Category-based and Target-based Data Augmentation for Dysarthric Speech Recognition Using Transfer Learning

Sarkhell Sirwan NAWROLY[1]*, Decebal POPESCU[1], Mariya Celin THEKEKARA ANTONY[2]

[1] Faculty of Automatic Control and Computer Science, National University of Science and Technology Politehnica Bucharest, 313 Splaiul Independentei, 060042, Bucharest, Romania
sarkhell.sirwan@spu.edu.iq (**Corresponding author*), decebal.popescu@cs.pub.ro

[2] School of Computing and Data Science, Sai University, Tamil Nadu, 603104, India
mariya.c@saiuniversity.edu.in

Abstract: Dysarthric speech recognition poses unique challenges in comparison with normal speech recognition systems due to the scarcity of dysarthric speech data. To address this data sparsity issue, researchers have developed data augmentation techniques. These techniques utilize either the original dysarthric speech examples or speech data pertaining to normal speakers to generate new dysarthric speech data, thereby improving the dysarthric speech recognition performance. This study uses dysarthric speech examples to create augmented examples for training purposes in order to retain the identity of the dysarthric speakers in terms of their speech errors. A two-stage transfer learning strategy is employed, in the first stage of which a category-specific low-frequency noise augmentation method is introduced, while in its second stage a dysarthric speaker-specific data augmentation approach is implemented. The proposed method blends the advantages of various data augmentation approaches in the literature to develop a fine two-stage model that can handle data augmentation without compromising on the quality of the target model. This two-stage approach achieved a notable Word Error Rate (WER) reduction of approximately 11.369%, especially among the severely affected dysarthric speakers, by contrast to the transfer learning method that relies only on normal speech-related data for training.

Keywords: Dysarthric speech recognition, Noise analysis, Transfer learning approach.

1. Introduction

Dysarthria, a neurological speech disorder, can be caused by conditions such as cerebral palsy, amyotrophic lateral sclerosis (ALS), stroke, or traumatic brain injuries (Darley et al., 1975). People with dysarthria struggle to articulate their words, which results in a slurred or unintelligible speech due to poorly coordinated articulatory movements.

Augmentative and alternative communication (AAC) devices have been developed to help individuals communicate effectively. Among the AAC aids, those utilizing automatic speech recognition (ASR) systems could significantly improve their quality of life without social support dependency. However, the current ASR systems in use perform very poorly for dysarthric speakers as they lack training data from this population. A substantial corpus of their speech data is essential to develop ASR systems tailored to dysarthric speakers. However, gathering such a large corpus is challenging. While certain corpora of dysarthric speech data exist (Kim et al., 2008; Rudzicz et al., 2012; Thekekara Antony et al., 2016), they are limited in size and available only for analysis. On the other hand, unlike normal speakers, acquiring a comprehensive corpus for dysarthric speakers is complicated because the process entails recording the speech of individuals with physical and intellectual impairments. Consequently, collecting data from these individuals is laborious and challenging.

Data augmentation techniques are used as a better alternative for handling the above-mentioned issue of data sparsity in dysarthric speech. These techniques have been thoroughly studied in the context of normal speech recognition tasks, employing different techniques, including tempo and speed variation (Geng et al., 2020), vocal tract length perturbation (VTLP), SpecAugment, cross-domain feature adaptation, and noisy and reverberant speech simulation. An increased augmented data for training enhances the overall performance of ASR systems.

The use of data augmentation to enhance dysarthric speech data has had a minimal effect on research because it is challenging to expand the number of instances while preserving the unique characteristics of the dysarthric speakers. Several approaches were outlined by adjusting the spectro-temporal disparities for a normal speech. Tempo-stretching (Vachhani et al., 2018; Xiong et al., 2019), VTLP (Jaitly & Hinton, 2013), and speed perturbation (Ko et al., 2015) are some techniques applied to normal speech recordings to obtain resultant speech data, characterized by attributes like reduced volume and slower rate, which are then utilized to augment the limited training data for the dysarthric speech.

https://doi.org/10.24846/v33i4y202408

Most of these approaches try to obtain the dysarthric speaker's speech characteristics like speech rate, volume, breathiness and hoarse voice. However, the most important and higher-level attributes, like articulatory imprecision or speech errors inherent to each dysarthric speaker, cannot be inferred or addressed through the data augmentation examples obtained from normal speakers as a source.

On the other hand, the approaches aimed at augmenting data by using the original dysarthric speech are employed to address these differences. Xiong et al. (2020) increases the size of the training dataset by directly modifying dysarthric speech samples. The study uses a threefold speed perturbation method to expand the data for model training. It incorporates factored time delay neural networks (TDNN-F) and convolutional neural networks (CNNs) to identify isolated words in the Universal Access (UA) speech corpus (Kim et al., 2008). This paper achieves an overall average Word Error Rate (WER) of 30.76% through data augmentation and transfer learning. In (Takashima et al., 2020), a two-step speaker adaptation process is performed. First, a speaker-independent (SI) model trained on normal speech is adapted to accommodate multiple dysarthric speakers. The modified SI dysarthria model is then tailored to align with the speech features of a specific dysarthric speaker. However, incorporating data from normal speakers for training or augmentation can introduce variability within the speaker's speech, particularly for those with mild to moderate dysarthria. Severe dysarthric speaker categories, especially moderate-to-severe and severe classes, encounter difficulties due to the substantial acoustic differences from normal speech. Consequently, these methods become less effective as the severity of dysarthria increases.

