Dysarthric Speech Detection and Severity Classification using Audio Spectrogram Transformer

Komal Bharti, Sandeep Agri, Pradip K. Das

Abstract—**Evaluating the level of dysarthria severity offers valuable insights into a patient's progress, helps pathologists in therapy planning, medication and supports the functionality of automated dysarthric speech recognition systems. In this study, we conducted experiments on dysarthric speech detection followed by severity classification for speaker-dependent and speaker-independent scenarios. Our findings highlight the effectiveness of Speech-Vision approaches, particularly those leveraging transformers and spectrograms. Audio Spectrogram Transformer (AST) has been taken as a base model in this paper, marking the development of a convolution-free, exclusively attention-driven model for audio classification. While various deep learning techniques have been explored in this domain, our paper distinguishes itself by introducing detection and classification using an audio spectrogram through a speech-vision approach. For all the experiments UASpeech database has been utilized and achieved state-of-the-art results of 99.64% accuracy for dysarthric speech detection and 78.97% accuracy for severity classification in a speaker-independent context. These outcomes surpass all previous results in the field.**

Index Terms—**Dysarthric speech detection, Severity classification, Intelligibility assessment, Audio-Spectrogram Transformer, Speech disorder.**

I. Introduction

HUMANS are social creatures by nature and cannot live alone. Mutual dependence is required for growth in this environment and hence communication is an essential aspect of life. There are different ways humans communicate, such as via speech-language, non-verbal gestures and electronic channels. In most cases, fully or partially, speech disorders affect the ability of an individual to communicate. According to the fact sheets of the World Health Organization (WHO) updated on 7th March 2023, 1.3 billion people are estimated to have significant disabilities, which equates to 16% of the world's population [1] [2] [3]. According to the Census 2011 by the Government of India, 2.21% of the total Indian population, which is around 21.9 million people, are suffering from speaking disability [4]. This is a significant number of individuals that need to be supported and taken care of. People with disabilities are twice as likely as the general population to suffer disorders such as anxiety, asthma, heart disease, stroke, or poor oral health [5]. In speech processing, speech disorders can be addressed in two ways: by helping individuals to improve their comprehension rate or by analyzing their speech patterns so that therapy can be adjusted based on severity [6].

This paper is focused only on dysarthria among several speech disorders. Dysarthria is a neurological motor speech disorder characterized by inadequate synchronization of speech production subsystems. Neurodegenerative diseases like cerebral palsy and Parkinson's disease typically cause it, or it can be acquired through neurological injuries such as stroke, brain injury or tumors. Consequently, speech quality deteriorates due to imprecise articulation, low audibility, atypical prosody, inter and intra-speaker variability and irregular speech rate [7]. Although dysarthric patients can form syntactically flawless sentences, they struggle to produce them phonetically or pronounce them correctly. Dysarthria is not a life-threatening disorder, but it affects the livelihood of patients in many aspects, including social, physical and emotional challenges [8] [9]. As severity increases, they are more likely to rely on others for their daily activities and household chores. Pathologists have tried to assist them using keyboard or joystick-based applications, but due to a lack of muscle coordination and trembling hands, these methods are not very effective.

Early detection of dysarthria is essential, as it allows timely therapy that can improve communication and reduce the disorder's impact on their lives. Assessing the severity of dysarthria is a crucial diagnostic step, as it provides valuable insights into the patient's condition, the progression of the disorder, and potential treatment options. It can also assist clinicians to determine the appropriate course of medication and therapy sessions. However, classifying the intelligibility and severity of dysarthric speech poses challenges due to variable speech features and subjective judgments. Figure 1 provides a visual comparison representation for clear differentiation between normal and dysarthric speech characteristics. It compares the utterances of the words 'Delete' and 'Zero' taken from the UASpeech corpus, representing normal and dysarthric speech, respectively. The utterances were taken from speaker F02, whose dysarthric speech had a low intelligibility score of 29%, indicating a high level of severity.

The primary goal of motor speech assessment is to determine the severity of a person's speech difficulty. In clinical practice and research, we frequently utilize severity ratings to explore speech difficulties. However, existing methods for determining the intensity of speech difficulty have not been properly evaluated and there is

Manuscript received March 13, 2025; revised April 6, 2025.

Komal Bharti is a research scholar in the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Assam 781039, India (e-mail: kbharti@iitg.ac.in).

Sandeep Agri is a master's student of the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Assam 781039, India (e-mail: sandeep.agri@alumni.iitg.ac.in).

Prof. Pradip K. Das is a Professor in the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Assam 781039, India (e-mail: pkdas@iitg.ac.in).

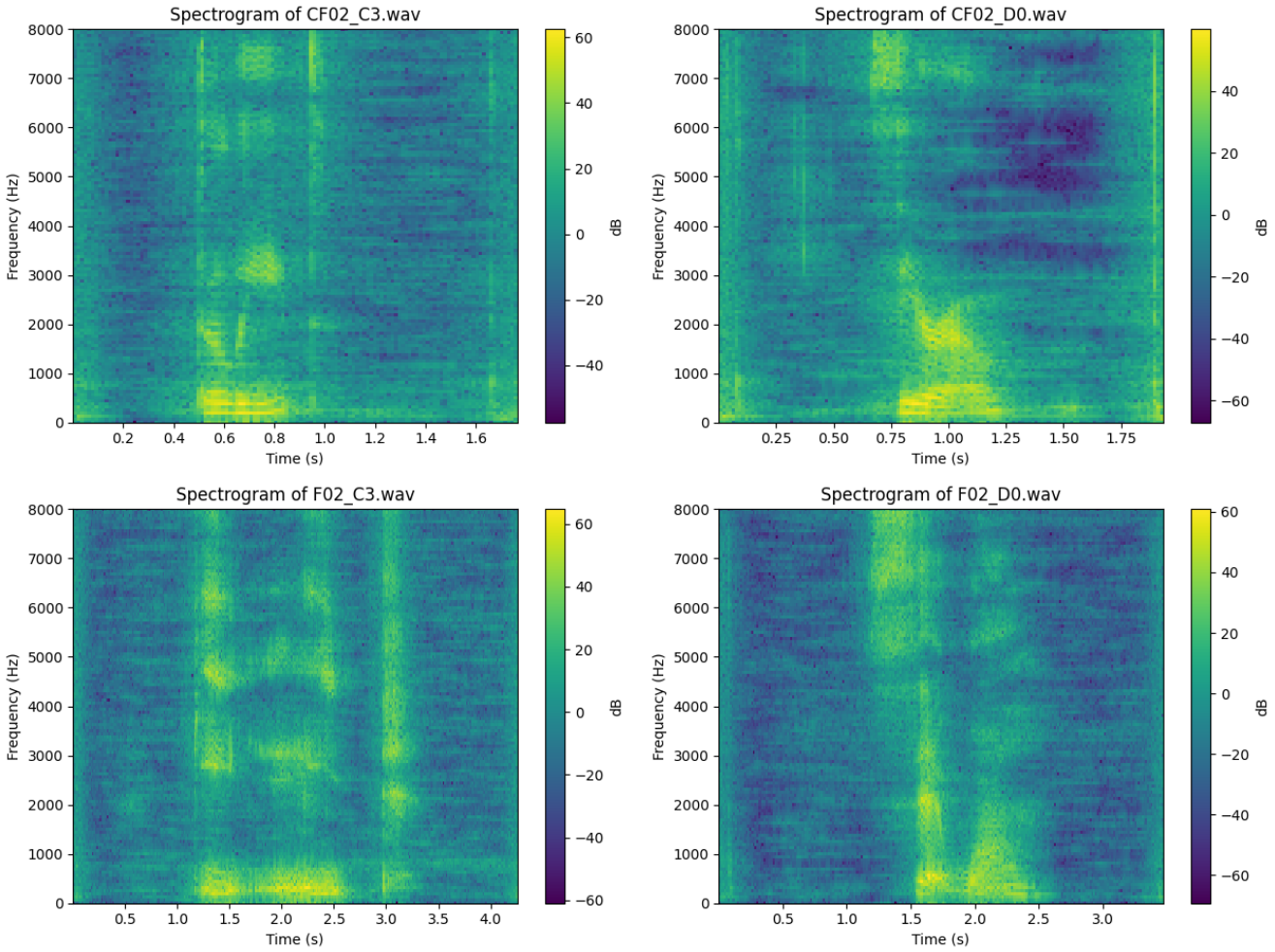

Fig. 1. Normal (top) and Dysarthric (down) utterance of word ”Delete” and ”Zero” respectively

