation characteristics) and vocal-tract features (resonant features). High-level features are learned traits like Prosody (Pitch, duration, and energy dynamics) and Idiolect (use of languages or words).

For computing acoustic features of a speech signal, this study utilised both MFCC and prosodic features. The MFCC features were computed using the Kaldi speech recognition toolkit [33], while the prosody features (mentioned in Section 2) were computed using the openSMILE toolkit [22]. The Munich open Speech and Music Interpretation by Large Space Extraction (openSMILE) toolkit is an open source modulator and audio feature extractor used in signal processing and machine learning applications. It is an extremely popular, fast, real time and versatile feature extraction tool used to extract acoustic features of an audio.

MFCCs are one of the most commonly used feature extraction techniques because they represent the relevant frequencies shaped by the vocal-tract while discarding redundant fundamental frequency information. The acoustic features of the human ear, which is more sensitive to variations at lower frequencies, have prompted the usage of MFCC. MFCC feature extraction is based on the assumption that speech is wide-sense stationery across short frames of 10 to 25 ms with an overlapping to prevent loss of information. After that windowing is done by a hamming window and then each frame is converted to the frequency domain using Discrete Fourier Transform (DFT) and the periodogram power spectral estimate of each frame is calculated. It is then applied to the mel

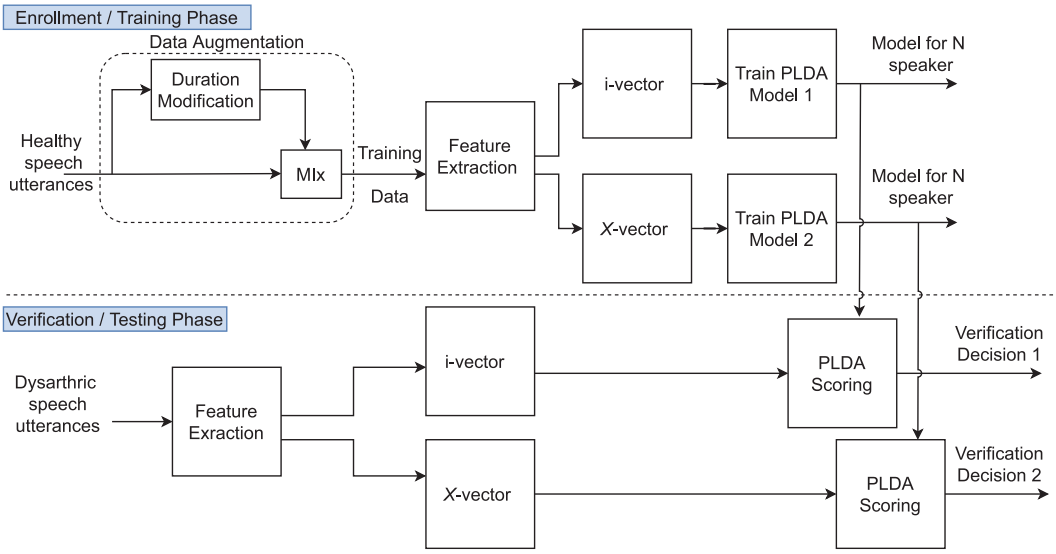

Fig. 7. Simplified block diagram of proposed automatic speaker verification system for dysarthria patients.

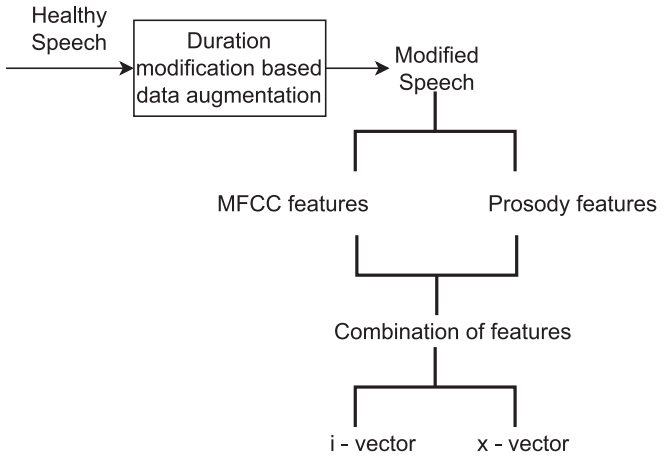

Fig. 8. Block diagram summarizing the training phase of the proposed system.

filterbank. Multiply each filterbank with the power spectrum, then add up the coefficients to get filterbank energies. Taking log of each of the energies will obtain log filterbank energies. Cepstral coefficients are obtained by applying the Discrete Cosine Transform (DCT) to the log filterbank energies. Cepstral coefficients c(n) is calculated as:

$$c(n) = \sum_{m=1}^{M} log_{10}(s(m)) cos\left(\frac{n\pi(m-\frac{1}{2})}{M}\right) \quad ; \qquad n = 1, 2, \ldots N \tag{7}$$

where M is the analysis order, N is the number of cepstral coefficients and s(m) is the mel spectrum. The resulting features are the Mel Frequency Cepstral Coefficients (MFCC).

MFCCs can effectively detect irregular vocal fold movements of a dysarthric speaker. Even with conventional machine learning classifiers such as GMM, HMM, SVM, LDA and KNN, MFCCs have been demonstrated to be accurate in detecting dysarthric speech. MFCCs are identified as suitable performing parameters to represent dysarthric acoustic features in automatic speech recognition [34]. In this work, we used MFCC features as the baseline acoustic features.

4.1.1. *i*-vector system

i-vector extraction is a speaker recognition method that eliminates the distinction between speaker and channel variability subspaces and models them in a single constrained low dimensional space, referred to as the total variability space [4]. A speech segment is represented in this approach by a low-dimensional "identity vector" (*i*-vector) extracted through Joint Factor Analysis. The main idea is to represent the session and channel-dependent supervectors of concatenated Gaussian Mixture Model (GMM) means as

$$s = m + Tw \tag{8}$$

where, m is the UBM supervector, T is the total variability matrix, and w is a standard normally distributed latent variable. For each utterance, *i*-vector is the Maximum A Posteriori (MAP) point estimate of the latent variable w. Our work employs a conventional *i*-vector system that follows the GMM-UBM recipe outlined in [5]. Specifically, we use a 2048-component full-covariance GMM as the Universal Background Model (UBM). The system utilizes a 600-dimensional *i*-vector extractor and employs Probabilistic Linear Discriminant Analysis (PLDA) for scoring.

