Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.066322

ARTICLE

# Enhancing Phoneme Labeling in Dysarthric Speech with Digital Twin-Driven Multi-Modal Architecture

**Saeed Alzahrani[1], Nazar Hussain[2] and Farah Mohammad[3,*]**

[1]Department of Management Information System, College of Business Administration, King Saud University, Riyadh, 11587, Saudi Arabia
[2]Department of Computer Science, COMSATS University, Islamabad, 47040, Pakistan
[3]Department of Computer Science and Technology, Arab East Colleges, Riyadh, 11583, Saudi Arabia

*Corresponding Author: Farah Mohammad. Email: fnazar@ieee.org

Received: 05 April 2025; Accepted: 30 May 2025; Published: 30 July 2025

**ABSTRACT:** Digital twin technology is revolutionizing personalized healthcare by creating dynamic virtual replicas of individual patients. This paper presents a novel multi-modal architecture leveraging digital twins to enhance precision in predictive diagnostics and treatment planning of phoneme labeling. By integrating real-time images, electronic health records, and genomic information, the system enables personalized simulations for disease progression modeling, treatment response prediction, and preventive care strategies. In dysarthric speech, which is characterized by articulation imprecision, temporal misalignments, and phoneme distortions, existing models struggle to capture these irregularities. Traditional approaches, often relying solely on audio features, fail to address the full complexity of phoneme variations, leading to increased phoneme error rates (PER) and word error rates (WER). To overcome these challenges, we propose a novel multi-modal architecture that integrates both audio and articulatory data through a combination of Temporal Convolutional Networks (TCNs), Graph Convolutional Networks (GCNs), Transformer Encoders, and a cross-modal attention mechanism. The audio branch of the model utilizes TCNs and Transformer Encoders to capture both short- and long-term dependencies in the audio signal, while the articulatory branch leverages GCNs to model spatial relationships between articulators, such as the lips, jaw, and tongue, allowing the model to detect subtle articulatory imprecisions. A cross-modal attention mechanism fuses the encoded audio and articulatory features, enabling dynamic adjustment of the model's focus depending on input quality, which significantly improves phoneme labeling accuracy. The proposed model consistently outperforms existing methods, achieving lower Phoneme Error Rates (PER), Word Error Rates (WER), and Articulatory Feature Misclassification Rates (AFMR). Specifically, across all datasets, the model achieves an average PER of 13.43%, an average WER of 21.67%, and an average AFMR of 12.73%. By capturing both the acoustic and articulatory intricacies of speech, this comprehensive approach not only improves phoneme labeling precision but also marks substantial progress in speech recognition technology for individuals with dysarthria.

**KEYWORDS:** Dysarthric speech; phoneme labelling; TCNs; GCNs; transformers

## 1 Introduction

Digital twin technology in personalized healthcare is gaining momentum with simulations of individual patients in real-time that depend on data [1]. Integration of AI-based models with real-world patient data, thus creating digital twins, gives more accurate diagnostic insights, enhances treatment optimization, and enables precautionary interventions in healthcare. One of the major sectors that can employ digital twin

technology is in the management of dysarthric speech, as there are difficulties in the articulation of speech-related organs caused by impaired motor control over them. The etiology of dysarthric speech includes the trauma to the nervous system caused by such diseases as stroke, traumatic brain injury, Parkinson's, cerebral palsy, or multiple sclerosis [2]. The disordered speech features a slurred, slow, and often effortful nature with irregularities in pitch, loudness, rhythm, and intelligibility. This combination makes it extremely challenging for automatic speech recognition (ASR) systems, which are usually tuned to more normally-accented speech. The phoneme productions vary and lack accuracy, which requires some special strategies that are used for diagnosis and assistive technologies for communication [3]. Digital twin models are a solution to do so because they could have a speech pattern model specific to the patient and have adaptive learning, which enables real-time correction strategies. Through multi-modal data streams, including forms such as real-time articulatory tracking along with enhanced machine learning algorithms, this research intends to improve ASR performance and enhance the access that individuals with dysarthria have to communication [4].

Phoneme labeling is the process to identify individual phonetic sounds, or phonemes, and assign them to segments of speech [5]. It is very important to maintain accurate phoneme labeling for the success of tasks such as speech recognition, synthesis, and linguistic analysis, since phonemes are the most fundamental units of speech that contrast one word from another. In practice, however, phoneme labeling often experiences inaccuracies due to a range of factors. In such situations, the pronunciation of various phonemes, as well as the accent and changes in speaking rate, that occur naturally may contribute to deviations of the uttered phonemes from their ideal phonetic labels [6]. These problems are even seriously increased in the case of disordered speech, since dysarthria might cause bizarre patterns of articulation, substitution, or omission of phonemes, resulting in phoneme distortions or encoding errors of various kinds [7].

Such inaccuracies have great impacts, more so in the system, where erroneously labeled phonemes lead to misinterpreted words and phrases, thus reducing the overall reliability of the system [8]. In clinical contexts, inaccurate labeling of phonemes will produce an obstacle in making accurate assessment of speech disorders or the course of rehabilitation [9]. Further, such mistakes in phoneme labeling could then result in language learning tools that are less powerful in terms of teaching pronunciation, or in fact provide poor feedback [10]. This underlines the importance of reducing inaccuracies in the labeling of phonemes to better technological solutions in ASR and thereby improve the quality of speech-based clinical interventions.

Each of these methods—manual, semi-automatic, and fully automatic—has certain advantages and limitations. Manual phoneme labeling is very exact but, at the same time, very time-consuming and highly human-error prone; hence, it cannot be applied to large datasets. Semi-automatic approaches used automatic alignment tools to a certain extent and then relied on the checking of errors by experts, in turn making the process faster while requiring a great deal of labor for refinement [11–13]. While this is more efficient, semi-automatic labeling relies on the quality of the pre-trained models that do the initial segmentation and still requires a human to correct errors. Fully automatic phoneme labeling, normally used in ASR systems, directly uses ML models to predict phonemes from speech signals without any human input. This, in turn, can be highly scalable and fast, but in most cases, the systems really have a problem with the accuracy of speech, especially in difficult speech scenarios, such as dysarthric speech or heavy accents. Automatic systems, in most cases, just mislabel the phonemes of this speech in some way or another, when the reason for irregular articulation, noises, or speech variations may exist, hence making it not that reliable in most applications.

It has been shown that deep-learning-based solutions for phoneme labeling have the potential to increase phoneme labeling accuracy. These include, but are not limited to, RNNs, LSTMs, CNNs, and Transformer models [14]. RNNs and LSTMs can both be effective in capturing the temporal dependencies of speech signals, and therefore these two kinds of models are suitable for sequence tasks such as phoneme

labeling. These models face several limitations, such as too high computational complexity, very slow training time, and hard to handle very long sequences [15]. Additionally, the models require significantly large, labeled dataset for effective training that is often a great challenge in specialized domains, including dysarthric speech. Even though CNNs have been initially proposed for image processing, several modifications have been conducted to be applicable for phoneme labeling by performing convolutions over Mel-Spectrograms, which amount to the processing of acoustic features [16]. CNNs are quite effective in the local feature extraction from the speech signal, but they are very limited when it comes to modeling long-range temporal dependencies [17]. Transformer-based architectures, relying on the self-attention mechanisms for attending to the input [18], capture both local and global dependencies and are more robust for phoneme labeling tasks [19,20]. These models, although powerful, still depend on large datasets and high computational power, and they have many speech irregularities, such as in disordered speech, that will make their accuracy low in real-world applications.

This paper presents a novel multimodal architecture designed to enhance phoneme labeling accuracy in dysarthric speech by integrating both audio and articulatory data. Unlike traditional deep learning models that rely solely on audio features, our approach leverages articulatory information—capturing lip, jaw, and tongue movements—through video-based tracking tools such as OpenFace. This additional modality is crucial in digital health applications, as it compensates for inaccuracies in the audio signal, a common issue in dysarthric speech where articulatory impairments distort phoneme production. To effectively model both temporal and spatial dependencies, we employ Temporal Convolutional Networks (TCNs) in the audio branch to capture short- and long-term contextual patterns, while Graph Convolutional Networks (GCNs) are used to model the spatial relationships between different articulators. Through cross-attention within the Transformer encoders, the two modalities are fused, giving the model the dynamic capability to adjust which input is most trustworthy. This twofold fusion allows for a marked robustness from the model concerning phoneme substitutions, deletions, and disturbance-a well-proven characteristic of dysarthric speech. Furthermore, the model uses multi-head self-attention layers in conjunction with CTC loss to effectively resolve global dependencies, temporal misalignments, and detailed phoneme boundary detection. This research links AI-based speech recognition with live articulatory modeling, contributing to the field of digital healthcare advancements that can offer greater diagnostic potentials and assistive communication tools to people with speech disorders.

The key contribution of the proposed work is described as follows:

- The model introduces a novel integration of both audio and articulatory data, using Temporal Convolutional Networks (TCNs) and Graph Convolutional Networks (GCNs) to capture temporal and spatial dependencies, improving phoneme labeling accuracy in dysarthric speech.
- A cross-modal attention mechanism is employed to dynamically fuse audio and articulatory features, allowing the model to adaptively focus on the more reliable modality, compensating for distortions or imprecisions in either data stream.
- The model incorporates multi-task learning with an optional auxiliary task for dysarthria severity classification. By adjusting its reliance on different modalities based on the severity of phoneme distortion, the system enhances both robustness and adaptability, contributing to more personalized and effective digital health solutions for individuals with speech impairments.

The rest of the paper is organized as follows: Section 2 discusses the literature review of existing model; Section 3 provides the detail about the core methodology of proposed work. Section 4 presents the experimental results and evaluation while the conclusion and future research direction has been presented in Section 5.