no commonly accepted definition or classification system [10] [11]. Speech difficulties caused by dysarthria are commonly described by both clinicians and researchers. Speech-Language Pathologists (SLPs) often use two common informal methods to measure speech intelligibility: (a) Estimating the percentage of a patient's speech that others can understand and (b) using descriptive labels like 'normal', 'mild', 'moderate', 'severe', or 'profound'. We require further research to establish the reliability and accuracy of these assessments and to understand the factors that influence how we perceive the severity of someone's speech issue.

There is often a correlation between speech intelligibility and speech severity in dysarthria literature, yet these two measures are unconnected. *Kaila et al.* highlighted the relationship between speech intelligibility and severity very elegantly [10]. Dysarthria severity pertains to the extent of motor impairment affecting speech production, while intelligibility refers to how well listeners comprehend the speaker. Speech intelligibility relies on speech efficiency, voice quality and speaking rate of patients. In common practice, severity is determined based on the intelligibility rate. Table I shows that researchers have used different cutoff points to define severity levels based on speech intelligibility scores. Nevertheless, inconsistencies exist not just in the assigned ranges of intelligibility for each category but also in the approaches used to measure intelligibility across various studies.

TABLE I
THE DISTRIBUTION OF SPEECH SEVERITY BASED ON THE INTELLIGIBILITY RANGE (%) IN THE LITERATURE

Article	Profound	Severe	Moderate	Mild	Normal
[12]	0-60	60-70	70-80	80-90	90-100
[13]	-	0-45	45-75	75-100	-
[14]	0-25	25-50	50-75	75-100	-
[15]	-	0-40	40-70	70-100	-
[16]	0-25	25-50	50-75	75-100	-
[17]	-	0-50	50-75	75-100	-
[18]	0-50	50-80	80-90	90-95	95-100

II. MOTIVATION AND RELATED WORK

To diagnose dysarthria, it is important to measure the severity of the disease, which helps Speech-Language Pathologists (SLPs) to determine the appropriate medication and to schedule speech therapy sessions if needed. In conventional SLP practice, the severity of speech disorders was typically evaluated using the standardized rating scales given by Frenchay Dysarthria Assessment (FDA) [19]. This evaluation process incorporates a combination of acoustic, physiological and perceptual measures. However, it is worth noting that while treating patients with dysarthria the use of physiological measurements can be demanding and require

specialized equipment and expertise and on the other hand, perceptual measures can vary considerably depending on the clinician's level of experience and listening skills. Additionally, this would be costly and time-consuming, limiting its use in remote rehabilitation. In order to maintain homogeneous interpretation across SLPs, it is necessary to conduct dynamic assessments to determine speech severity rates. It is, therefore, necessary to develop a system that automatically classifies dysarthria severity levels. Automated severity assessment methods are cost-effective, traceable, reliable and allow remote monitoring of rehabilitation progress for patients.

Researchers have been exploring various approaches to achieve accurate results for dysarthric speech severity classification [20] [21] [22]. Numerous studies have been conducted to investigate the objective evaluation of dysarthric speech intelligibility by capturing essential acoustic data related to prosody, vocal tract dynamics, and excitation source information [23]. Mel-Frequency Cepstral Coefficients (MFCCs) and spectral and temporal features have been used exhaustively for the feature selection process. Deep Belief Networks (DBNs) [24] were compared with MFCC giving a marginal improvement in dysarthric severity classification using a multi-layer perceptron neural network. The combination of Glottal-to-Noise Excitation Ratio (GNER) and Harmonics-to-Noise Ratio (HNR) with MFCC was pursued in [25], as both of these metrics can determine the degree of noise caused by the disorder.

Linear discriminant analysis and non-linear techniques based on self-organizing maps were studied for dysarthria classification, but the decision was based on human evaluations [26]. *Dahmani et al.* [27] introduced a novel method to differentiate dysarthric speech from healthy controlled speech using rhythm metrics based on vocalic and intervocalic intervals durations on the Nemours dataset [28]. They applied a Gaussian Bayes classifier for this classification task. However, the rhythm metrics they extracted did not yield promising results in expressing the severity level of dysarthria. *Garima et al.* used a genetic algorithm to select prosodic features and apply SVM to classify dysarthric speech severity [29]. Machine learning models such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks have been used to achieve high classification accuracy. While the results have generally been favorable, the training process often requires significant computation time *Chitralekha et al.* explored Bi-directional LSTM (BLSTM) for a binary classification of dysarthric and non-dysarthric speech using the transfer learning method and achieved an improvement of 6% compared to the traditional machine learning method [30]. *Kwanghoon et al.* used the Mel spectrogram as input to a CNN for the early detection of Amyotrophic Lateral Sclerosis (ALS). This approach was later extended to capture spectro-temporal variations for assessing dysarthric speech.

An et al. [31] employed CNN to automatically detect early stage ALS from highly intelligible speech. They utilize both time domain and frequency domain CNNs to categorize speech from a group of 13 patients with early stage ALS and healthy individuals. The frequency based CNN showed better performance at predicting ALS at the individual level, compared to the time based CNN. Researchers utilized joint spectro temporal features extracted from a mel scale spectrogram for dysarthria severity estimation as well [32]. Their findings demonstrated that a time frequency CNN that captures both spectral and temporal information outperforms CNN that captures only temporal or spectral information separately. This highlights the importance of jointly considering both aspects to achieve superior dysarthria severity estimation results. In the long run, CNN was exhausted and researchers started to add an attention layer with CNN for end-to-end audio classification.

Over the last decade, deep learning techniques have been extensively explored for end-to-end audio classification, emphasizing direct mapping from spectrogram to corresponding labels [33]. To effectively capture long range global context, some researchers have proposed hybrid models that combine CNNs with self attention mechanisms. The proposed method builds upon the Transformer architecture, which has been previously explored for language and audio processing, but in this case, it is combined with CNN. Researchers have experimented with various combinations of transformer and CNN, such as implementing Transformer on top of CNN and incorporating a Transformer within each block of CNN. Hybrid models that combine CNNs with attention mechanisms have demonstrated remarkable effectiveness in delivering precise outcomes across various tasks like audio event classification, emotion recognition and command recognition. In the vision domain, purely attention based models have demonstrated remarkable success, leading to the question of whether CNNs are still necessary for audio classification. The approach proposed in this study is purely attention based, eliminating the need for convolution and offering a distinctive solution for the task.

