# Investigating the Impact of Feature Extraction techniques for Classification and synthesis of Dysarthric Speech

## Abstract

Dysarthria, a motor speech disorder marked by irregular acoustic patterns and rapid articulatory shifts, poses significant challenges for Dysarthric Speech Recognition (DSR) systems. Although spectrogram-based methods are extensively used in current research and have demonstrated strong performance, their fixed windowing approach may not fully capture the transient dynamics of dysarthric speech. To overcome this limitation, the Hilbert Spectrum is introduced as a novel feature extraction approach that offers adaptive time-frequency analysis to capture instantaneous amplitude and frequency variations. An end-to-end system was designed to integrate this approach. Classification evaluation using Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) revealed that ViTs achieve 91% accuracy with spectrograms, which drops to 81% with Hilbert Spectrum features, while CNNs attain 95% accuracy with the Hilbert Spectrum, closely matching spectrogram-based CNN's accuracy(96–97%). Additionally, HuBERT, which processes raw audio signals, reaches 97.27% accuracy, demonstrating the strength of direct waveform modeling. The end-to-end assistive system integrating a fine-tuned Whisper ASR for transcription with Google Text-to-Speech (gTTS) for synthesis further supports the approach. These results highlight the Hilbert Spectrum as a competitive alternative, offering enhanced adaptability for dysarthric speech analysis and promising improvements in assistive communication technologies.

**Keywords:** Dysarthria, Automatic Speech Recognition, Hilbert Spectrum, Convolutional Neural Networks, Vision Transformers, HuBERT, Whisper, Google Text-to-Speech, Assistive Technology

## Introduction

Dysarthria is a motor speech disorder caused by nervous system damage that impairs the muscles for speech. Conditions like stroke, traumatic brain injury, Parkinson's disease, or ALS result in weakness, poor coordination, and reduced control over the tongue, lips, and vocal cords. Consequently, speech becomes slurred, slow, or monotonous, significantly reducing intelligibility and hindering effective communication. While speech therapy and assistive technologies help, these challenges necessitate specialized Dysarthric Speech Recognition (DSR) systems. Traditional DSR methods use spectrogram-based techniques such as Mel-frequency cepstral coefficients (MFCCs) and short-time Fourier transforms (STFTs) that assume stationary signals but fail to capture the non-linear, variable acoustic features of dysarthric speech, leading to reduced recognition accuracy (Sahu & Pradhan, 2022).

To address these limitations, the Hilbert Spectrum, derived from the HHT, provides an adaptive time-frequency representation that enables the precise capture of instantaneous frequency and amplitude variations. Unlike traditional spectrograms, which impose predefined frequency bins and may obscure transient articulatory patterns, the Hilbert Spectrum dynamically decomposes speech signals into intrinsic mode functions (IMFs), allowing for a more detailed analysis of speech distortions and rhythm irregularities (Deng et al., 2010) . This adaptability makes the Hilbert Spectrum particularly well-suited for analyzing the complex articulatory distortions and variable phonetic features characteristic of dysarthric speech, enabling more robust classification and recognition performance.

Additionally, HuBERT—a self-supervised model that processes raw audio directly—learns robust speech representations, yielding high accuracy in dysarthric speech recognition. Complementing this, a fine-tuned Whisper model is used for speech-to-text transcription; when combined with gTTS synthesis, it converts dysarthric input into clear, intelligible speech, thereby improving overall assistive communication.

This paper explores the potential of Hilbert Spectrum-based feature extraction in DSR by comparing its performance against conventional spectrogram-based methods using deep learning models including CNNs, HuBERT, and ViTs. An end-to-end assistive communication system, integrating the fine-tuned Whisper ASR model with gTTS synthesis—is also developed to enhance speech intelligibility for dysarthric individuals. The competitive performance of the Hilbert Spectrum-based approach highlights the promise of adaptive signal processing techniques in developing more inclusive and effective ASR technologies.

The structure of this paper is organized as follows: Section 2 provides a comprehensive review of related work, highlighting existing studies and comparative analyses of pertinent research. Section 3 details the methodology employed in this study, including dataset descriptions, feature extraction processes, and data preprocessing techniques. Section 4 discusses the end-to-end systems implemented, covering CNN architectures, the Whisper model, and TTS systems. Section 5 presents the proposed approach in detail, outlining the custom CNN architecture and modifications made to the Whisper model. Section 6 describes the experimental analysis, including the setup, training parameters, and evaluation metrics. Section 7 presents the results and their evaluation, comparing the proposed models against baseline models through various visualizations. Finally, Section 8 concludes the paper by summarizing the main findings, discussing their implications, and suggesting potential directions for future research.

## 2. Related Work

Dysarthric speech, characterized by irregular and non-linear acoustic variations, has long posed a significant challenge to conventional ASR systems. Early studies primarily focused on the extraction of traditional features such as MFCCs and spectrogram-based representations. Zhang et al., (2022) introduced an improved MFCC variant, DMFCC, which incorporated signal decomposition techniques to enhance the detection of Wilson's disease dysarthria, achieving notable improvements over standard MFCC approaches. Complementing this, Lauraitis et al.,(2020) demonstrated that the fusion of cepstral coefficients, auditory spectrograms, and wavelet-based features within deep learning frameworks could yield classification accuracies exceeding 96%, underscoring the benefits of integrating multiple feature domains. Spectrogram-based features, particularly mel-spectrograms, have been shown to be highly effective; Igarashi et al. (2023) reported that mel-spectrograms outperformed MFCCs in binary classification tasks for dementia-related speech impairments, while rhythm metrics derived from spectrograms achieved classification rates near 90% as evidenced by Dahmani et al. (2013). In parallel, articulatory-based approaches have provided additional insights; studies by Yunusova et al. (2011) and Mendoza Ramos et al. (2021) highlighted that articulatory kinematics and acoustic cues, such as sentence accent identification, offer robust classification cues, achieving accuracies around 80% and demonstrating high performance across varying severity levels. Further, research by Fonville et al. (2008) and Bhattacharjee et al. (2023) emphasized the importance of dynamic source–filter attributes and perceptual analysis, while the role of listener expertise and the choice of speech tasks were noted as influential factors by Pernon et al. (2022) and Mucha et al. (2017).