4.1.2. *x*-vector system

Deep Neural Networks (DNNs) bring in a new phase in the growth of automatic speaker recognition technology, allowing for the extraction of highly discriminative speaker-specific features from a voice sample. A speaker discriminative DNN is trained to produce speaker embeddings, called as *x*-vectors [5]. *x*-vectors are fixed length vectors. Extraction of *x*-vectors involves the following steps: First, short-time front-end acoustic features are computed for each utterance. The acoustic features are then fed into a Time Delay Neural Network. DNN is composed of 3 components: frame-level, statistics-level, and segment-level components. In frame-level component (layer 1 to 5), the input features cascade across the layers, capturing temporal information by increasing the time context of the frames being modelled. The statistics-level component convert a variable length speech input to a single fixed-dimensional vector. The statistics-level is made up of one layer: statistics-pooling, which aggregates over the DNN's frame-level output vectors and calculates their mean and standard deviation. The segment-level component assigns speaker identities to the segment-level vector. The mean and standard deviation are

concatenated and passed to two additional hidden levels (layers 7 and 8), followed by a softmax output layer (layer 9). Layer 6 is set as the speaker embedding, which converts the information from the previous layer into a low-dimensional representation.

4.1.3. PLDA Scoring

Given two per-utterance embeddings e_i, e_j, the PLDA calculates a log-likelihood ratio (LLR) that measures the likelihood of the two embeddings. LLR is given by

$$LLR(e_i, e_j) = log\left[\frac{P\left(\frac{e_i, e_j}{H_1}\right)}{P\left(\frac{e_i, e_j}{H_0}\right)}\right] \tag{9}$$

where H_1 and H_0 denote the same-speaker and different-speaker hypothesis, respectively. The log-likelihood ratio for the test hypotheses corresponding to the two *i*-vectors or *x*-vectors can be computed to verify whether or not the same speaker generated the utterance.

In an *x*-vector system where there is no enrollment stage of the target speaker, a trial list is used to evaluate the performance of the speaker verification system. A trial list contains pairs of utterances from different test speakers, along with their labels indicating whether the speakers are the same or different. During testing, the *x*-vector system processes each utterance in the trial list to obtain its corresponding *x*-vector representation. The *x*-vector representations for the two utterances in each pair are then used as inputs to the PLDA model for scoring. The PLDA model computes a log-likelihood ratio for each pair, which reflects the degree of similarity between the two speakers. If the two speakers in a pair have the same label, the system is expected to output a high score, indicating that the two speakers are the same. If the two speakers have different labels, the system is expected to output a low score, indicating that the two speakers are different. The performance of the system is evaluated using metrics, such as Equal Error Rate (EER) and Detection Cost Function (DCF), which measure the ability of the system to correctly classify the pairs of speakers as same or different.

The *x*-vectors of the training data are used to estimate the parameters of the PLDA model. Specifically, the training *x*-vectors are used to estimate the mean and covariance matrices of the between-class and within-class variability models in the PLDA model. During scoring, the *x*-vectors of the target speakers are projected onto the discriminative subspace learned by the PLDA model. This subspace is defined by the eigenvectors of the between-class and within-class covariance matrices estimated from the training *x*-vectors. Therefore, the training *x*-vectors are needed to estimate the parameters of the PLDA model and to define the discriminative subspace used for scoring. Without the training *x*-vectors, it would not be possible to estimate the parameters of the PLDA model or to project the *x*-vectors of the target speakers onto the discriminative subspace. Same is the case for *i*-vector also.

Proposed ASV system is trained on a training dataset that comprises only control speakers and no dysarthria speakers. Therefore, the system is optimized to recognize speech from control speakers, and it may not perform as well on speech from dysarthric speakers. This is because dysarthric speech has different acoustic characteristics than control speech, and the system may not have learned to capture these differences during training. As a result, the system may have difficulty distinguishing between dysarthric and control speech during testing. To improve the performance of the system on dysarthric speech we have adopted duration modification based data augmentation with MFCC and prosody features.

4.2. Performance evaluation

Equal Error Rate (EER) and minimum of normalized Detection Cost Function (DCF) at p-target=0.01 were used as the evaluation metrics. The EER is the value at which the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) are equal.

$$FRR = \frac{\text{number of rejective true speaker}}{\text{total number of true speaker}} \tag{10}$$

$$FAR = \frac{\text{number of accepted impostor}}{\text{total number of impostor}} \tag{11}$$

Detection Cost Function (DCF) is defined as the weighted sum of FAR and FRR

$$DCFs = C_{miss} \times P\left(\frac{miss}{target}\right) \times P_{target} + C_{falsealarm} \times P\left(\frac{falsealarm}{nontarget}\right) \times P_{nontarget} \tag{12}$$

where C_{miss} and $C_{falsealarm}$ are the cost of miss and false alarm error. P_{target} and $P_{nontarget}$ are the prior probabilities of true and impostor speaker ($P_{nontarget} = 1 - P_{target}$), where P_{target}= 0.01, C_{miss} = 10 and $C_{falsealarm}$ = 1 [5]. The Detection Error Tradeoff (DET) curve is used to visualize the performance of the ASV system.

4.3. Summarizing the ASV system approaches

The speaker verification systems described in this study are summarized below. The following acronyms will be used in the rest of this work.

- **(a)** Baseline: ASV system trained only using speech data from the VoxCeleb1 database serves as the baseline. MFCC features were used as acoustic baseline features. The number of speakers in the training data is 1211.
- **(b)** Data Augmentation + MFCC: The amount of speech data used for training was increased by duration modification based data augmentation technique. Three way duration modification was performed on the control speech. The perturbed and unperturbed data were then mixed to create a final training set. MFCC features were extracted in the feature extraction stage.
- **(c)** Data Augmentation + Prosody: Duration modified versions were pooled with the unperturbed speech training data. The amount of training data is increased by four times. Prosody features were extracted during the feature extraction stage.
- **(d)** Data Augmentation + MFCC + Prosody: In this case also amount of data used for training was increased by a factor of four. A combination of MFCC and prosody features were used as acoustic features.