The main objective of this work is to address the tasks of dysarthric speech detection in a speaker-dependent manner, and severity classification in both speaker-dependent and speaker-independent contexts, specifically:

- Speaker-dependent dysarthric speech detection: This involves detecting whether a speaker's speech is affected by dysarthria or not.
- Speaker-dependent dysarthric speech severity classification: After detecting dysarthric speech, the system aims to classify the severity level of dysarthria for the specific speaker.
- Speaker-independent dysarthric speech severity classification: This task shares the same objective as the second one; however, it aims to classify the severity of dysarthric speech from speakers who were not included in the training set, thereby enhancing the model's generalizability.

The proposed architecture for these tasks has been designed to outperform all previous methods in terms of accuracy for dysarthric speech detection and severity classification.

A. Datasets

The primary focus of this paper is detection and assessment of dysarthric speech severity. The experiments utilize the publicly available UASpeech dataset for all

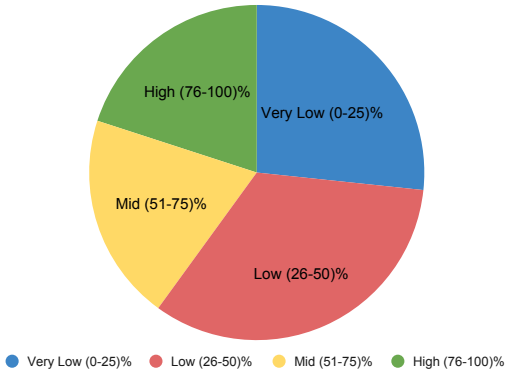

Fig. 2. UASpeech database graphical representation

objectives. The UASpeech dataset consists of speech samples from 19 individuals affected by cerebral palsy, exhibiting a wide range of intelligibility level. Speakers consist of both males and females and their ages range from 18 to 58 years. The data was recorded using an eight-microphone array, with 1.5 inches of spacing between each microphone and complemented by video recordings to capture visual features. Throughout the recording sessions, participants were seated comfortably in front of a laptop and they were instructed to read isolated words displayed on slides. The entire recording process was divided into three blocks, each containing 255 words. Among these, 155 words were repeated across the blocks and 100 words were selected from uncommon categories. Uncommon words comprised 10 digits, 26 radio alphabets, 19 computer commands and 100 words sourced from the Brown Corpus. Subsequently, each of the 19 speakers recorded a total of 765 isolated words.

To assess the intelligibility of dysarthric patients, five impartial listeners, proficient in American English aged between 18 and 40, were selected for each speaker. Listeners were instructed to listen to real words spoken by an individual with a speech disorder. The speech files were presented on a web page and listeners used headphones in a quiet room. Each listener transcribed the words and the percentage of correct responses were calculated. Speaker intelligibility was determined by averaging the correct percentages across five listeners. The classification of speakers intelligibility into four categories (very low, low, mid and high) was based on the average percent accuracy for each speaker. The recorded speech files were saved in the .wav format, and the entire dataset is publicly available for further research and analysis. They found that as dysarthria severity increased, listeners' confidence in transcribing dysarthric speech decreased [34]. So, we can say that as severity increases, speech intelligibility decreases. The distribution of all speakers in the UASpeech database is illustrated in Fig. 2. The database comprises 60% of speakers with very low and low intelligibility, where inter-speaker variability is notably higher. In Fig. 3, the word 'November' from the UASpeech database is annotated at the phoneme level by both a naive listener and a speech pathologist. The start and end boundaries of each phoneme differ, reflecting the annotator's listening experience. Similarly, all words from speakers with low and very low intelligibility are annotated and labeled accordingly. These annotations serve as ground truth for the model, enabling detection and intelligibility assessment by providing both the ideal reference and the originally annotated data.

The entire UASpeech database is utilized for the detection and severity classification experiments in which data is split in a random fashion for training, testing and validation. All experiments are conducted using the UASpeech dataset, which includes audio recordings from 16 individual speakers. Each speaker contributed 765 isolated words, with seven recordings available for each word. For the detection task, we included both controlled and dysarthric speech samples from these 16 speakers, resulting in a total of 171,360 audio files ($2\times16\times765\times7$ = 171,360). However, for the severity classification task, we exclusively used dysarthric speech data, totaling 85,680 audio files ($16\times765\times7$ = 85,680). The following section elaborates on the research methodology and model architecture, which is then followed by a comprehensive report detailing the experimental procedures and the outcomes achieved for each specific objective.

III. Methodology

Lately, the Transformer architecture has gained significant popularity in the realm of image processing. To adapt it for audio processing, a modification has been made wherein the Audio Spectrogram Transformer (AST) is designed in such a way that instead of taking an image as an input AST utilizes logarithmic Mel spectrograms derived from speech signals. The baseline of AST draws inspiration from the architecture of the Visual Transformer (ViT) [35] [36].

A. Audio Spectrogram Transformer

To achieve this transformation, the input audio waveform is initially converted into a sequence of 128-dimensional vectors using librosa, called log Mel filterbank (fbank) features [37]. The process of Mel filterbank feature extraction involves handling the input audio waveform with a duration of 't' seconds. This process includes dividing the audio input into manageable chunks every 10 ms and performing a Short-Time Fourier Transform (STFT) with a 25 ms hamming window on each chunk generating a sequence of 128-dimensional vectors that illustrate the evolution of the input audio waveform [38]. The STFT determines the power spectrum of each chunk, which is then passed through a set of filters. These filters are designed to mimic human hearing and are spaced non-linearly in frequency to better capture the properties of speech and other sounds. The Mel filterbanks use non-linear spacing to emphasize frequency regions that are important for distinguishing phonemes and sounds, which also makes the system more robust to noise. The distribution of the filters in the Mel scale tends to concentrate more on the lower frequencies, which often contain critical information for speech understanding. The filters in the Mel filterbank are designed to capture the distribution of energy across different frequency bands. Human perception of sound is not linear with respect to frequency, and the Mel scale is a perceptual scale that approximates the human ear's response to different frequencies. The set of filters helps to map the raw frequency content of the audio signal into a representation that aligns better with human perception. The resulting sequence of