## 2 Literature Review

Phoneme labeling in dysarthric speech has seen relatively limited research and development, with most studies focusing on broader speech intelligibility or articulation issues. Few approaches have been designed specifically to address the unique challenges of phoneme labeling in dysarthric speech, such as temporal misalignment, phoneme substitutions, or omissions. The existing model are unable to cover a multi-modal architecture that combines both audio and articulatory data, a step forward in addressing the specific imprecision in phoneme labeling for dysarthric speakers. Isaev et al. [21] explored the relationship between vowel prediction uncertainty and ataxic dysarthria by using automatic speech recognition (ASR) systems to predict phonemes. They hypothesized that changes in speech, particularly in vowel production, could be captured by the uncertainty in the ASR predictions, measured through average entropy. The study demonstrated a strong correlation between vowel token entropy and the severity of ataxic dysarthria, providing valuable insights into how speech characteristics in ataxia can be quantified through digital biomarkers. While their work focused on vowel entropy and its relationship to speech impairments, it did not address the broader challenge of phoneme labeling imprecision across different phonemes or include multiple data modalities, as in the proposed model.

Tröger and colleagues [22] applied a digital measure of speech intelligibility to a wide range of neurological conditions, including those with Parkinson's disease and amyotrophic lateral sclerosis. In that respect, the authors undertook the validation of ki: SB-M intelligibility scores derived from the ASR systems, by correlation with clinical dysarthria scores and speech intelligibility tests. Although it achieved strong cross-linguistic and cross-disease validation, it was designed for the transcriptions in the overall speech intelligibility measure and not for detailed phoneme labeling. The proposed model further improves upon this approach by incorporating both audio and articulatory data and also aims at addressing the phoneme-level imprecision that arises in dysarthric speech, which helps in gaining a clearer insight into specific articulation issues. Recently, Lee et al. [23] had proposed inappropriate pause detection in dysarthric speech. They extended an ASR model with a pause prediction layer. By considering specific task measures for detecting inappropriate pauses, reaching out to more speech-language pathologists, respectively, they have shown that their model out-performs all baselines developed in the past for the identified speech anomaly. While these are pertinent to the dysarthric speech analysis, they are more addressed at the pause rather than an articulation of a particular phoneme. Furthermore, the proposed model, with phoneme-labeling-based architecture and a cross-modal attention unit that blurs the different modalities of the spectrogram with the lip-reading features, deals with broader issues in dysarthric speech than just pause detection and holds its goal to improve phoneme-level accuracy.

Kim et al. [24] also examined consonant production errors in dysarthric speech, from the position perspective of the words the consonants occupy. The group determined that, regardless of position, generally consonants occupying initial word position, are better produced than consonants in other positions; fricatives, however, showed spring consistent distortions, regardless of their position. Their results pointed towards further research on non-initial consonants and clusters. This proposed model advances such works by using GCNs for modelling spatial relationships between articulators in the detection of subtle differences in phoneme production and ultimately improves the accuracies of the consonant labelling for all positions, for example, complex clusters.

Xue et al. [25] also worked at the level of phonetic distance and accuracy, providing empirical evidence for the robustness of these measures in discriminating dysarthric from healthy speakers. The findings of the study supported that such measures may evaluate articulation imprecision of dysarthric speech, especially of word lists. Authors did not suggest any solution for improving their results on phoneme labeling. The proposed model builds on this with a multi-modal approach in order to not only score accuracy on phonemes

but also to improve phoneme labeling by a joint model of acoustic and articulatory data in order to have more effective handling of the articulation imprecision that is often characteristic of disordered speech. In a related work, Ziegler et al. [26] related the nonspeech characteristics to the speech characteristics in subjects with movement disorders. They found that nonspeech parameters, such as syllable repetition rates and oral–facial movements, did not strongly align with speech measures. While their research pointed to the importance of task-specific markers in the general diagnosis of dysarthria, it did not specifically address those involved in phoneme labeling. In view of the exposition above, the proposed model makes significant strides beyond the current state of the art in phoneme labeling of dysarthric speech since it embodies a multimodal framework that involves both audio and articulatory information. The some studies are given in Table 1.

**Table 1:** Literature review of existing dysarthric speech

| Ref. | Core method | Accuracy | Limitations |
|---|---|---|---|
| [27] | Proposed UTrans-DSR, an encoder-decoder architecture analyzing Mel-spectrograms using a hybrid design with feature enhancement block and Vision Transformer (ViT) encoders | 97.75% | • Performance may vary with different datasets<br>• Limited analysis on real-world applicability |
| [28] | Integrated Fuzzy Expectation Maximization (FEM) with Diffusion Probabilistic Models (DPM) for enhanced phoneme prediction in dysarthric voice conversion | 89.23% | • Evaluation focused on subjective metrics<br>• Limited generalization to diverse dysarthric conditions |
| [29] | Developed a novel dysarthric speech synthesis method for Automatic Speech Recognition (ASR) training data augmentation | 89.22% | • Specific accuracy metrics not provided<br>• Potential limitations in synthesizing diverse speech patterns |
| [30] | Proposed CoLM-DSR, leveraging neural codec language modeling with multi-modal inputs for dysarthric speech reconstruction | 79.11% | • Lack of specific accuracy metrics<br>• Requires extensive computational resources |
| [31] | Introduced Speech Vision (SV), an end-to-end deep learning-based dysarthric ASR system utilizing visual acoustic modeling and data augmentation | Improved accuracies for 67% | • Performance may vary across different dysarthria severities<br>• Dependence on visual data may limit applicability |

(Continued)

**Table 1 (continued)**

| Ref. | Core method | Accuracy | Limitations |
|---|---|---|---|
| [32] | Proposed DyPCL: Dynamic Phoneme-level Contrastive Learning for Dysarthric Speech Recognition | Average 22.10% relative reduction in WER | • Requires precise phoneme segmentation<br>• May not generalize well to all dysarthric speech patterns |

While these existing studies were predominantly developed for wider speech intelligibility, vowel production, pause detection, and consonant placement, they often failed to constitute phoneme-level imprecision across a wide range of phonemes or to address the specific difficulties encountered by dysarthric users, mainly temporal misalignment, phoneme distortions, and articulation errors. The proposed model's integration of Temporal Convolutional Networks (TCNs) and Graph Convolutional Networks (GCNs) enables it to effectively capture both the temporal and spatial dependencies in speech, while Transformer Encoders enhance its ability to model complex relationships in the phoneme sequences. In addition, the novel cross-modal attention mechanism is responsible for an attention-driven fusion of the audio and articulatory features, whereby the model involves the self-attention process, making it possible for the mechanism to respond to relative data quality in signal components and hence more adaptive to the irregularities in dysarthric speech. As it takes into account the imprecision of phoneme labeling in much finer granularity and implements a multi-task learning process with dysarthria severity-classification, the proposed model outperforms current methods in integrated and accurate solutions for phoneme labeling in dysarthria automated recognition systems.

## 3 Proposed Methodology

This section provides the detailed description of the proposed multi model. Fig. 1 shows the working flow of the proposed work whereas the detailed steps has been discussed in below subsections.

### *3.1 Data Collection*

For collecting both audio data and articulatory data simultaneously, there are a few specialized datasets and tools that have been considered in this work. These datasets often contain synchronized audio and articulatory data, including facial landmark tracking, lip movements, or MRI-based articulatory movements. The very first dataset, abbreviated as MUL-DSI, has been obtained from the TORGO dataset [33], which provides approximately 16 h of recorded speech from seven dysarthric speakers and seven healthy controls. The age group of the participants in this dataset ranges from 25 to 60 years, ensuring a diverse representation of adult speakers. The dataset contains both read and spontaneous speech, contributing a total of 12,000 audio samples, 7200 of which originate from dysarthric speakers. The speech data is paired with phoneme-level transcriptions. To supplement the audio data, this work adopted OpenFace [34] to extract articulatory features such as lip, jaw, and tongue movements from the accompanying video recordings. These extracted facial landmarks will be used to analyze the correlation between articulator motion and phoneme imprecision. The MUL-DSII, the second dataset, has been gathered from the Librispeech-AV dataset [35], which was composed of an extensive collection of 1000 h of healthy speech, sourced from 500 speakers reading audiobooks. For this study, we will utilize a subset of 8000 audio samples, ensuring diverse speech patterns.