In (Thekekara Antony et al., 2020), the authors present a data augmentation technique exclusively using dysarthric speech data. They introduce a VM-MRFE (Virtual Microphone-Multi Resolution Feature Extraction) method which is a data augmentation approach used by transforming the original dysarthric speech data. Initially, dysarthric speech samples are produced based on the original using virtual microphone array synthesis. This expands the original three examples to 21 new examples through a 7-set virtual microphone array synthesis. This set is further augmented using a multi-resolution feature extraction technique by applying five different resolutions to the virtual microphone signals resulting in 105 new examples. The augmented examples are then utilized to train a hybrid ASR system for isolated words based on DNN-HMM. Assessments on the UA corpus indicate average Word Error Rates (WERs) of 5.82%, 11.62%, and 50.36% for the mild, moderate, and severe dysarthric speaker categories, respectively. Similarly, the assessments on the SSN-TDSC corpus demonstrate average WERs of 20.51%, 29.71%, and 54.04% for the mild, moderate, and severe dysarthric speaker categories. This approach was also applied for continuous dysarthric speech in (Thekekara Antony et al., 2023) using the transfer learning approach. It was suggested that the size of the augmented data could not be further increased, as increasing the array length reduces the signal's energy. Apart from microphone arrays and multiple resolutions, data augmentation through noise has been a focus in (Nawroly et al., 2023).

Noisy data is one source that is available in abundance in the literature (Muthu Philominal et al., 2020; Borrie et al., 2017). The intuitive concept of the work of Nawroly et al. (2023) is to introduce noise in the dysarthric speech by analysing it and selecting the appropriate noise frequency ranges that affect the dysarthric speech examples less even after augmenting it on them. On that note, low-frequency noises that affect the intelligibility less at the dysarthric speech frequency range are chosen, and only noises with a low-frequency range are augmented to the dysarthric speech examples. This work is conducted solely for individual words, and further examination is required to evaluate its efficacy in improving the intelligibility of continuous dysarthric speech.

It is understood from the data augmentation approaches using the original dysarthric speech that data augmentation is performed either at the isolated word level (Nawroly et al., 2023), where there are no limitations on the expansion of augmented speech examples, or at the sentence level (Thekekara Antony et al., 2023), where there are constraints on the number of examples generated. To handle this, the proposed method blends the advantages of various data augmentation approaches to create a data augmentation method suitable for continuous dysarthric speech recognition that can also expand the number of examples with fewer constraints. For this purpose, a two-stage transfer learning

(TL) approach is proposed in the current work. In the first stage, a dysarthric speech recognition system is trained using a set of dysarthric speech data based on the speaker category which is augmented with noisy data to serve as a pre-trained (source) model. The target model is then trained using a data augmentation approach based on VM-MRFE, as proposed in (Thekekara Antony et al., 2020) for the target dysarthric speaker alone.

Noise augmentation is performed for the source model to increase the data volume of the source model in transfer learning. As the model parameters highly depend on the volume of the source data, the noise data augmentation approach that has no limit with regard to the augmented examples as mentioned in (Nawroly et al., 2023) is used at the source side. For augmenting the target dysarthric speech data, the data augmentation approach in (Thekekara Antony et al., 2020) is applied as VM-MRFE has the ability to retain the identity of the target dysarthric speaker to its best which is a highly required feature for the target model.

The paper is organized as follows. Section 2 provides details about the dysarthric speech corpus and the utilization of noisy data. Section 3 details the techniques used for data augmentation in the two-stage transfer learning approach. Section 4 explains the training process for an ASR system based on two-stage transfer learning using augmented speech data. Section 5 compares the proposed approach with the data augmentation approaches in the literature. Section 6 concludes this paper.

2. Speech Corpora Used

This study validated the carried out experiments using the Nemours dysarthric speech corpus (Menendez-Pidal et al., 1996). The speech data in this database is from 10 speakers with dysarthria who uttered 74 sentences. The first 37 of them have two nouns and a verb, followed by the next 37 sentences with the nouns in reverse order. This means that each word in the entire corpus has two examples which were recorded at a sampling rate of 16 kHz. Furthermore, it includes transcriptions aligned at the word and phoneme levels.

For noise augmentation, the NOISEX-92 database (Varga & Steeneken, 1993) is used to supplement the noisy data with samples from Babble, Factory, Pink, Benz, Car, Bus, Volvo, and Train noises, covering a broad frequency range.

3. Data Augmentation Techniques for the Two-Stage Transfer Learning Approach

This work has used two data augmentation approaches to train the source and target TL models. Both approaches utilize the dysarthric speech data itself to synthesize new examples. Data augmentation through noise is performed on the first level for the dysarthric speaker category-based data, and data augmentation through the VM-MRFE approach is performed on the second level for the target dysarthric speaker model as shown in Figure 1.

