EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

UTran-DSR: a novel transformer-based model using feature enhancement for dysarthric speech recognition

Usama Irshad[1], Rabbia Mahum[1*], Ismaila Ganiyu[3,4], Faisal Shafique Butt[2], Lotfi Hidri[3,4], Tamer G. Ali[3] and Ahmed M. El-Sherbeeny[3,4]

Abstract

Over the past decade, the prevalence of neurological diseases has significantly risen due to population growth and aging. Individuals suffering from spastic paralysis, brain attack, and idiopathic Parkinson's disease (PD), among other neurological illnesses, commonly suffer from dysarthria. Early detection and treatment of dysarthria in these patients are essential for effectively managing the progression of their disease. This paper provides UTrans-DSR, a novel encoder-decoder architecture for analyzing Mel-spectrograms (generated from audios) and classifying speech as healthy or dysarthric. Our model employs transformer encoder features based on a hybrid design, which includes the feature enhancement block (FEB) and the vision transformer (ViT) encoders. This combination effectively extracts global and local pixel information regarding localization while optimizing the mel-spectrograms feature extraction process. We keep up with the original class-token grouping sequence in the vision transformer while generating a new equivalent expanding route. More specifically, two unique growing pathways use a deep-supervision approach to increase spatial data recovery and expedite model convergence. We add consecutive residual connections to the system to reduce feature loss while increasing spatial data retrieval. Our technique is based on identifying gaps in mel-spectrograms distinguishing between normal and dysarthric speech. We conducted several experiments on UTrans-DSR using the UA speech and TORGO datasets, and it outperformed the existing top models. The model performed significantly in pixel's localized and spatial feature extraction, effectively detecting and classifying spectral gaps. The Tran-DSR model outperforms previous research models, achieving an accuracy of 97.75%.

Keywords Dysarthria speech recognition, Transformer encoder, Deep learning, Vision transformer, Feature enhancement block

1 Introduction

Speech is an essential mode of interaction among individuals. Impairment of this mode significantly complicates the communication process. Additionally, individuals with neurological conditions often experience a speech disorder known as dysarthria [1]. It is described as a motor speech disease characterized by ineffective control of the motor systems involved in speech production [2]. It is characterized as "mild, moderate as well as severe, or extremely severe." Extreme situations may impede speech function and communication, even with speech therapy [3, 4].

*Correspondence:
Rabbia Mahum
rabbia.mahum@uettaxila.edu.pk
[1] Department of Computer Science, University of Engineering and Technology Taxila, Taxila 47050, Pakistan
[2] Department of Computer Science, COMSATS University Islamabad, Wah Campus, Wah Cantt, Pakistan
[3] Industrial Engineering Department, College of Engineering, King Saud University, PO Box 800, Riyadh 11421, Saudi Arabia
[4] King Salman Center for Disability Research, PO Box 94682 , Riyadh 11614, Saudi Arabia

Dysarthric speech is stated to be 15 times slower than normal because the lips, tongue, and jaw are difficult to move [5, 6]. The difficulty pronouncing words may create physical and psychological suffering, limiting the afflicted person's capacity to participate fully in society [7]. Effective rehabilitation remedies are required to assist people to enhance their communication skills and continue productive lives. Without proper supervision and early rehabilitative training, treating the disease may become difficult and worsen over time. Diagnosing this speech obstacle is subjective but often expensive and time-consuming [8].

The distinctive characteristics of dysarthric speech make automated recognition difficult. Dysarthric speech abnormalities harm phoneme formation and pronunciation, significantly complicating automatic processing and identification. However, the total amount of public dysarthric speech databases is small. This constraint occurs because gathering a significant volume of speech from people with dysarthria is difficult owing to muscular exhaustion produced by the condition. As a result, the lack of dysarthric speech mockups stances a substantial encounter to the effective progress of automated speech recognition (ASR) methods for dysarthric speech. Still, the existing techniques may fail due to the unavailability of huge and real data for training. Therefore, when executing dysarthric automated speech recognition (ASR) systems, it is crucial to address the following challenges: (1) addressing the variability and inaccuracy of phonemes in such speech, (2) mitigating the limited availability of data, and (3) improving the accuracy of dysarthric speech recognition.

Traditional ASR techniques like Mel-frequency cepstral coefficients (MFCCs) work with short-term power spectra and cannot handle dysarthric speech's unpredictable nature. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) with long short-term memory (LSTM) and bidirectional LSTMs (BiLSTMs) enhance feature learning and temporal patterns' processing but lack the long-term relationships' consideration. Self-attention mechanisms in the transformer models are more effective at capturing long-range relationships but, at the same time, are computationally intensive. Improving the system's transportability includes data amplification distortion masking, speaker adaptation, and transfer learning. However, they may fail in highly fluctuating conditions. The use of supplementary videos to the audio signal increases the recognition accuracy and the model's complexity, and it requires a strong synchronization of the data from several modalities. Therefore, these methods seem not to adequately address the variability and distortion incorporated in the speech of people with dysarthria.

To resolve the issues of dysarthric speech recognition, we present our model named UTrans-DSR, an advanced end-to-end dysarthric speech recognition (DSR) model that employs U-shaped and transformer architectures. Transformers and neural self-attention methods [9] are among the most effective ASR approaches, achieving considerable outcomes. A transformer is a special deep-learning architecture that is normally applied when developing natural language processing-based systems. Transformer self-attention is a mechanism that allows each position in a sequence to attend to every other position in the sequence. This process enables the model to weigh the importance of each word when encoding a given word. This makes it possible for the model to capture the relations between the different words in a sentence, hence enhancing the understanding of a context.

Our model uses a hybrid encoder that includes feature enhancement block (FEB) and vision transformer (ViT) encoders to examine Mel-spectrograms and categorize speech as healthy or dysarthric. Mel spectrogram is a graphic illustration of an audio signal spectrum in the form of frequencies as it varies over time and is used for analyzing speech patterns. Thus, the UTran-DSR model stands out as a major improvement in the field of automatic speech recognition (ASR) for people with dysarthria. Traditional ASR systems, heavily reliant on Mel-frequency cepstral coefficients (MFCCs), struggle to accurately process the distorted and variable nature of dysarthric speech. To address this limitation, we introduce a novel feature enhancement block (FEB) designed to extract crucial speech features, capturing both fine-grained details and broader speech patterns. Identifying the dependencies is a critical task in tracking the speech signal over longer periods. We achieve this by utilizing transformer and vision transformer encoders, which enhance our model's coverage and depth in capturing these dependencies. Moreover, the deep supervision method used here fuses multiple layers' features to speed up the training and boost the overall performance. This study's contributions are outlined below:

- The creation of Utrans-DSR, a unique encoder-decoder architecture specifically designed for dysarthric speech recognition.
- Integration of the feature enhancement block (FEB) and vision transformer (ViT) to improve both local and global pixel information recovery.
- A deep-supervision method and sequential residual connections are used to increase spatial data recovery and convergence in models.
- Addressing the lack of dysarthric data using a multi-phase transfer learning strategy and significant audio data augmentation approaches.

- A detailed scrutiny of the influence of various architectural configurations and parameter changes on model performance.
- A comprehensive examination of UTrans-DSR's performance on the UA speech and TORGO datasets revealed considerable increases in detection accuracy and precision.
- A detailed per-speaker performance comparison with leading models demonstrates UTrans-DSR's better efficacy in dysarthric speech categorization.

The leftover segments of the paper are coordinated as follows: the "Related work" section describes the associated work. The "Methodology" section covers projected procedures and materials. The "Results and discussion" section deliberates the experiments' outcomes, whereas the "Conclusion" section offers the conclusion.

2 Related work

To provide objective and precise diagnoses for persons with dysarthria, a growing number of researchers are using deep learning algorithms to identify the disorder automatically. Several researchers rely on linguistic clues for speech recognition and use a range of feature extraction techniques to discover features in speech signals.

Zaidi et al. [8] used the TORGO dataset to explore dysarthric automatic speech recognition (ASR) with a similar goal of improving it. The combination of the hybrid hidden Markov model (HMM) framework was used to build their approach. The required parameters were adjusted by changing Gaussian mixture models (GMM), and the resultant configuration was applied for training a deep learning (DL-HMM) ASR model. Feifei et al. [10] Investigated a technique for non-linearly adjusting speech tempo. They employed an automatic speech recognition (ASR) structure to survey speech beats at the phonetic level by consuming a forced-alignment approach from the outmoded Gaussian mixture model and hidden Markov model (GMMHMM). Rather than utilizing time-domain signals, the anticipated tempo changes were applied directly to the acoustic characteristics. The trials revealed that modifying normal speech to mimic dysarthric speech was more successful for data augmentation in customizing dysarthric ASR training. This resulted in roughly a 7% improvement over the baseline speaker-dependent system tested using the UA-Speech corpus. In recent years, research has switched to deep neural networks (DNNs). Dong et al. [11] developed transformer and attention-based ASR, proposing and evaluating a 2-D attention mechanism on the Spatialized Multi-Speaker Wall Street Journal (SMS-WSJ) normal speech corpus. This investigation resulted in much-decreased training expenses and an exceptional word error rate (WER), confirming the speech transformer's efficiency and effectiveness. Since their work, transformers have been utilized in different ASR frameworks.

Yilmaz et al. [12] proposed a model using "bottleneck features and pronunciation features" to decrease the auditory space variance instigated by dysarthric speakers' unfortunate capabilities of pronunciation. This approach aims to improve automatic speech recognition (ASR) accuracy for dysarthric speech. Additionally, researchers have undergone improvements in acoustic feature extraction by employing speaker-adaptive models to minimize discrepancies between dysarthric speech's acoustic spaces and mapped acoustic characteristics to phonemes. Narendra [13] used the UA-Speech information base to identify dysarthria and utilized the CNN-LSTM hybrid model as the arrangement model, accomplishing a precision of 77.57%. To enhance the correctness of the dysarthria identification model, this investigation applied a short-time Fourier transform (STFT) to speech signals collected from dysarthria patients and healthy individuals before converting the signals into spectrograms. Subsequently, the signals were distorted into a map of spectrals, and features were selected using mel-frequency cepstral coefficients (MFCC). At last, the arranged CNN-GRU (gated repetitive unit) deep learning model was tried for dysarthria recognition exactness against three extra models (CNN, LSTM, and CNN-LSTM).