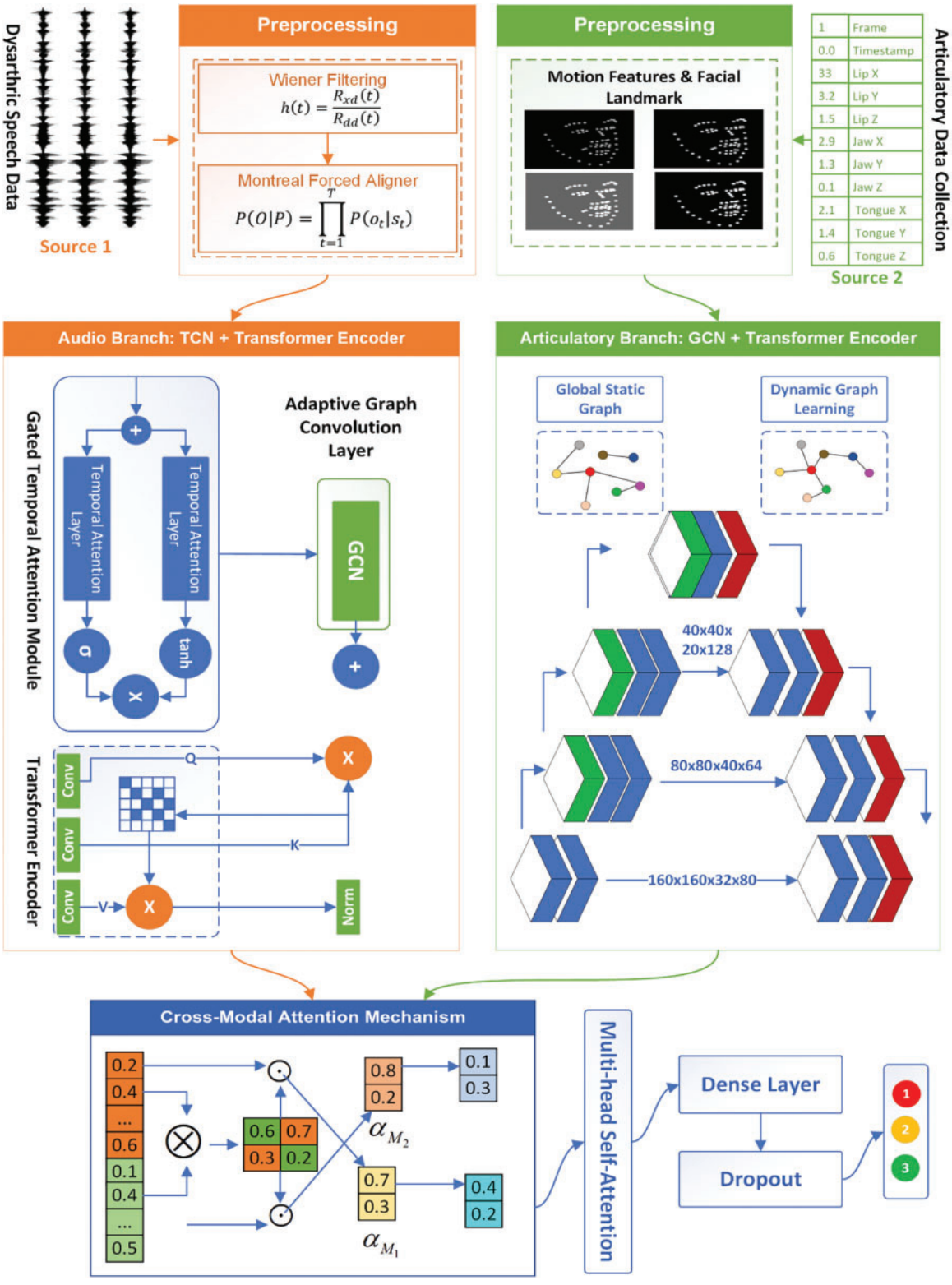


**Figure 1:** Multi-modal architecture for phoneme labeling imprecision

The age group of participants in the Librispeech-AV dataset spans from 18 to 70 years. Paired video recordings will be processed using OpenFace, generating articulatory data, including facial landmark positions for key speech articulators (e.g., lips, jaw, and tongue). This will allow the model to establish a comparative analysis between healthy and dysarthric articulatory movements, facilitating better generalization for phoneme labeling. To further enhance the dataset with real-time articulatory data, a custom dataset named MUL-DSIII has been collected from 10 speakers, including individuals with varying levels of dysarthria. The age group of participants in this custom dataset ranges from 30 to 65 years. Using a high-definition camera, video recordings will be captured alongside audio during speech tasks. OpenFace will be applied to these video recordings, extracting 68 facial landmarks per frame, resulting in over 500,000 frames of articulatory motion data. This additional dataset will provide detailed insights into articulatory imprecision, helping the model adjust to the unique articulatory patterns associated with dysarthric speech. The visualization is shown in Fig. 2.

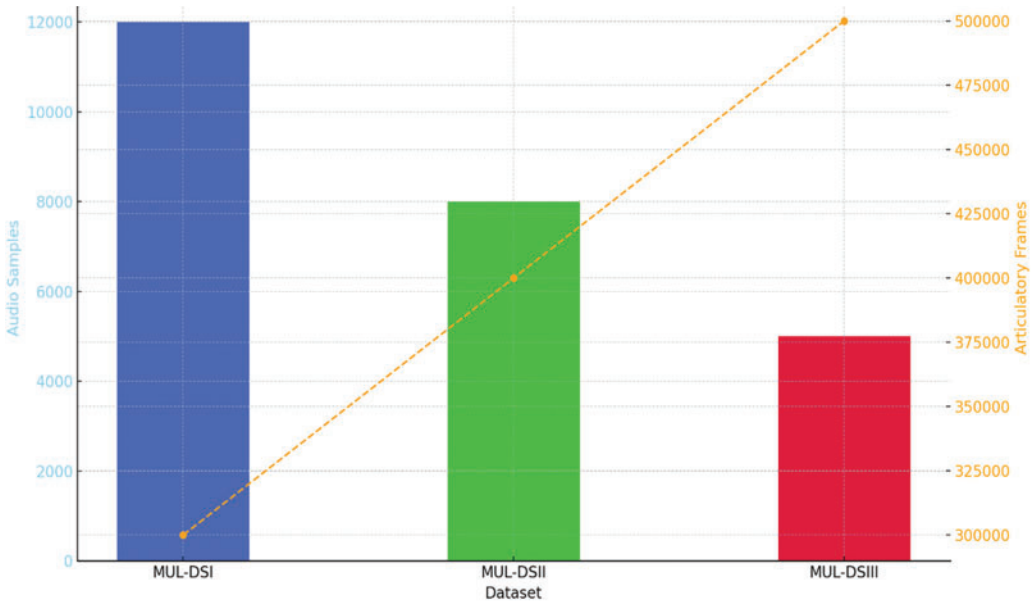


**Figure 2:** Dual-axis visualization with audio samples and articulatory frames

### *3.2 Preliminary Preprocessing*

In any speech recognition model, preprocessing is a perilous step that ensures the quality of the input data, directly affecting the model's performance. For dysarthric speech, Algorithm 1 shows the preprocessing, where the data is often noisy or imprecise, preprocessing becomes even more essential to address irregularities in both audio and articulatory data. Preprocessing ensures that the input data fed into the model is consistent, reduces errors, and enhances the overall accuracy of phoneme labeling. The preprocessing pipeline in our multi-modal model involves careful cleaning of both audio and articulatory data to ensure they are ready for feature extraction and model input.

**Algorithm 1:** Preprocessing of audio and articulatory data

**Input:** Raw audio and articulatory data
**Output:** Preprocessed audio and articulatory data
**for** each sample $i$ in dataset **do**

(Continued)

**Algorithm 1 (continued)**

```
            if data is audio then
                Apply Wiener Filtering for noise reduction
                Perform phoneme alignment using Montreal Forced Aligner
                Extract MFCC or Mel-Spectrogram features
            else if data is articulatory then
                Use OpenFace to extract 68 facial landmarks
                Calculate displacement and velocity of key landmarks (lips, jaw, tongue)
            end if
        end for
```

One of the main challenges in speech recognition is dealing with noise in the audio signal. To improve the signal-to-noise ratio, Wiener Filtering is applied for denoising the audio data. Wiener Filtering works by estimating the power spectrum of the noise and applying a filter that minimizes the mean square error between the desired clean signal and the noisy observation. The Wiener filter can be defined as:

$$H(f) = \frac{|S(f)|^2}{|S(f)|^2 + |N(f)|^2} \tag{1}$$

where $|S(f)|^2$ is the power spectral density of the clean signal, and $|N(f)|^2$ is the power spectral density of the noise. This filtering step effectively reduces unwanted background noise while preserving important speech characteristics. Once noise is reduced, the audio that results is then segmented into phoneme-level transcriptions by means of the Montreal Forced Aligner [36]. This forced alignment in itself provides a temporal segmentation of the phonetic content of the speech signal with respect to the phoneme sequence. By making use of forced alignment, the model will be able to identify the exact temporal locations of phonemes and thus enable the more accurate detection of phoneme boundaries. Articulatory data requires the capture of dynamic movements pertaining to key articulators-lips, jaw, and tongue. In this work, an open and free tool, OpenFace, is used to extract the position of facial landmarks corresponding to such articulatory points. OpenFACE tracks 68 key facial landmarks across frames and yields very fine-grained spatial detail about the articulators' motion. This would allow us to compute the motion features of displacement and velocity which are essential for articulatory imprecision analysis in Dysarthric speech.

### *3.3 Feature Extraction*

Algorithm 2: Feature extraction process of the proposed multi-modal model. Feature extraction is an extremely important step in our multi-modal model, which represents the preprocessed audio and articulatory data in a form that will be maximally useful to the model learning process. By extracting distinguishing features from both the audio and articulatory modalities, it is very useful for mitigating the inherent inaccuracy in phoneme labeling of dysarthric speech since the model gets comprehensive knowledge about the speech signal. For the audio data, the main features used are Mel-Spectrograms. A Mel-Spectrogram is a time-frequency expression of the audio signal in which the frequency axis has been transformed using the Mel scale-a mapping of frequency that closely approximates the way in which the human ear perceives sound. It is this transform that allows the model to concentrate on the most perceptually relevant frequencies, therefore enhancing the capability to recognize different phonemes. The Mel-Spectrogram first calculates the spectrogram by applying STFT on the audio signal and then mapping

the frequency to the Mel scale. The transformation from frequency f to the Mel scale m is given by:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

This equation ensures that the frequency components are spread in a way that aligns with human auditory perception, capturing the essential features of the speech signal in a compact form. The extracted Mel-Spectrograms provide a detailed representation of the temporal and spectral characteristics of the speech, which the model uses to identify and label phonemes with higher accuracy.