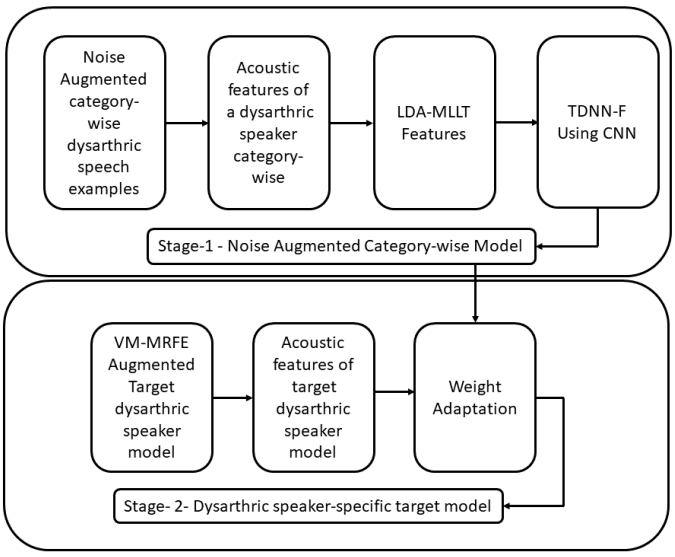

Figure 1. Block Diagram of the two-stage transfer learning approach

For any target dysarthric speaker, the source model is decided based on the category of the target dysarthric speaker. If the dysarthric speaker is from a severe category, then the data for source model training includes all the dysarthric speech data from the severe category along with its noise-augmented examples. The corresponding target data for training the target model is the dysarthric speech of the target speaker itself augmented by using the VM-MRFE approach as shown in Figure 2. Hence, initially category-wise models are trained leaving the target dysarthric speaker to avoid overlapping of the source and target data.

The Nemours dysarthric speech corpus is utilized for both the source and target training data, with 37 utterances allocated for testing and the other 37 used for training. Data augmentation methods are used on the training set. The source model is trained in a category-wise manner (mild, moderate, and severe) using noise-based data augmentation. This choice is rooted in the fact that it pre-trains the initial layers of the neural network based on the common characteristics inherent to the noise-based data-augmented examples. The shared characteristics specific to a category (mild, moderate, and severe) in the context of the noise-based data augmentation method are primarily centered on the acoustic information for each category rather than the diverse noise characteristics. This is because, in this approach, low-frequency analysis-based noises are specifically added to the dysarthric speaker's speech, as opposed to introducing a generic noise. Consequently, the impact and nature of the influence on each original example vary, precluding it from being a generalizable feature. Furthermore, this approach does not limit the number of examples to be augmented (Nawroly et al., 2023). The following sub-section discusses dysarthric speech augmentation using noise as a source.

3.1 Stage 1: Noise-based Data Augmentation Approach for Source Model Training

Noise categories such as "Volvo," "Golf," "Car," "Benz," "Babble," "Train," "Bus-i," and "Bus-j" are chosen for noise-based data augmentation. By contrast to the approach mentioned in (Nawroly et al., 2023), noises such as "train" and "babble" are also being considered in this work. The reason for this is that the original model is a pre-trained version that is tailored to the specific category of the dysarthric speaker being targeted, rather than being a model dependent on the individual speaker used for recognizing dysarthric speech.

Signal-to-Noise ratio (SNR) dB levels ranging from -5dB to +20dB in steps of 5dB were used for augmenting noise data to dysarthric speech data. Eight different noise conditions across five SNR dB levels (-5, 5, 10, 15, and 20dB) were applied, and the noise data was added to dysarthric speech data on a frame-by-frame basis. Figures 2(a) and 3(a) show the original dysarthric speech signal and its spectrogram for the speaker BB with mild dysarthria, and Figures 2(b) and 3(b) show the augmented version with noise data for the same speaker. It's apparent that the augmented example

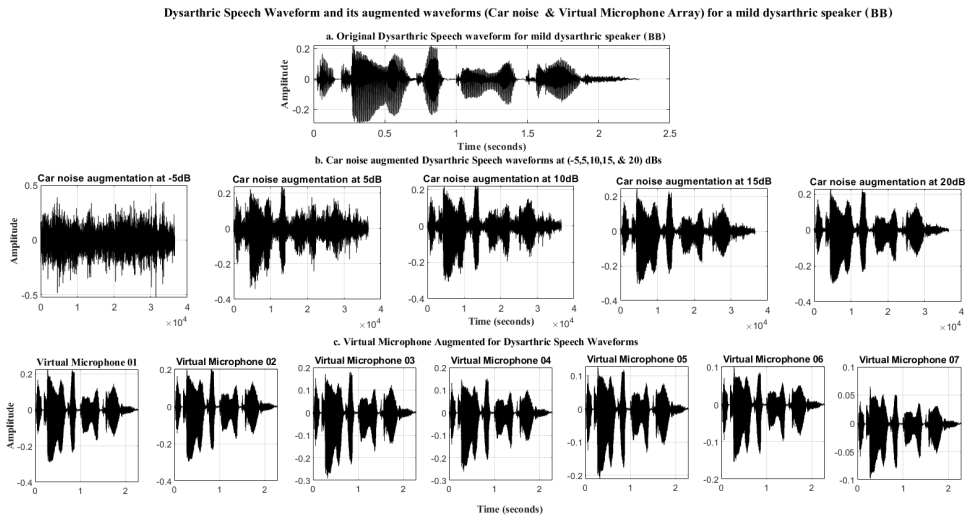

Figure 2. Dysarthric speech signal for a mild dysarthric speaker (BB) and its augmented versions