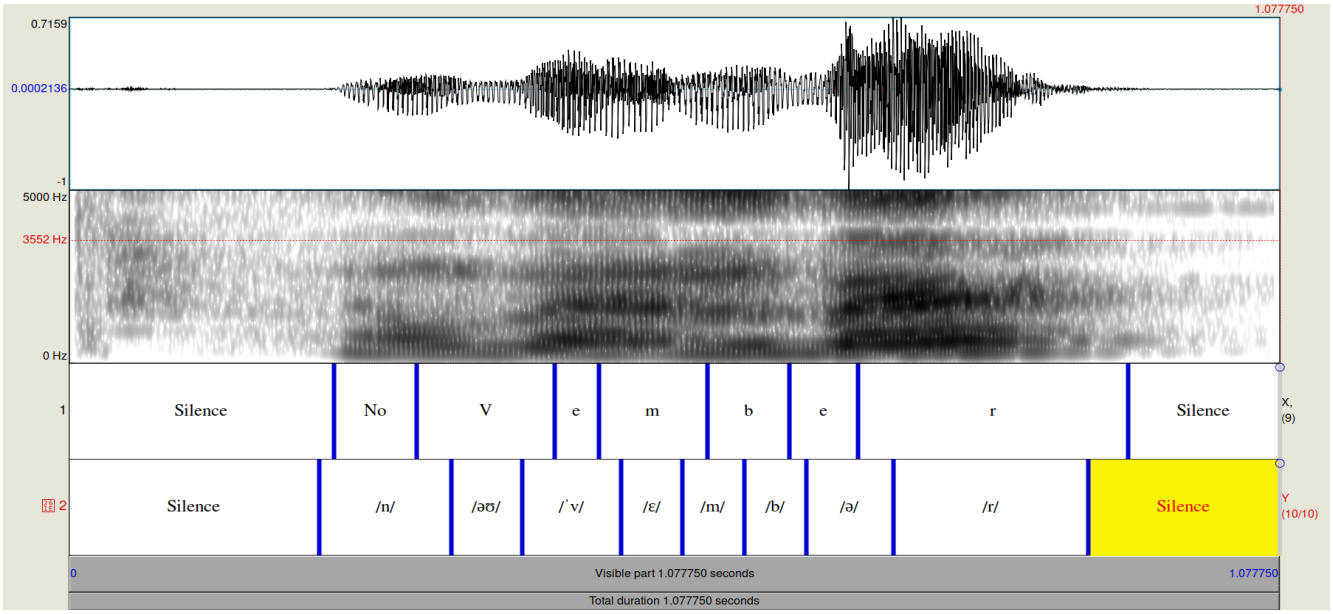

Fig. 3. Annotations of word "November" from UASpeech database

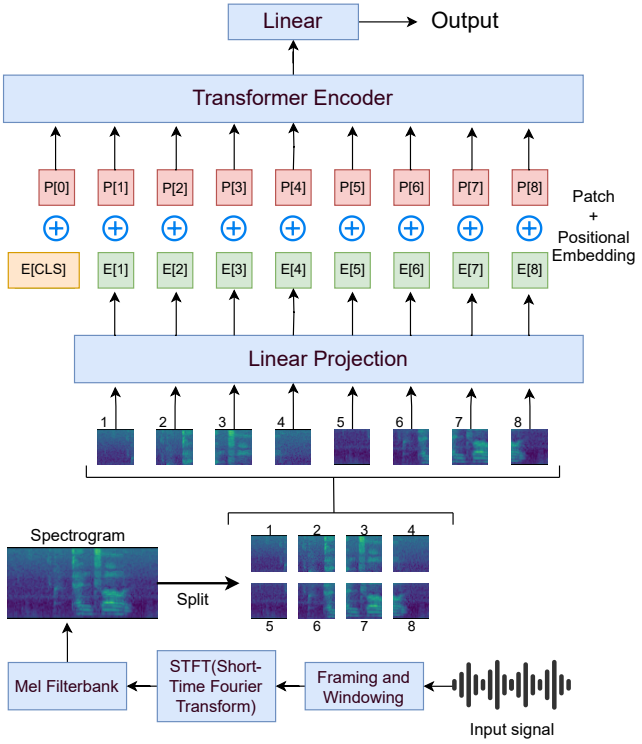

Fig. 4. Audio Spectrogram Transformer with dysarthric speech

vectors forms a spectrogram, which serves as the input to the AST (Audio Spectrogram Transformer) model.

Next, the spectrogram is split into a sequence of smaller N patches, each having a size of 16 by 16. These patches are extracted with a 6-step overlap in both time and frequency dimensions. The overlap refers to the degree of overlap between consecutive patches derived from the spectrogram. A 6-step overlap in both time and frequency dimensions means that when extracting patches from the spectrogram, each patch is offset by 6 steps in both the time and frequency directions compared to the previous patch. This overlap is used to ensure that information from adjacent patches is shared, which is beneficial for capturing temporal and frequency related patterns in the data. To determine the total number of patches, denoted as N, the following formula is utilized:

$$N = 12 * (100 * t - 16)/10 \quad (1)$$

Here, 't' represents the duration of the audio waveform in seconds. The value of N corresponds to the effective input sequence length for the transformer at that particular stage. Each patch is then treated as an individual input token and the transformer processes them separately.

Thereafter, using a linear projection layer, each 16x16 patch that was extracted from the spectrogram is flattened into a 1D patch embedding of size 768. This dimensionality reduction technique helps retain essential features while reducing the complexity of the input data. To preserve the spatial structure of the original 2D audio spectrogram, a trainable positional embedding of size 768 is added to each patch embedding. A trainable positional embedding of size 768 means that for each patch, there is a learnable vector of 768 elements that represents its position in the 2D space. These embeddings are trainable, signifies that the model can adjust them during training to best capture the spatial relationships in the data. By incorporating positional embeddings, the model retains information about the spatial location of each patch, which is otherwise lost during the flattening process. The positional embeddings are required since the model does not naturally capture the order information of the input data, making them necessary to capture the spatial structure of the input. By encoding the spatial structure, the model becomes capable of distinguishing between patches that are close together and those that are far apart. Along with the other model parameters, the positional embeddings are learned during training and updated during backpropagation.

The sequence begins by appending a unique token named [CLS]. In classification tasks, the [CLS] token is a special symbol that is used to represent the entire sequence and

allows the model to make predictions based on the full input sequence. The sequence is then fed to the transformer. In this case, only encoder layers are being utilized as we are doing detection and classification instead of recognition tasks. The output of the Transformer encoder, specifically the hidden state of the [CLS] token, serves as the representation of the spectrogram. To generate the final prediction, a linear layer with a sigmoid activation function is applied to map the audio spectrogram to the labels. The combination of the linear layer and sigmoid activation allows the model to learn a mapping from the features extracted from the audio spectrogram to a prediction of positive or negative class. For the detection task, label 0 corresponds to normal voice, while label 1 corresponds to dysarthric voice.

B. Adaptation of ViT in AST

In Visual Transformer (ViT), the input image is divided into non overlapping patches, and each patch is treated as an individual token. Similarly, in audio processing, the spectrogram is divided into overlapping segments, treating each segment as a token. This allows the transformer to capture local patterns in both vision and audio. ViT uses positional embeddings to provide the model with information about the spatial arrangement of patches. Similarly, in audio processing, positional embeddings are used to convey information about the temporal order of spectrogram tokens. In this way, the architecture of the ViT serves as inspiration for the baseline of the Audio Spectrogram Transformer (AST).

The AST model is designed in such a way that it is able to transfer the 2D spatial knowledge from a pre-trained ViT to the AST even when the input shapes are different. The AST takes advantage of transfer learning by utilizing pre-trained weights from the ViT architecture, which enables it to use expertise gained from a sizable dataset of images to improve its performance on the audio classification job. Additionally, since the network has already picked up useful features from the images it was trained on, using pre-trained weights reduces the amount of training data required for the AST to perform well. Given the limited availability of dysarthric speech in UASpeech, transfer learning enables the model to acquire valuable representations from a larger visual dataset and subsequently adapt these representations to the audio domain.

The positional embedding of ViT architecture is fixed in size since it employs a fixed size input image, but while dealing with audio data it can be of variable length. As audio signals vary in length, maintaining the sequential nature of the data becomes essential for capturing temporal relationships within the audio signal. To accommodate variable length audio sequences, the model employs padding, where shorter audio sequences are padded with zeros to match the length of the longest sequence in the dataset. The AST, analyses 16x16 pixel patches in variable length audio spectrograms.

The adaptation of positional embedding from the ViT to AST architecture involves the utilization of cut and bi-linear methods [39]. These techniques enhance the model's capacity to adeptly handle audio data having diverse sequence lengths. By effectively capturing temporal relationships, mitigating the impact of padding, and enhancing generalization across varying sequence lengths, these methods contribute significantly to the model's performance in processing audio data.

In a more technical sense, the patch embedding layer is likened to a single convolution layer with an extensive kernel and stride size and the projection layer within each transformer block is equivalent to a 1x1 convolution. It is important to note that the design diverges from conventional CNNs, which typically employ multiple layers having smaller kernel and stride sizes. Transformer models are often labeled as "convolution-free" to distinguish them from traditional CNN architecture.