5. Experiments

Experiments were carried out in this study to explore the effectiveness of prosodic features in ASV systems for dysarthria patients. The verification decision outcomes were evaluated, and performance evaluations were conducted for two separate classifiers based on *i*-vector and *x*-vector models trained on prosody features, MFCC features, and their combinations.

5.1. Databases

The dysarthric speaker verification system developed in this study is evaluated using two different dysarthric speech corpora, namely Torgo database and the Universal Access (UA-Speech)

dysarthric speech corpus. The VoxCeleb1 database, created by Chung et al., is used to train the speaker model.

5.1.1. VoxCeleb1 database

Voxceleb1 database is created by Chung et al. in [35]. It consists of short clips of human speech, extracted from interview videos uploaded to YouTube, where speakers are speaking in a natural setting. Over 140 K utterances for 1211 speakers are included in the VoxCeleb1 development set. Majority of the speech samples in this database are from the healthy speakers.

5.1.2. Torgo database

The Torgo database [36] was developed by the departments of Computer Science and Speech-Language Pathology at the University of Toronto, in collaboration with the Holland Bloorview Kids Rehabilitation Hospital in Toronto. It is made up of aligned acoustics from speakers who have cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS). It contains approximately 23 h of English speech data from 8 dysarthric speakers (3 females and 5 men) of varying speech intelligibility and 7 control speakers (3 females and 4 males). Approximately 500 utterances have been recorded by each of the eight dysarthric speakers.

5.1.3. UA-Speech Corpus

The University of Illinois provided a Universal Access (UA)-Speech database [37] publicly available for experiments related to neuromotor disability. Speech samples from 15 dysarthric speakers (4 females and 11 males) with cerebral palsy and 13 healthy control speakers (4 females and 9 males) are collected and stored in the database. Each speaker has 765 isolated words in the database. All speech signals are sampled at a rate of 16 kHz.

5.2. Experimental setup

Data from all the 23 dysarthric speakers of the Torgo and UA-Speech databases are used for evaluating the system performance (Test-DS). These databases includes speech intelligibility ratings for each dysarthric speaker, in terms of severity-level. Based on the ratings, 23 dysarthric speakers in the current study were classified into 3 severity level categories namely low, medium and high as shown in Table 4. Torgo database is structured in such a way that it consist of non words, short words, restricted sentences and unrestricted sentences. In our work, we selected unrestricted sentences as they are naturally spoken speech with disfluencies and syntactic diversity, and we need to develop an ASV system capable of accepting novel sentences. As these sentences are fewer in number we selected 10 utterances per speaker such that they ensured enough phoneme coverage. In UA-Speech corpus each speaker has 765 isolated words in the database, it includes unique uncommon words, three repetitions of numerals, computer commands, the radio alphabet, and common words. We chose to select uncommon words to ensure that the network gets evaluated on unknown words. We used 10 uncommon words from each of the 15 dysarthric speakers. Therefore test set (Test-DS) has a total of 230 utterances, ten utterances from each dysarthric speaker in the two databases. We randomly selected 23 control speakers from the Voxceleb1 test set (Test-CS) to evaluate the system with healthy speakers. The experiment was evaluated using a total of 52670 trials, 2070 of which were genuine trials and 50600 imposter trials.

Table 4
Severity wise description of 23 dysarthric speakers.

Severity level	Torgo	UA-Speech
Low	M03, M05, F01, F03, F04	M05, M08, M09, M10, M11, M14, F04, F05
Medium	M01, M02, M04	M07, M16, F02
High		M01, M04, M12, F03

ASV system development and evaluation were performed using the Kaldi speech recognition toolkit. For MFCC and prosody feature extraction, the signal is framed into short frames of length 25 ms with 10 ms frame step using overlapping Hamming windows. A 25-channel log Mel-filterbank was employed for spectral warping. Finally, the features are 20 dimensional MFCC features and 3 dimensional prosody features, that are mean normalized over a sliding window of 3 second. LDA project 512 dimensional *x*-vectors extracted using TDNN to 150 dimensional subspace and 600 dimensional *i*-vectors to 200 dimensional subspace. Scoring was done using PLDA.

5.3. Results and Discussion

As a baseline, an ASV system was trained only using speech data from the training set of Voxceleb1 database, i.e., ASV system trained without data augmentation. The system performance was evaluated using the VoxCeleb1 test set (Test-CS) and dysarthric speech test set (Test-DS). This was done to investigate how a dysarthric speech test set impacts the performance of an ASV system trained on a large amount of speech data from control speakers. Performance of the baseline system was evaluated with 20 dimensional MFCC and 60 dimensional MFCC (MFCC with delta and double delta) features. The obtained EER and the minDCF values for control (Test- CS) and dysarthric speech test sets (Test-DS) are given in Table 5. Table 5 shows that the performance improvement obtained by combining MFCC features with their derivatives in the *i*-vector system is only minimal. Moreover, in the case of the *x*-vector system, the results are almost similar. This is because in DNN-based speaker recognition systems, the neural network is designed to learn the relevant acoustic features directly from the raw waveform data. This is done using a time-delay neural network (TDNN) architecture that processes the waveform data at multiple time scales, allowing it to capture both short-term and long-term acoustic dynamics. The TDNN model uses several layers of convolutional and fully connected neural network units to extract high-level features from the raw waveform data. These features are then aggregated using a global pooling layer to obtain a fixed-length *x*-vector representation for each utterance. Therefore in *x*-vector systems, both static and dynamic characteristics of the Mel-Frequency Cepstral Coefficients (MFCCs) are captured, without the need for delta and double delta coefficients. This is because the frame-level layers in the DNN capture the dynamic characteristics of the speech signal, allowing for temporal information to be encoded directly into the model. As a result, delta and double delta coefficients are not required as additional input features to capture the temporal dynamics of the speech signal [38]. Therefore, we selected only the 20-dimensional MFCC features as acoustic baseline features. Same front end and back end for both *i*-vector and *x*-vector allows for more direct comparison between the two modelling approaches.