is not simply a copy of the original but it also represents a new instance of speech produced by the dysarthric speaker. The spectrogram in Figure 3 shows that the general shape and characteristics of the spectrum are largely preserved across the augmented dysarthric speech examples with respect to the original dysarthric speech retaining the dysarthric speakers' identity and speech errors, that is represented by drawing a black box as it is also shown in Figure 3. This fact is also represented as a line graph in Figure 4 which was plotted using the Mel Frequency cepstral coefficients (MFCCs) features derived from the speech signal for a mild dysarthric speaker (BB), along with the noise-augmented MFCC versions at various SNR levels and for different noise types. MFCCs are the set of coefficients that capture the shape of the power spectrum of each sound unit, here the MFCCs of each sound are calculated at the frame level for each sound and plotted in a line graph to compare the characteristics of the original dysarthric speech data with its noise-augmented versions across various Signal-to-Noise Ratio (SNR) levels. The Mel-Frequency Cepstral Coefficients (MFCCs) were chosen as they are used for feature representations to analyze the impact of added noise on the speech data's spectral properties. From Figure 4, it is evident

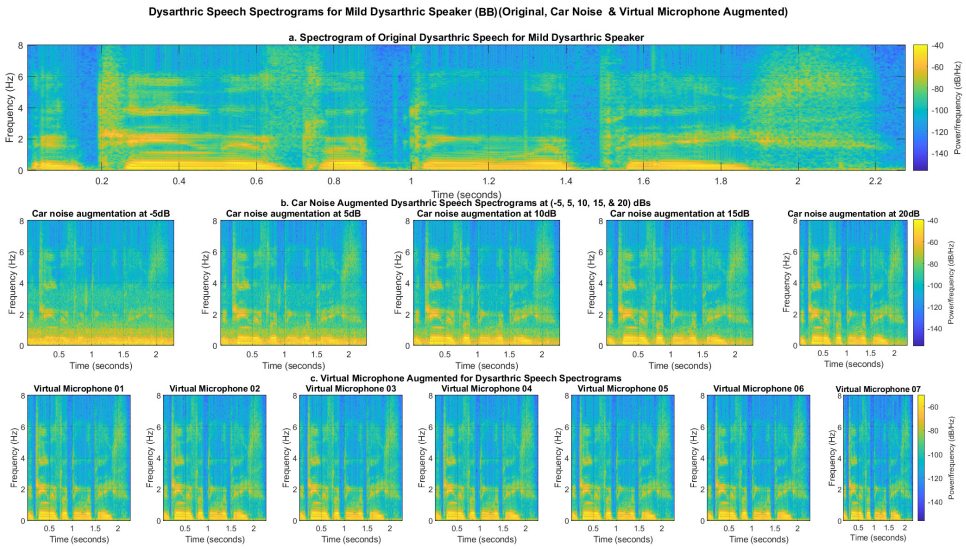

Figure 3. Spectrogram of dysarthric speech signal for a mild dysarthric speaker (BB) and its augmented versions

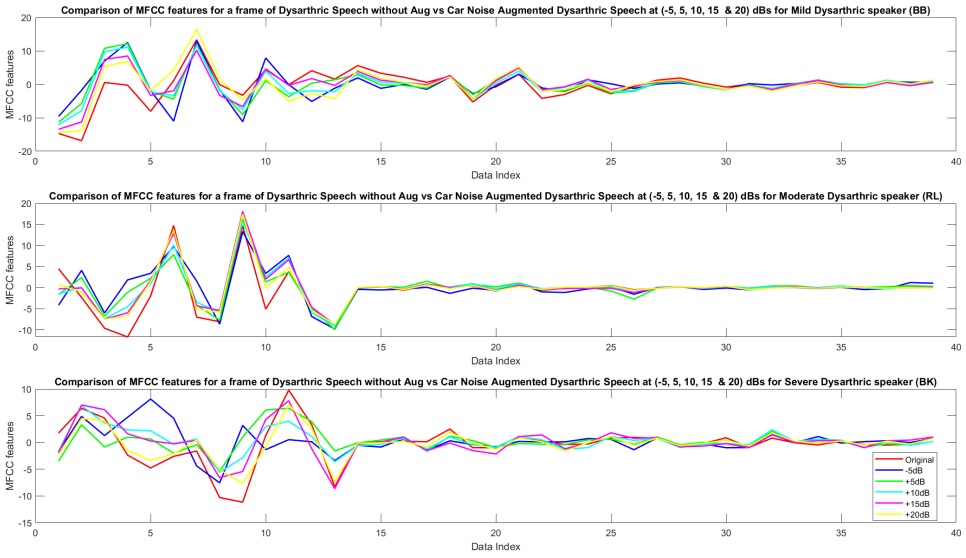

Figure 4. Line graph comparing the features of original dysarthric speech data for mild, moderate, and severe dysarthric speakers

that not all the lines are identical or completely overlapping as it is shown through the black box in this figure. Complete overlapping would indicate that the augmented features are mere copies of the original ones, potentially introducing bias during training. Instead, the observed variations confirm that the augmented examples are transformed versions of the original data.