The limitations of conventional feature extraction methods have driven researchers to explore alternative representations and adaptive systems. Recent efforts in dysarthric ASR have investigated state-of-the-art models such as OpenAI's Whisper. Early comparisons by Doyle et al. (1997) revealed that conventional systems struggled under severe dysarthria conditions, though subsequent adaptive techniques, including speaker-dependent approaches as demonstrated by Calvo et al. (2021), significantly improved recognition accuracy. Mengistu and Rudzicz (2011) provided a direct comparison between ASR systems and human listeners, reporting that even with monophone-based systems, there remains a noticeable gap in performance, with human listeners achieving superior recognition accuracy. Advances in acoustic analysis, leveraging tools such as Praat and MATLAB, have further contributed to refining classification techniques [14], and fine-tuning on dysarthric speech databases, as explored by Sun et al. (2023), has been shown to markedly reduce word error rates.

An emerging alternative to fixed-resolution spectrograms is the Hilbert Spectrum, derived from the Hilbert–Huang Transform. This approach offers an adaptive, instantaneous frequency analysis that is particularly suited to capturing the transient and non-stationary characteristics of dysarthric speech. Agarwal and Kumar (2022) employed the Hilbert Transform to successfully recognize imagined word pairs from EEG signals, achieving high accuracy, while Liu et al. (2020) applied the Hilbert–Huang Transform to extract energy-frequency features for depression recognition in speech. Additional applications, such as EEG classification of imagined syllable rhythm (Deng et al., 2010) and temporal envelope analysis in Mandarin sentence recognition (Guo et al., 2017), have further illustrated the Hilbert Spectrum's ability to capture subtle spectral variations. In clinical contexts, De Boer et al. (2023) demonstrated that Hilbert-derived features, when integrated into machine learning frameworks, could effectively diagnose schizophrenia-spectrum disorders, and Wang et al. (2020) reported remarkable classification performance in sEMG-based speech recognition. These studies collectively suggest that the Hilbert Spectrum not only serves as a viable alternative to conventional spectrogram-based methods but also provides enhanced flexibility in capturing dynamic speech features.

Recent advancements have also pushed the boundaries toward end-to-end systems tailored specifically for dysarthric speech. Traditional ASR systems, typically trained on unimpaired speech, often yield high word error rates when confronted with dysarthria. Data augmentation techniques, including the synthesis of dysarthric speech via multi-speaker text-to-speech systems, have been shown to mitigate these challenges (Soleymanpour et al.,2024). Moreover, innovative training strategies, such as staged knowledge distillation (Lin et al.,2020) and the integration of visual speech features in a unified framework (S. R. Shahamiri., 2021), have demonstrated considerable promise in addressing the intrinsic variability of dysarthric speech. The integration of adaptive representations like the Hilbert Spectrum into these end-to-end architectures represents a promising direction, as it may provide a more nuanced understanding of non-stationary features, ultimately advancing the robustness of dysarthric ASR systems.

***Table 1***: *Summary of Classification Approaches in Dysarthric Speech Processing*

| **Reference** | **Dataset** | **Features Used** | **Model/Architecture** | **Evaluation Metrics** | **Performance/Results** |
|---|---|---|---|---|---|
| **Zhang et al.,2022 [1]** | 60 WD patients, 60 HC | DMFCC | Custom Classifier | Accuracy | 86.10% |
| **Lauraitis et al., 2020 [2]** | 339 samples from 15 subjects | Cepstrum, Auditory Spectrogram, Wavelet Scattering | BiLSTM, WST-SVM | Accuracy | 94.5% (BiLSTM), 96.3% (WST-SVM) |
| **Igarashi et al., 2023[3]** | 29 participants (7M, 22F) | Mel-Spectrogram, MFCC | Machine Learning Classifier | Accuracy, Precision, Recall, F1-score | Mel-Spectrogram: 93.2%, MFCC: 50.2% |
| **Dahmani et al., 2013[4]** | Nemours database | Rhythm Metrics | Feature Space Classification | Classification Rate | 90% classification accuracy |

| | | | | | |
|---|---|---|---|---|---|
| **Yunusova et al., 2011[5]** | 19C, 7PD, 8ALS speakers | Articulatory Kinematics | Linear Discriminant Analysis | Classification Accuracy | 80% accuracy |
| **Fonville et al., 2008[7]** | 100 patient speech samples | Perceptual Analysis | Perceptual Classification | Inter-observer Agreement | 35% correct classification |
| **Bhattacharjee et al., 2023[8]** | 80 ALS, 80 HC | Static and Dynamic Source-Filter | Feature Analysis | Accuracy | 76.66% (/i/), 66.29% (/a/) accuracy |
| **Pernon et al., 2022[9]** | 29 neurotypical, 14 HD, 10 post-stroke | Perceptual Classification | Expert & Student SLPs Classification | Perceptual Classification Accuracy | 72% accuracy |

***Table 2**: Summary of Whisper Fine-Tuning Approaches*

| **Reference** | **Features Used** | **Model/Architecture** | **Evaluation Metrics** | **Performance/Results** |
|---|---|---|---|---|
| **P. Doyle et al., 1997 [11]** | IBM VoiceType, Adaptive Learning | Computerized Speech Recognition | Recognition Accuracy | IBM VoiceType: 50-70% accuracy |
| **I. Calvo et al., 2020[12]** | mPASS Speaker-Dependent ASR | Speaker-Dependent ASR | Word and Sentence Recognition Accuracy | mPASS: 85-90% word accuracy, 80-88% sentence accuracy |
| **K. T. Mengistu & F. Rudzicz, 2011 [13]** | Monophone ASR | Monophone-Based ASR | Word Recognition Accuracy | Monophone ASR: 68.39%, Human Listener: 79.78% |
| **M. G. Thoppil et al., 2017[14]** | Praat & MATLAB Acoustic Analysis | Machine Learning Acoustic Feature Classifier | Classification Accuracy | Praat & MATLAB: 80-90% classification accuracy |
| **M. Sun et al., 2023 [15]** | CDSD Speech Database, Fine-tuned ASR | Whisper-Like ASR with Fine-Tuning | Word Error Rate (WER) | Fine-tuned ASR: WER reduction from 50% to 20% |

A comprehensive summary of classification techniques applied in dysarthric speech processing is presented in Table 1. This table details the datasets utilized, feature extraction methodologies, model architectures, and evaluation metrics employed across various studies. Following this, Table 2 provides an overview of Whisper fine-tuning strategies implemented in different studies, highlighting the datasets, models, evaluation metrics, and performance outcomes. Collectively, these tables illustrate the advancements in dysarthric speech recognition and contextualize the improvements introduced by the current project.