The results shown in Table 5 shows a significant reduction in system performance when the dysarthric speech test set is used, confirming the challenging nature of the dysarthric speech ASV task. Performance degradation is related to the significant differences in acoustic properties between the training and test data. To overcome the issue of stark differences in the acoustic attributes present in the training and test data, duration-modification-based data augmentation was performed in which we extended the duration of the training data from the VoxCeleb1 database by a modification rate α ($\alpha < 1$), varying from 0.3 to 0.9 in steps of 0.1. Modified speech data corresponding to each value of α was used as training data to construct distinct ASV systems for each value

Table 5
Equal Error Rate and minimum DCF values with respect to an *i*-vector and *x*-vector- based baseline ASV system when evaluated using control (Test-CS) speech from Voxceleb1 test set, and dysarthric speech (Test-DS) test sets from Torgo database and UA-Speech corpus with both MFCC features alone and MFCC with its derivatives.

Test Set	Feature Set	*i*-vector		*x*-vector	
		EER (%)	min DCF	EER (%)	min DCF
Test-CS	MFCC	5.41	0.475	3.43	0.675
	MFCC + Δ + Δ^2	4.01	0.325	3.31	0.668
Test-DS	MFCC	17.18	0.703	15.56	0.675
	MFCC + Δ + Δ^2	16.94	0.702	15.52	0.671

of α using MFCC as baseline features. Performances of each of these ASV systems was evaluated with dysarthric spech test set (Test-DS) and is shown in Table 6, and separately evaluated using low, medium and high severity dysarthric speech test set. The variation of EER with variation in α for the 3 severity level is shown in Fig. 9. Decreasing α beyond 0.3 will reconstruct a training set with an excessively long duration. This might impact the performance of the low severity dysarthric speech test set. Increasing the phone duration of training speech improves the performance of high and medium severity dysarthric speech while degrading low severity dysarthric speech.

Based on the observations, optimal values of modification factor chosen was 0.3, 0.4 and 0.8. Data obtained using these three scaling factors were all merged into training data. A final ASV system was trained utilizing this out of domain data augmentation. The training set for the proposed system is the Voxceleb1 dataset augmented with its duration modified variants, and the training set of the baseline ASV system uses the original Voxceleb1 dataset.

The proposed method is compared to the baseline system for the three feature sets with *i*-vector and *x*-vector classifiers in terms of EER and min DCF and is shown in Table 7. From the table it can be observed that the proposed method of data augmentation based on duration modification is effective for both the classifiers. To demonstrate the role of prosody features with DM-based data augmentation in dysarthria speaker verification, we tested the ASV system with both control (Test-CS) and dysarthria speech test sets (Test-DS), with and without the addition of prosody features. Table 7 shows the performance of the baseline system and the proposed ASV systems with MFCC features alone, prosody features alone, and with the combination of MFCC and prosody features when evaluated with both test sets for *i*-vector and *x*-vector based classifiers. The final column in the Table 7 labeled relative improvement (%) for system (d) indicates the percentage of improvement in performance of system (d) compared to systems (a), (b), and (c) individually. Both test sets performed better when the prosody features were combined with the MFCC acoustic features, showing the complementary nature of the prosody features. As a result, a single ASV system can efficiently serve dysarthria and control speakers. The relative improvement in performance for dysarthric speakers is slightly higher than that of control speakers. As a result, the proposed method of duration modification-based data augmentation using a combination of MFCC and prosody features improves the ASV performance of dysarthria speakers.

DET curve showing the performance of baseline system and the proposed ASV system for the three feature sets with *i*-vector and *x*-vector-based classifiers is shown in Fig. 10. It is evident that signif-

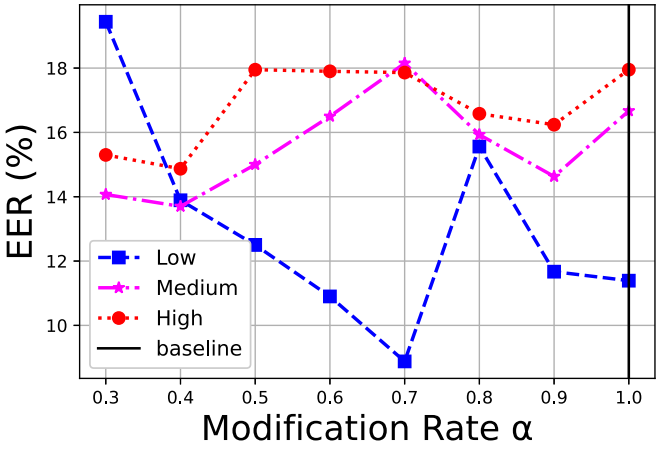

Fig. 9. EER profile showing the impact of duration-modification on dysarthric speaker verification task for 3 severity levels, low, medium, and high.

icantly reduced EER and minDCF values are obtained by performing out-of-domain duration modification based data augmentation. Among the 3 feature sets, the performance of the proposed classifiers developed using combination of MFCC and prosody features is higher than classifiers developed using either of these feature sets alone.

To validate the effectiveness of the proposed Duration Modification (DM) based data augmentation technique, we compared the system's performance with 3 standard data augmentation methods used in the Kaldi recipe, which include reverberation, noise, music, and babble from MUSAN and the RIR datasets, speed perturbation, and volume perturbation. When evaluated with a dysarthric speech test set, Table 8 compares the performance of ASV systems trained using the proposed duration modification-based data augmentation technique with the data augmentation method described in [5], 3-way speed perturbation and volume perturbation with baseline MFCC features. The table shows that the DM-based data augmentation results in a clear improvement for both systems relative to the other data augmentation method. *x*-vectors may benefit from the data augmentation more than the baseline system. The *x*-vector system achieves slightly lower equal error rates than the *i*-vector systems. The duration modification based data augmentation maintain an advantage over the existing data augmentation techniques.

EER and min DCF values were computed considering speech data from each of the 3 severity levels and the performance of

Table 6
Performance in terms of EER and min DCF of the ASV system created with modified speech data corresponding to each value of α when evaluated with dysarthric speech test set (Test-DS).