Furthermore, while these transformations alter the feature values, increasing or decreasing them, the overall MFCC feature shape of the signal in the line graph remains unchanged. This preservation of the MFCC feature shape indicates that the speech characteristics of every phoneme (rightly articulated or misarticulated by the dysarthric speaker), including the inherent errors typical of the dysarthric speaker, are retained. Specifically, the frequency components corresponding to each sound unit remain intact, ensuring that the augmented examples represent the original data in a transformed but meaningful way.

Additionally, the correlation coefficient in Table 1 shows the extent to which the speech characteristics are retained from the original in the noise-augmented version as this correlation is obtained using the MFCC features.In a similar way, each dysarthric speech example was augmented with noise data across various SNR ranges, resulting in 40 additional unique examples (8 types of noise and 5 dB levels). Therefore, after augmentation, each word has a minimum of 80 examples.

Table 1. Correlation Between the MFCC of the Original Dysarthric Speech Data (Speaker (BB)) and Noise-Augmented Speech Data for the Same Speaker (BB)

Category of Noise	Correlation Coefficient
Volvo	0.79
Gulf	0.62
Car	0.73
Benz	0.68
Train	0.61
Bus-i	0.7
Bus-j	0.74
Babble	0.58

The noise-based augmentation approach is quite flexible regarding the number of augmented examples, as it is not limited to a specific number. By contrast to the previous work of (Nawroly et al., 2023, the current study demonstrates the flexibility of incorporating additional low-frequency noises for augmentation by utilizing different noises. This noise-based dysarthric speech augmentation technique is applied to train the source model.

This study incorporates data from three categories of dysarthric speakers: individuals with mild, moderate, and severe symptoms. Training the source model for a specific dysarthric speaker follows a category-based approach. For example, when training the source model for a mild dysarthric speaker *"X"* all the examples from the mild category except for *"X" are* used. As a result, new source models need to be trained for each dysarthric speaker based on a "leave 1-out" approach. Further details on training the model will be provided in Section 4.

3.2 Stage 2: VM-MRFE-based Data Augmentation Approach for Target Model Training

In the second stage, virtual microphone array (MA) signals are synthesized using the original sample from the dysarthric speaker, followed by the application of multi-resolution feature extraction (MRFE) to these synthesized MA signals (Thekekara Antony et al., 2020; Thekekara Antony et al., 2023). A linear array configuration with 7 microphones generates the virtual microphone array (MA) signals. To prevent spatial aliasing, the microphones are positioned 0.02 meters apart. The virtual MA signals are produced using the phase spectrum, incorporating a phase shift e^{jkd} (Arcienega et al., 2000), corresponding to a time delay denoted by d. In this context, $k = 2\pi f/c$, where c represents the speed of sound (343 m/s).

Let M represent the spectrum of the source signal, and M_n represent the spectrum of the n^{th} element, which is given by:

$$M_n = Me^{jk(n-1)d}; \qquad n = 1,2,.....,N \tag{1}$$

where N is the total number of elements in the array.

In Figures 2(c) and 3(c), the virtual MA signals and their corresponding original dysarthric speech signal are shown. These virtual MA signals are used for multi-resolution feature extraction (MRFE) (Priyanka et al., 2013) to increase the number of examples at the feature level. The signals undergo extraction of 39-dimensional MFCC features, which consist of 13-dimensional static MFCCs, and 13-dimensional (delta and

acceleration) coefficients. Various resolutions of these characteristics are derived using different window sizes, ranging from 10 ms to 20 ms at 2 ms intervals, with each window size operating at a 50% frame rate, resulting in 5 unique resolutions.

This paper uses a starting window size of 10 ms for MRFE to guarantee that the frame size exceeds one pitch period for every speaker. Consequently, by using the VM-MRFE method, each source example is expanded by a factor of *35 (7 (virtual MA examples*5 (MRFE))* through this VM-MRFE data augmentation technique.

An interesting aspect of these two data augmentation methods is that they use the original dysarthric speech samples to create augmented data while preserving the identity and speech errors of the dysarthric speaker. These techniques are used with continuous dysarthric speech data, and the approaches for expanding the pool of augmented dysarthric speech samples are considered. As it was discussed in the previous section, a two-stage transfer learning approach is then trained using a specific dysarthric speech example and its corresponding source model.

4. Training Two-Stage Transfer-Learning

4.1 Approach for Dysarthric Speech Recognition System

The concept of transfer learning entails commencing with a pre-trained model using a vast dataset and subsequently adjusting the parameters with a smaller dataset. This approach allows the model to learn general features from the source model and then adapt them to the target model, improving its performance with less labelled data. The model used in this study is a pre-trained model that is trained on specific categories using a noise-data augmentation technique. To perform this, a pre-trained model is trained using a category of dysarthric speakers' speech data and its corresponding noise-augmented examples, leaving the target dysarthric speaker out of this pre-training process. Thus, for each target dysarthric speaker a separate source model is involved. For the Nemours corpus, as shown in Table 2 (sourced from (Thekekara Antony et al., 2019)), dysarthric speakers are classified into mild, moderate, and severe categories.