## 3 Theoretical Aspects of the Approach

This section outlines the comprehensive methodology adopted for the proposed dysarthric speech recognition system. The workflow integrates advanced techniques across various stages, including data collection, preprocessing, feature extraction, classification, and synthesis which involves both text transcription using our fine-tuned model and voice generation through gTTS. The following sections will delve into the theoretical aspects underpinning this work.

**Time-Frequency Feature Extraction:**

Spectrograms have long been the de facto standard for representing speech signals in the time-frequency domain. They are generated by applying the STFT to audio signals, which enables the visualization of formants, harmonics, and phoneme transitions. The STFT-based spectrogram is mathematically expressed as:

$$X(m,k) = \sum_{n=0}^{N-1} x(n) \cdot w(n-mH) \cdot e^{-j2\pi kn/N} \tag{1}$$

where x(n) is the input signal, w(n) is a window function, H denotes the hop size, N is the frame length, and X(m,k) represents the time-frequency representation. Spectrograms have been widely applied in speech recognition and classification tasks, proving effective in extracting frequency-domain features crucial for phoneme recognition and acoustic modeling Li, Pi, & Xiao, (2018). However, their reliance on fixed window lengths can introduce information loss, particularly when attempting to capture the nonlinear distortions present in dysarthric speech, limiting their ability to preserve dynamic speech variations accurately.

To address these limitations, the proposed Hilbert Spectrum is introduced as a novel feature representation. Unlike traditional spectrograms, the Hilbert Spectrum offers an adaptive, nonlinear representation that captures instantaneous amplitude and frequency variations, making it well-suited for modeling the dynamic and non-stationary characteristics of dysarthric speech. The Hilbert Spectrum is derived using the Hilbert–Huang Transform. Initially, the analytic signal $z(t)$ is computed as:

$$z(t) = x(t) + j \cdot H\{x(t)\} \tag{2}$$

where $Hx(t)$ denotes the Hilbert Transform of $x(t)$, and $j$ is the imaginary unit. The Hilbert Transform is given by:

$$H\{x(t)\} = \frac{1}{\pi} \text{P.V.} \int_{-\infty}^{\infty} \frac{x(\tau)}{t-\tau} d \tag{3}$$

Here, P.V. denotes the Cauchy principal value of the integral. The analytic signal $z(t)$ allows for the computation of the instantaneous amplitude A(t)A(t)A(t) and instantaneous phase $\phi(t)$ as follows:

$$A(t) = |z(t)| \tag{4}$$

$$\phi(t) = \arg\big(z(t)\big) \tag{5}$$

To obtain the instantaneous frequency f(t)f(t)f(t), the instantaneous phase φ(t)\varphi(t)φ(t) is first unwrapped to prevent discontinuities:

$$\phi_{\text{unwrapped}}(t) = \text{unwrap}\big(\phi(t)\big) \tag{6}$$

The instantaneous frequency is then derived from the unwrapped phase:

$$f(t) = \frac{1}{2\pi}\frac{d\phi_{\text{unwrapped}}(t)}{dt} \tag{7}$$

This adaptive representation is particularly advantageous for dysarthric speech, where rapid articulatory impairments and dynamic phoneme distortions are present (Kumar et al., 2022; Vazhenina & Markov, 2020).

While both spectrograms and the Hilbert Spectrum convert raw audio signals into two-dimensional representations, these transformations may cause specific forms of data loss critical for dysarthric speech analysis. Spectrograms, due to their fixed time-frequency resolution determined by the STFT, can miss transient speech features like rapid formant transitions, brief consonant bursts, and subtle perturbations in harmonic frequencies essential for accurate dysarthria characterization. Similarly, the Hilbert Spectrum, despite its adaptive ability to depict instantaneous frequency and amplitude via Empirical Mode Decomposition (EMD), can omit intricate temporal nuances such as brief pauses, irregular amplitude modulation patterns, or minor frequency fluctuations typical in dysarthric speech. Consequently, utilizing raw audio waveforms directly as inputs to self-supervised learning models, becomes advantageous. This ensures the preservation of subtle phonetic variations, rhythm irregularities, and transient speech dynamics, enabling models to learn nuanced representations directly from the original signal without information loss due to preprocessing transformations (Hsu et al., 2021). Figures 1, 2, and 3 illustrate representative visualizations of the three feature types for dysarthric speech.

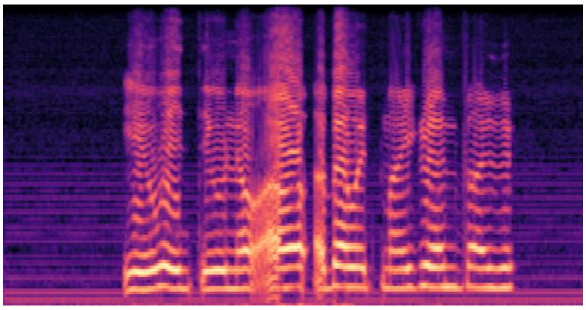


***Fig 1.*** *Spectrogram Dysarthric speech*

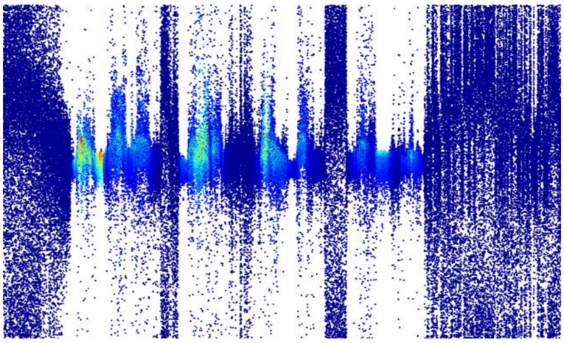


***Fig 2.*** *Hilbert Spectrum Dysarthric speech*

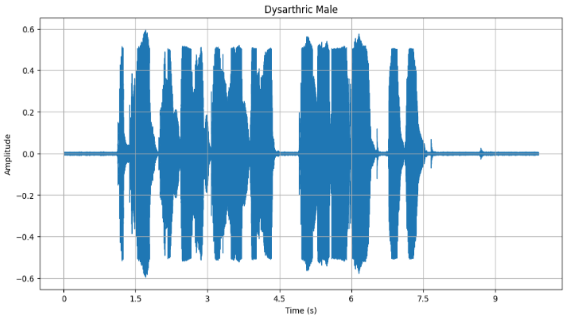


***Fig 3** Raw Audio for Dysarthric speech*

**Model Selection :**

Selecting an appropriate model is critical, as each feature representation captures unique aspects of speech impairments. The choice of model must align with the properties of the input data, ensuring optimal learning of local, global, and temporal dependencies within dysarthric speech patterns.