	α = 0.3	α = 0.4	α = 0.5	α = 0.6	α = 0.7	α = 0.8	α = 0.9	α = 1 (Baseline)
EER(%)	13.33	13.53	15.65	14.78	15.94	13.54	14.4	15.56
min DCF	0.649	0.658	0.679	0.674	0.685	0.662	0.668	0.675

Table 7
Performance of the ASV sytems (a), (b), (c) and (d) in terms of EER and min DCF for the *i*-vector and *x*-vector based classifiers when evaluated with dysarthric speech test set (Test-DS) and control test set (Test-CS). Rel. Impr(%) column represents the relative improvement of system "(d)" with respect to each of the other systems.

ASV system	Duration of training data (in hours)	Classifier	Test Set	EER(%)	min DCF	Rel. Impr(%) of (d)	
						EER(%)	min DCF
Baseline (a)	340	*i*-vector	Test-CS	5.41	0.475	25.54	3.15
			Test-DS	17.18	0.703	27.47	9.10
		x-vector	Test-CS	3.43	0.355	15.18	15.77
			Test-DS	15.56	0.675	25.51	21.48
Data Aug + MFCC (b)	2751	*i*-vector	Test-CS	4.63	0.470	13.04	2.12
			Test-DS	16.14	0.680	22.80	6.02
		x-vector	Test-CS	3.02	0.343	3.86	12.82
			Test-DS	12.17	0.655	4.76	19.08
Data Aug + Prosody (c)	2751	*i*-vector	Test-CS	7.06	0.701	42.91	34.37
			Test-DS	23.77	0.983	47.58	34.99
		x-vector	Test-CS	5.24	0.450	44.65	33.55
			Test-DS	22.05	0.802	47.43	33.91
Data Aug + MFCC + Prosody (d)	2751	*i*-vector	Test-CS	4.03	0.460		
			Test-DS	12.46	0.639		
		x-vector	Test-CS	2.90	0.299		
			Test-DS	11.59	0.530		

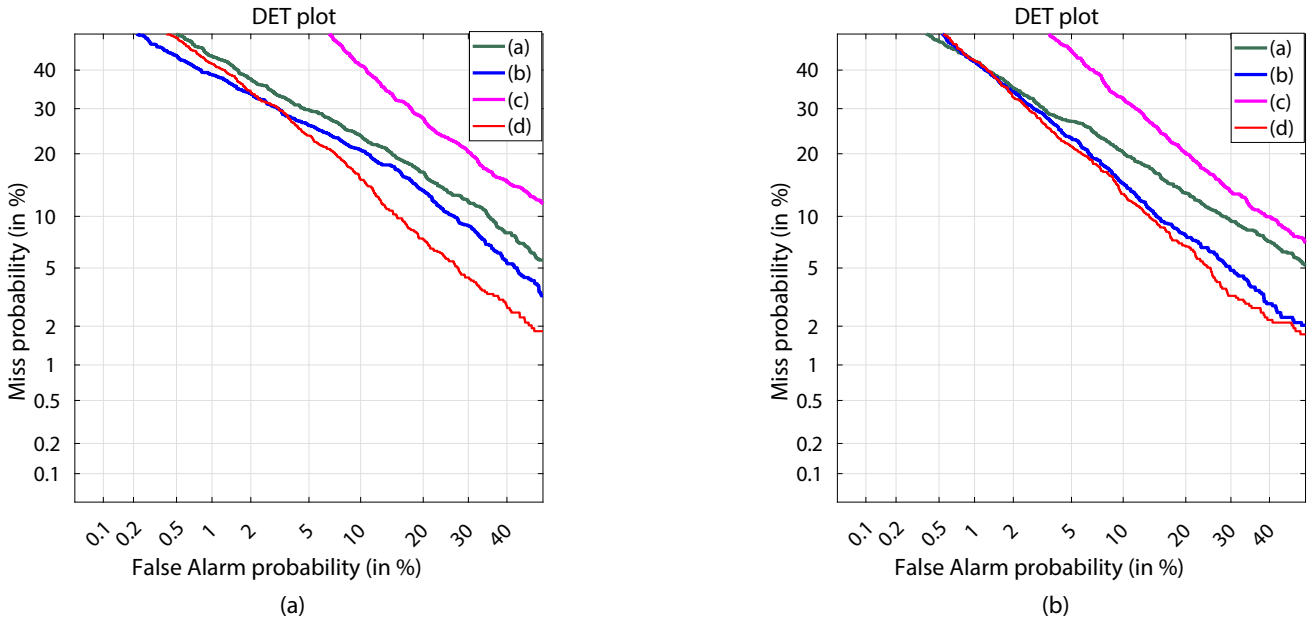

Fig. 10. DET curve demonstrating the performance of the baseline system and proposed ASV system for the three feature sets using an (a) *i*-vector based classifier, and (b) *x*-vector based classifier.

Table 8
Equal Error Rate and minimum DCF values with respect to an *i*-vector and *x*-vector- based ASV systems with 3 standard data augmentation methods and proposed data augmentation method with MFCC features when evaluated with dysarthric speech (Test-DS) test set

Data Augmentation	*i*-vector		*x*-vector	
Type	EER (%)	min DCF	EER (%)	min DCF
Data Aug in [5]	16.91	0.688	15.27	0.665
Speed Perturbation	17.17	0.703	15.52	0.669
Volume Perturbation	17.10	0.695	15.54	0.670
DM based Data Aug	**16.14**	**0.680**	**12.17**	**0.655**

the classifiers were shown in Table 9. In comparison to solely using MFCC features, there is an improvement in performance when combining MFCC features with prosody features for both classifiers. Fig. 11 is the bar chart created to demonstrate the improvement in the performance of the classifiers while combining MFCC and prosody feature sets. These evaluation results show that the proposed ASV system performed significantly better even when the speech data was severely impaired due to dysarthria. The *x*-vector model had the best verification performance when MFCC features were combined with prosody features.