In Table 2, the distribution example for pre-training the TL model while leaving out the target is shown. The original training of the source model includes extracting 13-dimensional MFCC features. These features are later transformed into a 40-dimensional vector using linear discriminant analysis and maximum likelihood linear transform, as it was shown in Figure 1. Additionally, a speaker adaptive training along with a Gaussian mixture model – Hidden Markov model (GMM-HMM) training, using feature space maximum likelihood linear regression as described in (Xiong et al., 2019) is performed. Next, a DNN architecture, referred to as factored time delay neural networks (TDNN-F), is employed, combined with convolutional neural networks (CNNs). The core of this architecture consists of five CNN layers that receive input from 40-dimensional log-Mel spectrogram features. Following these layers are nine TDNN-F layers and one linear layer, culminating in the output layer. The TDNN-F training process utilizes

Table 2. Number of dysarthric speakers in each category and number of examples for pre-training the source model in stage 1

Dysarthric Speaker ID	Category	No. of Examples for pre-training the source model
FB	Mild	40 (8 noises * 5dB levels) * 37 (training utterances) * 3 (speakers leaving the target) = 4440 examples
BB		
MH		
LL		
RL	Moderate	40 (8 noises * 5dB levels) * 37 (training utterances) * 2 (speakers leaving the target) = 2960 examples
JF		
BV		
SC	Severe	40 (8 noises * 5dB levels) * 37 (training utterances) * 2 (speakers leaving the target) = 2960 examples
BK		
RK		

the lattice-free maximum mutual information criterion. The linear layer is an additional hidden layer incorporated specifically for speaker adaptation purposes. The learning rates initially set at 0.0002 for training the source model over five epochs gradually decrease at 0.0005.

Since the source model pre-trains the initial layers of the neural network based on the common characteristics inherent to the noise-based data-augmented examples, the common characteristics are specific to a dysarthric category (mild, moderate, and severe) in the noise-based data augmentation method, which is centered on the acoustic information for each category rather than the diverse noise characteristics captured in the source model. This is because, for this approach, low-frequency analysis-based noises are specifically added to the dysarthric speaker, as opposed to introducing a generic noise. Hence, the speaker characteristics are preserved over the noise characteristics, which have a very poor influence in the dysarthric speech recognition systems.

With the pre-trained source model, the target model with VM-MRFE examples is fine-tuned at the second stage of the TL approach. The examples for each dysarthric speaker, as given in Table 1, are used to fine-tune each speaker individually, making this system a speaker-dependent one. For fine-tuning, for each dysarthric speaker there are 1295 dysarthric speech examples (35 * 37), 35 of them being related to VM-MRFE-based data augmentation, and the other 37 examples being allocated for training. In stage two, three epochs were used, and the learning rate was set at 0.0005, which is half of the value used by the original model.

The hyper-parameters are transferred from the original noisy dysarthric speech data to the VM-MRFE data by linearly adjusting the weights of the original model within the final TDNN-F layer. This is tailored to align with the characteristics of the augmented target dysarthric speech data. This two-stage transfer learning process is conducted separately for ten dysarthric speakers. The original model is selected for each speaker based on the category of the target dysarthric speaker.

As it was noted in Section 3, the remaining 37 utterances from each dysarthric speaker as test data without applying data augmentation are used. Table 3 shows the WER performance of the dysarthric speech recognition system based on two-stage transfer learning using this test data. Since it is a continuous dysarthric speech recognition system, word error rate (WER) would be a more appropriate scheme of evaluation. It can be observed from the table that for the mild category of dysarthric speakers, the reduction in the value of WER is almost 16.47%, while for the moderate category, it is 24.66%, and for the severe one, it is 34.54%. A greater reduction of WER is observed as the severity of the condition increases, which could also be attributed to the flexibility of the utterance syntax related to the analysed corpus. A paired t-test is conducted with regard to the WER obtained before and after data augmentation. The results of the paired t-test indicated that there is a significantly large difference between the Before (*Mean* = 31.8,

Table 3. WER performance of the Two-Stage TL-based dysarthric speech recognition system

Category	Dysarthric Speaker ID	WER for dysarthric speech recognition without data augmentation	WER for dysarthric speech recognition using two-stage TL-based data augmentation
Mild	FB	14.31	4.89
	BB	15.87	4.18
	MH	10.98	6.9
	LL	22.11	5.64
Moderate	RL	23.12	8.55
	JF	24.21	10.85
	BV	34.12	9.46
Severe	SC	52.33	22.4
	BK	58.12	26.05
	RK	63.19	38.65

SD = 19.2) and After (*Mean* = 13.8, *SD* = 11.5) category based TL-approach, where $t(9) = 6.2$, with $p < .001$. Since the p-value < α, the Null hypothesis is rejected, indicating that the alternate hypothesis of accepting the TL-model has shown a better improvement.