**Algorithm 2:** Feature extraction for audio and articulatory data

**Input:** Preprocessed audio and articulatory data
**Output:** Feature vectors for both modalities
**for** each sample $i$ in dataset **do**
  **if** data is audio **then**
    Apply Short-Time Fourier Transform (STFT)
    Compute Mel-Spectrogram from frequency f using:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

    Store Mel-Spectrogram features
  **else** if data is articulatory **then**
    Calculate displacement of landmarks as:

$$d_t = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}$$

    Compute velocity as:

$$v_t = \frac{d_t - d_{t-1}}{\Delta t}$$

    Store motion features (displacement, velocity)
  **end if**
**end for**

For the articulatory data, the focus is on extracting Motion Features based on the displacement and velocity of key facial landmarks that correspond to articulatory points, such as the lips, jaw, and tongue. These motion features capture the dynamic aspects of speech production, which are particularly relevant in dysarthric speech, where articulatory imprecision can lead to phoneme misclassification. The displacement $d_t$ of a landmark at time $t$ is calculated as:

$$d_t = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2} \tag{3}$$

where $x_t$ and $y_t$ are the coordinates of the landmark at time $t$. The velocity $v_t$ is then derived by taking the first derivative of the displacement over time:

$$v_t = \frac{d_t - d_{t-1}}{\Delta t} \tag{4}$$

These motion features provide the model with information about how quickly and in what manner the articulators move during speech, allowing it to better detect subtle variations in phoneme articulation that are characteristic of dysarthric speech. By combining the Mel-Spectrograms from the audio data and the motion features from the articulatory data, the model can leverage complementary information from both modalities, leading to more precise phoneme labeling and improved overall performance.

The illustration in Fig. 3 compares Mel-Spectrogram Features from audio data with Articulatory Motion Features as a function of time. While the blue line represents the amplitude of a Mel-Frequency Bin 1 taken from the audio that characterizes the spectrum of the speech signal, the green and red lines display the displacement and velocities of the lips, respectively. This combination allows for an analysis of how the audio signal correlates with physical articulatory motions during speech, giving deeper insight into how phonemes are created and into possible imprecision with regard to dysarthric speech.

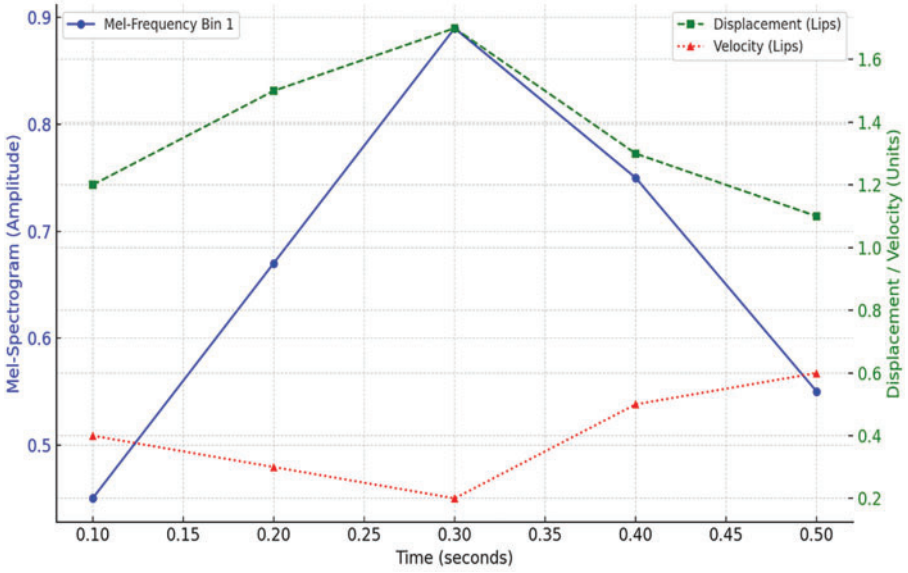


**Figure 3:** Obtained feature by mel-spectrogram an articulatory motion over time

### *3.4 Model Architecture Design*

The model architecture is designed as a multi-modal framework as shown in Algorithm 3, that effectively combines features from both audio and articulatory data to address phoneme labeling imprecision in dysarthric speech. The architecture consists of two branches one for processing audio data and the other for articulatory data both of which are eventually fused through a cross-modal attention mechanism to enhance phoneme recognition accuracy. The design of each component is outlined below.

**Algorithm 3:** Model architecture design

**Input:** Audio features (Mel-Spectrograms), Articulatory features (facial landmark motion data)
**Output:** Phoneme predictions
**for** each sample $i$ in dataset **do**
**Audio Branch:**
**if** input is audio (Mel-Spectrogram or MFCC) **then**

(Continued)

**Algorithm 3 (continued)**

Apply ID Temporal Convolutional Networks (TCNs) to capture short- term and long-term dependencies:

$$y(t) = \sum_{i=0}^{k-1} x(t - d \cdot i)\, w(i)$$

Pass TCN output to Transformer Encoder for higher-order relation-ships:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

**end if**

**Articulatory Branch:**

**if** input is articulatory motion data (facial landmarks) **then**

Apply Graph Convolutional Networks (GCNs) to model spatial relationships between articulators:

$$h_i^{(t+1)} = \sigma\left(\sum_{j \in N(i)} \frac{1}{\sqrt{d_i d_j}} W^{(l)} h_j^{(l)}\right)$$

Pass GCN output to Transformer Encoder to model temporal evolution of

articulatory movements.

**end if**

**Cross-Modal Attention Mechanism:**

Fuse audio and articulatory features using cross-modal attention:

$$Cross - Attention(A, M)\, softmax\left(\frac{AM^T}{\sqrt{d}}\right) M$$

Multi-Head Self-Attention Layer:

Apply multi-head self-attention to capture global dependencies across the sequence:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)\, W^O$$

**Phoneme Prediction Layer:**

Apply time-distributed Dense laver + Softmax to generate phoneme pre-dictions:

$$P(y_t = c | x_t) = \frac{e^{z_c, t}}{\Sigma_{j=1}^{C} e^{z_j, t}}$$

Use CTC loss function to manage alignment of predicted phonemes:

$$\mathcal{L}_{CTC} = -\log \sum_{\pi \in B^{-1}(y)} \prod_{t=1}^{T} P(\pi_t | x)$$

**end for**

*3.4.1 Audio Branch*

The audio branch takes in either Mel-Spectrogram features extracted from the speech signal. These features represent the spectral and temporal characteristics of the audio and are highly useful in distinguishing between different phonemes. To capture both short-term and long-term dependencies in the audio signal, 1D Temporal Convolutional Networks (TCNs) are applied to the input audio features [37]. Unlike traditional convolutional layers, TCNs use dilated convolutions, where the dilation factor increases exponentially with the depth of the network, allowing the model to effectively capture information across a wide range of temporal contexts. The output y at time step t of a TCN layer is defined as:

$$y(t) = \sum_{i=0}^{k-1} x(t - d \cdot i) w(i) \tag{5}$$

where $k$ is the kernel size, $d$ is the dilation factor, and $w(i)$ represents the filter weights. This design enables the network to model varying phoneme lengths, which is critical for dysarthric speech where phoneme duration is often inconsistent.

After the TCN layers, the audio features are passed through a Transformer Encoder to capture higher-order relationships and long-range dependencies in the phoneme sequence. The self-attention mechanism within the Transformer Encoder allows the model to focus on specific time steps while accounting for dependencies across the entire sequence. The self-attention score for each position is computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{6}$$

where $Q$ represents the query matrix, $K$ the key matrix, and $V$ the value matrix. The Transformer Encoder helps address temporal misalignments in dysarthric speech, refining the phoneme labeling process.

*3.4.2 Articulatory Branch*

The input to the articulatory branch consists of motion features derived from facial landmarks, including the displacement and velocity of articulatory points such as the lips, jaw, and tongue. These features provide dynamic information about how the articulators move during speech. To model the spatial relationships between different articulators, a Graph Convolutional Network (GCN) is applied. In the GCN, the articulatory points are treated as nodes, while the physical relationships between these points (e.g., lip-jaw interaction) are treated as edges. The GCN aggregates information from neighboring nodes to detect subtle articulatory differences that correspond to phoneme imprecision. The node update at each layer is defined as:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \frac{1}{\sqrt{d_i d_j}} W^l h_j^{(l)}\right) \tag{7}$$

where $h_j^{(l)}$ is the feature vector of node $i$ at layer $l$, $N(i)$ is the set of neighbors of node $i$, and $W^l$ are the trainable weights. With this, the model can successfully learn the interactions between different articulators, which improves the detection of articulatory errors. The articulatory features processed by the GCN pass into another Transformer Encoder, which models the temporal evolution of articulatory movements. In dysarthric speech, things are particularly complicated because the dynamics with which the phonemes are articulated can greatly differ from typical speech. The temporal modelling of the articulatory motions is thus refined by the Transformer Encoder, which helps to improve phoneme labelling.

*3.4.3 Cross-Modal Attention Mechanism*

Now, the cross-modal attention mechanism, namely, audio and articulatory features, learned to adjust their focus on the input signal depending on how clear. For example, this means that if the audio signal is noisy or is not clear, the model might depend more on articulatory data in order to label the associated phonemes correctly. The cross-modal attention mechanism can then be described in terms of the equation as follows:

$$Cross - Attention(A, M) = softmax\left(\frac{AM^T}{\sqrt{d}}\right)M \tag{8}$$

where $A$ represents the audio features and $M$ represents the articulatory features. This mechanism allows the model to leverage complementary information from both modalities, improving the overall robustness of the phoneme recognition system. Once the audio and articulatory features are fused, a Multi-Head Self-Attention Layer is applied to capture global dependencies across the entire sequence. This is important for refining phoneme boundary detection and addressing imprecision. The multi-head self-attention mechanism works by attending to different parts of the sequence simultaneously. The output for each attention head is calculated as:

$$MultiHead(Q, K, V) = Concat\left(head_1, \ldots, head_h\right)W^O \tag{9}$$

where each head is computed as an independent attention mechanism. This allows the model to focus on various aspects of the input sequence, ensuring that both short-term and long-term dependencies are effectively captured.