Initial seq-to-seq ASR structures were often created by consuming recurrent neural networks (RNNs). Bahdanau et al. [14] used a deep bi-directional RNN to encode the signal of voice into an appropriate feature depiction, followed by an attention-based recurrent sequence generator RNN to interpret this demonstration into a character arrangement. Hussain and Alaa [15] discovered that plain deep neural networks (DNNs) were ineffective for dysarthric speech recognition, so they developed a hybrid model (CRNN) that merged recurrent neural networks (RNNs) with convolutional neural networks (CNNs) and trained it on samples from the TORGO database [16]. The findings revealed that adding a convolution layer to the conventional RNN enhanced performance, surpassing the regular CNN and potentially increasing dysarthric speech recognition accuracy to 40.6%, compared to 31.4% for CNN. Takashima and colleagues [17] developed a speech recognition system using the attend, hear, and spell technique. This system offered different models for the English and Japanese languages. Their research highlighted the benefits of employing several databases for speech recognition in this scenario. It is important to emphasize that advancements in language extraction of features and the application of linguistic methodologies are not mutually exclusive. Instead, they collaborate to enhance the overall effectiveness of DSR.

M.S. Yakoub et al. [18] dysarthric speech recognition by combining empirical mode decomposition and Hurst-based mode selection (EMDH) with a CNN. The EMDH approach preprocesses the speech, enhancing its quality, and Mel-frequency cepstral coefficients generated from the treated speech are fed into the CNN-based recognizer. Our EMDH-CNN technique, evaluated on the Nemours corpus, shows considerable accuracy increases of 20.72% and 9.95% over baseline HMM-GMM systems and CNNs without augmentation, respectively, as shown by k-fold cross-validation. T. Takiguchi et al. [19] Addressed the problem of local overfitting by merging convolutionally constrained Boltzmann machines with a CNN that had a limiting structure for previously trained models. To address the large variation in phoneme among different talkers with dysarthria, Hahm et al. [20] used Procrustes matching (a physiological articulatory approach), vocal tract length normalization (VTLN), and maximum likelihood linear regression (MLLR) were employed. The study employed the amyotrophic lateral sclerosis (ALS) database and discovered that training the deep neural network-hidden Markov model (DNN-HMM) using auditory and articulatory data and normalizing three ways produced the best results. The best combination in the reference had a phoneme error rate of 30.7%, which was 15.3% lower than the baseline approach "triple phoneme Gaussian mixture model-hidden Markov model (GMM-HMM)" trained on audio samples. These discoveries recommend that adding data from both the hearable and articulatory spaces may fundamentally further develop ASR exactness for dysarthric speech.

The Korean phonetically optimized words (KPOW) database, Korean phonetically rich words (KPRW) database, Korean phonetically balanced words (KPBW) database, and SI dysarthria adaption were utilized to recognize dysarthria speech using Kullback–Leibler hidden Markov model (KL-HMM) and compared to GMM-HMM and DNN-HMM. The KL-HMM structure was shown to be useful in improving dysarthric speaker's performance [21]. The bimodality of phonetic perception, as well as the proficient utilization of audio-visual speech recognition (AVSR), perceives natural speech [22, 23]. Allow the exploration of the local area an opportunity to utilize visual portrayal to expand dysarthric discourse identification. Contrasted to ASR frameworks that depend just on sound figures, AVSR is more precise and tough. Liu [24] fostered an audio-visual speech recognition (AVSR) system for speech jumble utilizing a Bayesian gated brain system to join visual and sound information sources, surpassing the early-level DL-based automated speech acknowledgment model. Miyamoto and colleagues [25] fostered an imaginative strategy for distinguishing dysarthric speech. Their cycle involved the assessment of various acoustic edges in procuring acoustic data, which tended to the hardships in identifying dysarthric speech prompted by muscular strain. The absence of audio-visual datasets is really difficult when AVSR is involved in dysarthric speech. To address this test, another methodology for cross-modular synthesis of visual attributes was created [26], which utilized the LRS2 lip-reading dataset and UA Speech sound accounts to develop a varying audio-visual reversal structure. This technique empowered the making of visual attributes. The productivity of this cross-modular synthesis approach was additionally affirmed by Liu et al. [27] by effectively lessening the number of word mistakes. In this research, Javanmardi et al. [28] used the pre-trained models, namely, wav2vec2-BASE, wav2vec2-LARGE, and HuBERT, examined in terms of their capability of serving as feature extractors for dysarthric speech recognition using two datasets: UA-Speech and TORGO. SVM and CNN classifiers were used to assess the performance of these proposed features compared to three benchmarks (MFCC, openSMILE, and eGeMAPS). It was evident that based on features obtained from previous models, improvements were recorded and HuBERT was the best performing achieving higher accuracy. This superior performance is attributed to HuBERT's utilization of context network embeddings to create hidden units that improve the quality of targets. Nevertheless, the study identifies that there is scope for enhancement when working with dysarthric speech to achieve optimum system performance. However, the complexity of these pre-trained models requires a lot of computational resources, which may be a constraint when it comes to real-time or low-resource-based applications. Sajiha et al. [29] introduced automatic dysarthria detection (ADD) and automatic dysarthria severity level assessment (ADSLA) models based on the layered CWT-CNN approach. Such wavelets included Amor, Morse, and Bump, which were used to compare the performance and effectiveness of the models on the two datasets, TORGO and UA-Speech. Several tests carried out suggested that among the investigated wavelets, the Amor wavelet provided the most accurate reconstruction of the signal after compression and decompression, suppressed the noise more effectively, and gave higher accuracy. Thus, for the UA-Speech dataset, the ADD's accuracy was 97.00%, and ADSLA's accuracy was 93.70%. This result emphasized the fact that the selection of wavelets for signal processing plays a vital role, and hence, the Amor wavelet yielded better accuracy for both datasets. However, the general workflow of the CWT-layered CNN model involves a great number of computations, which might prove to

be a disadvantage in terms of real-time usage. Also, the study is based solely on two datasets, and it is required to confirm the obtained approach's effectiveness on a greater number of dysarthria databases. Redha et al. [30] presented a convolutional neural networks (CNN) model that incorporates STFT layers with the technique adapted for the detection of dysarthria and the determination of its severity using datasets, namely, the TORGO and UA-Speech datasets. It examines all the basic changes to the first layer of CNNs, namely spectrogram, log spectrogram, and pre-emphasis filtering (PEF). However, it was found that with five learnables, PEF had the highest accuracy at 99%. Eighty-nine percent on the UA-Speech dataset. The application of log spectrogram and different PEF variants increased the discrimination capability by increasing the focus on important acoustic features. It is worth noting that with learnable parameters, the PEF has several disadvantages. It resulted in the creation of a more complex model that requires more computationally intensive procedures and takes more time to train. The likelihood of overfitting was also present and may increase when working with a small sample of data, which, in turn, could decrease the model's generalizability.

Traditional ASR techniques like Mel-frequency cepstral coefficients (MFCCs) work with short-term power spectra and are incapable of handling dysarthric speech's unpredictable nature. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) with long short-term memory (LSTM) and bidirectional LSTMs (BiLSTMs) enhance feature learning and temporal patterns' processing but lack the long-term relationships' consideration. Self-attention mechanisms in the transformer models are more effective at capturing long-range relationships but, at the same time, are computationally intensive. Some methods of improving the system's transportability include data amplification distortion masking, speaker adaptation, and transfer learning. However, they may fail in highly fluctuating conditions. The use of supplementary videos to the audio signal increases the recognition accuracy but also increases the model's complexity, and it requires a strong synchronization of the data from several modalities. Therefore, these methods seem not to adequately address the variability and distortion incorporated in the speech of people with dysarthria.

3 Methodology

The proposed method to deal with dysarthric speech is introduced in this segment. To separate between solid and dysarthric speech, our pipeline incorporates feature extraction, preprocessing, data growth, and a grouping stage. In any case, sound examples are assembled and changed into Mel spectrograms, which give an exhaustive portrayal of the speech patterns of individual speakers.

To help the assortment of the dataset and reinforce the flexibility of the model, the data is exposed to different data augmentation methods after standardization, for example, time-shifting, noise addition, and pitch variety. The U-Trans-DSR model is then given augmented data. Figure 4 illustrates how the model architecture uses U-Net's skip connections to maintain fine-grained information and transformer-based layers for effective feature extraction. The resultant classification layers achieve a high degree of accuracy and reliability by analyzing the spectrogram characteristics to discriminate between dysarthric and healthy speech. Figure 1 provides a graphical representation of this process, showing the whole flow from raw audio input to output for categorization.

TransUNet is the base of our dysarthric speech recognition model due to its capacity to fuse the advantages of transformers and convolutional neural networks (CNNs). TransUNet offers an efficient framework for tackling the challenge of dysarthric speech detection by skillfully combining these two approaches. TransUNet's architecture is flexible and may be used to assess speeches by varying speakers, requiring global context awareness as well as local feature extraction. Due to its adaptability, we were able to add our own feature enhancement block (FEB) to further enhance the model's dysarthric speech optimization. Through the utilization of TransUNet's hybrid structure and its customization to our particular requirements, we have established a resilient and efficient model for this demanding field.