The source model comprises noisy dysarthric speech data, while the target model consists of VM-MRFE data, which primarily offers diverse feature examples without significantly altering the speech acoustics. This setup allows for the effective fine-tuning of general dysarthric features from the source model to the target model. Conversely, if the data augmentation approaches for the source and target model were reversed then the system would be more noise-resistant, rather than showing a reduction in word error rate. Nevertheless, the system could have been robust enough to handle noise.

5. Comparison of the Proposed Approach with the Data Augmentation Approaches in the Literature

The current research work is derived from the limitations of the works of Thekekara Antony et al. (2023) and Nawroly et al. (2023). The data augmentation approach in (Nawroly et al., 2023), provides the objective of augmenting the category-wise source model, as it inherits the speech characteristics specific to the three categories of dysarthric speakers, without any limits in creating a number of augmentation examples. The data augmentation approach in (Thekekara Antony et al., 2023) is used for the target model to make the target model specific to the speech characteristics of the target dysarthric speaker.

Hence, it would be fair to compare the results of the current work with the results obtained by these two approaches. Table 4 provides the WER performance comparison for the approach proposed in this paper with the data augmentation approach using VM-MRFE in (Thekekara Antony et al., 2023) for continuous dysarthric speech recognition and noise augmentation in (Nawroly et al., 2023).

The Table 4 shows that the approach proposed by the current paper continues to perform well with a low WER compared to both previous studies, especially when dealing with severe dysarthric speeches. Compared to the work of Thekekara Antony et al. (2023), that used normal speech data as the source model, the current work has used category-based dysarthric speech data, that has highly supported the reduction of WER to up to 11.369% for the severe dysarthric speaker category. Additionally, a comparative analysis with the works of Shahamiri et al. (2023) and Shah et al. (2023) is provided, where a dysarthric speech Transformer model is used for training.

Table 4. WER Comparison for the proposed approach and two other approaches in the literature

Category	**Dys. Speaker ID**	**Dysarthric speech recognition system augmented with VM-MRFE** (Thekekara Antony et al., 2023)	**Dysarthric speech recognition system using noise augmentation** (Nawroly et al., 2023)	**Dysarthric Speech transformer** (Shahamiri et al., 2023)	**The Proposed Approach**	**Average of the proposed approach**
Mild	FB	5.284	3.618	12.0	4.89	5.4025
	BB				4.18	
	MH				6.9	
	LL				5.64	
Moderate	RL	14.91	11.394	35.0	8.55	9.62
	JF				10.85	
	BV				9.46	
Severe	SC	40.402	50.69	43.0	22.4	29.03
	BK				26.05	
	RK				38.65	

In (Shahamiri et al., 2023), an attention model was trained on the UA corpus, utilizing transfer learning with the dysarthric speech dataset to address data sparsity. While this approach showed an improved accuracy for certain mild to moderate dysarthric speakers in the UA corpus, the overall average performance was comparatively lower. Hence, it can be understood from this comparison that the reduction in WER can be related to an increase in quality and in the number of unique examples from data augmentation. An increase in the low-frequency noise-based data augmentation for the category-based source model and in the target-specific data augmentation for the target model contributes to reducing the value of WER to approximately 11.369%.

6. Conclusion

This research introduces a category-based two-stage transfer learning method to enhance dysarthric speech recognition accuracy by addressing sparse data conditions through data augmentation. In the first stage, a category-based dysarthric speech recognition model for the mild, moderate, and severe dysarthric speaker categories using noise-augmented data is trained.

In the second stage, based on the category-wise source modelling, the target dysarthric speaker which is augmented by employing the VM-MRFE approach is trained using its corresponding category-based source model. Hence, each target dysarthric speaker uses his/her category-based source model for weight updating using transfer learning.

This paper uses two levels of data augmentation to create new dysarthric speech examples from the original dysarthric speech data and maintaining the dysarthric speaker's acoustic characteristics is important. The proposed two-stage approach led to more benefits for the severe dysarthric speaker category compared to the approaches in other similar works because it used speakers from the same category to train the source model.

REFERENCES

Arcienega, M., Drygajlo, A. & Malsano, J. F. (2000) Robust phase shift estimation in noise for microphone arrays with virtual sensors. In: *Proceedings of the 10-th IEEE European Signal Processing Conference, 4-8 September 2010, Tampere, Finland*. IEEE. pp. 1–4.

Borrie, S. A., Baese-Berk, M., Van Engen, K. & Bent, T. (2017) A relationship between processing speech in noise and dysarthric speech. *The Journal of the Acoustical Society of America.* 141(6), 4660-4667. doi: 10.1121/1.4986746.

Darley, F. L., Aronson, A. & Brown, J. R. (1975) *Motor Speech Disorders.* 1-st ed. Philadelphia, PA, WB Saunders.

Geng, M., Xie, X., Liu, S., Yu, J., Hu, S., Liu, X. & Meng, H. (2020) Investigation of data augmentation techniques for disordered speech recognition. In: *Proceedings of Interspeech 2020, 25-29 October 2020, Shanghai, China*. International Speech Communication Association (ISCA). pp. 696-700.