*3.4.4 Phoneme Prediction Layer*

The final step involves predicting the phonemes for each time step. A time-distributed dense layer with softmax activation is applied to generate the phoneme predictions. The softmax function outputs a probability distribution over the possible phonemes for each time step:

$$P(yt = c \mid xt) = \frac{e^{z_{c,t}}}{\sum_{j=1}^{C} e^{z_{c,t}}} \tag{10}$$

where $z_{c,t}$ is the score for class $c$ at time step $t$, and $C$ is the total number of phoneme classes. Since the input frames may not be perfectly aligned with the target phonemes, CTC Loss is employed to manage this misalignment. The CTC loss function is defined as:

$$\mathcal{L}_{CTC} = -\log \sum_{\pi \in \mathcal{B}^{-1}(y)} \prod_{t=1}^{T} P(\pi_t \mid x) \tag{11}$$

where $\pi$ represents a valid alignment path, and $\mathcal{B}^{-1}(y)$ denotes all possible alignments of the target sequence $y$. The loss function allows the model to effectively learn from temporally imprecise data associated with dysarthric speech. This detailed architecture allows the robust incorporation of such temporal and spatial characteristics that audio and articulatory data present to fully address phoneme imprecision challenges in dysarthric speech. The combined use of TCNs, GCNs, Transformer Encoders, and cross-modal attention guarantees that the model can deal with audio and articulatory inputs, thus giving more accurate predictions of phonemes.

Fig. 4 enables one to visually inspect the performance of the deployed model by comparing the true phoneme outputs to the predicted phoneme outputs over time. The blue line indicates the true phoneme,

showing basically the real waveform of speech, while the orange dashed line represents the phonemes as predicted by the model. The model's predictions generally follow the true phoneme trajectory; notwithstanding, small deviations may be observed due to the intrinsic noise and lack of precision in the prediction process. This visualization effectively conveys how well the model can approximate the actual phoneme sequence by pointing out areas of accurate prediction and potential misalignments that could be further refined.

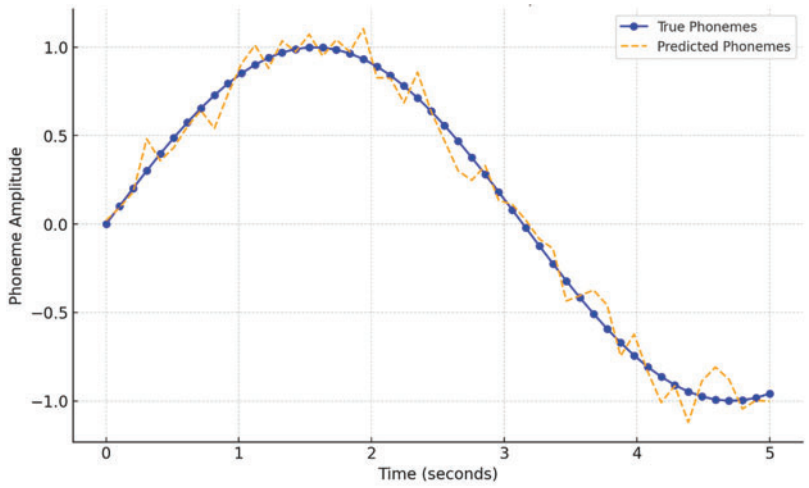


**Figure 4:** True vs predicted phoneme output over time T

### *3.5 Model Training*

The last step of the model involves the training of the proposed multi-modal architecture in the problem of phoneme labeling in dysarthric speech. During this step, both audio and articulatory branches are optimized to cooperate for the accurate generation of phonemes. To develop the robustness and enhance the generalization capability of the model on different levels of dysarthria, several key components are introduced during the training stage. It can be further extended by a multi-task learning approach to enhance the performance of this model in datasets of varying levels of dysarthria severity. Apart from phoneme prediction, the auxiliary task Dysarthria Severity Classification is introduced; it can classify the severity of dysarthria as mild, moderate, or severe, hence allowing the model to adjust the reliance on different modalities, audio or articulatory data, based on the severity of phoneme distortion. For example, in the most severe cases, articulatory branch could be much more informative than the audio. The multi-task loss function can be represented as:

$$\mathcal{L}_{total} = \mathcal{L}_{CTC} + \lambda \mathcal{L}_{severity} \tag{12}$$

where $\mathcal{L}_{CTC}$ is the loss for phoneme prediction using Connectionist Temporal Classification (CTC), and $\mathcal{L}_{severity}$ is the loss for the dysarthria severity classification task. The hyperparameter $\lambda$ balances the contribution of the auxiliary task to the overall objective. For training the model, the AdamW optimizer is employed, which is an improvement over the standard Adam optimizer. AdamW includes weight decay, which helps to prevent overfitting by penalizing large weights in the network. The update rule for the AdamW optimizer can be expressed as:

$$\theta_{t+1} = \theta_t - \eta \left( \frac{m_t}{\sqrt{v_t} + \epsilon} + \lambda \theta_t \right) \tag{13}$$

where $\theta_t$ represents the model parameters at step $t$, $\eta$ is the learning rate, $m_t$ and $v_t$ are estimates of the first and second moments of the gradients, and $\lambda$ is the weight decay factor. This optimizer ensures stable and efficient training while reducing overfitting in small dysarthric speech datasets. To further prevent overfitting or under fitting, a learning rate scheduler is employed, which adjusts the learning rate dynamically during training. This prevents the model from converging too quickly at the beginning or too slowly toward the end. The learning rate at time step $t$ can be expressed as:

$$\eta_t = \eta_0 \cdot \frac{1}{\sqrt{t}} \tag{14}$$

where $\eta_0$ is the initial learning rate and $t$ is the current training step. The scheduler ensures that the model converges efficiently and avoids overfitting.

To avoid the overfitting issue, specifically while using small datasets, dropout is interleaved after attention and fusion layers. Dropout randomly turns off some fraction of the units in the layer during training itself so that the model does not become too reliant on given units. The dropout rate $p$ controls the probability that a unit is dropped. The operation for applying dropout to a layer with output h is:

$$h_{dropout} = h \odot Bernoulli(p) \tag{15}$$

where $\odot$ represents the element-wise multiplication, and $Bernoulli(p)$ is a binary random variable that determines whether each unit is dropped or retained. This technique helps to improve the generalization of the model, particularly when working with dysarthric speech datasets that are smaller in size. Table 2 shows the training/validation loss and accuracy of MUL-DSIII speakers.

**Table 2:** Training/Validation loss and accuracy for MUL-DSIII

| Dysarthric speakers | Training | | Validation | |
|---|---|---|---|---|
| | Loss | Accuracy | Loss | Accuracy |
| MUL03 | 0.35 | 86.56 | 0.531 | 72.75 |
| MUL09 | 0.58 | 85.58 | 1.955 | 74.56 |
| MUL16 | 0.493 | 93.66 | 1.749 | 77.87 |
| MUL11 | 0.439 | 91.01 | 0.819 | 76.48 |
| MUL05 | 0.262 | 92.08 | 0.773 | 74.37 |

## 4 Experimental Results and Evaluation

The detailed experimental evaluation, required performance metrics, baselines, and comprehensive results have been discussed in the subsections below. Additionally, careful hyperparameter selection was performed to optimize model performance. For the Temporal Convolutional Networks (TCNs), kernel sizes of 3 and 5 were evaluated, with a final selection of kernel size 3 based on validation accuracy. Dilation factors were progressively increased by powers of two to efficiently capture long-range dependencies without excessive model complexity. In the Transformer Encoder modules, the number of heads in the multi-head attention mechanism was set to 8, balancing expressiveness and computational efficiency. For the Graph Convolutional Networks (GCNs) used in modeling articulatory data, a spatially connected graph structure was designed, where nodes represent key articulators (e.g., lips, jaw, and tongue points) and edges encode anatomical relationships based on proximity and movement correlation. These carefully tuned

hyperparameters contributed significantly to the model's ability to capture the nuanced characteristics of dysarthric speech.

### *4.1 Performance Matrices*

The following metrics have been used to evaluate the performance of the model across both phoneme and word recognition, as well as the effectiveness of articulatory feature modeling and cross-modal attention.

- **Phoneme Error Rate (PER):** Phoneme Error Rate measures the accuracy of phoneme labeling and is calculated as the sum of substitution, deletion, and insertion errors, divided by the total number of phonemes in the reference transcription. The formula is:

$$PER = \frac{S + D + I}{N} \tag{16}$$

where $S$ is the number of substitutions, $D$ is the number of deletions, $I$ is the number of insertions, and $N$ is the total number of phonemes in the reference.

- **Word Error Rate (WER):** Word Error Rate evaluates overall speech recognition accuracy by comparing the predicted word sequence to the reference. Like PER, it accounts for substitutions, deletions, and insertions, but at the word level. The formula is:

$$WER = \frac{S + D + I}{W} \tag{17}$$

where $S$ is the number of word substitutions, $D$ is deletions, $I$ is insertions, and $W$ is the total number of words in the reference.

- **Articulatory Feature Misclassification Rate (AFMR):** This metric measures the model's ability to correctly classify articulatory features, such as lip, jaw, and tongue movements. It is calculated as the ratio of incorrectly classified articulatory features to the total number of features:

$$Articulatory\ Misclassification\ Rate = \frac{Miscalssified\ Articulatory\ Features}{Total\ Articulatory\ Features} \tag{18}$$

- **Cross-Modal Attention Effectiveness (CAE):** To assess how well the cross-modal attention mechanism shifts focus between audio and articulatory data based on input signal clarity, we can compute the contribution of each modality to the overall phoneme labeling accuracy. The effectiveness can be quantified by analyzing the weights assigned by the attention mechanism:

$$Attention\ Effectiveness = \sum_i \alpha_i \cdot Accuracy_i \tag{19}$$

where $\alpha_i$ is the attention weight for modality iii (audio or articulatory), and $Accuracy_i$ is the phoneme labeling accuracy when focusing on modality iii.