So the ViT model was chosen as the first model for dysarthric speech classification due to its ability to capture global and temporal dependencies in speech representations. ViT's self-attention mechanism enables it to process spectrograms and the Hilbert Spectrum by identifying long-range relationships between time-frequency components, which is crucial for detecting subtle speech impairments (Gong et al., 2021). Unlike CNNs, ViT treats image patches as sequential tokens, allowing a context-aware analysis of dysarthric speech patterns (Baade et al., 2022). This tokenization approach enhances its ability to detect articulatory distortions that span across broad spectral regions. Additionally, ViT's strong generalization across variable speech patterns and efficient parameterization (~2.09M parameters) make it a highly effective candidate for dysarthric speech classification (Atito et al., 2022).

While ViT models global dependencies, CNNs specialize in extracting local speech features, making them highly effective for detecting fine-grained spectral variations in spectrograms and Hilbert Spectrum. CNNs leverage spatial hierarchies in spectrograms, enabling the detection of localized articulatory distortions characteristic of dysarthric speech (Marma et al., 2023). Unlike ViT, CNNs require less training data, making them effective in small dysarthric speech datasets (Musaev et al., 2019). Additionally, CNNs are computationally efficient, making them preferable for real-time processing and deployment on edge computing devices (Dayal et al., 2022). The custom CNN (3,304,769 parameters) used in this study provides a balanced trade-off between computational cost and performance, serving as a strong comparative baseline against ViT.

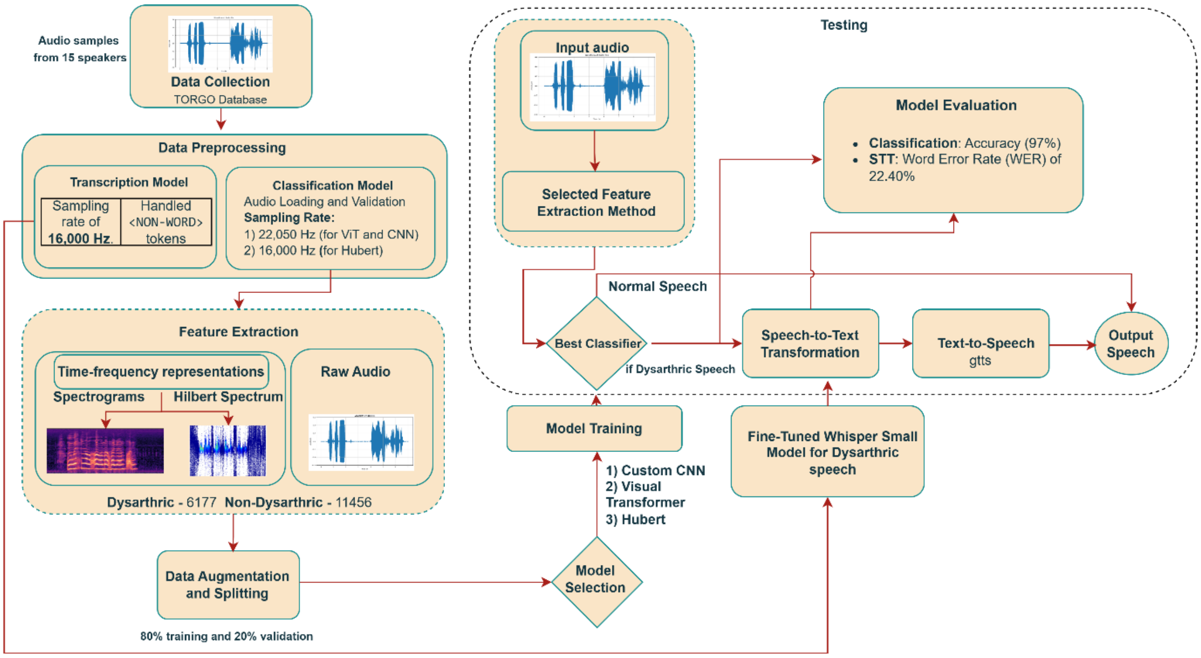
Since both ViT and CNN rely on spectrogram-based representations, which may lead to information loss, HuBERT-Large-LS960-FT, a self-supervised model, was introduced to process raw waveforms directly. HuBERT learns speech representations from unlabeled audio by predicting hidden unit assignments, making it effective for capturing phonetic and prosodic features of dysarthric speech (Hsu et al., 2021). Unlike spectrogram-based models, HuBERT processes waveforms without predefined acoustic transformations, preserving full spectral and temporal information (Dimitriadis et al., 2023). Additionally, HuBERT has been pretrained on 960 hours of speech, allowing it to generalize well even with limited labeled dysarthric speech data, reducing the need for extensive fine-tuning (Yadav et al., 2024).

For speech transcription, Whisper-Small, an ASR model optimized for handling dysarthric speech variability, was selected. Dysarthric speech presents irregular articulation, inconsistent phoneme production, and variable speech rates, making transcription particularly challenging. Whisper-Small, trained on a large and diverse multilingual dataset, generalizes well across impaired speech, accents, and noisy environments, outperforming models trained on more uniform datasets (V. R et al., 2024). Unlike models like DeepSpeech, which rely on hand-engineered spectrogram features, Whisper's end-to-end transformer-based architecture enables it to learn acoustic and linguistic patterns jointly, improving transcription accuracy (Leung et al., 2024).