Jaitly, N. & Hinton, G. E. (2013) Vocal Tract Length Perturbation (VTLP) Improves Speech Recognition. In: *ICML Workshop on Deep Learning for Audio, Speech, and Language Processing (WDLASL 2013), 16 June 2013, Atlanta, USA.*

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T. S., Watkin, K., & Frame, S. (2008) Dysarthric speech database for universal access research. In: *9th Annual Conference of the International Speech Communication Association 2008 (INTERSPEECH 2008), 22-26 September 2008, Brisbane, Australia*. International Speech Communication Association (ISCA). pp. 1741–1744.

Ko, T., Peddinti, V., Povey, D. & Khudanpur, S. (2015) Audio Augmentation for Speech Recognition. In: *Proceedings of Interspeech 2015, 6-10 September 2015, Dresden, Germany.* pp. 3586-3589.

Menendez-Pidal, X, Polikoff, J. B., Peters, S. M., Leonzio, J.E. & Bunnell, H.T., (1996) The Nemours database of dysarthric speech. In: *4th International Conference on Spoken Language Processing (ICSLP 96*), 3-6 October 1996, Philadelphia, PA, USA. pp. 1962–1965.

Muthu Philominal, A. J., Nagarajan, T., Vijayalakshmi, P. (2020) Adaptive multi-band filter structure-based far-end speech enhancement. *IET Signal Processing*. 14(5), 288-299. doi: 10.1049/iet-spr.2019.0226.

Nawroly, S. S., Popescu, D. G., Thekekara Antony, M. C. & Muthu Philominal, and A. J. (2023) SNR-Selection-Based-Data Augmentation for Dysarthric Speech Recognition. *Studies in Informatics and Control.* 32(4), 129-140. doi: 10.24846/v32i4y202312.

Priyanka, M. A. S., Solomi, V. S., Vijayalakshmi, P. & Nagarajan, T. (2013) Multiresolution feature extraction (MRFE) based speech recognition system.

In: *Proceedings of IEEE International Conference on Recent Trends in Information Technology, 25-27 July 2013, Chennai, India*. IEEE. pp. 152–156.

Rudzicz, F., Namasivayam, A. K. & Wolff, T. (2012) The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*. 46(4), 523-541. doi: 10.1007/s10579-011-9145-0.

Shah, D., Lal, V., Zhong, Z., Wang, Q. & Shahamiri, S. R. (2023) Dysarthric Speech Recognition: A Comparative Study. In: *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), 25-27 October 2023, Bucharest, Romania*. IEEE. pp. 89-94.

Shahamiri, S. R., Lal, V. & Shah, D. (2023) Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. 31, 3407 – 3416. doi: 10.1109/TNSRE.2023.3307020.

Takashima, R., Takiguchi, T. & Ariki, Y. (2020) Two-step acoustic model adaptation for dysarthric speech recognition. In: *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020), 4-8 May 2020, Barcelona, Spain*. IEEE. pp. 6104–6108.

Thekekara Antony, M. C., Nagarajan, T. & Vijayalakshmi, P. (2016) Dysarthric speech corpus in Tamil for rehabilitation research. In: Proceedings of the IEEE Region 10 Conference (TENCON), 22-25 November 2016, Singapore. IEEE. pp. 2610–2613.

Thekekara Antony, M. C., Anushiya Rachel, G., Nagarajan, T. & Vijayalakshmi, P. (2019) A weighted speaker-specific confusion transducer-based augmentative and alternative speech communication aid for dysarthric speakers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27(2), 187–197.

Thekekara Antony M. C., Nagarajan, T. & Vijayalakshmi, P. (2020) Data Augmentation Using Virtual Microphone Array Synthesis and Multi-Resolution Feature Extraction for Isolated Word Dysarthric Speech Recognition. *IEEE Journal of Selected Topics in Signal Processing*. 14(2), 346-354. doi: 10.1109/JSTSP.2020.2972161.

Thekekara Antony M. C., Nagarajan, T. & Vijayalakshmi, P. (2023) Data Augmentation Techniques for Transfer Learning-Based Continuous Dysarthric Speech Recognition. *Circuits, Systems, and Signal Processing*. 42(1), 601-622. doi: 10.1007/s00034-022-02156-7.

Vachhani, B., Bhat, C. & Kopparapu, S. K. (2018) Data Augmentation Using Healthy Speech for Dysarthric Speech Recognition. In: *Proceedings of Interspeech 2018, 2-6 September 2018, Hyderabad, India.* pp. 471-475.

Varga, A. & Steeneken, H. J. M. (1993) Assessment for automatic speech recognition: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech communication*. 12(3), 247-251.

Xiong, F., Barker, J. & Christensen, H. (2019) Phonetic Analysis of Dysarthric Speech Tempo and Applications to Robust Personalised Dysarthric Speech Recognition. In: *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 12-17 May 2019, Brighton, UK*. IEEE. pp. 5836-5840.

Xiong, F., Barker, J., Yue, Z. & Christensen, H. (2020) Source domain data selection for improved transfer learning targeting dysarthric speech recognition. In: *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, *4-8 May 2020, Barcelona, Spain*. IEEE. pp. 7424–7428.