3.1 Mel-spectrograms

Traditional acoustic models, which largely depend on phonetic details to effectively map a link between speech

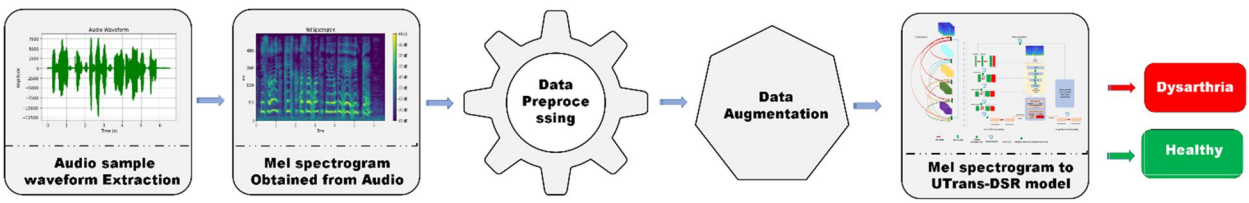

Fig. 1 Block diagram for the detection of speech dysarthria

voices and phonetic patterns, are challenged by the heterogeneity in phoneme generation across persons with dysarthria. Our study changed its emphasis from just phonetic data to incorporating the visual features of speech in response to this challenge. We attempted to solve the lack of dysarthric speech data by using visual-data augmentation techniques. Moreover, we investigated if there are any observable differences in the visual patterns of dysarthric and non-dysarthric speech using the UTran-DSR model.

We found connections in the Mel-spectrograms from a variety of dysarthric and non-dysarthric utterances throughout our research. The frequency spectrum of sounds and their variations over time are represented visually in a Mel-spectrogram, which is useful for analyzing speech patterns in both dysarthric and normal speakers. Figure 2 displays mockup waveforms for a dysarthric speaker (Fig. 2a) and a healthy speaker (Fig. 2b), collected with the supplementary Mel-spectrograms. The Mel-spectrograms vividly label the power of sound incidences at dissimilar stages by exhausting a color pattern where navy blue indicates low strength and yellowish-green signifies high strength. These graphic representations highlight the variances among the speech patterns of healthy and dysarthric speakers, proposing information that may aid in growing healthier speech recognition and assessment practices.

3.2 Data preprocessing

We preprocessed [31] the Mel-spectrograms prior to moving further with the extension. This included denoising, histogram equalization, and climbing them to a typical size of 256×256 pixels. These activities were vital to advance the input data's quality and constancy and offer deep learning models with additional reliable training. We also altered the contrast sceneries to showcase the unique elements inside every spectrogram and improve the model's performance. By sensibly arranging the data, we condensed biases and errors that may arise during the training procedure, which improves the exactness and flexibility of our model's forecasts.

3.3 Data augmentation

We augmented our training dataset by consuming the visual data augmentation methods [32] to resolve the issue of inadequate dysarthric voice data. We used pre-existing

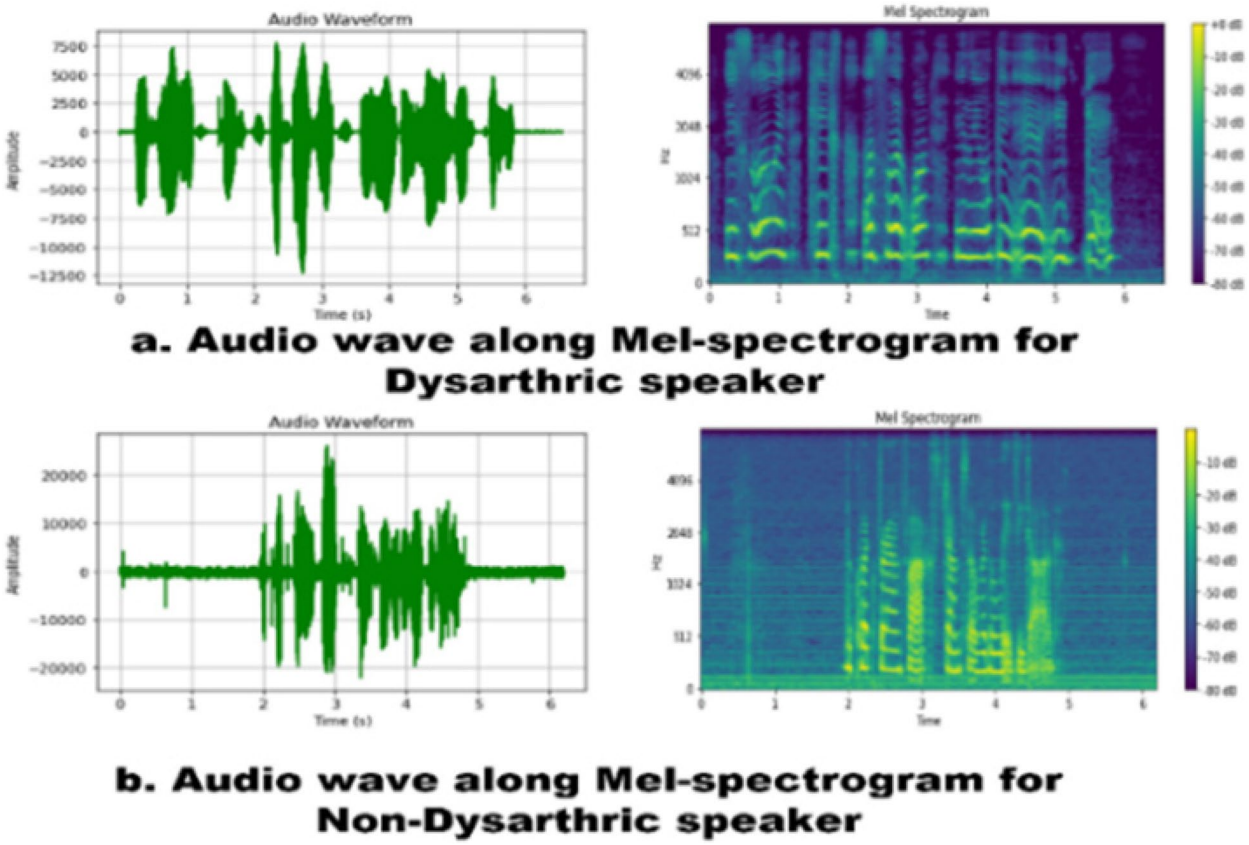

Fig. 2 Mel-spectrogram along with audio wave

samples to generate new Mel-spectrograms and applied numerous alterations such as rotation, scaling, horizontal shifting, shearing, zooming, and magnifying. We were able to significantly upsurge the variety of dysarthric mel-spectrograms for model training. Figure 3 displays the fallouts of these augmentation methods, which exemplify the numerous modifications that were completed. These augmented samples are vital for enhancing our models' flexibility and applicability in identifying and comprehending dysarthric speech.

3.4 UTran-DSR model

We propose a dysarthric speech recognition model, namely UTran-DSR, which attempts to improve TransUnet's architecture [33, 34] by efficiently converting raw audio to Mel-spectrograms. To make the classic TransUNet architecture acceptable for processing Mel-spectrograms of dysarthric speech, our model adds four major architectural alterations. To process the spectrogram features, we first swapped out TransUNet's basic ResNet50 module combined with the feature enhancement block (FEB). The characteristics of dysarthric speech, which may be less obvious than those of regular speech patterns, are especially well captured by this adaption.

Secondly, we implemented two distinct extended pathways: the first one for class-token characteristics and another for gathering spectrogram patterns. The class-token patterns within the transformer encoding work as a compact representation of the whole speech sample. This enables the model to process and recover both localization information—such as the solid gaps in Mel-spectrograms caused by slurred speech, and global semantic information analyzing overall speech at the same time, which is essential for effective speech recognition.

Thirdly, we use a deep supervision technique to aggregate the results of the two growing routes. By providing more direct gradients during backpropagation, this method not only speeds up the training process but also combines the benefits of both original spectrogram features and class-token features in an integrated manner to improve model performance for dysarthric speech recognition tasks.

Fourthly, we added consecutive residual connections from the early layers to the output to reduce the loss of significant spectrogram features throughout the network. To extract specific elements from complicated dysarthric speech patterns, these links aid in maintaining important information and enhancing gradient flow. See Fig. 4 for a comprehensive illustration of how these improvements are included in our model.

Figure 4 provides an overview of the UTran-DSR architecture alongside a detailed representation of the feature enhancement block (FEB). The UTran-DSR schematic illustrates the integration of the FEB blocks within the encoder and decoder paths, highlighting the dual-path strategy for processing Mel-spectrograms. The depiction of the FEB block showcases its internal structure, emphasizing the residual and dense connections that enhance feature utilization and stability throughout the network.

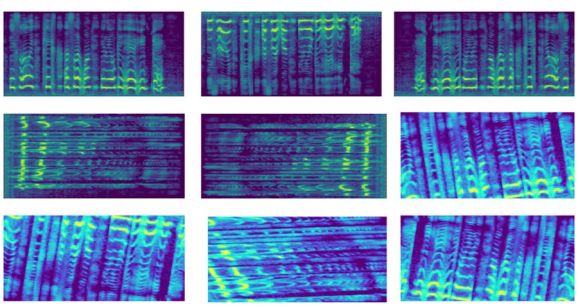

Fig. 3 Augmented samples of the TORGO speech dataset

3.5 Feature enhancement block (FEB) module

To significantly improve the extraction and representation of features from dysarthric speech, our model substitutes the feature enhancement block (FEB) for the baseline ResNet50 module in TransUNet. Distinct from ResNet50, which utilizes a conventional collection of convolutional layers and residual connections, FEB is designed to precisely tackle the distinct obstacles presented by dysarthric speech. By concentrating on both local and global speech patterns, FEB improves feature extraction and makes it possible for it to recover contextual information and minor nuances that ResNet50 could overlook. The block's ability is increased by the combination of sophisticated convolutional processes and spatial attention methods. Because FEB's design offers multi-scale feature extraction and enhanced resilience, it is capable of addressing the variability and distortions present in dysarthric speech. Due to this, FEB is especially good at identifying and deciphering the many phonetic abnormalities that may be seen in dysarthric speech. As a result, as compared to the conventional ResNet50, FEB's integration produces enhanced feature representations and improved overall performance in dysarthric speech detection.

By using the feature enhancement block (FEB) [35, 36], which is intended to optimize the processing of Mel-spectrograms for dysarthric speech recognition, our model considerably improves the TransUNet architecture. To enhance feature flow and reduce information loss throughout the network, this block combines dense and residual connections.