## 4 Proposed Approach:

Key components of the pipeline include the development of a CNN-based classifier for distinguishing dysarthric and non-dysarthric speech, the fine-tuning of the Whisper architecture for speech-to-text transformation, and the utilization of gTTS for text-to-speech synthesis. The pipeline is evaluated using metrics such as classification accuracy and Word Error Rate (WER).

**Fig 4** provides an overview of the workflow, depicting the sequential steps involved from data preprocessing to final output. The following subsections delve into each stage of the methodology, starting with a detailed description of the dataset, followed by feature extraction techniques, and a discussion of the models and architectures employed in this study.



Audio samples from 15 speakers
Data Collection
TORGO Database
Data Preprocessing
Transcription Model
Sampling rate of 16,000 Hz.
Handled <NON-WORD> tokens
Classification Model
Audio Loading and Validation
Sampling Rate:
1) 22,050 Hz (for ViT and CNN)
2) 16,000 Hz (for Hubert)
Feature Extraction
Time-frequency representations
Spectrograms
Hilbert Spectrum
Raw Audio
Dysarthric - 6177 Non-Dysarthric - 11456
Data Augmentation and Splitting
80% training and 20% validation
Model Selection
1) Custom CNN
2) Visual Transformer
3) Hubert
Model Training
Fine-Tuned Whisper Small Model for Dysarthric speech
Testing
Input audio
Selected Feature Extraction Method
Best Classifier
Normal Speech
if Dysarthric Speech
Speech-to-Text Transformation
Text-to-Speech gtts
Output Speech
Model Evaluation
Classification: Accuracy (97%)
STT: Word Error Rate (WER) of 22.40%

***Fig.4** Proposed Methodology*

### Dataset and Preprocessing

The TORGO dataset was selected for this study due to its well-structured nature and widespread use in dysarthric speech research. Developed by the University of Toronto, this corpus comprises approximately 23 hours of recorded speech from 15 speakers, including eight individuals with speech impairments resulting from Cerebral Palsy (CP) or Amyotrophic Lateral Sclerosis (ALS) and seven control speakers without speech disorders. With a total of 16,552 utterances encompassing a variety of speech tasks—ranging from non-words and short words to restricted and unrestricted sentences—the TORGO dataset offers a comprehensive representation of dysarthric speech characteristics, thereby facilitating robust model training and ensuring that the developed models can generalize across diverse speech conditions.

Prior to model training, the audio recordings from the TORGO dataset underwent extensive preprocessing to ensure consistency and reliability. Given that the recordings were made in various environments and with different hardware setups, standardization was essential to minimize inconsistencies. All audio files were initially resampled to 22,050 Hz to achieve a uniform dataset, and then further resampled to 16 kHz for tasks involving HuBERT and Whisper-based transcription to match their respective pretraining configurations. In addition, corrupted or invalid audio files were systematically identified and removed, and background noise reduction techniques were applied when necessary to enhance the signal-to-noise ratio, thereby improving the extraction of phonetic and acoustic features.

For transcription tasks, the text transcripts associated with the audio recordings were meticulously cleaned. Non-verbal annotations, such as phonetic representations, were either removed or replaced with standardized tokens, and punctuation, capitalization, and numerical formatting were normalized to ensure uniformity across the dataset. The processed dataset was then managed using the Hugging Face Datasets library, which facilitated efficient loading, transformation, and integration into the subsequent training pipelines.

The dataset was partitioned into training and validation sets with an 80:20 split for classification tasks involving ViT ,CNN and Hubert, ensuring a sufficient amount of data for both model learning and performance evaluation. For Whisper-based transcription tasks, an 80:10:10 split was employed for training, validation, and testing to maintain an optimal balance between learning and evaluation. Additionally, tokenization for ASR models was applied using Whisper's built-in tokenizer, which converted text transcripts into numerical token sequences while accommodating non-standard symbols and phonetic markers as special tokens.

**Model Selection for Classification: ViT vs. CNN vs. Hubert**

Selecting an appropriate model is critical, as each feature representation captures unique aspects of speech impairments. In this study, three distinct architectures—ViT, CNNs, and HuBERT—were evaluated to determine which model best leverages the different representations of dysarthric speech, such as spectrograms, Hilbert Spectrum images, and raw audio waveforms.

**Vision Transformer (ViT) Architecture**

The ViT model was selected for its ability to capture global and temporal dependencies through self-attention mechanisms. In our implementation, each 224×224 input image is divided into non-overlapping 16×16 patches, which are then flattened and linearly embedded into a 128-dimensional vector. These tokens, augmented with positional embeddings, are processed through six transformer layers—each comprising eight attention heads and a multi-layer perceptron (MLP) of dimension 256. The core self-attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{8}$$

where Q, K, and V are the query, key, and value matrices, respectively, and $d_k$ is the key dimension. This configuration allows ViT to effectively capture long-range dependencies among patches, thus modeling subtle, distributed features in dysarthric speech. Our experiments revealed that when using spectrograms, ViT achieved reasonable performance; however, its effectiveness diminished notably when using Hilbert Spectrum images, highlighting an architectural sensitivity to the type of feature representation.

**Proposed Convolutional Neural Network (CNN) Architecture**

In contrast, CNNs excel at learning localized features through convolutional operations. Our custom CNN architecture, designed for dysarthric speech classification, comprises three convolutional layers with increasing filter counts (32, 64, and 128), each followed by ReLU activation and max-pooling layers. The convolution operation is mathematically represented as:

$$\text{Conv}(X) = \sigma(W * X + b) \tag{9}$$

where W denotes the convolutional kernel, X is the input feature map, b is the bias term, and $\sigma$ is the ReLU activation function. Max-pooling further reduces spatial dimensions via:

$$\text{MaxPool}(X) = \max_{i,j \in \text{pool\_size}} X_{i,j} \tag{10}$$

The proposed model architecture for classifying dysarthric and non-dysarthric speech is represented in Fig. 5. This architecture is particularly effective in capturing fine-grained spectral variations inherent in both spectrogram and Hilbert Spectrum representations. Our CNN experiments demonstrated that while spectrogram-based CNNs achieved the highest accuracy, those using the Hilbert Spectrum still delivered competitive results, suggesting that adaptive representations can capture important transient features in dysarthric speech.