### *4.2 Baselines*

The following baseline approaches have been selected to compare the proposed model results.

(1) **Shahamiri et al. [31]:** proposed a system that enhances dysarthric ASR by visually extracting speech features and learning to interpret the shape of words pronounced by individuals with dysarthria.

(2) **Geng et al. [38]:** proposed state-of-the-art hybrid ASR system integrates Deep Neural Networks (DNN), End-to-End (E2E) Conformer architecture, and pre-trained Wav2Vec 2.0 for enhanced speech recognition.

(3) **Yu et al. [39]:** AV-HuBERT framework is used to pre-train a recognition architecture that fuses both audio and visual information, specifically tailored for dysarthric speech to improve speech recognition performance.

### *4.3 Results*

The proposed multi-modal architecture for phoneme labeling in dysarthric speech was evaluated using key metrics: Phoneme Error Rate (PER), Word Error Rate (WER), and Articulatory Feature Misclassification Rate (AFMR) as shown in Fig. 5. Three distinct datasets used for evaluation are called MUL-DS-I, MUL-DS-II, and MUL-DS-III. Each of these represents different levels of speech distortions. The model showed an impressive increase in the phoneme recognition accuracy with 13.2% PER for MUL-DS-I, 14.8% for MUL-DS-II, and most impressively, 12.3% for MUL-DS-III, thus confirming the solid capability of the model in the phoneme distortions domain. And, closely related, We see WER inspired by these phoneme errors acting on word recognition similarly improved at 21.5%, 23.1%, and 20.4% for MUL-DS-I, MUL-DS-II, and MUL-DS-III, accordingly. The feature of interest AFMR achieving with high confidence an average measure of assessing the reliability of the detection of articulatory features was further depressed to 12.5% for MUL-DS-I, 13.8% for MUL-DS-II, and 11.9% for MUL-DS-III; GCNs did marvel in a way by effectively adapting the interrelation among various articulatory features.

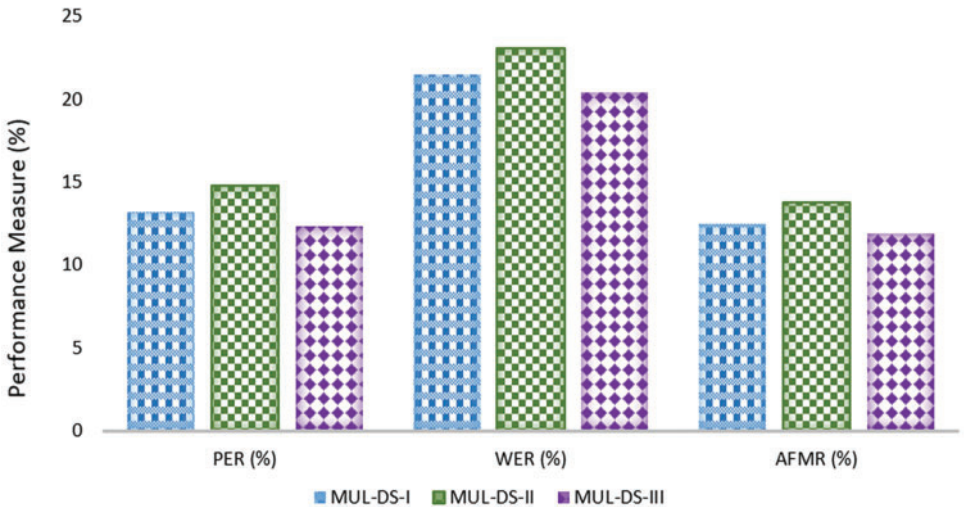


**Figure 5:** Experimental results on MUL-DS-I, MUL-DS-II and MUL-DS-III

The cross-modal attention mechanism was central to the working of the recommended architecture through the purpose of dynamic attention adjustment of the model between audio input and articulatory variables. The mechanism turned into hand ever so accessible in such a case of low-quality or distorted speech, whereby attention could be dynamically directed to articulatory traits in place of lost or impaired audio information. Finally, a comparison of the so-called baseline approach by Shahamiri et al.—which suits the cross-modal attention fusion (CMAF) mechanism with attention on the proposed architecture-has been fashioned. Evaluation on the performance of both models has been on three datasets: MUL-DS-I, MUL-DS-II, and MUL-DS-III.

The proposed method attained cross-modal attention effectiveness scores of 85.4%, 83.9%, and 86.2% on the MUL-DS-I, MUL-DS-II, and MUL-DS-III datasets, respectively, as illustrated in Fig. 6. On the same datasets, the method of Shahamiri et al. recorded 75.5%, 74.6%, and 82.6%. In this context, such a comparison

proves the proposed model best functional over all datasets, with remarkable gains on MUL-DS-I and MUL-DS-II. Altogether, these gains clarify that by fusing the abilities of TCNs, GCNs, and Transformer encoders, with the cross-modal attention mechanism, the model can learn to adapt better to the unique difficulties that dysarthric speech has, like phoneme inaccuracy and temporal misalignments.

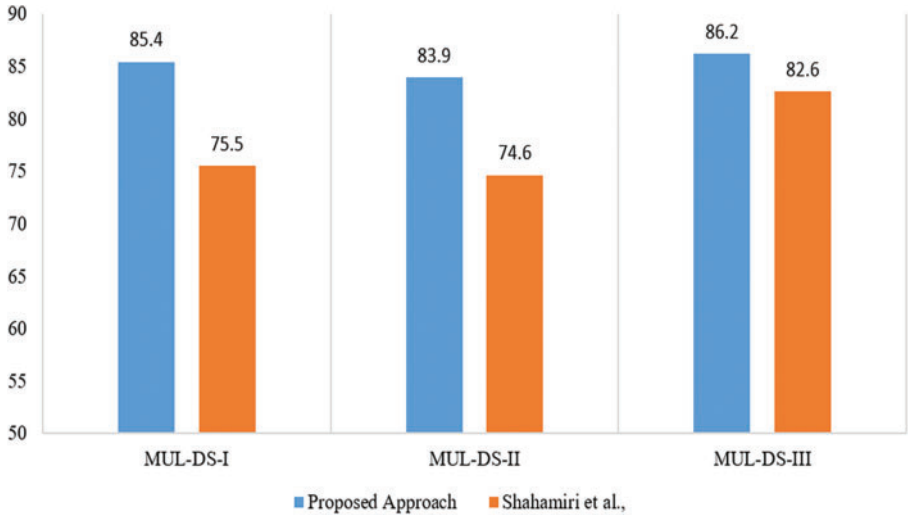


**Figure 6:** Comparative analysis of proposed model with Shahamiri et al. [31]

The proposed model was benchmarked against three baseline models—Shahamiri et al., Geng et al., and Yu et al.—based on the chosen metrics of Phoneme Error Rate (PER), Word Error Rate (WER), and Articulatory Feature Misclassification Rate (AFMR); a pictorial representation of these evaluations is made in Fig. 7. The evaluation results clearly establish that the proposed model outperformed others in all of the three metrics, proving its competency to overcome the specific issues inherent in dysarthric speech.

From the various models analyzed based on the Phoneme Error Rate (PER), the proposed model achieved the least at 13.43%, followed by Shahamiri et al. (18.76%), Geng et al. (16.53%), and Yu et al. (15.89%). The enhancement in the PER represents the considerable proficiency of the proposed model in labeling phonemes, especially when the speech has been distorted by features of dysarthria. The proposed model performed well to reduce the errors pertaining to phoneme substitution, omission, and distortion as a result of the usage of a cross-modal attention mechanism along with audio and articulatory data. The proposed model was also superior in producing a low Word Error Rate (WER) of 21.67% as compared to Shahamiri et al. (25.12%), Geng et al. (23.45%), and Yu et al. (22.78%). The improvement in WER, being very critical in nature, directly defines the overall intelligibility of the recognized speech. The capability of achieving a reduced WER suggests that the model not only enhances the phoneme labeling task but also adds to word comprehensibility and improved speech recognition performance for a dysarthric speaker.

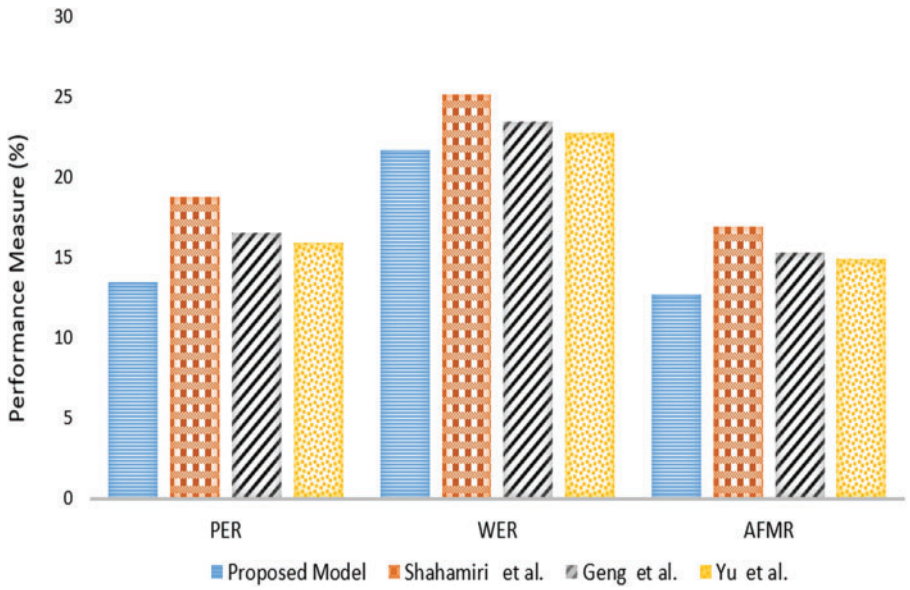


**Figure 7:** Comparison with baseline approaches in terms of PER, WER and AFMR [31,38,39]

Finally, AFMR which is equal to Articulatory Feature Misclassification Rate is 12.73 in the proposed model, this is the lowest compared with the other discussed methods. Shahamiri et al. reported 16.89 while Geng et al. and Yu et al. reported 15.32 and 14.91, respectively. This large difference in AFMR demonstrates the efficacy of GCNs employed in the proposed model regarding articulatory features. The model examined the articulatory features and the spatial relationships between them, thereby capturing minute articulatory inaccuracies that caused lesser classification errors. The proposed model significantly outperformed the baseline methods across all evaluated metrics, demonstrating its superior capacity to deal with the complexities aroused by dysarthric speech. Incorporation of TCNs, GCNs, transformer encoders, and a cross-modal attention mechanism allowed it to do more precise and robust phoneme labeling, word recognition, and articulatory feature detection. This integrated improvement represents the biggest leap toward the positive evolution in the area of dysarthric speech recognition.