The following combinations are used by the FEB block:

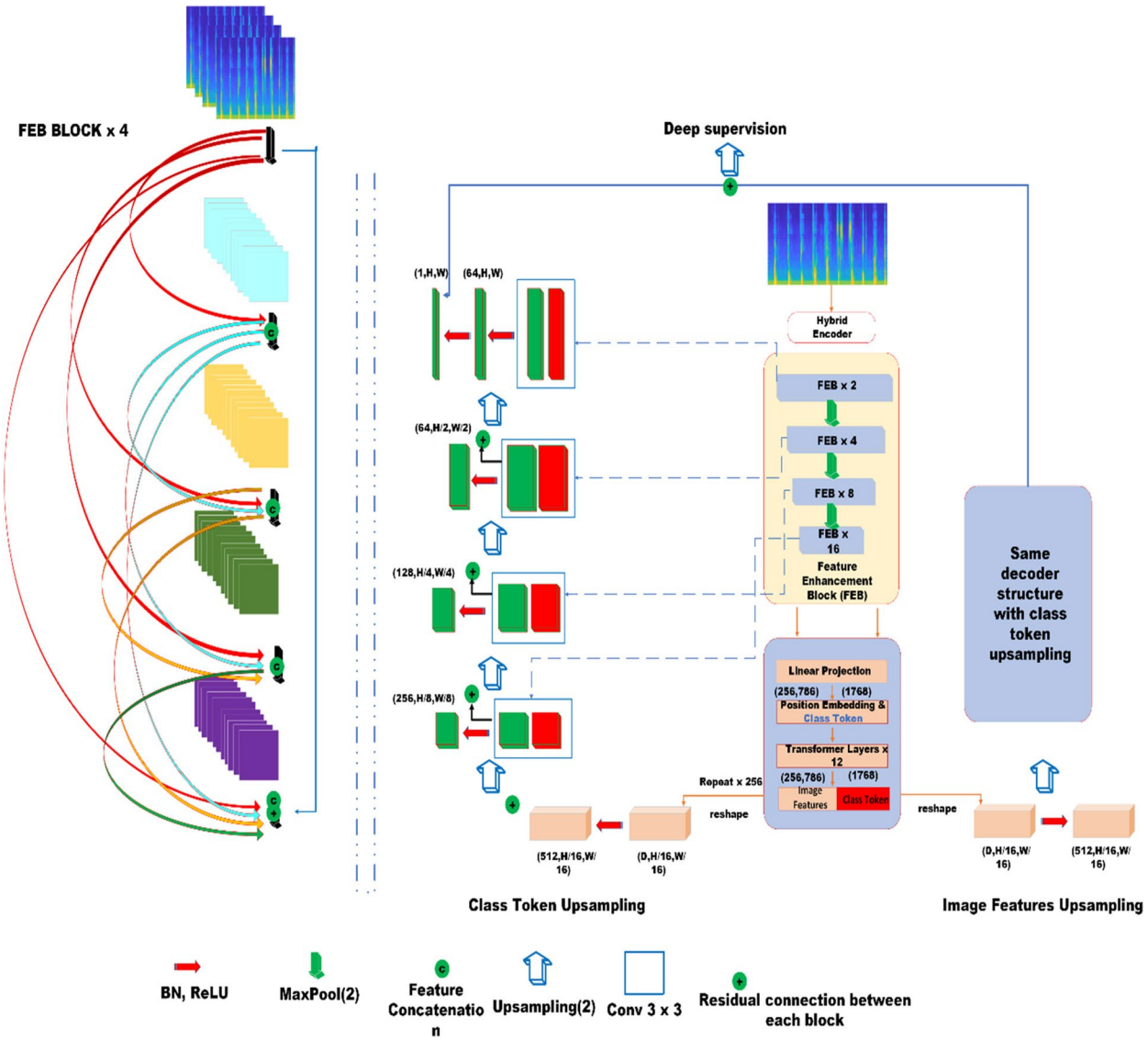

Fig. 4 Overview of the UTran-DSR architecture

Residual connection:

$$\boldsymbol{H_L}(\boldsymbol{X}) + \boldsymbol{X_L} - 1, \tag{1}$$

Dense connection:

$$\boldsymbol{X_L} = \boldsymbol{H_L}(\boldsymbol{X}_0, \boldsymbol{X_i} \cdots \boldsymbol{X}_{L-1}), \tag{2}$$

Combined Connection:

$$\boldsymbol{X_L} = \boldsymbol{H_L}(\boldsymbol{X}_0, \boldsymbol{X_i}, \cdots \boldsymbol{X}_{L-1}) + \boldsymbol{X}_{L-1}, \tag{3}$$

In this case, robust feature extraction and reusability is promoted by the non-linear transformation represented by $\boldsymbol{H_L}(\bullet) + \boldsymbol{X}_{L-1}$, which involves 1×1, 3×3, and 1×1 convolutional sequences with batch normalization and ReLU activations $\mathbf{X_0}$ through $\boldsymbol{X}_{L-1}$. The concatenated feature maps from previous layers are represented as $[\boldsymbol{X}_0, \boldsymbol{X_i}, \cdots \boldsymbol{X}_{L-1}]$, improving the model's ability to learn from a large feature collection. These layers' channel counts, or "feature increment," of FEB block for our model we set at 32 and 64 to maximize the network's flexibility and feature processing effectiveness. The impact of varying feature increments is observable in Fig. 6.

3.5.1 Encoding and decoding

The encoder had four feature enhancement block (FEB) components and a Transformer component. The former

reduces image resolution by 256×256 to 16×16. Each FEB component has a distinct quantity of modules and parameters according to its growth pattern. We extracted the outputs of the four modules and employed skip connections based on matching sampling features.

The latter extracts 16×16 image features and generates a class-token pattern that represents the image's overall significance. The transformer component will generate both the class-token arrangement and its initial image characteristics independently. The decoding unit has two distinct expansion routes and could upscale abstract image characteristics and class-token patterns to their unique image area. In addition, multiple skipped connections were employed among each down-sampling and up-sampling layer, as visible via the dashed lines in Fig. 4.

3.6 Vision transformer (VIT)

3.6.1 Input version

2D Mel-spectrograms were processed using the feature enhancement block (FEB) and successfully condensed into feature-rich representations. The FEB block's output features were linearly projected on flattened 2D sequence patches $\boldsymbol{x_p} \in \boldsymbol{R^{N\times P^2\cdot C}}$, indicated as preparing the data for input to the Transformer block. C denotes the channel number for each patch, while P determines the patch's resolution. The total number of patches, N, is calculated as $\frac{\boldsymbol{HW}}{\boldsymbol{P^2}}$, where H and W are the height and breadth of the feature map produced by the FEB block.

3.6.2 Class-token utilization

The transformer encoder layers maintain a learnable class-token sequence with a constant length of $\mathrm{D}=\boldsymbol{P}^2\times\mathrm{C}$. This sequence collected global semantic information that was essential for thorough speech interpretation. The transformer extracted the class-token sequence, concatenated its copies, and reshaped them from $\boldsymbol{z_L} \in \boldsymbol{R^{N\times D}}$ To $\frac{\boldsymbol{H}}{\boldsymbol{p}} \times \frac{\boldsymbol{w}}{\boldsymbol{p}} \times \boldsymbol{D}$. This reshaped tensor was subsequently sent via two independent expanding channels inside the decoder, improving the model's capacity to refine both global and localized features.

3.6.3 Spatial information encoding

Positional embedding [37] was critical for preserving the sequence's spatial context, which was naturally destroyed during the flattening of 2D patches. These embeddings, known as Epos, were added to the patch embedding to represent relative and absolute positioning information inside the spectrogram.

$$z_0 = \left[x_{class}x_p^1E; x_p^2E; \cdots x_p^NE\right] + E_{pos}, E \in R^{(P^2,C)\times D}, E_{pos} \in R^{(\mathrm{N}+1)\times D} \tag{4}$$

3.6.4 Transformer encoding

The transformers, which are encoded in our model, have a similar architecture as its novel vision transformer (ViT) with L layers, where L is a hyper-parameter. Equations (5) and (6) demonstrate that every layer includes an MLP block and Multi-Head Self-Attention (MSA). After each block, a residual connection and dropout are applied. Layer normalization (LN) comes first, followed by residual connection and dropout. The MLP block is made up of two linear segments coupled by a (Gaussian Error Linear Unit) GELU instigation function. The equations are given below:

$$\boldsymbol{z_L^{'} = MSA(LN(z_{l-1})) + z_{l-1}))LN_{z_{l-1}}, l = 1 \ldots L} \tag{5}$$

$$\boldsymbol{z_l = MLP\left(LN\left(z_L^{'}\right)\right) + z_L^{'}, l = 1 \ldots L} \tag{6}$$

3.7 Feature enhancement block (FEB)–Utransnet

3.7.1 Advanced supervising

Class token [38] represents global semantic information that encompasses all mel-spectrograms and captures primary aspects for thorough interpretation. However, using a single class-token combination collects semantic data from the backdrop and further asymmetrical objects, leading to non-discriminatory and chaotic localization. To correctly retrieve semantic and geographical data using our approach, we used deep supervision to collect the class-token series and mel-spectrogram features, upsampling outputs [39].

Deep supervision is employed to combine outputs from two different routes, which improves training speed and accuracy. To direct the training process, deep supervision involves implementing auxiliary loss functions at the network's intermediary levels. This method combines outputs from the two expanding paths, one emphasizing global context perceptions and the other local feature extraction. Furthermore, deep supervision gives these intermediary layers direct gradients via the introduction of the intermediate loss functions, which help them to train more efficiently. This method makes sure that every route makes a significant contribution to the final prediction by imposing uniform learning weights throughout the network. Moreover, this leads to more immediate input for the model during training, which improves overall convergence and enhances feature integration. In the end, the deep supervision method helps develop a more reliable and accurate model for dysarthric speech recognition.

3.7.2 Residual upsampling

The mel-spectrogram segmentation challenge requires pixel-level classification, with the label having the same resolution as the actual image. To overcome this, we added two separate expansion routes to the decoder that up-sample both the image's characteristics and the sequence in which the tokens are classified. Each extended route comprised many elements, including two 3×3 convolutions, activation of the ReLU functions, and double-scale upsampling levels. The quick connections were utilized to maintain vital characteristics as well as ease the up-sampling.