The ASV system verifies the identity of the dysarthric speaker regardless of the severity of the condition, and the performance

Table 9
Severity wise comparison of the ASV systems for the 3 feature sets in terms of EER and min DCF for the *i*-vector and *x*-vector classifiers when evaluated with dysarthric speech test set (Test-DS)

ASV System		Severity Level					
		Low		Medium		High	
		EER (%)	min DCF	EER (%)	min DCF	EER (%)	min DCF
Baseline (a)		15.56	0.605	19.26	0.696	19.66	0.655
Data Aug + MFCC (b)		14.17	0.596	18.52	0.692	17.09	0.597
Data Aug + Prosody (c)	*i*-vector	23.59	0.982	26.85	0.987	30.56	0.995
Data Aug + MFCC + Prosody (d)		**11.97**	**0.519**	**14.63**	**0.689**	**13.89**	**0.583**
Baseline (a)		11.39	0.488	16.67	0.614	17.95	0.743
Data Aug + MFCC (b)		11.11	0.477	13.52	0.496	13.33	0.639
Data Aug + Prosody (c)	*x*-vector	14.26	0.580	18.46	0.750	20.74	0.760
Data Aug + MFCC + Prosody (d)		**11.09**	**0.462**	**13.26**	**0.481**	**11.97**	**0.599**

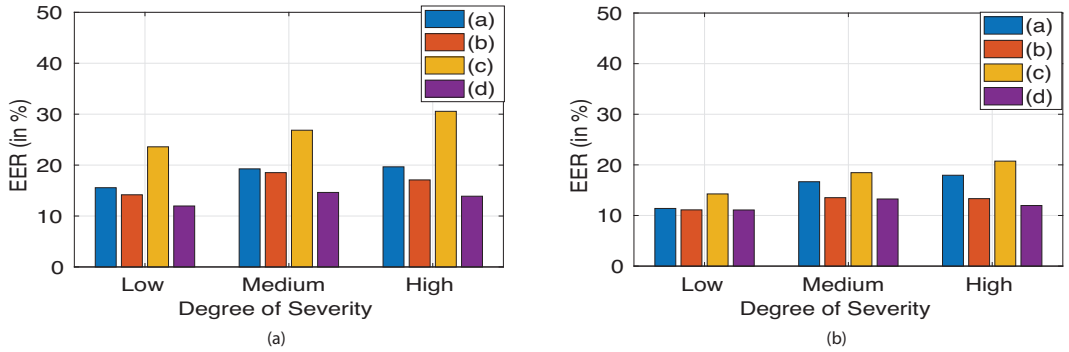

Fig. 11. Severity wise EER for an (a) *i*-vector model and (b) *x*-vector model for the 4 ASV systems.

of the system may depend on the severity of the disease. Proposed ASV system can be used to verify the identity of the same dysarthric speaker in different phases of the dysarthria. Dysarthria is not a progressive disorder by itself, but it can be caused by underlying conditions that are progressive in nature [39]. Some causes of dysarthria may lead to a progressive deterioration of speech function over time, while other causes may not necessarily result in progressive symptoms. However, the rate of progression and the severity of symptoms can vary widely depending on the individual case. Since the available dysarthric speech databases do not contain speech recordings of a single speaker in different stages of dysarthria, we cannot evaluate the performance of our system under such conditions. People with dysarthria typically speak at a slower rate than those without the condition, and this rate tends to worsen as the dysarthria severity increases. To account for this, we modified the duration of the test utterances using various scaling factors to generate dysarthric speech samples from a single speaker at different stages of dysarthria. We then compared the accuracy of our proposed system to the baseline system as is shown in Table 10.

This approach helps to simulate the variation in speaking rate that is observed in dysarthric speech and enables the evaluation of the proposed ASV system under such conditions. The proposed system shows promising results for all modification rates compared to the baseline system. The Table 3 shown above does provide some evidence that the proposed ASV system is more effective than the baseline system in verifying speakers even when the duration of the test utterances varies. The EER and min DCF values for the proposed system are consistently lower than those for the baseline system across different rates of duration modification. This suggests that the proposed system is indeed able to learn extended phoneme durations from the control data and become more robust for dysarthria patients. It is important to note that while the proposed ASV system may not rely heavily on duration

Table 10
Equal Error Rate and minimum DCF values with respect to *x*-vector based Baseline and Proposed ASV systems when evaluated on a dysarthric speech test set (Test-DS) that has been modified with different rates of duration modification denoted by α

Modification Rate	Baseline		Proposed	
α	EER (%)	min DCF	EER (%)	min DCF
0.3	19.23	0.781	15.75	0.678
0.4	18.16	0.715	13.72	0.643
0.5	17.10	0.671	12.66	0.641
0.6	17.50	0.672	15.6	0.661
0.7	18.07	0.732	15.07	0.646
0.8	15.02	0.627	11.59	0.608
0.9	15.46	0.639	11.98	0.627
1.1	16.62	0.636	12.17	0.629
1.2	17.10	0.650	12.46	0.647
1.3	17.39	0.653	12.85	0.648

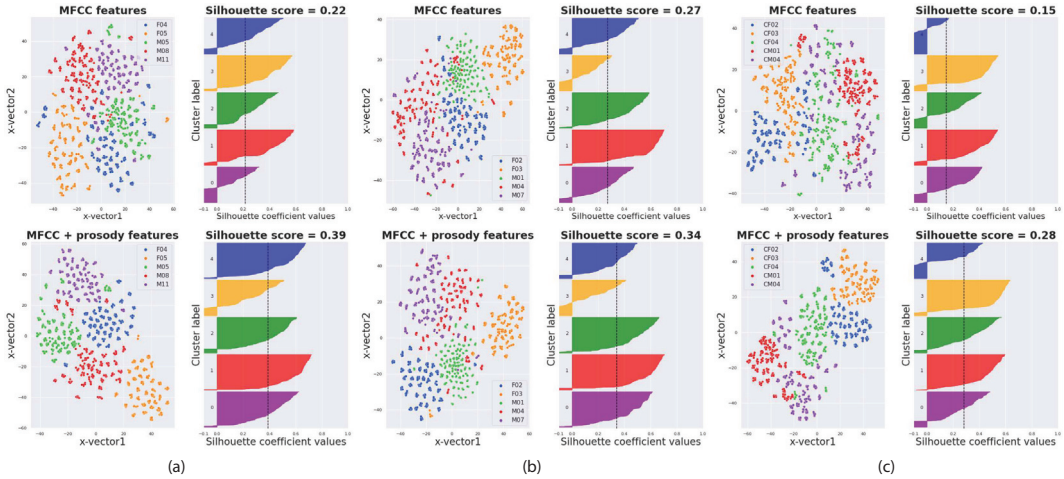

Fig. 12. t-SNE plot with the silhouette score produced by the *x*-vectors of (a) low severity, (b) medium and high severity dysarthric speakers, and (c) healthy control speakers for 1250 utterances from UA-Speech corpus.