TABLE II
EXPERIMENTAL SETUP AND MODEL PARAMETERS USED FOR EXPERIMENTS

Parameter	Value
Input Normalisation	Dataset mean -4.268 and std 4.569
Number of Classes	2 and 4
Frequency Stride	10
Time Stride	10
Loss Function	BCE and CE
Learning Rate Scheduler	MultiStepLR with 0.5 decay
Training Device	cuda
Total Parameter Number	87.728 million
Optimizer	Adam and SGD
Input Method	JSON file

TABLE III
CLASSIFICATION RESULT WITH GLOTTAL FEATURES AND CNN+LSTM MODEL BY NARENDRA ET AL. [40] FOR DYSARTHRIC SPEECH DETECTION

Input	Accuracy	Sensitivity	Specificity
Raw Speech	74.19	69.26	81.48
Glottal flow	**77.57**	73.13	82.48

IV. EXPERIMENTS AND RESULTS

The experimental setup involved using spectrograms, which are visual representations of the frequency content of a signal over time. The Audio Spectrogram model is employed for all tasks. To execute the experiments, we utilized a Kubernetes cluster, specifically utilizing an Nvidia A100 GPU with 40GB RAM for training the model for all scenarios. We summarize our experimental setup in Table II. The subsequent subsection provides an in-depth description of each objective.

A. Speaker-dependent dysarthric speech detection

Initially, we conducted dysarthric speech detection using a limited subset of the UASpeech dataset, specifically encompassing 16% of its content. The training set consisted of 11,486 audio files, while the testing and validation sets contained 5,744 and 5,743 audio files, respectively. After training the model for 3 epochs, we achieved an accuracy of 94%. To further improve the results, we extended training to 10 epochs, and the accuracy significantly improved to 96.86%. These outcomes demonstrated better performance compared to some early experiments conducted by Narendra et al. [40]. They applied CNN+LSTM model on raw speech and glottal flow as shown in Table III.

Subsequently, we conducted the same experiment using the entire UASpeech dataset. Remarkably, the accuracy improved to an impressive **99.64%**, surpassing the performance presented by Dong-Her et al. [41]. The detailed results are documented in Table IV and the loss curve is visualized in Figure 5. Additionally, Figure 6 showcases a comparison graph illustrating the accuracies of the previous state-of-the-art CNN-GRU model alongside our proposed AST model. The graph clearly demonstrates the superior accuracy achieved by the AST model over the previous approach.

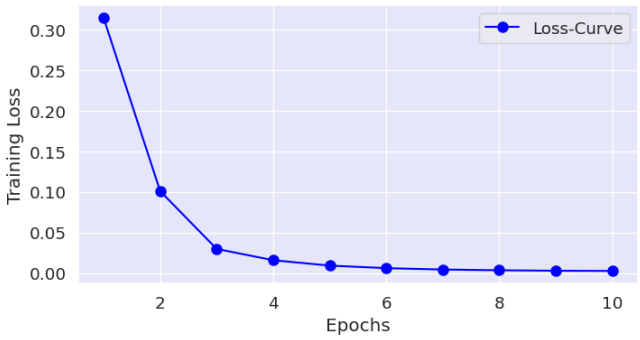

Fig. 5. Loss-curve for dysarthric speech detection by AST

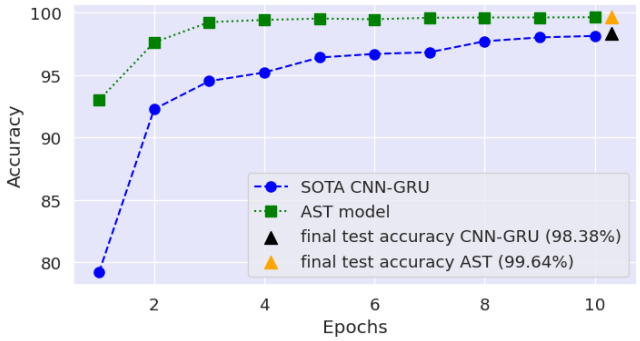

Fig. 6. Comparison between the SOTA CNN-GRU results [41] and current method results for dysarthric speech detection

Table V presents a comprehensive comparison of various approaches for dysarthric speech detection on the UASpeech corpus.

Various authors have employed different approaches for dysarthric speech detection on the UASpeech corpus. In this paper, we utilized the AST approach, achieving the highest accuracy of 99.64% for dysarthric speech detection. The AST model is a neural network architecture that is specifically designed to process audio spectrograms for speech recognition tasks. Overall, the results in Table V demonstrate the effectiveness of deep learning approaches, particularly those based on CNNs and their variations, are effective for dysarthric speech detection on the UASpeech corpus. Our model surpassed all previous dysarthric speech detection accuracies, highlighting its superiority in this task. These approaches have the potential to improve the accuracy and efficiency of dysarthria screening in clinical settings.

B. Speaker-dependent severity classification

After dysarthric speech detection, we proceeded with severity classification experiments. The severity levels for dysarthria in the UASpeech dataset are grouped into four categories, which are based on the assessment of speech-language pathologists. Table VI includes the severity categories of UASpeech along with corresponding speaker IDs. These severity levels form the foundation for assessing the effectiveness of dysarthric speech severity classification models in accurately predicting the level of dysarthria exhibited by the speaker.

Initially, we performed severity classification using Binary Cross Entropy (BCE) as the loss function, Automated Dynamic Analysis of Mechanical Systems(ADAMS) as the optimizer, and a batch size of 16 for 30 epochs, resulting in an accuracy of 84%. In pursuit of better performance, we switched to Stochastic Gradient Descent (SGD) as the optimizer, which significantly improved the accuracy to 93.6%. Continuing our efforts to enhance the model, we implemented a dynamic learning rate strategy, reducing it after every 3rd epoch. Additionally, we opted for the Cross-Entropy (CE) loss function. The choice of the Cross-Entropy (CE) loss function is motivated by its suitability for classification tasks, including severity classification models. This is derived from the principle of maximum likelihood estimation. It encourages the predicted probability distribution to be close to the true distribution of the labels. These combined adjustments resulted in the highest accuracy achieved so far, reaching an impressive **95.6%** classification accuracy. However, further experiments were conducted to enhance the model, such as experimenting with the learning rate by decreasing it after every 4 epochs. Unfortunately, this alteration did not yield the desired results, and the accuracy dropped to 89%. Figure 7 shows the accuracy versus loss graph plotted for 30 epochs during the classification experiments. Additionally, we evaluated the model's performance on both the validation and test sets and the corresponding confusion matrices are provided in Figure 8 and 9.

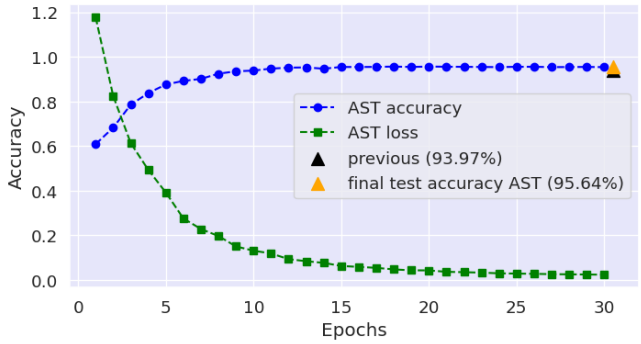

Fig. 7. Accuracy and loss curve for severity level classification by AST

C. Speaker-independent binary severity classification

Initially, the severity classification was carried out using a speaker-dependent model, where the training and testing data included only one specific speaker. However, to enhance practicality and applicability, it was desirable to develop a speaker-independent model. The goal is to build a speaker-independent model that can accurately classify speech severity without prior knowledge of the speaker. Therefore, we aimed to create a speaker-independent model.