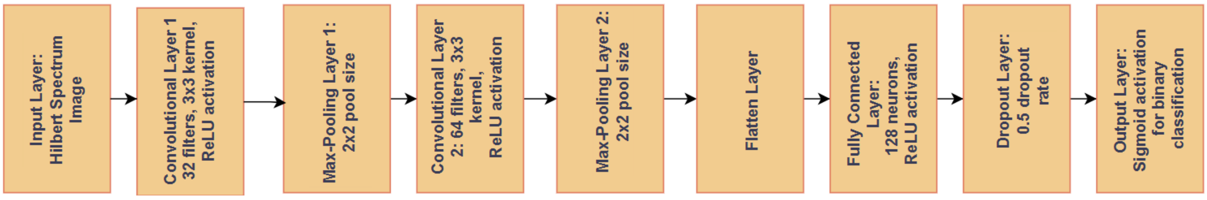

Input Layer: Hilbert Spectrum Image
Convolutional Layer 1 32 filters, 3x3 kernel, ReLU activation
Max-Pooling Layer 1: 2x2 pool size
Convolutional Layer 2: 64 filters, 3x3 kernel, ReLU activation
Max-Pooling Layer 2: 2x2 pool size
Flatten Layer
Fully Connected Layer: 128 neurons, ReLU activation
Dropout Layer: 0.5 dropout rate
Output Layer: Sigmoid activation for binary classification

***Fig 5*** *Proposed CNN Architecture*

## HuBERT Model Fine-Tuning

Transitioning from CNNs, HuBERT (Hidden-Unit BERT), a self-supervised learning model that processes raw audio waveforms directly, was further evaluated. HuBERT leverages extensive pre-training on unlabeled speech to learn robust acoustic representations. Its architecture employs a multi-layer Transformer encoder, where the self-attention mechanism (as defined earlier for ViT) is used to extract contextual features from the audio. Unlike spectrogram-based models, HuBERT avoids the potential data loss associated with transforming raw audio into two-dimensional representations, thereby preserving subtle phonetic details, minute rhythmic variations, and transient acoustic cues essential for characterizing dysarthric speech.

The fine-tuning of HuBERT for our classification task involved adapting a pre-trained HuBERT-Large-LS960-FT model by adding a sequence classification head. The audio was preprocessed by resampling to 16 kHz and normalizing to a fixed duration, ensuring compatibility with the model's training configuration. During fine-tuning, the model was optimized using the Adam optimizer with a learning rate of $2 \times 10^{-5}$ over ten epochs with mixed precision training. This process enabled HuBERT to capture both long-term dependencies and fine temporal details without relying on hand-engineered features, making it particularly adept at distinguishing between dysarthric and non-dysarthric speech.

Empirical results demonstrated that HuBERT achieved a validation accuracy comparable to the CNN with Hilbert Spectrum features, underscoring its effectiveness in capturing the nuanced acoustic patterns inherent in dysarthric speech. Its capacity to leverage raw waveform inputs directly allows HuBERT to maintain full spectral and temporal information, which is often compromised in spectrogram-based transformations. This advantage positions HuBERT as a compelling alternative, especially in scenarios where preserving the integrity of the original signal is paramount.

## Model Architecture for Transcription

The proposed model architecture for fine-tuning the Whisper-small model for dysarthric speech transcription is illustrated in Fig 6. This architecture builds upon the original encoder-decoder transformer framework of Whisper, incorporating key adaptations to enhance its performance for recognizing and transcribing dysarthric speech patterns.

The architecture begins with an input layer that processes audio data resampled to a consistent 16kHz sampling rate, ensuring compatibility with the Whisper-small model. The feature extraction module within the Whisper processor transforms this audio input into a Spectrogram -like representation, emphasizing key spectral features vital for transcription tasks.

The encoder layer, consisting of 12 transformer blocks, utilizes multi-head self-attention mechanisms to analyze temporal and spectral dependencies in the input. These mechanisms enable the model to capture both short-term acoustic variations and long-term speech patterns, which are often irregular in dysarthric speech. The encoder's positional encoding module further enriches the input by embedding sequential information, crucial for maintaining the temporal structure of the speech data.

The decoder layer, also composed of 12 transformer blocks, integrates cross-attention mechanisms to align the encoded features with the tokenized text output. This alignment facilitates accurate transcription of speech by effectively mapping acoustic patterns to linguistic representations. The architecture's token embedding layer is extended to include special tokens, such as <noise>, which are introduced during fine-tuning to handle non-verbal sounds and annotations present in the TORGO dataset. These customizations ensure that the model can appropriately manage speech irregularities without misclassifying them as verbal content.

Finally, the output layer produces the transcriptions as tokenized sequences, leveraging the enhanced decoder and token embedding layers to achieve improved transcription accuracy. The model's fine-tuned parameters, adjusted for dysarthric speech, ensure robustness in handling the phonetic and acoustic variations inherent in the dataset.

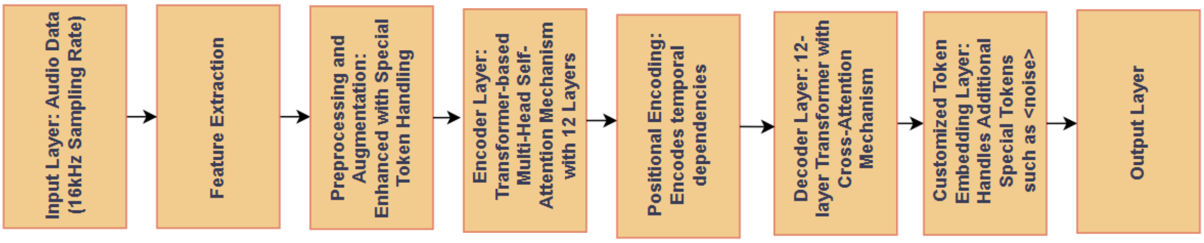

Input Layer: Audio Data (16kHz Sampling Rate)
Feature Extraction
Preprocessing and Augmentation: Enhanced with Special Token Handling
Encoder Layer: Transformer-based Multi-Head Self-Attention Mechanism with 12 Layers
Positional Encoding: Encodes temporal dependencies
Decoder Layer: 12-layer Transformer with Cross-Attention Mechanism
Customized Token Embedding Layer: Handles Additional Special Tokens such as <noise>
Output Layer