4 Results and discussion

In this segment, we demonstrate the dataset, metrics, investigational details, and results attained by our proposed model.

4.1 Dataset

For training and experimentation, we used publicly available datasets, such as UA-Speech [40], ASVSpoof 2019 [41], and TORGO [42]. Our primary training dataset was the UA-Speech database, which comprises around 11,475 audio samples from 15 dysarthria speakers. To effectively overcome class imbalance, the normal speech dataset from the ASVSpoof2019 dataset was utilized in our study.

For cross-validation, we used the TORGO database, which contains both dysarthric and healthy speech samples, allowing us to evaluate the model's performance in a variety of speech circumstances. We used visual-data augmentation methods to create additional spectrograms by altering them using width shifting, shearing, and zooming, resulting in a much larger training set and improved model performance. Tables 1 and 2 contains information about the datasets utilized.

We obtained a total of 89,875 speech files, including 57,375 dysarthric speech samples from UA-Speech, 7500 normal speech samples from ASVSpoof 2019, and 25,000 samples from TORGO. To train, we separated the data into 80% training and 20% testing sets, yielding 71,900 training samples (45,900 dysarthric and 26,000 normal) and 17,975 testing samples (11,475 dysarthric and 6500 normal).

Table 2 The characteristics of the UA speech dataset

Participants	Speech comprehension (%)	Age	Level of comprehension	Level of dysarthria
M08	95	28	High	Low
F05	95	22		
M10	93	21		
M09	86	18		
M14	90	44		
M11	62	48	Mild	Mild
F04	62	18		
M05	58	21		
M07	28	58	Low	High
F02	29	30		
F16	43	40		
M01	17	18	Very Low	Very High
M04	2	18		
M12	7	19		
F03	6	51		

The TORGO dataset, used for cross-validation, has both dysarthric and healthy speech samples, totaling 25,000 (12,500 dysarthric and 12,500 healthy).

It is important to mention that we converted these audios into 256×256 visual Mel spectrograms to capture their time–frequency properties, which are required for successful speech recognition and analysis for classification.

4.2 Evaluation metrics

We assessed the UTrans-DSR utilizing various measurements, including accuracy, exactness, review, and F1 score. Accuracy is the proportion of accurately projected occurrences to total instances. Precision is the extent to which genuine positive forecasts add up to positive expectations (true positives plus false positives). Recall is the proportion of precise positive estimates to add up to genuine up-sides (true positives plus false negatives). F1 Score is the consonant mean of exactness and review, which gives solitary data to demonstrate execution for managing unequal classes. The equations are presented below.

Table 1 Details of the dataset utilized both before and after augmentation

Dataset	Participants	No. of Audio clips	Mel spectrogram from clips(before augmentation)	After augmentation
UA Speech	15	11,475	11,475	57,375
ASVSpoof 2019 dataset (Normal)	4	1500	1500	7500
TORGO	8	5000	5000	25,000

$$\textbf{Accuracy} = \frac{\textbf{True Positive} + \textbf{True Negative}}{\textbf{Total Positive} + \textbf{Total Negative}} \quad (7)$$

$$\textbf{Percision} = \frac{\textbf{True Positive}}{\textbf{True Positive} + \textbf{FalsePositive}} \quad (8)$$

$$\textbf{Recall} = \frac{\textbf{True Positive}}{\textbf{True Positive} + \textit{False}\ \textbf{Negative}} \quad (9)$$

$$\textbf{F1Score} = 2 \times \frac{\textbf{Percision Rate} \times \textbf{Recall Rate}}{\textbf{Percision Rate} + \textbf{Recall Rate}} \quad (10)$$

4.3 Training overview

For our model's training to categorize healthy and dysarthria audio samples, we employed the AdamW optimizer [43], with default hyperparameters except for weight decline and the rate of learning. We determined the learning ratio to $2e-5$ and used CyclicLR to schedule it. The size of the batch was fixed to 32, and the model was trained across 15 epochs. We used binary cross entropy with dice loss as the loss function. During training, each batch received basic data augmentation such as normalization, randomized rotation (degree = 15), and random horizontal flipping ($p = 0.5$), but not throughout the inference phase. Rather than loss values, we monitored parameter adjustments throughout the training and validation procedures using the validation set's mean intersection over union (mIoU). In addition, we employed early stopping criteria with a threshold of 5 epochs for no progress in mIoU to discontinue the learning. The pixel identification threshold has been placed at 0.4, which means that anticipated pixels larger than or equal to 0.5 were classified as one and pixels less than 0.5 as 0.

We ran multiple tests on a high-performance computer system to ensure efficient processing and accurate findings. The testing was done on a machine equipped with an NVIDIA GEFORCE GTX graphics card with 4 GB of RAM. The system also included a computer GPU server with four NVIDIA GEFORCE GTX GPUs, 16 GB of RAM, and an Intel Core i5 CPU, making it more capable of handling large amounts of work. This arrangement offered enough processing capacity to tackle complicated data augmentation and model training tasks successfully. Table 3 shows that the approach outperformed other approaches in identifying healthy as well as dysarthria audio samples. The mIoU index improved by 1.4 to 3.8 pts. Such better results are due to the use of the feature enhancement block (FEB), which may significantly decrease audio feature loss throughout the encoding. In contrast, many lower-resolution characteristics and multiple expanding paths using a deep supervising approach may yield more exact localization data, i.e., the minor gaps in mel-spectrograms. When comparing FEB-UTrans32 to FEB-UTrans-DSR, the latter delivers somewhat better results while requiring fewer parameters.

Table 3 The experimental results acquired using the UA Speech database ("FEB UTrans 32" denotes that the feature increment used in the FEB UTrans model had been set to 32.)

Approach	Coefficient (dice)	Coefficient (mIoU)	Recall	Precision	Parameters
U Net	97.7	87.2	96.5	98.1	28.9 M
U Net++	97.9	87.4	95.9	98.5	47.1 M
U Net3++	98.1	88.1	98.0	98.2	25.4 M
FEB 32 UTran-DSR	98.7	90.2	97.4	97.7	104.3 M
FEB 64 UTran-DSR	98.9	91.2	98.7	98.5	101 M

Table 4 The experimental results acquired using the TORGO database

Approach	Coefficient (dice)	Coefficient (mIoU)	Recall	Precision	Parameters
U Net3++	96.8	82.3	95.3	96.4	25.4 M
U Net	94.5	86.5	92.6	93.4	28.9 M
U Net++	96.6	83.4	92.3	94.2	47.1 M
FEB 32 UTran-DSR	98.2	94.3	97.3	97.5	104.3 M
FEB 64 UTran-DSR	98.3	96.4	98.3	98.4	101 M

To verify our model's generalization capacity, we tested it on a different dysarthria detection dataset, as shown in Table 4. In many circumstances, dysarthria identification is a class-imbalanced problem, with dysarthria samples comprising just a tiny part of mIoU (mean Intersection over Union) representing a common Intersection over Union (IoU) across every class, and each one is affected based on the total quantity and count of classes. The approach we construct concentrates on dysarthria locations and may overlook certain regular speech segments. Thus, the majority of forecasts for normal speech have a large influence on mIoU, resulting in a decreased mIoU %. Therefore, as a consequence, the dysarthria detection process returns a below mIoU value. The coefficient dice measures segmentation accuracy by dividing the intersection of expected outcomes and facts by their union. According to its definition, the dice is more responsive in dealing with class imbalances, making it especially ideal for measuring minority class segmentation performance. The findings in Table 4 indicate that our

approach surpasses others and has excellent generalization capabilities.

To show our model's enhanced detection findings intuitively, we performed qualitative comparisons using audio samples of healthy speech, mild dysarthria, moderate dysarthria, and severe dysarthria from the left side to the right side: original audio spectrograms, ground truth, TransUNet, U-Net++, U-Net, and Attention U-Net and FEB-UTransNet with 64 and 32 feature increment, respectively. Our qualitative comparison findings effectively reveal that the proposed approach has higher classification margins and identifies more dysarthria-specific factors than other models. Figure 5 shows the input and output map of features of the model produced by FEB-UTrans-DSR and UTrans-DSR to show the use of the FEB block and class-token order. The FEB block has four feature maps, while the ResNet50 block in TranUNet has three basic feature maps. In the beginning, we can view each of the basic levels feature representations from FEB-UTrans-DSR, concentrate on dysarthria-related features, and generate highlighted pixel values when contrasted with UTrans. Even with the most abstract acoustic features (the center and lower sections), the proposed model can excerpt useful data. It indicates that the FEB block may keep higher audio information than the ResNet50 block throughout the encoding procedure. The feature of output maps of both FEB-UTrans-DSR produces more dysarthria-related borderline properties, taking advantage of the class-token sequence's extra up-sampling approach.

4.4 Ablation study

Based on our baseline algorithm, we performed an ablation study on the UA Speech database. The fallouts are presented in Table 5. The original baseline model has a convolutional block, a transformer block, and a cascaded up-sampler (CUP) decoder block (Conv-Trans-CUP). Initially, we replace the original convolutional module with a feature enhancement block (FEB) that has a different feature increment. We found that FEB32-Trans-CUP and FEB64-Trans-CUP increased the mIoU score by 1.5 and 1.1 points, in that order in comparison to the baseline model. It also preserved the convolutional and CUP blocks; however, it added a class-token order to the transformer block and its expansion track (CEP), leading to a 0.6-point increase in mIoU. Lastly, to show the use of the global residuals, we included the remaining link into two extension pathways depending on CEP to achieve a one mIoU enhancement (CEPR).