TABLE IV
TRAINING AND VALIDATION METRICS FOR DYSARTHRIC SPEECH DETECTION USING AST

Epoch	Validation				Training loss	Validation loss	learning rate
	Accuracy(%)	AUC	Avg Precision	Avg Recall			
1	93.0039	0.988149	0.890938	0.966523	0.314101	0.55059	0.001
2	97.6181	0.998203	0.937057	0.997080	0.100899	0.517632	0.0005
3	99.2443	0.999622	0.943930	0.999259	0.029803	0.508477	0.00025
4	99.4219	0.999764	0.948494	0.999717	0.015913	0.532678	0.000125
5	99.5299	0.999825	0.967261	0.997639	0.009256	0.506424	6.25e-05
6	99.4742	0.999758	0.976602	0.993125	0.006116	0.506301	3.125e-05
7	99.5960	0.999848	0.986126	0.999197	0.004430	0.505693	1.5625e-05
8	99.6169	0.999820	0.992635	0.997011	0.003522	0.505428	7.8125e-06
9	99.6204	0.999800	0.995619	0.996684	0.002949	0.505475	3.90625e-06
10	99.6378	0.999750	0.995623	0.993877	0.002666	0.505362	1.953125e-06

TABLE V
PERFORMANCE COMPARISON FOR DYSARTHRIC SPEECH DETECTION FOR UASPEECH CORPUS

Author	Classification Method	Accuracy
Hernandez et al. (2019) [42]	SVM	72%
Narendra et al. (2019) [43]	SVM	96.38%
Narendra et al. (2020) [40]	CNN-LSTM	77.57%
Rajeswari et al. (2022) [44]	CNN	95.95%
Dong-Her et al. (2022) [41]	CNN-GRU	98.38%
Present work	**AST**	**99.64%**

TABLE VI
UASPEECH DATASET DISTRIBUTION ACCORDING TO SEVERITY LEVEL

Severity	UASpeech
Very Low	F05, M08, M09, M10, M14
Low	F04, M05, M11
Medium	F02,M07, M16
High	F03, M04, M12, M01

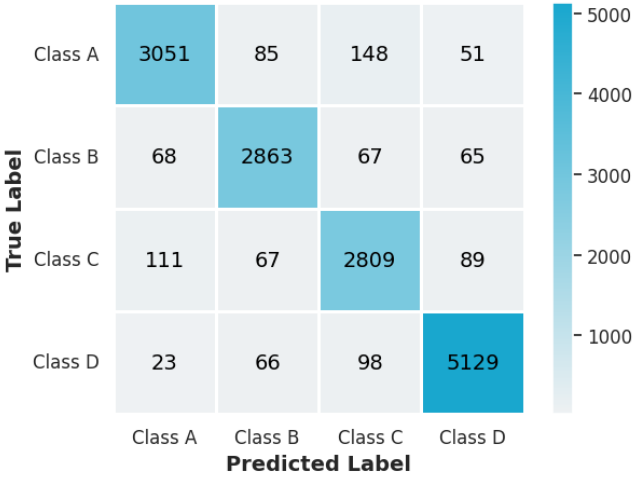

Fig. 8. Validation confusion matrix for speaker-dependent severity classification

Table VI shows that the intermediate classes have fewer speakers than the border classes, indicating a class imbalance in the UASpeech database. In pursuit of creating a speaker-independent model capable of accurately classifying severity, we merged the "low" and "very low" severity classes, as well as the "high" and "medium" severity classes into one. This decision was motivated by the limited availability of data for the "low" and "medium"

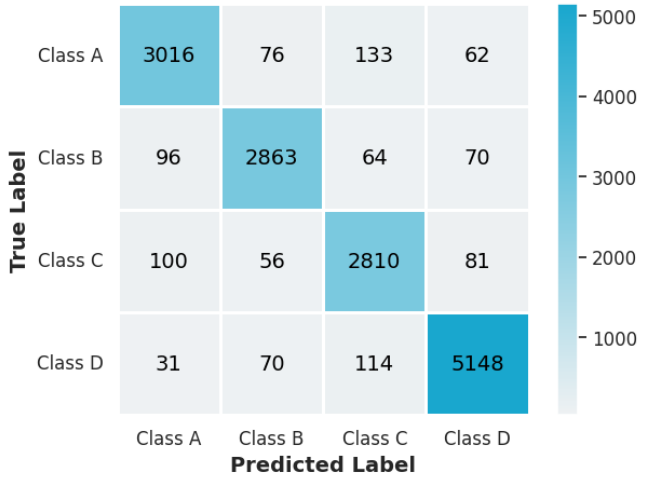

Fig. 9. Test confusion matrix for speaker-dependent severity classification

TABLE VII
SPEAKER DISTRIBUTION FOR TRAIN AND TEST SETS BASED ON SEVERITY LEVEL

Severity	Speakers	
	Train	Test
High	F02,M16,F03,M12,M01,M07	M04
Low	F05,M09,M10,M14,F04,M11,M05	M08

severity classes, which could potentially lead to reduced model accuracy if these classes were treated separately. By combining them, we effectively increased the amount of data available for training, enabling us to build a more robust model capable of classifying severity independently of the speaker's voice. The speaker's data taken for training and testing can be seen in Table VII. This approach allowed us to improve the model's performance across different speakers.

For speaker-independent classification, we trained the model for 10 epochs, using Binary Cross Entropy (BCE) as the loss function and utilizing the ADAM optimizer. The resulting accuracy was 62.5%. To enhance the model's performance, we switched to the Stochastic Gradient Descent (SGD) optimizer and reduced the learning rate every two epochs during training. These modifications had a significant impact on the model's accuracy, which increased the accuracy to **78.97%**. The dynamic learning rate strategy plays a crucial role in enhancing the accuracy of the

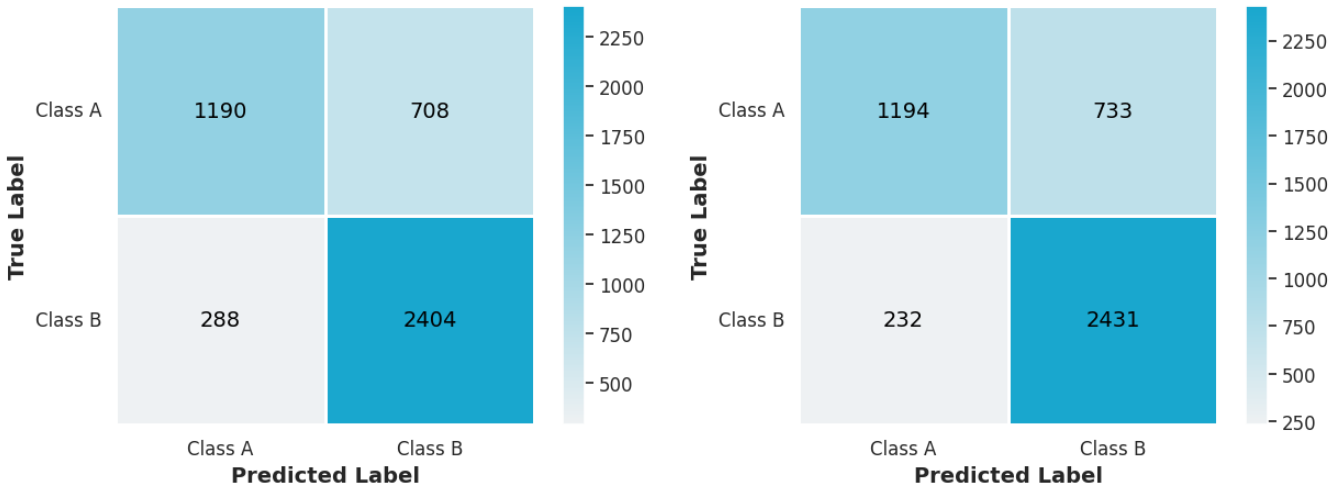

Fig. 10. Validation and Test confusion matrix for speaker-independent severity-level classification

speaker-independent severity classification model. Adjusting the learning rate allows the optimization process to converge more efficiently. Initially, a higher learning rate helps the model make large updates to its parameters, potentially escaping from local minima. As the optimization progresses, reducing the learning rate helps the model to converge more precisely to the optimal solution. By reducing the learning rate, the model becomes more sensitive to smaller gradients and makes finer adjustments to its parameters. This is particularly useful in later stages of training when the model is close to convergence. Figure 11 presents the validation accuracy and loss curves, along with a marker indicating the final test accuracy. This is compared against the highest reported accuracies from [45] and [22].