as an acoustic cue for verification, it is still able to effectively verify the identity of both control and dysarthric speakers, as demonstrated by the results presented in the study. By using a duration-based augmentation method on control speakers, the system is able to learn to recognize and accommodate for extended phoneme durations, which can be a common feature of dysarthric speech.

Fig. 12 visualizes silhouette analysis for t-SNE (t-distributed Stochastic Neighbor Embedding) clustering on *x*-vector data for 1250 dysarthric and 1250 normal speech utterances randomly selected from low, medium, and high severity dysarthric speakers and healthy speakers from UA-Speech corpus. t-SNE is a nonlinear dimensionality reduction technique that can be used to understand high dimensional data by projecting it into low dimensional space. In Fig. 12, the t-SNE displays the 500-dim *x*-vectors using 2 dimensions. From the figure, it can be seen that there is a good clustering and general separation of *x*-vectors from different speakers for all the 3 severity levels as well as healthy speakers when prosody features are combined with the MFCC features compared to MFCC features alone. The separation distance between the generated clusters can be investigated using silhouette analysis. Silhouette score for each clustering improved by including the prosody features with the proposed duration modification-based data augmentation approach compared to the other features in both classifiers.

Dysarthric speakers speak more slowly than healthy speakers due to weaker muscles, and the rate of elongation of duration depends on the severity of the disease. Because of the differences in phone duration, there is some acoustic mismatch between training data and testing data where duration adjustment can help. Duration adjustment helps improve performance for each severity level. As a result, by appropriately modifying the duration of the speech, a single ASV system may be efficiently employed for verification of both normal and dysarthric speakers with variable speech intelligibility. Since prosodic deficiency is one of the hallmark characteristics of dysarthria, including prosody features with the MFCC features in the feature extraction stage can significantly improve the performance of the proposed system. In comparison with the baseline system with MFCC feature set alone, the proposed system with MFCC and prosody feature set had a relative improvement of 27.13% and 2.15% for EER and min DCF respectively for *i*-vector model and 25.51% and 6.66% for EER and min DCF respectively for *x*-vector model. Higher improvement was found for both models at the high severity level, with about 29.34 % and 9.53 % for EER and min DCF, respectively, for the *i*-vector model, and 33.4 % and 19.38 % for EER and min DCF, respectively, for the *x*-vector model over baseline system with MFCC.

To the best of our knowledge, the current study is the first detailed investigation of the prosody features for automatic speaker verification, particularly in the context of dysarthric speakers, and evaluation results showed that proposed ASV system achieved a satisfactory performance even when the speech is severely impaired due to dysarthria.

6. Conclusion

Dysarthria is a neurological disorder which can cause impairment in speech intelligibility, quality, articulation and prosody. Therefore speaker recognition is a challenging problem for dysarthria patients. In this paper we proposed an automatic speaker verification (ASV) system for dysarthria patients based on prosody features. For that an *i*-vector model based on GMM-UBM recipe and *x*-vector model based on DNN was trained using individual or combination of MFCC fetures and prosody features. Due to muscle weakness and exhaustion dysarthric speakers find it very difficult to speak for long period of time. In order to overcome the issue of data scarcity and diversity we opted for data augmentation of the speech utterances from healthy speakers. Based on the analysis, we found that the average phoneme duration of a dysarthric speaker is relatively longer than that of the control speakers. So to avoid the differences in acoustic features of training data and testing data, we incorporated a duration-modification-based data augmentation module in the front-end of the ASV system to improve the performance of the baseline ASV system for dysarthric speakers. The experimental results showed that the proposed approach of duration modification based data augmentation was found to be very effective for both *i*-vector model as well as *x*-vector model. Performance improvement was observed when prosody features were combined with MFCC features. Best verification performance was achieved with *x*-vector model trained with a combination of MFCC features and prosody features. The performance of the system was computed while considering speech data from each of the three severity levels, namely low, medium, and high. The proposed method also helps to improve the performance for each severity level. This approach opens the door to a variety of exciting possibilities for using DNN in speaker verification and biometric applications for dysarthria patients.

CRediT authorship contribution statement

Shinimol Salim: Conceptualization, Methodology, Software, Writing - original draft. **Syed Shahnawazuddin:** Conceptualization, Methodology, Writing - review & editing. **Waquar Ahmad:** Conceptualization, Methodology, Writing - review & editing.