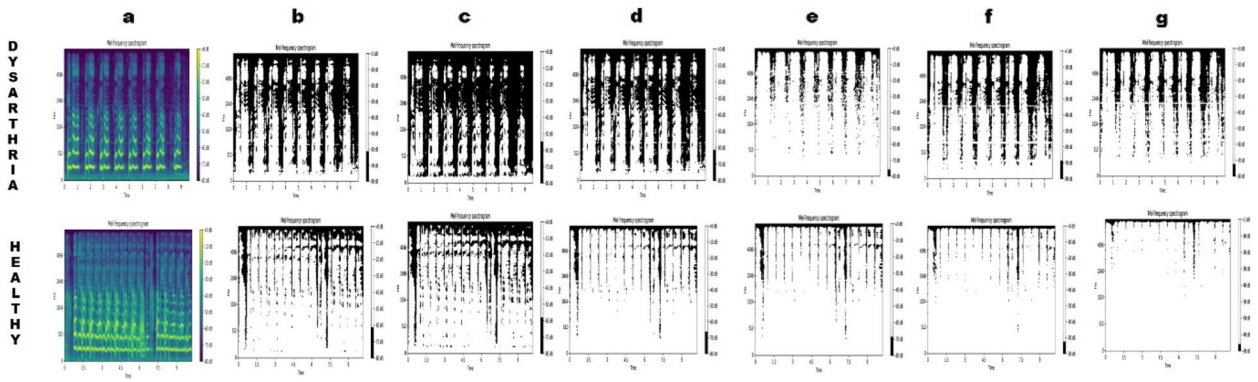

Fig. 5 A comprehensive comparison is constructed on **a** image, **b** ground truth, **c** U Net, **d** U Net++, **e** UNet3++, **f** FEB 64Tran-DSR, **g** FEB 32 UTran-DSR from left to right

Table 5 Ablation studies on various module implementations

Approach	Coefficient (dice)	Coefficient (mIoU)	Recall	Precision	Parameters
Baseline Model	93.6	81.5	95.4	96.2	105.3 M
FEB32-UTrans-CUP	94.3	83.1	95.6	96	101.1 M
FEB64-UTrans-CUP	96.5	82.6	96.0	97.4	97.9 M
Baseline-CEP	95.9	81.2	96.4	97.1	105.3 M
Baseline-CEPR	96.3	81.5	96.6	96.8	105.6 M
FEB-UTrans 32	98.2	94.3	97.3	97.5	104.3 M
FEB-UTrans 64	98.3	96.4	98.3	98.4	101 M

A comparison between feature maps produced by FEB-UTrans-DSR and UTrans-DSR provides strong proof of the FEB block's ability to capture characteristics unique to dysarthria. Subtle phonetic nuances, inconsistent speech patterns, and changes in voice volume and pitch—all of which are often seen in dysarthric speech—are all superbly extracted by the FEB block. The FEB block improves feature representations using sophisticated convolutional procedures, which helps the model distinguish between normal and dysarthric speech patterns. Prolonged phonemes, decreased pronunciation, and abnormal prosodic patterns are among the dysarthria-related traits that are more clearly defined in the feature maps produced by FEB-UTrans-DSR, as shown in Fig. 6. On the other hand, UTrans-DSR, which uses the standard ResNet50 module, often generates feature maps for these important characteristics of dysarthric speech that are less clear and informative. The model's overall accuracy and robustness in identifying dysarthric speech is mainly due to the FEB block's higher performance in capturing these dysarthria-specific elements in detail.

5 Results

The model we presented worked significantly, attaining an accuracy of 97.70%. This shows that the model can correctly identify dysarthric speech. Furthermore, the model displayed the ability to correctly identify positive cases while limiting false negatives, with a precision of 97.75% and 97.43% recall, respectively. The F1 score attained was 97.41%, indicating a balanced performance. Table 6 presents the results of the proposed model.

Figure 7 depicts a confusion matrix that was formed to acquire further information about the performance of the model. For every class, the confusion matrix provides a thorough split of projected and true labels. The training of the model was assessed for accuracy and loss. Figures 8 and 9 show how accuracy and loss evolve over training epochs. The accuracy plot indicates that the model's performance improves with each consecutive training period. The misfortune plot shows a comparable pattern, showing that the model took in the fundamental examples with the entry of preparing and diminished its loss.

Table 6 Performance evaluation of the proposed model

Parameters	Value obtained
Accuracy	0.9770
Precision	0.9775
Recall	0.9743
F1-score	0.9741

We also provide the receiver operating characteristic (ROC) curve for evaluating the effectiveness of the model. The ROC curve computes the true positive rate (TPR) and false positive rate (FPR), which helped to assess our model's classification ability. Figure 10 depicts the ROC curve for our model.

5.1 Evaluation with existing DSR systems

The effectiveness of UTrans-DSR was compared to numerous techniques employed in earlier works, with a summary provided in Table 7. Hernandez et al. [44] employed a machine-learning model to identify dysarthria by extracting fricative sounds from speech. They employed auditory fricatives as input features for an SVM model. This technique produced an accuracy of 72%. Rajeswari et al. [45] improved speech sounds using variational mode breakdown, then fed the resultant signals through a CNN to train. Their method achieved an accurate score of around 95.95. 95.95%. Narendra et al. [13] trained an SVM using auditory and glottic variables taken out from encoded speech sounds, as well as dysarthria tags. Their approaches produced an outstanding

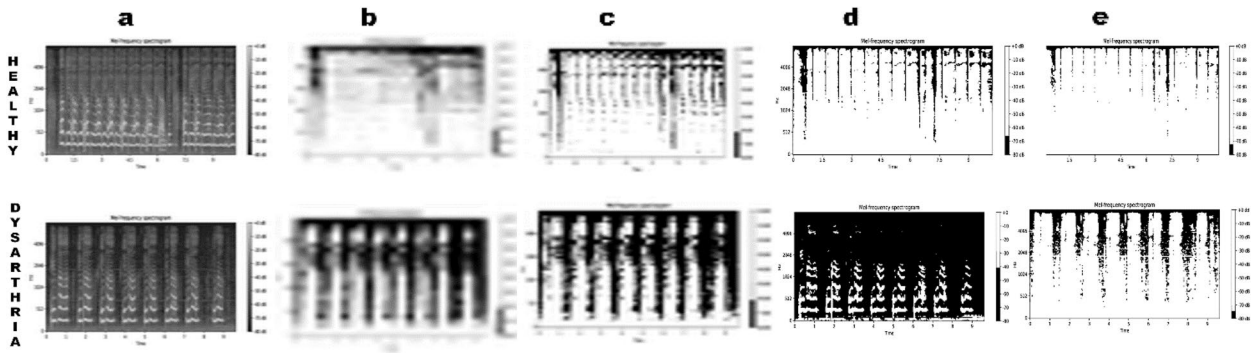

Fig. 6 Illustration of feature maps. From left to right: FEB-UTrans-DSR 32 and FEB-UTrans-DSR64. **a** The original picture is shown in the left corner, followed by 16×16, 32×32, 64×64, 128×128 feature maps, as well as the model-generated feature maps