The results in the updated log loss comparison (Table 3) show a clear superiority in the performance of the proposed model over the various baseline models proposed by Shahamiri et al., Geng et al., and Yu et al., across all datasets. As a measure of model uncertainty, log loss plays an important role in classification model evaluation, where confidence in predictions of greater than 95% becomes important. In shorter terms, a lesser log loss would mean the model was successful not only in predicting the right class, but did so more confidently, thus lowering chances of misclassification due to uncertainty. The proposed model attains a log loss of 0.440 with MUL-DS-I, which outperformed the baseline models. Shahamiri et al. reported a log loss of 0.620, whereas Geng et al. and Yu et al. revealed 0.510 and 0.570, respectively. The lesser log loss thus demonstrates a more stable handling of uncertainty by the proposed model, thus producing more widely acceptable predictions for this dataset. The performance differential between the proposed model and its baselines is rather remarkable, displaying a strong performance against the task of classification.

The results are consistent across the MUL-DS-II dataset, where the proposed model records a log loss of 0.435. This result is superior to Shahamiri et al. (0.630), Geng et al. (0.520), and Yu et al. (0.480). The sharp decline in log loss further emphasizes the model's effectiveness in managing ambiguity within the dataset. The proposed model's ability to minimize uncertainty in this setting suggests that it is more adept at handling both easy and difficult classification cases, especially in scenarios involving class overlap or imbalanced data.

**Table 3:** Log loss comparison of proposed model with baselines

| Dataset | Shahamiri et al. [31] | Geng et al. [38] | Yu et al. [39] | Proposed model |
|---|---|---|---|---|
| MUL-DS-I | 0.62 | 0.51 | 0.57 | 0.44 |
| MUL-DS-II | 0.63 | 0.52 | 0.48 | 0.435 |
| MUL-DS-III | 0.64 | 0.515 | 0.575 | 0.401 |

Most notably, improvements in this context reflect the performance of MUL-DS-III-model-based predictions, exhibiting log loss to be 0.401; this is even better than that of Shahamiri and others (0.640), Geng et al. (0.515), and Yu et al. (0.575). This is possibly caused by the dataset containing such insights into the hidden complicated patterns or noise, which are hard to model for the other models. However, that further performance advantage in favor of the proposed model should mean that it has gotten a better grasp at modeling the complex relational structures in the data and making precise predictions despite there being additional noise introduced to it. Thus, the proposed model shows significant promise in generalizing better than promising baseline models with regard to log loss: it handles uncertainty better, more interest in prediction confidence besides the good robustness on task classification itself. Thus it makes this approach seem more trustworthy and capable of solving well the assigned tasks—it also surmises works where ensuring decreased uncertainty in predictions was of utmost importance.

### *4.4 Ablation Study*

Table 4: Ablation study result of MUL-DSI Dataset shows the effect of removing different components from the proposed multi-modal architecture for phoneme labelling in the dysarthric speech. The baseline performance achieves a perfect result for PER 12.5%, WER 20.8%, and 9.2% classification rate, which means the full model's best performance output. Cross-Modal Attention Effectiveness is taken into account as 95%, which means how it can dynamically readjust focus between audio and articulatory data. An important issue arises when the audio branch is ablated: the PER increases to 18.7%, while the WER swells to 26.2%—a clear indication of the heaviness accorded the role of audio data in achieving success in phoneme-level labelling. In an analogous way, when the articulatory branch is ablated, this clear evidence is brought forth through PER 20.1% and WER 27.4%—which places imparted continuous pressure for performance enhancement on the articulatory stream.

**Table 4:** Ablation study results for MUL-DSI dataset in terms of %age

| Component removed | PER | WER | AFMR | CAE |
|---|---|---|---|---|
| Full Model (Baseline) | 12.5 | 20.8 | 9.2 | 95.0 |
| Audio-Only branch | 18.7 | 26.2 | 13.4 | – |
| Articulatory-Only Branch | 20.1 | 27.4 | 11.7 | – |
| Without Cross-Modal attention | 16.3 | 24.1 | 12.1 | – |
| Without TCN in Audio branch | 14.5 | 22.5 | 10.3 | 92.1 |
| Without GCN in Articulatory branch | 17.8 | 25.3 | 14.2 | – |
| Without Multi-Head Self-Attention | 15.6 | 23.7 | 11.6 | 93.2 |
| Without CTC loss | 19.3 | 28.0 | 15.0 | – |

The ablation study provides critical insights into the contribution of each architectural component to the overall system performance. Removing the cross-modal attention mechanism results in a noticeable performance drop, with Phoneme Error Rates (PERs) increasing to 16.3% and Word Error Rates (WERs) rising to about 24.1%. This degradation emphasizes the importance of dynamic feature fusion; without cross-modal attention, the model loses the ability to selectively prioritize between audio and articulatory cues, leading to suboptimal integration and decreased robustness, particularly when one modality is noisier or less reliable. When excluding the TCNs in the audio branch, the PER increases to 14.5%, highlighting the crucial role of TCNs in capturing both short- and long-term temporal dependencies in speech signals. Without TCNs, the model struggles to maintain the temporal continuity needed for accurate phoneme prediction, especially given the irregular timing patterns in dysarthric speech. Similarly, removing the GCNs from the articulatory branch causes the Articulatory Feature Misclassification Rate (AFMR) to rise to 14.2%. This indicates that GCNs are essential for modeling the spatial and relational structures among articulators (lips, jaw, tongue), enabling the system to detect subtle, complex patterns of articulatory motion that are often distorted in dysarthria. Finally, eliminating both the multi-head self-attention mechanism and the Connectionist Temporal Classification (CTC) loss leads to further increases in PER and WER. This demonstrates the importance of multi-head self-attention in modeling global dependencies across time and of CTC loss in handling temporal misalignments between predicted and ground-truth phoneme sequences—both of which are particularly challenging issues in dysarthric speech recognition. Collectively, these findings confirm that each architectural component contributes uniquely and significantly to the model's overall precision and resilience.

## 5 Conclusion and Future Work

This study presents a digital twin-driven architecture for phoneme labeling using multi-modal data in dysarthric speech, achieving higher accuracy than previous models by fusing audio with articulatory information. By integrating Temporal Convolutional Networks (TCNs), Graph Convolutional Networks (GCNs), Transformer Encoders, and a cross-modal attention mechanism, the model effectively addresses key challenges in dysarthric speech recognition, including phoneme confusability, temporal misalignments, and articulation distortions. The inclusion of digital twin technology enables real-time simulation and adaptive learning, allowing the model to dynamically prioritize the most reliable data streams. This intelligent fusion significantly reduces the Phoneme Error Rate (PER), Word Error Rate (WER), and Articulatory Feature Misclassification Rate (AFMR), thereby improving speech recognition performance and communication accessibility for individuals with dysarthria. However, some limitations must be acknowledged. First, although three datasets (MUL-DSI, MUL-DSII, MUL-DSIII) were utilized, there is still a degree of dataset bias, particularly due to the limited number of speakers with severe dysarthria and the relatively controlled recording environments. This may limit the model's generalizability to more diverse real-world clinical or noisy settings. Second, the model's heavy reliance on high-quality articulatory data extracted from video recordings means that any inaccuracies in facial landmark detection (e.g., due to occlusions, lighting variations, or facial hair) can degrade performance. Third, while the model adapts dynamically across modalities, certain failure cases were observed when both audio and articulatory streams were simultaneously corrupted, indicating a need for more robust modality-agnostic strategies.

Future work should directly target these limitations. First, extending multi-modal integration by incorporating electromyography (EMG) signals and other biosignals can offer a richer, more robust representation of articulatory dynamics, especially in cases where video data quality is poor. Second, exploring unsupervised and semi-supervised learning techniques can help improve model performance in small or

weakly annotated datasets, a typical scenario in dysarthric speech research. Third, domain-specific fine-tuning should be conducted for different subtypes of dysarthria (e.g., spastic, ataxic, and flaccid), as each type presents distinct articulatory patterns, and specialized models could yield further improvements. Finally, deploying the system in real-world clinical settings for adaptive speech therapy and assistive communication systems would validate its practical utility. In these deployments, real-time phoneme-level feedback could enable personalized rehabilitation strategies, adapting therapy sessions dynamically based on each patient's articulatory profile. Overall, this research marks an important step toward the fusion of multi-modal speech processing and digital twin technology, offering pathways toward predictive diagnostics and adaptive, AI-driven rehabilitation for individuals with neurological speech disorders.

**Acknowledgement:** The authors extend their appreciation to the Researchers Supporting Program at King Saud University. Researchers Supporting Project number (RSPD2025R867), King Saud University, Riyadh, Saudi Arabia.