Data availability

The authors do not have permission to share data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Xiao Qinghan. Technology review - Biometrics-Technology, Application, Challenge, and Computational Intelligence Solutions. IEEE Comput Intell Mag 2007;2(2):5–25.
[2] Kinnunen Tomi, Li Haizhou. An overview of text-independent speaker recognition: From features to supervectors. Speech Commun 2010;52 (1):12–40.
[3] Reynolds Douglas A, Quatieri Thomas F, Dunn Robert B. Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Process 2000;10 (1):19–41.
[4] Dehak Najim et al. Front-End Factor Analysis for Speaker Verification. IEEE Trans Audio, Speech, Language Process 2011;19(4):788–98.
[5] David Snyder et al. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018, pp. 5329–5333.
[6] Sergey Ioffe. Probabilistic linear discriminant analysis. In: Computer Vision-ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, 2006, Proceedings, Part IV 9. Springer, Berlin Heidelberg. 2006, pp. 531–542.
[7] Farhadipour Aref et al. Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks. ETRI J 2018;40.
[8] KamilKadi et al. Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge. In: Biocybernetics and Biomedical Engineering 36 (Nov. 2015).
[9] B.E. Murdoch. Dysarthrias associated with extra-pyramidal syndromes. In: Acquired Speech and Language Disorders: A neuroanatomical and functional neurological approach. Boston, MA: Springer, US, 1990, pp. 234–254.
[10] Mounira Chaiani et al. Dysarthric speaker identification with constrained training durations. In: 2018 International Conference on Signal, Image, Vision and their Applications (SIVA). 2018, pp. 1–6.
[11] Mohammed Sidi Yakoub, Sid-Ahmed Selouani, and Douglas O Shaughnessy. Improving dysarthric speech intelligibility through re-synthesized and grafted units. In: 2008 Canadian Conference on Electrical and Computer Engineering. 2008, pp. 001523–001526.
[12] Ren Jun, Liu Mingzhe. An Automatic Dysarthric Speech Recognition Approach using Deep Neural Networks. Int J Adv Computer Sci Appl 2017;8.
[13] Senoussaoui Mohammed et al. 1. State-of-the-art speaker recognition methods applied to speakers with dysarthria. In: Voice Technologies for Speech Reconstruction and Enhancement. p. 7–34.
[14] Kamil Kadi et al. Discriminative Prosodic Features to Assess the Dysarthria Severity Levels. In: Lecture Notes in Engineering and Computer Science. Vol. 3. July 2013.
[15] Vance James E. Prosodic deviation in dysarthria: a case study. European journal of disorders of communication: the journal of the College of Speech and Language Therapists, 29. London; 1994. p. 61–76.
[16] G.H. Monrad-Krohn. The Third Element of Speech: Prosody in the Neuro-Psychiatric Clinic. In: Journal of Mental Science 103.431 (1957), 326–331.
[17] Hernandez Abner, Kim Sunhee, Chung Minhwa. Prosody-Based Measures for Automatic Severity Assessment of Dysarthric Speech. Applied Sciences Oct. 2020;10:6999.
[18] Hernandez Abner et al. Dysarthria Detection and Severity Assessment Using Rhythm-Based Metrics. In: Interspeech, ISCA, London.
[19] Kim Jangwon et al. Automatic intelligibility classification of sentence-level pathological speech. Computer Speech Language 2014;29.
[20] Leena Mary and B. Yegnanarayana. Prosodic features for speaker verification. In: Ninth International Conference on Spoken Language Processing. Jan. 2006.
[21] Kent RD, Rosenbek John C. Prosodic disturbance and neurologic lesion. Brain Lang 1982;15(2):259–91.
[22] Florian Eyben et al. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: MM 2013 - Proceedings of the 2013 ACM Multimedia Conference. Oct. 2013, pp. 835–838.
[23] Seshadri Guruprasad, Yegnanarayana Bayya. Perceived loudness of speech based on the characteristics of glottal excitation source. J Acoust Soc Am 2009;126(4):2061–71.
[24] Tjaden Kris, Wilding Gregory E. Rate and loudness manipulations in dysarthria: acoustic and perceptual findings. J Speech, Language, Hearing Res.: JSLHR 2004;47(4):766–83.
[25] Tjaden Kris, Wilding Greg. The Impact of Rate Reduction and Increased Loudness on Fundamental Frequency Characteristics in Dysarthria. Folia phoniatrica et logopaedica: official organ of the International Association of Logopedics and Phoniatrics (IALP) 2010;63:178–86.
[26] Esther Ramdinmawii, Abhijit Mohanta, and Vinay Kumar Mittal. Emotion recognition from speech signal. In: TENCON 2017–2017 IEEE Region 10 Conference. 2017, pp. 1562–1567.
[27] Bhavik Vachhani, Chitralekha Bhat, and Sunil Kumar Kopparapu. Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition. In: Proc. Annu. Conf. Int. Speech Commun. Assoc. Sept. 2018, pp. 471–475.
[28] Syed Shahnawazuddin, Nagaraj Adiga, and Hemant Kathania. Effect of Prosody Modification on Children's ASR. In: IEEE Signal Processing Letters PP (Sept. 2017), pp. 1–1.
[29] SRM Prasanna et al. Fast prosody modification using instants of significant excitation. In: Speech Prosody 2010-Fifth International Conference. 2010.
[30] Sri Rama Murty K, Yegnanarayana B. Epoch Extraction From Speech Signals. IEEE Trans Audio, Speech, Lang Process 2008;16(8):1602–13.
[31] Rao KS, Yegnanarayana B. Prosody modification using instants of significant excitation. IEEE Trans Audio, Speech, Language Process 2006;14(3):972–80.
[32] Anusha Prakash, M. Reddy, and Hema Murthy. Improvement of Continuous Dysarthric Speech Quality. In: Proceedings of SLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies. Sept. 2016, pp. 43–49.
[33] Daniel Povey et al. The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (Jan. 2011).
[34] Gupta Shikha et al. Feature Extraction Using Mfcc. Signal Image Process: Int J Aug. 2013;4:101–8.
[35] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Vox- Celeb: A Large-Scale Speaker Identification Dataset. In: INTERSPEECH. June 2017.
[36] Rudzicz Frank, Namasivayam Aravind, Wolff Talya. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Language Resour Eval Jan. 2010;46:1–19.
[37] Heejin Kim et al. Dysarthric speech database for universal access research. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. Jan. 2008, pp. 1741–1744.
[38] Finnian Kelly et al. Deep Neural Network Based Forensic Automatic Speaker Recognition in VOCALISE using x-Vectors. In: Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics, Audio Engineering Society.
[39] Soliveri P et al. Progressive dysarthria: Definition and clinical follow-up. Neurolog Sci Nov. 2003;24:211–2.
[40] S. Salim, G. Deekshitha, A. George and L. Mary, ‘‘Automatic Spotting of Vowels, Nasals and Approximants from Speech Signals,’’ 2018 International CET Conference on Control, Communication, and Computing (IC4), Thiruvananthapuram, India, 2018, pp. 272–277
[41] Salim S, Shahnawazuddin S, Ahmad W. Automatic Speaker Verification System for Dysarthria Patients. Proc. Interspeech 2022;2022:5070–4. https://doi.org/10.21437/Interspeech.2022-375.