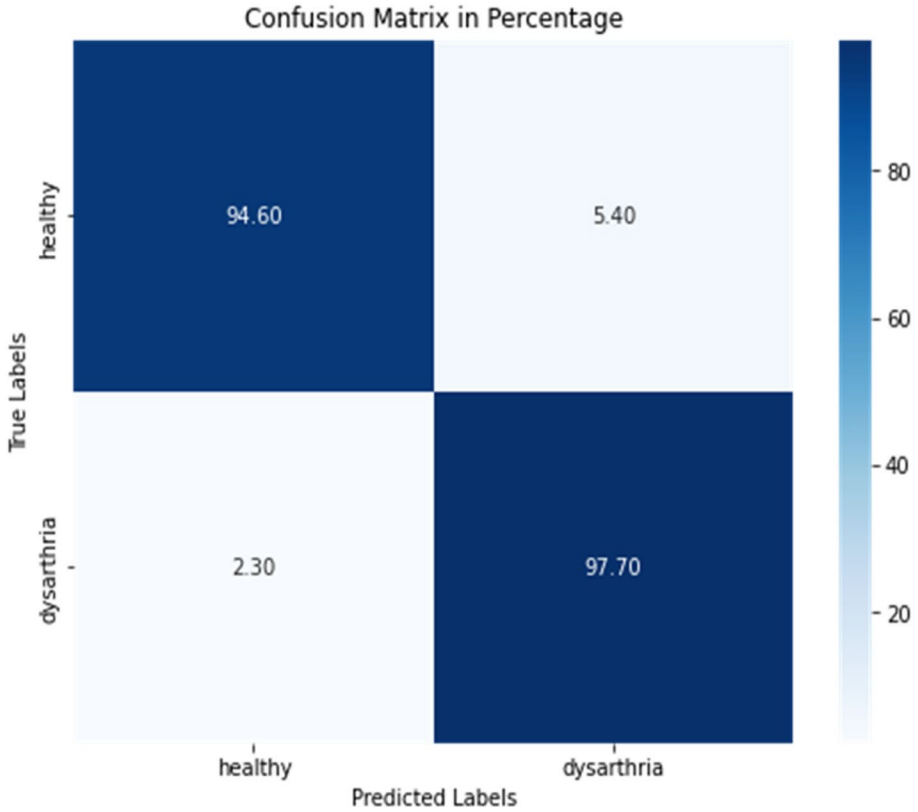

Fig. 7 Confusion matrix

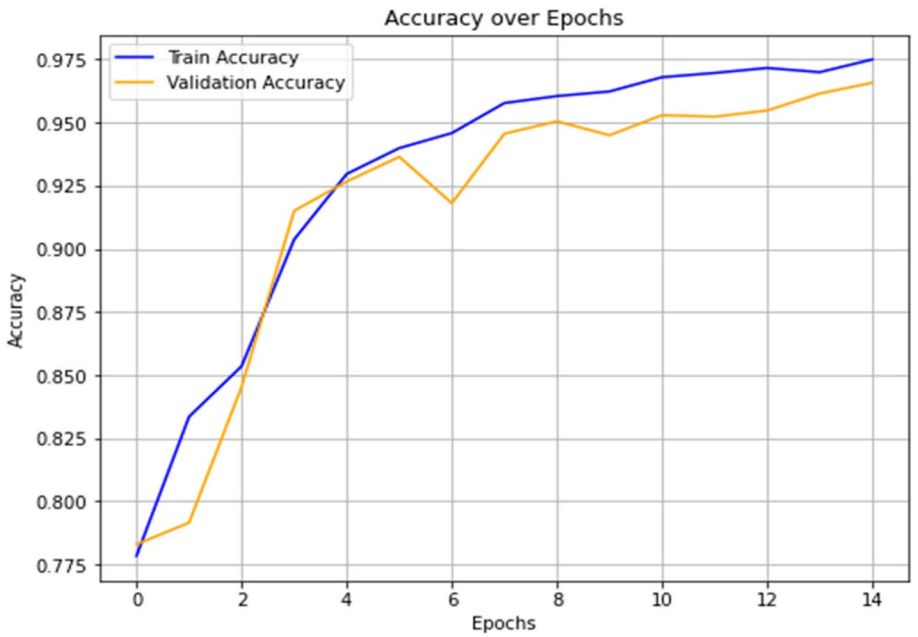

Fig. 8 Accuracy of the model over epochs

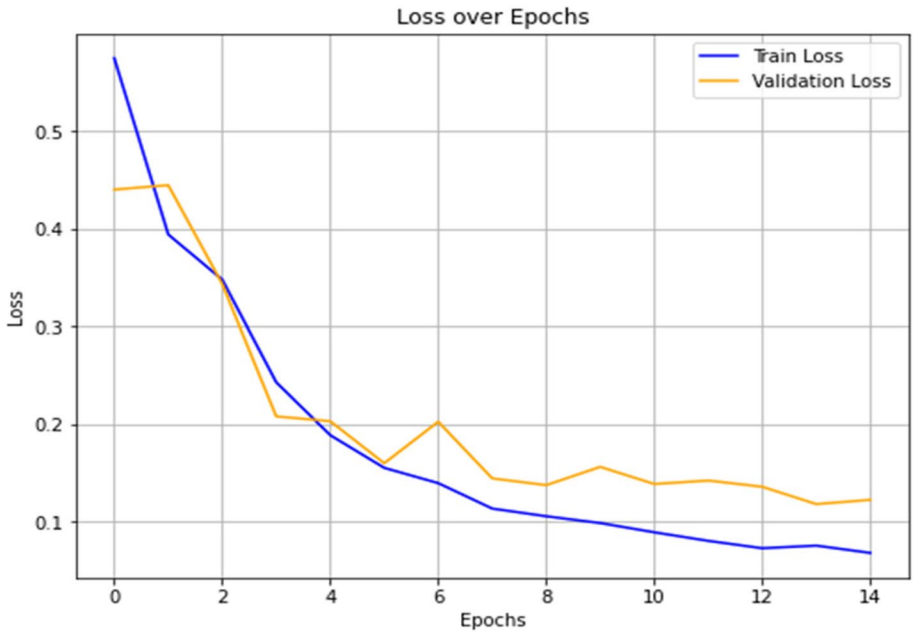

Fig. 9 Loss of the model over epochs

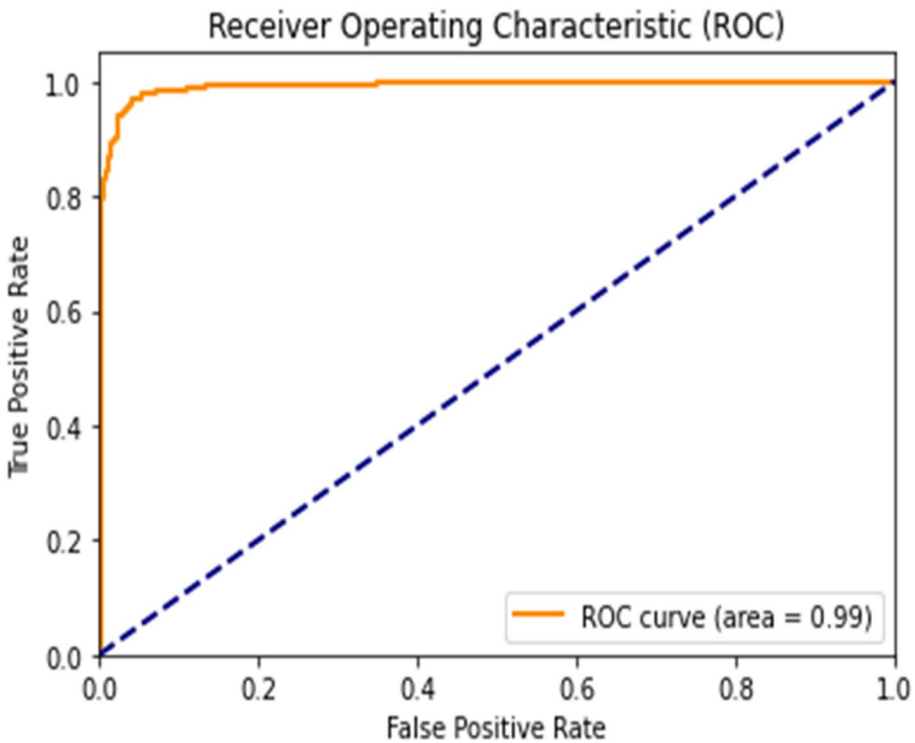

Fig. 10 ROC curve of UTrans-DSR

96.38% accuracy for the SVM model. Narendra et al. [46] created an end-to-end network for dysarthria detection that was mainly reliant on unprocessed sound and glottal flowing patterns. They compared two deep learning architectures: CNN with MLP and CNN with LSTM. The findings indicated that using the initial glottal flowing waveforms was more useful for model training than real speech. CNN-MLP attained an accuracy of 87.93%,

Table 7 Comparison of the proposed model with existing techniques

Study	Dataset	Year	Approach	Accuracy (%)
Hernandez [44]	UA speech	2019	SVM	72%
Narendra [46]	UA speech	2019	SVM	96.38%
Narendra [13]	UA speech	2020	CNN-MLP CNN-LSTM	87.93% 77.57%
Rajeswari [45]	UA speech	2022	Convolutional neural network	95.95%
Shahamiri [47]	UA speech	2023	Transformer-encoder	68%
Almadgor [48]	UA speech	2023	CNN-transformer	64.79%
Sunakshi [49]	-	2024	BiLSTM_GRU	97.64%
UTrans-DSR	UA speech + TORGO	2024	Improvement in U-net	97.75%

whilst the CNN-LSTM got 77.57%. Shahamiri et al. [47] created a pair of approaches that included attention-based and transformer techniques. They examined 45 dysarthric models with different transformer-encoder designs and found that augmenting audio data enhanced detection performance. Their method resulted in an estimated recognition accuracy of 68%. Almadgor et al. suggested a transformer-based DSR system with CNN in an end-to-end architecture [48]. They used the UA speech sample to do trials at varying levels of dysarthric speech, from mild to severe. Their approach has a maximum recognition accuracy of 90.75% at the intermediate level. Similarly, among these models, UTrans-DSR had the best accuracy, precision, and recall. Table 7 provides a comparison study, demonstrating that our suggested UTrans-DSR obtains the greatest accuracy of 97.75 on the UA speaking + TORGO dataset. Figure 11 shows a comparison of all suggested models to the state-of-the-art using the UA speech dataset.

6 Conclusion

The UTrans-DSR is a more advanced encoder-decoder architecture designed specifically for assessing Mel spectrograms and identifying speech as healthy or dysarthric. The encoder had adjustable block depths that were employed using different feature increments, enabling it to effectively reduce feature loss. We greatly enhanced pixel localization feature removal and

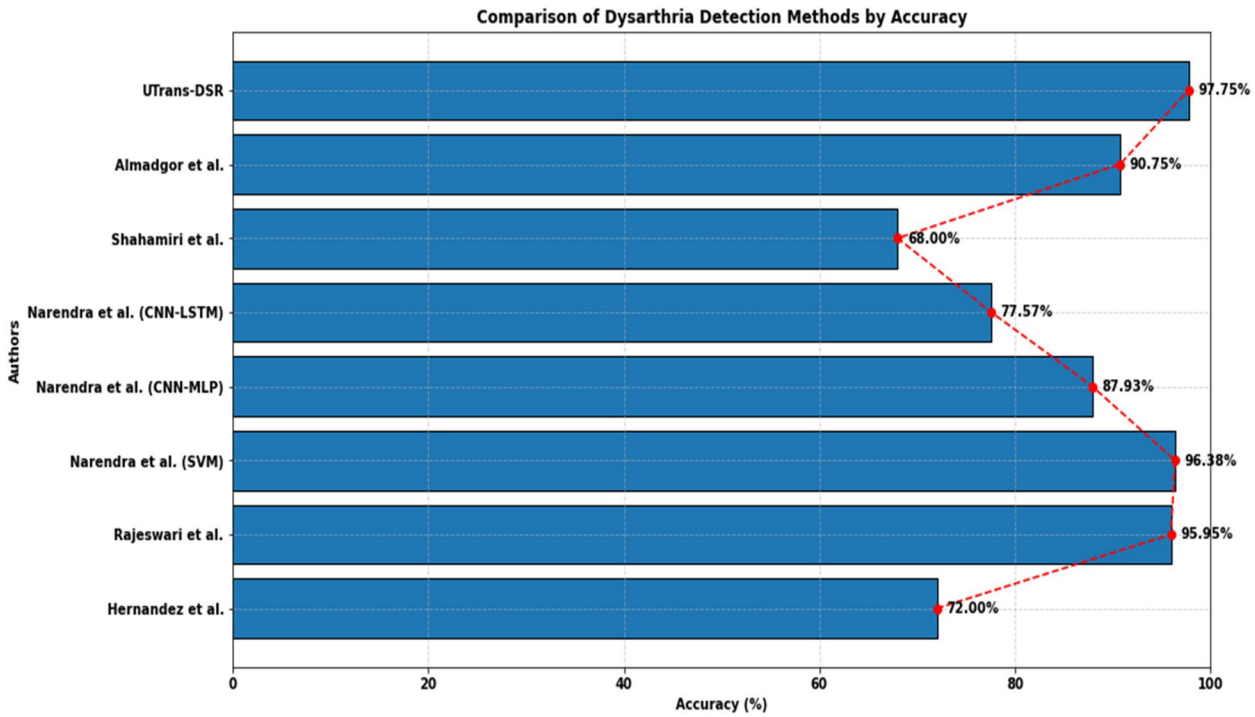

Fig. 11 The corresponding plot between proposed models and known methodologies

spatial data retrieval by using the feature enhancement block (FEB), ViT, dual-path decoder, fully supervised approach, and consecutive residual connections. Our approach focuses on recognizing gaps in Mel spectrograms, which are essential for discriminating between normal and dysarthric speech. More specifically, if the gaps are minimal, the speech is considered healthy; if the gaps are more frequent, the speech is characterized as dysarthric. The experimental findings on two datasets, UA Speech, and TORGO, demonstrated that our model could accurately identify and categorize gaps, resulting in the precise and reliable classification of speech. However, we understand that our technique may have limitations and raise issues. It is dependent on data fluctuations and needs significant computing power, and the selection of parameters is critical to efficiency. Looking forward, we want to update our model to address these limits and difficulties. We will extend the model to include a wider range of speech datasets and continue to study effective techniques to increase our model's robustness so that it can be used in real-time speech classification applications to aid diagnosis.