***Fig. 6*** *Fine-tuned Whisper*

## End-to-End Framework

The End-to-End Framework as depicted in Fig. 1, begins with the TORGO dataset, which comprises over 16,500 audio samples from dysarthric and non-dysarthric speakers. Initially , audio is first passed through the same feature extraction procedure (spectrogram, Hilbert spectrum, or raw waveform processing) used in training. The resulting features are then fed to the best-performing classifier (e.g., ViT, CNN, or HuBERT), which determines whether the speech is normal or dysarthric. If the classifier detects normal speech, the pipeline simply outputs the classification result. If dysarthric speech is detected, the audio is sent to a fine-tuned Whisper-Small model for transcription, converting the impaired speech into text. This text will then be passed through gTTS, generating an enhanced or more intelligible audio output. Model evaluation focuses on classification accuracy (how well the system differentiates dysarthric from non-dysarthric speech) and the Word Error Rate (WER) of the transcription step.

## 5.1. Performance evaluation metrics

Evaluating the performance of both the dysarthric speech classification model and the fine-tuned Whisper model for transcription necessitates the application of specific metrics tailored to each task.

**Classification Model Evaluation:**

The effectiveness of the dysarthric speech classification model is assessed using several key metrics:

- **Accuracy** quantifies the overall correctness of the model by calculating the ratio of correctly predicted instances to the total number of predictions. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

where TP represents true positives, TN true negatives, FP false positives, and FN false negatives.

- **Precision** measures the model's ability to correctly identify dysarthric speech among all instances it predicts as dysarthric. It is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

- **Recall**, or sensitivity, evaluates the model's capability to detect all actual cases of dysarthric speech. It is expressed as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{13}$$

- **F1-Score** provides a harmonic mean of precision and recall, offering a balanced measure of the model's performance, especially in scenarios with class imbalance. The F1-Score is given by:

$$F1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

- **Confusion Matrix** offers a comprehensive breakdown of the model's predictions, displaying the counts of true positives, true negatives, false positives, and false negatives. This matrix facilitates the identification of specific areas where the model may be underperforming, thereby guiding targeted improvements.

**Whisper Transcription Model Evaluation:**

For the fine-tuned Whisper model, transcription accuracy is primarily evaluated using:

- **Word Error Rate (WER)**, a standard metric in speech recognition, which calculates the proportion of errors in the transcribed text compared to a reference transcript. It encompasses substitutions (S), deletions (D), and insertions (I) relative to the total number of words in the reference (N). WER is computed as:

$$\text{WER} = \frac{S + D + I}{N} \tag{15}$$

In addition to these quantitative metrics, conducting a qualitative assessment of the transcriptions is essential. This involves reviewing the transcribed text in conjunction with the original audio to identify systematic errors or patterns of misrecognition. Such an analysis can uncover nuances in dysarthric speech that quantitative metrics might overlook, thereby offering deeper insights into the model's performance and areas necessitating refinement.

By employing these evaluation metrics, the performance of both the classification and transcription models can be rigorously assessed, ensuring they meet the desired standards for accuracy and reliability in handling dysarthric speech.

## Results and Discussion

This section presents the comparative results of different feature extraction techniques within the end-to-end pipeline. **Fig 7** illustrates the performance of the ViT model when trained on spectrogram inputs, achieving a validation accuracy of 90.53%. In contrast, **Fig 8** presents the corresponding results for the ViT model using Hilbert Spectrum images, which attained a validation accuracy of 80.61%. Although the spectrogram-based model demonstrates higher overall accuracy, the Hilbert Spectrum variant exhibits competitive performance in dysarthric speech classification. In particular, the Hilbert-based model achieves a dysarthric speech precision of 0.91, which is only marginally lower than the 0.95 precision obtained with spectrogram inputs. However, as depicted in Fig 8, its recall for dysarthric speech is 0.65 compared to 0.77 for the spectrogram-based approach, leading to an F1-score of 0.76 rather than 0.85. For non-dysarthric speech, the Hilbert model shows a precision of 0.75 and a recall of 0.95, yielding an F1-score of 0.84, while the spectrogram-based model, as shown in Fig 7, attains a precision of 0.89, recall of 0.98, and F1-score of 0.93. These findings indicate that although the Hilbert Spectrum representation results in slightly lower recall, its competitive precision and robust performance in classifying non-dysarthric speech underscore its potential for further optimization.

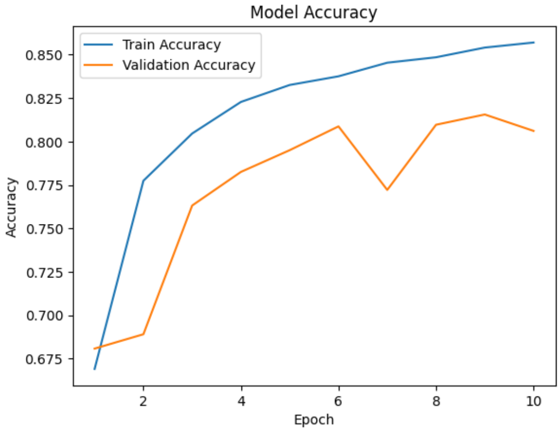


**Fig 7.** *ViT + Hilbert Accuracy*

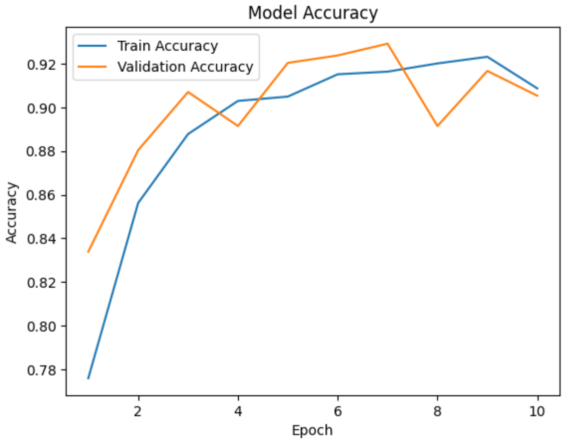


**Fig 8.** *ViT + Spectrogram Accuracy*

The convergence behavior of the models is detailed in Fig 9 and 10, which show the training and validation loss curves for the ViT models. The spectrogram-based ViT model converges to a final training loss of 0.2318 and a validation loss of 0.2399, reflecting effective generalization. In comparison, the Hilbert-based model converges at higher loss values, with a training loss of 0.3288 and a validation loss of 0.4169, as evident in Fig 10. These higher loss values and the observed fluctuations in the validation curve for the Hilbert model suggest that the current architecture faces challenges in consistently capturing the detailed temporal and spectral features inherent in dysarthric speech. The results imply that further refinement, such as incorporating advanced normalization techniques or modifying the attention mechanism, could better exploit the rich time-frequency information provided by the Hilbert Spectrum.