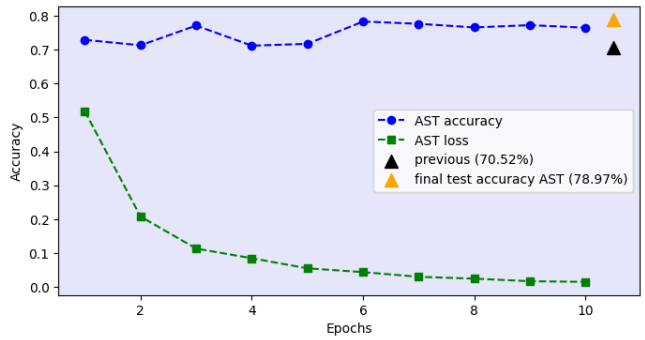

Fig. 11. Accuracy and loss-curve for speaker-independent dysarthric speech detection by AST

Table VIII shows the results for speaker-independent severity level classification for dysarthric speech and Table IX presents a comparison between the current results and the prior findings, highlighting that our outcomes demonstrate superior performance compared to the earlier results. Fig. 10 shows the validation and test confusion matrices which provide a visual representation of the model's performance in classifying severity levels.

V. Conclusion and Future Work

This study represents a comprehensive exploration of various deep-learning models employing the detection and classification of dysarthria severity levels. We use the Speech Vision approach with AST for the various tasks and achieve 99.64% accuracy for the detection surpassing the performance of previous state-of-the-art models. For severity classification, our model achieved an accuracy of 95.6%. Additionally, we developed a speaker-independent model, which demonstrated a notable accuracy of 78.97%. In summary, our research highlights the effectiveness of speech-vision approaches, specifically those leveraging transformers and spectrograms, for dysarthric speech detection and severity classification. These findings emphasize the potential of these advanced techniques to improve the accuracy and performance of dysarthria-related tasks significantly.

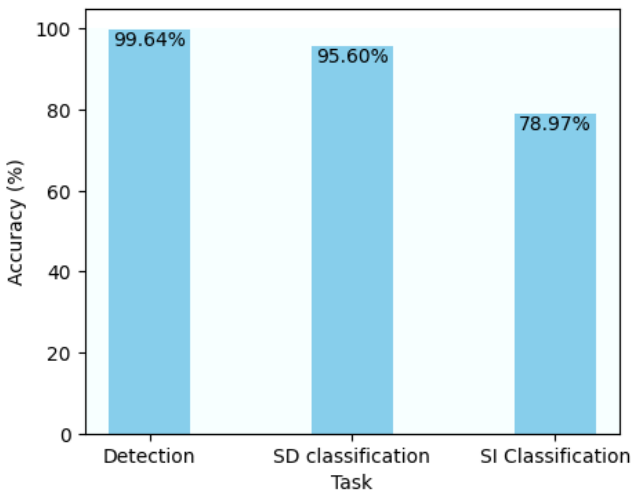

Fig. 12. Accuracy of different objectives performed

The model's capabilities can be extended by incorporating additional features such as phonetic and prosodic information. This inclusion would enable capturing more intricate details regarding the speech patterns of individuals with dysarthria, enhancing the model's overall performance and accuracy. To address the challenge of limited data, data augmentation and speech synthesis techniques offer

TABLE VIII
TRAINING AND VALIDATION METRICS FOR SPEAKER-INDEPENDENT SEVERITY-LEVEL CLASSIFICATION OF DYSARTHRIA USING AST

Epoch	Validation				Training loss	Validation loss	Learning rate
	Accuracy(%)	AUC	Avg Precision	Avg Recall			
1	72.8976	0.8015	0.6404	0.8753	0.51792	0.6454	0.001
2	71.2854	0.8212	0.7163	0.8088	0.20765	0.6436	0.001
3	77.1023	0.8467	0.6818	0.8963	0.11326	0.6191	0.0005
4	71.1764	0.8172	0.7134	0.8127	0.08471	0.6438	0.0005
5	71.7211	0.8253	0.7157	0.8085	0.05499	0.6398	0.00025
6	78.3006	0.8623	0.7199	0.8676	0.04382	0.6110	0.00025
7	77.6252	0.8559	0.7204	0.8601	0.02992	0.6133	0.000125
8	76.5577	0.8461	0.7144	0.8506	0.02458	0.6181	0.000125
9	77.2331	0.8456	0.7253	0.8429	0.01694	0.6159	0.0000625
10	76.4705	0.8445	0.7236	0.8351	0.01497	0.6178	0.0000625

TABLE IX
PERFORMANCE COMPARISON FOR DYSARTHRIC SPEECH SPEAKER-DEPENDENT AND SPEAKER-INDEPENDENT SEVERITY CLASSIFICATION

Work	Approach	Results
A.Tripathi, S.Bhosale, and S.K.Kopparapu [45]	Deep Speech posteriors with SVM	97.40%(SD) 65.20%(binary)
Amlu Anna Joshy and Rajeev Rajan [22]	i_MFCC with DNN	93.97%(SD) 70.52%(binary)
Current work	Spectrogram and Transformers, (AST), Speech-vision	95.6%(SD) **78.97%**(binary)

valuable solutions. By artificially creating new data through variations applied to existing data, we can substantially expand the dataset. Common augmentation techniques, such as pitch shifting, time stretching, noise addition and spectrogram manipulation prove highly beneficial in this context. Embracing these techniques opens avenues for advancing dysarthria research and developing more robust models for speech-related tasks.

This research can serve as a valuable resource for both patients and clinicians in accurately identifying the exact level of speech severity, thereby enabling them to track and assess progress toward improvements. In severe cases, where individuals often encounter significant difficulty in articulating words, such advancements can be transformative and have a life-changing impact.

REFERENCES

[1] W. H. Organization. (2023) Disability and health. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/disability-and-health

[2] U. D. of Health and H. Services. (2016, May) Statistics on voice, speech, and language. [Online]. Available: https://www.nidcd.nih.gov/health/statistics/quick-statistics-voice-speech-language

[3] L. Khan and M. Baig, "Automated cryptanalysis of plaintext xors of waveform encoded speech." *IAENG International Journal of Computer Science*, vol. 35, no. 2, pp. 234–241, 2008.

[4] G. Office of the Registrar and I. Census Commissioner. (2011, May) Disabled population by type of disability, age and sex. [Online]. Available: https://censusindia.gov.in/census.website/data/census-tables

[5] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 947–960, 2011.

[6] M. S. Hawley, S. P. Cunningham, P. D. Green, P. Enderby, R. Palmer, S. Sehgal, and P. O'Neill, "A voice-input voice-output communication aid for people with severe speech impairment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 21, no. 1, pp. 23–31, 2013.