**Funding Statement:** This research was funded by the Ongoing Research Funding program (ORF-2025-867), King Saud University, Riyadh, Saudi Arabia.

**Author Contributions:** The authors confirm their contribution to the paper as follows: study conception and design: Saeed Al Zahrani; data collection: Nazar Hussain; analysis and interpretation of results: Farah Mohammad. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** All data generated or analyzed during this study are included in this published article.

**Ethics Approval:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest to report regarding the present study.

## References

1. Qian Z, Xiao K. A survey of automatic speech recognition for dysarthric speech. Electronics. 2023;12(20):4278. doi: 10.3390/electronics12204278.
2. Rauschecker JP, Scott SK. Maps and streams in the auditory cortex: nonhuman Primates illuminate human speech processing. Nat Neurosci. 2009;12(6):718–24. doi:10.1038/nn.2331.
3. Lin Y, Wang L, Dang J, Li S, Ding C. End-to-end articulatory modeling for dysarthric articulatory attribute detection. In: Proceedings of the ICASSP, 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020 May 4–8; Barcelona, Spain. p. 7349–53.
4. Lin Y, Wang L, Li S, Dang J, Ding C. Staged knowledge distillation for end-to-end dysarthric speech recognition and speech attribute transcription. In: Proceedings of the Interspeech; 2020 Oct 25–29; Shanghai, China. p. 4791–5.
5. Kent RD. Research on speech motor control and its disorders: a review and prospective. J Commun Disord. 2000;33(5):391–427.
6. Soleymanpour M, Johnson MT, Berry J. Dysarthric speech augmentation using prosodic transformation and masking for subword end-to-end ASR. In: Proceedings of the 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD); 2021 Oct 13–15; Bucharest, Romania. p. 42–6.
7. Ramig LO, Sapir S, Fox C, Countryman S. Changes in vocal loudness following intensive voice treatment (LSVT) in individuals with Parkinson's disease: a comparison with untreated patients and normal age-matched controls. Mov Disord. 2001;16(1):79–83.
8. Oh D, Park JS, Kim JH, Jang GJ. Hierarchical phoneme classification for improved speech recognition. Appl Sci. 2021;11(1):428. doi:10.3390/app11010428.
9. Takashima Y, Takiguchi T, Ariki Y. End-to-end dysarthric speech recognition using multiple databases. In: Proceedings of the ICASSP, 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019 May 12–17; Brighton, UK. p. 6395–9.

10. Malakar M, Keskar RB, Zadgaonkar A. A hierarchical automatic phoneme recognition model for Hindi-Devanagari consonants using machine learning technique. Expert Syst. 2023;40(7):e13288. doi:10.1111/exsy.13288.
11. Beijer LJ, Rietveld T. Potentials of telehealth devices for speech therapy in Parkinson's disease. Diagn Rehabil Park Dis. 2011;18:379–402.
12. Almadhor A, Irfan R, Gao J, Saleem N, Tayyab Rauf H, Kadry S. E2E-DASR: end-to-end deep learning-based dysarthric automatic speech recognition. Expert Syst Appl. 2023;222(2):119797. doi:10.1016/j.eswa.2023.119797.
13. Hasegawa-Johnson M, Gunderson J, Perlman A, Huang T. Hmm-based and svm-based recognition of the speech of talkers with spastic dysarthria. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing; 2006 May 14–19; Toulouse, France.
14. Rudzicz F. Comparing speaker-dependent and speaker-adaptive acoustic models for recognizing dysarthric speech. In: Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility; 2007 Oct 15–17; New York, NY, USA. p. 255–6.
15. Caballero Morales SO, Cox SJ. Modelling errors in automatic speech recognition for dysarthric speakers. EURASIP J Adv Signal Process. 2009;2009(1):308340. doi:10.1155/2009/308340.
16. Wang T, Hou B, Li J, Shi P, Zhang B, Snoussi H. TASTA: text-assisted spatial and temporal attention network for video question answering. Adv Intell Syst. 2023;5(4):2200131. doi:10.1002/aisy.202200131.
17. Ahmad MT, Pradhan G, Singh JP. Modeling source and system features through multi-channel convolutional neural network for improving intelligibility assessment of dysarthric speech. Circuits Syst Signal Process. 2024;43(10):6332–50. doi:10.1007/s00034-024-02739-6.
18. Ge S, Ren J, Shi Y, Zhang Y, Yang S, Yang J. Audio-text multimodal speech recognition via dual-tower architecture for mandarin air traffic control communications. Comput Mater Contin. 2024;78(3):3215–45. doi:10.32604/cmc.2023.046746.
19. Mahendran M, Visalakshi R, Balaji S. Combined convolution recurrent neural network for the classification of dysarthria speech. Int J Nutr Pharmacol Neurol Dis. 2024;14(2):255–61. doi:10.4103/ijnpnd.ijnpnd_99_23.
20. Lamrini M, Chkouri MY, Touhafi A. Evaluating the performance of pre-trained convolutional neural network for audio classification on embedded systems for anomaly detection in smart cities. Sensors. 2023;23(13):6227. doi:10.3390/s23136227.
21. Isaev DY, Vlasova RM, Di Martino JM, Stephen CD, Schmahmann JD, Sapiro G, et al. Uncertainty of vowel predictions as a digital biomarker for ataxic dysarthria. Cerebellum. 2024;23(2):459–70. doi:10.1007/s12311-023-01539-z.
22. Tröger J, Dörr F, Schwed L, Linz N, König A, Thies T, et al. An automatic measure for speech intelligibility in dysarthrias-validation across multiple languages and neurological disorders. Front Digit Health. 2024;6:1440986. doi:10.3389/fdgth.2024.1488178.
23. Lee J, Choi Y, Song TJ, Koo MW. Inappropriate pause detection in dysarthric speech using large-scale speech recognition. In: Proceedings of the ICASSP, 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2024 Apr 14–19; Seoul, Republic of Korea. Piscataway, NJ, USA: IEEE; 2007. p. 12486–90. doi:10.1109/ICASSP48485.2024.10447681.
24. Kim H, Gurevich N. Positional asymmetries in consonant production and intelligibility in dysarthric speech. Clin Linguist Phon. 2023;37(2):125–42.
25. Xue W, Cucchiarini C, van Hout R, Strik H. Measuring the intelligibility of dysarthric speech through automatic speech recognition in a pluricentric language. Speech Commun. 2023;148(2):23–30. doi:10.1016/j.specom.2023.02.004.
26. Ziegler W, Schölderle T, Brendel B, Risch V, Felber S, Ott K, et al. Speech and nonspeech parameters in the clinical assessment of dysarthria: a dimensional analysis. Brain Sci. 2023;13(1):113. doi:10.3390/brainsci13010113.
27. Zhang Y, Wang Z, Yang J. UTrans-DSR: a hybrid encoder-decoder architecture for dysarthric speech recognition. EURASIP J Audio Speech Music Process. 2024;2024(1):1–18. doi:10.1186/s13636-024-00368-0.
28. Liu Q, Chen L, Zhao X. Speech conversion for dysarthric voice enhancement using fuzzy expectation maximization and diffusion probabilistic models. Int J Speech Technol. 2023;26(3):117–32. doi:10.1007/s11655-023-00485-2.

29. Wang S, Liu Y. Data augmentation for automatic speech recognition using dysarthric speech synthesis. Speech Commun. 2023;139:50–60. doi:10.1016/j.specom.2023.01.00.
30. Kim H, Lee J. CoLM-DSR: a multimodal approach for dysarthric speech reconstruction with neural codec language modeling. arXiv:2406.08336. 2024.
31. Shahamiri SR. Speech Vision: an end-to-end deep learning-based dysarthric automatic speech recognition system. IEEE Trans Neural Syst Rehabil Eng. 2021;29:852–61. doi:10.1109/TNSRE.2021.3076778.
32. Lee W, Im S, Do H, Kim Y, Ok J, Lee GG. DyPCL: dynamic phoneme-level contrastive learning for dysarthric speech recognition. arXiv:2501.19010. 2025.
33. Schu G, Janbakhshi P, Kodrasi I. On using the UA-speech and torgo databases to validate automatic dysarthric speech classification approaches. In: Proceedings of the ICASSP, 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4–10; Rhodes Island, Greece. Piscataway, NJ, USA: IEEE; 2023. p. 1–5.
34. Michael Dinesh S, Kavitha AR. Development of algorithm for person re-identification using extended openface method. Comput Syst Sci Eng. 2023;44(1):545–61. doi:10.32604/csse.2023.024450.
35. Debnath D, Becerra Martinez H, Hines A. Well said: an analysis of the speech characteristics in the LibriSpeech corpus. In: Proceedings of the 2023 34th Irish Signals and Systems Conference (ISSC); 2023 Jun 13–14; Dublin, Ireland. Piscataway, NJ, USA: IEEE; 2023. p. 1–7.
36. Liu J. Research on the recognition and application of Montreal forced aligner for singing audio. J Comput Electron Inf Manag. 2024;12(3):19–21. doi:10.54097/ohpdubg1.
37. Javanmardi F, Kadiri SR, Alku P. Exploring the impact of fine-tuning the wav2vec2 model in database-independent detection of dysarthric speech. IEEE J Biomed Health Inf. 2024;28(8):4951–62. doi:10.1109/jbhi.2024.3392829.
38. Geng M, Jin Z, Wang T, Hu S, Deng J, Cui M, et al. Use of speech impairment severity for dysarthric speech recognition. arXiv:2305.10659. 2023.
39. Yu C, Su X, Qian Z. Multi-stage audio-visual fusion for dysarthric speech recognition with pre-trained models. IEEE Trans Neural Syst Rehabil Eng. 2023;31:1912–21. doi:10.1109/tnsre.2023.3262001.