Code availability
The code can be provided on demand.

Authors' contributions
All authors contributed to the study's conception and design. All authors read and approved the final manuscript.

Funding
The authors extend their appreciation to the King Salman Center for Disability Research for funding this work through Research Group no KSRG-2022–136.

Availability of data and materials
Data sharing does not apply to this article as authors have used publicly available datasets, whose details are included in the "experimental results and discussions" section of this article.

Declarations

Ethics approval and consent to participate
Not applicable. All authors gave their consent to participate.

Consent for publication
All authors agreed to submit the manuscript for publication in the journal.

Competing interests
The authors declare that they have no competing interests.

Received: 12 June 2024 Accepted: 29 August 2024
Published online: 11 October 2024

References
1. R. Mahum, A. Irtaza, A. Javed, EDL-Det: A Robust TTS Synthesis Detector Using VGG19-Based YAMNet and Ensemble Learning Block. IEEE Access **11**, 134701–134716 (2023)
2. S. Sapir, A.E. Aronson, The relationship between psychopathology and speech and language disorders in neurologic patients. J. Speech Hear. Disord. **55**(3), 503–509 (1990)
3. A.B. Kain et al., Improving the intelligibility of dysarthric speech. Speech Commun. **49**(9), 743–759 (2007)
4. F. Rudzicz, G. Hirst, P. van Lieshout, *Vocal tract representation in the recognition of cerebral palsied speech* (2012)
5. M.J. Kim, J. Yoo, H. Dim, Dysarthric speech recognition using dysarthria severity-dependent and speaker-adaptive models, in *In Interspeech*. (2013)
6. G. Van Nuffelen et al., Speech technology-based assessment of phoneme intelligibility in dysarthria. Int. J. Lang. Commun. Disord. **44**(5), 716–730 (2009)
7. M. Ali, P. Lyden, M. Brady, Aphasia and dysarthria in acute stroke: recovery and functional outcome. Int. J. Stroke **10**(3), 400–406 (2015)
8. B.F. Zaidi et al., Deep neural network architectures for dysarthric speech analysis and recognition. Neural Comput. Appl. **33**(15), 9089–9108 (2021)
9. R. Mahum, A. Irtaza, A. Javed et al., DeepDet: YAMNet with Bottle-Neck Attention Module (BAM) for TTS synthesis detection. J AUDIO SPEECH MUSIC PROC **2024**, 18 (2024). https://doi.org/10.1186/s13636-024-00335-9
10. F. Xiong, J. Barker, H. Christensen, Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (IEEE, 2019, May), pp. 5836–5840
11. L. Dong, S. Xu, B. Xu, Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition, in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. (IEEE, 2018, April), pp. 5884–5888
12. E. Yılmaz et al., Articulatory and bottleneck features for speaker-independent ASR of dysarthric speech. Comput. Speech Lang. **58**, 319–334 (2019)
13. N. Narendra, P. Alku, Glottal source information for pathological voice detection. IEEE Access **8**, 67745–67755 (2020)
14. D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, Y. Bengio, End-to-end attention-based large vocabulary speech recognition, in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. (IEEE, 2016), pp. 4945–4949
15. H. Albaqshi, A. Sagheer, Dysarthric speech recognition using convolutional recurrent neural networks. Int. J. Intell. Eng. Syst. **13**(6), 384–392 (2020)
16. F. Rudzicz, A. Namasivayam, T. Wolff, The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang. Resour. Eval. **46**, 1–19 (2010)
17. Y. Takashima, T. Takiguchi, Y. Ariki, End-to-end dysarthric speech recognition using multiple databases, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (IEEE, 2019), pp. 6395–6399
18. M. Sidi Yakoub et al., Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network. EURASIP J. Audio Speech Music Proc. **2020**, 1–7 (2020)
19. R. Takashima, T. Takiguchi, Y. Ariki, Two-step acoustic model adaptation for dysarthric speech recognition, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (IEEE, 2020), pp. 6104–6108
20. S. Hahm, D. Heitzman, J. Wang, Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization, in *Proceedings of SLPAT 2015: 6th workshop on speech and language processing for assistive technologies*. (2015)
21. M. Kim et al., Regularized speaker adaptation of KL-HMM for dysarthric speech recognition. IEEE Trans. Neural Syst. Rehabil. Eng. **25**(9), 1581–1591 (2017)
22. S. Dupont, J. Luettin, Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimedia **2**(3), 141–151 (2000)
23. J. Yu, B. Wu, R. Gu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, D. Yu, X. Liu, H. Meng, *Audio-visual multi-channel recognition of overlapped speech* (2020). arXiv preprint arXiv:2005.08571
24. S. Liu et al., Exploiting visual features using bayesian gated neural networks for disordered speech recognition, in *INTERSPEECH*. (2019)
25. C. Miyamoto, Y. Komai, T. Takiguchi, Y. Ariki, I. Li, Multimodal speech recognition of a person with articulation disorders using AAM and MAF, in *2010 IEEE International Workshop on Multimedia Signal Processing*. (IEEE, 2010), pp. 517–520
26. S. Liu et al., Exploiting cross-domain visual feature generation for disordered speech recognition, in *Interspeech*. (2020)

27. S. Liu et al., Recent progress in the cuhk dysarthric speech recognition system. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 2267–2281 (2021)
28. F. Javanmardi, S.R. Kadiri, P. Alku, Pre-trained models for detection and severity level classification of dysarthria from speech. Speech Commun. **158**, 103047 (2024)
29. S. Sajiha et al., Automatic dysarthria detection and severity level assessment using CWT-layered CNN model. EURASIP J. Audio Speech Music Process. **2024**(1), 33 (2024)
30. K. Radha, M. Bansal, V.R. Dhulipalla, Variable STFT layered CNN model for automated dysarthria detection and severity assessment using raw speech. Circuits Syst. Signal Process. **43**(5), 3261–3278 (2024)
31. Å. Rinnan et al., Data pre-processing. Infrared Spectrosc. Food Qual. Anal. Ctl. **2009**, 29–50 (2009)
32. L. Perez, *The effectiveness of data augmentation in image classification using deep learning* (2017). arXiv preprint arXiv:1712.04621
33. B. Chen, Y. Liu, Z. Zhang, G. Lu, A.W.K. Kong, TransAttUnet: Multi-Level Attention-Guided U-Net With Transformer for Medical Image Segmentation, in *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 1, (2024), pp. 55–68. https://doi.org/10.1109/TETCI.2023.3309626
34. M.Z. Alom et al., Recurrent residual U-Net for medical image segmentation. J. Med. Imaging **6**(1), 014006–014006 (2019)
35. K. He et al., Deep residual learning for image recognition, in *In Proceedings of the IEEE conference on computer vision and pattern recognition*. (2016)
36. G. Huang et al., Densely connected convolutional networks, in *In Proceedings of the IEEE conference on computer vision and pattern recognition*. (2017)
37. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of naacL-HLT*, vol. 1, (2019), p. 2
38. L. Xu et al., Multi-class token transformer for weakly supervised semantic segmentation, in *in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (2022)
39. C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in *Artificial intelligence and statistics*. (Pmlr, 2015), pp. 562–570
40. H. Kim et al., Dysarthric speech database for universal access research, in *In Interspeech*. (2008)
41. X. Wang et al., ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. Comput. Speech Lang. **64**, 101114 (2020)
42. F. Rudzicz, A.K. Namasivayam, T. Wolff, The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang. Resour. Eval. **46**, 523–541 (2012)
43. I. Loshchilov, *Decoupled weight decay regularization* (2017). arXiv preprint arXiv:1711.05101
44. A. Hernandez, H.-Y. Lee, M. Chung, Acoustic analysis of fricatives in dysarthric speakers with cerebral palsy. Phon. Speech Sci. **11**(3), 23–29 (2019)
45. R. Rajeswari, T. Devi, S. Shalini, Dysarthric speech recognition using variational mode decomposition and convolutional neural networks. Wireless Pers. Commun. **122**(1), 293–307 (2022)
46. N. Narendra, P. Alku, Dysarthric speech classification from coded telephone speech using glottal features. Speech Commun. **110**, 47–55 (2019)
47. S.R. Shahamiri, V. Lal, D. Shah, Dysarthric Speech Transformer: A Sequence-to-Sequence Dysarthric Speech Recognition System, in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, (2023), pp. 3407–3416. https://doi.org/10.1109/TNSRE.2023.3307020
48. A. Almadhor et al., E2E-DASR: End-to-end deep learning-based dysarthric automatic speech recognition. Expert Syst. Appl. **222**, 119797 (2023)
49. S. Mehra, V. Ranga, R. Agarwal, A deep learning approach to dysarthric utterance classification with BiLSTM-GRU, speech cue filtering, and log mel spectrograms. The Journal of Supercomputing **80**, 14520–14547 (2024)

Publisher's Note