[7] B. Sy and D. Horowitz, "A statistical causal model for the assessment of dysarthric speech and the utility of computer-based speech recognition," *IEEE Transactions on Biomedical Engineering*, vol. 40, no. 12, pp. 1282–1298, 1993.

[8] J. Mueller, G. Wenning, M. Verny, A. Mckee, K. Ray Chaudhuri, K. Jellinger, W. Poewe, and I. Litvan, "Progression of dysarthria and dysphagia in postmortem-confirmed parkinsonian disorders," *Archives of neurology*, vol. 58, pp. 259–64, 02 2001.

[9] R. FAAP and R. Ruben, "Redefining the survival of the fittest: Communication disorders in the 21st century," *The Laryngoscope*, vol. 110, pp. 241 – 241, 02 2000.

[10] K. L. Stipancic, Y. Yunusova, J. D. Berry, and J. R. Green, "Minimally detectable change and minimal clinically important difference of a decline in sentence intelligibility and speaking rate for individuals with amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 61, no. 11, pp. 2757–2771, 2018. [Online]. Available: https://pubs.asha.org/doi/abs/10.1044/2018_JSLHRS170366

[11] M. Jin, Y. Song, and I. V. McLoughlin, "End-to-end DNN-CNN classification for language identification," *IAENG International Journal of Computer Science*, vol. 1, pp. 119-203, 2017.

[12] B. Blaney and N. Hewlett, "Dysarthria and friedreich's ataxia: What can intelligibility assessment tell us?" *International journal of language & communication disorders / Royal College of Speech & Language Therapists*, vol. 42, pp. 19–37, 01 2007.

[13] K. Connaghan and R. Patel, "The impact of contrastive stress on vowel acoustics and intelligibility in dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 60, pp. 1–13, 01 2017.

[14] S. dos Santos Barreto and K. Zazo Ortiz, "Protocol for the Evaluation of Speech Intelligibility in Dysarthrias: Evidence of Reliability and Validity," *Folia Phoniatrica et Logopaedica*, vol. 67, no. 4, pp. 212–218, 01 2016. [Online]. Available: https://doi.org/10.1159/000441929

[15] P. Doyle, H. Leeper, A. Kotler, N. Thomas-Stonell, C. O'Neill, M. Dylke, and K. Rolls, "Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility," *Journal of rehabilitation research and development*, vol. 34, pp. 309–16, 08 1997.

[16] K. Cote-Reschny and M. Hodge, "Listener effort and response time when transcribing words spoken by children with dysarthria," *Journal of Medical Speech-Language Pathology*, vol. 18, pp. 24–34, 12 2010.

[17] S. Fager and J. Burnfield, "Speech recognition for environmental control: Effect of microphone type, dysarthria and severity on recognition results," *Assistive Technology*, vol. 27, p. 150626083720004, 06 2015.

[18] K. Stipancic, Y. Yunusova, J. Berry, and J. Green, "Minimally detectable change and minimal clinically important difference of a decline in sentence intelligibility and speaking rate for individuals with amyotrophic lateral sclerosis," *Journal of Speech Language and Hearing Research*, vol. 61, p. 1, 10 2018.

[19] P. M. Enderby, *Frenchay Dysarthria Assessment*. Pro-Ed, 1983. [Online]. Available: https://books.google.co.in/books?id=0uCEswEACAAJ

[20] K. Stipancic, K. Palmer, H. Rowe, Y. Yunusova, J. Berry, and J. Green, "You say severe, i say mild": Toward an empirical classification of dysarthria severity," *Journal of Speech, Language, and Hearing Research*, vol. 64, pp. 1–18, 11 2021.

[21] E. J. Yeo, K. Choi, S. Kim, and M. Chung, "Automatic severity classification of dysarthric speech by using self-supervised model with multi-task learning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

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

[23] R. M. Sriranjani R, Umesh S, "Automatic severity assessment of dysarthria using state-specific vectors." *Biomed Sci Instrum*, vol. 51, pp. 99–106, 2015.

[24] A. Farhadipour, H. Veisi, M. Asgari, and M. Keyvanrad, "Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks," *ETRI Journal*, vol. 40, 07 2018.

[25] M. Paja and T. Falk, "Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech," *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, vol. 1, pp. 62–65, 01 2012.

[26] E. Guerra and D. Lovely, "Suboptimal classifier for dysarthria assessment," 11 2003, pp. 314–321.

[27] H. Dahmani, S. A. Selouani, D. O'shaughnessy, M. Chetouani, and N. Doghmane, "Assessment of dysarthric speech through rhythm metrics," *Journal of King Saud University - Computer and Information Sciences*, vol. 25, p. 43–49, 01 2013.

[28] X. Menendez-Pidal, J. Polikoff, S. Peters, J. Leonzio, and H. Bunnell, "The nemours database of dysarthric speech," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, vol. 3, 1996, pp. 1962–1965 vol.3.

[29] G. Vyas, M. K. Dutta, J. Prinosil, and P. Harár, "An automatic diagnosis and assessment of dysarthric speech using speech disorder specific prosodic features," in *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, 2016, pp. 515–518.

[30] C. Bhat and H. Strik, "Automatic assessment of sentence-level dysarthria intelligibility using blstm," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 322–330, 2020.

[31] A. K, L. T N, S. U. Bhat, S. R, and C. H M, "Automatic early detection of dysarthria using deep neural network," in *2023 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)*, 2023, pp. 1–4.

[32] H. M. Chandrashekar, V. Karjigi, and N. Sreedevi, "Spectro-temporal representation of speech for intelligibility assessment of dysarthria," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 390–399, 2020.

[33] J. Xiao and S. Xiaolin, "Research on a deep learning method for speech recognition." *IAENG International Journal of Computer Science*, vol. 51, no. 9, pp=1272-1280, 2024.

[34] K. Hustad, "Effects of speech stimuli and dysarthria severity on intelligibility scores and listener confidence ratings for speakers with cerebral palsy," *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)*, vol. 59, pp. 306–17, 02 2007.

[35] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," 08 2021, pp. 571–575.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 10 2020.

[37] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 01 2015, pp. 18–24.

[38] P. Podder, T. Khan, M. Khan, and M. Rahman, "Comparative performance analysis of hamming, hanning and blackman window," *International Journal of Computer Applications*, vol. 96, pp. 1–7, 06 2014.

[39] G. Gallo and A. Ülkücü, "Bilinear programming: An exact algorithm," *Mathematical Programming*, vol. 12, pp. 173–194, 12 1977.

[40] N. P. Narendra and P. Alku, "Glottal source information for pathological voice detection," *IEEE Access*, vol. 8, pp. 67 745–67 755, 2020.

[41] D.-H. Shih, C.-H. Liao, T.-W. Wu, X.-Y. Xu, and M.-H. Shih, "Dysarthria speech detection using convolutional neural networks with gated recurrent unit," *Healthcare*, vol. 10, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252859008

[42] A. Hernandez and M. Chung, "Dysarthria classification using acoustic properties of fricatives," 11 2019.

[43] N. N P and P. Alku, "Dysarthric speech classification from coded telephone speech using glottal features," *Speech Communication*, vol. 110, 04 2019.

[44] R. Rajeswari, D. Thirupathi, and S. Selvaraj, "Dysarthric speech recognition using variational mode decomposition and convolutional neural networks," *Wireless Personal Communications*, vol. 122, 01 2022.

[45] A. Tripathi, S. Bhosale, and S. K. Kopparapu, "Improved speaker independent dysarthria intelligibility classification using deepspeech posteriors," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6114–6118.