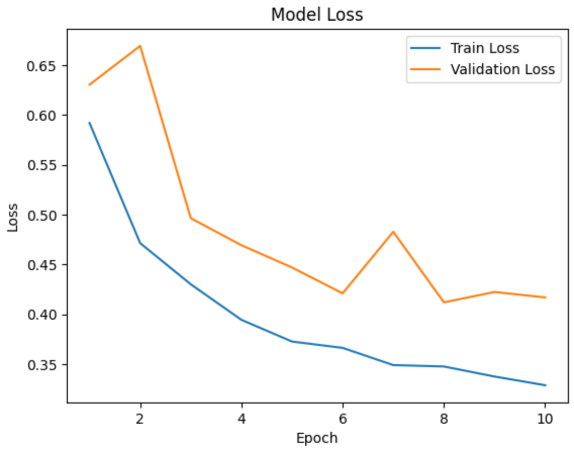


**Fig 9.** *ViT + Hilbert loss*

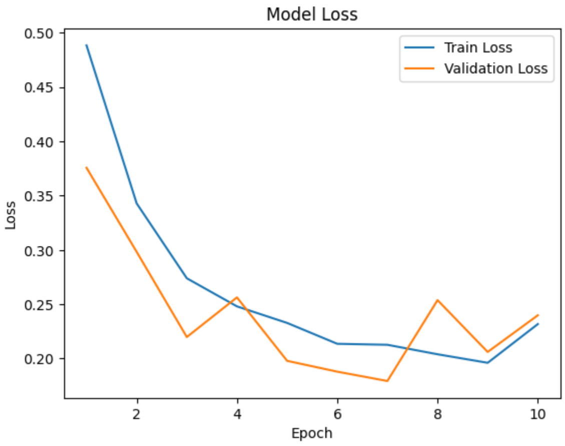


**Fig 10.** *ViT + Spectrogram loss*

**Fig 11 and 12** display the confusion matrices for the ViT models. The spectrogram-based model correctly classified 952 dysarthric instances while misclassifying 284 instances as non-dysarthric, as shown in Fig 12. Conversely, Fig 11 indicates that the Hilbert Spectrum model correctly identified 1270 dysarthric cases, albeit misclassifying 698 as non-dysarthric. While the spectrogram approach exhibits superior sensitivity, the ability of the Hilbert model to correctly classify a substantial number of dysarthric cases highlights its promise. These outcomes suggest that with targeted improvements to enhance recall, the Hilbert Spectrum representation could achieve a more balanced performance.

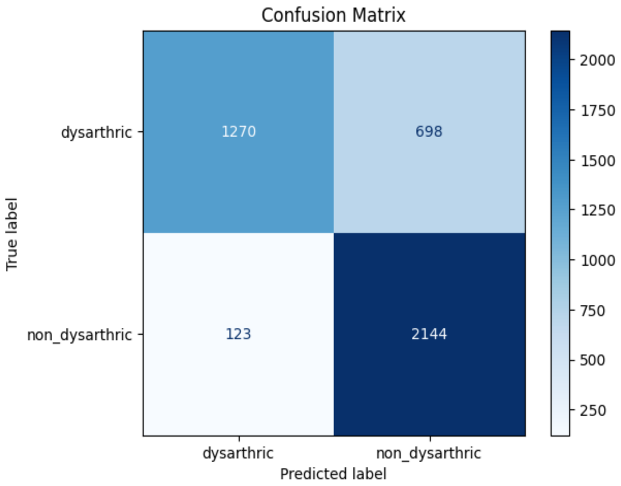


**Fig 11.** *ViT + Hilbert Confusion Matrix*

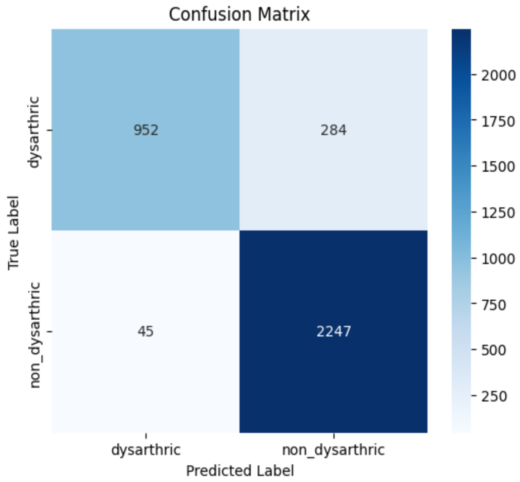


**Fig 12.** *ViT + Spectrogram Confusion Matrix*

The evaluation of CNN models establishes a robust benchmark for comparison. As demonstrated in **Fig 14**, the CNN model trained on spectrogram inputs achieved the highest validation accuracy of 97%, while Fig 13 shows that the CNN model utilizing Hilbert Spectrum images attained a validation accuracy of 95%. Despite the slight reduction in performance, the CNN-Hilbert model achieves an F1-score of 0.93 for dysarthric speech, compared to 0.97 for the CNN-Spectrogram model. The loss curves depicted in **Fig 15 and 16** indicate strong convergence for both CNN models, with final validation losses of 0.2052 and 0.2256 for the spectrogram and Hilbert Spectrum models, respectively. Moreover, the confusion matrices in **Fig 17 and 18** further confirm the competitive performance of the CNN-Hilbert approach, with only a marginal increase in misclassification rates relative to the spectrogram-based model.

(Shabber & Sumesh, 2024) employed an Amplitude-Frequency Modulation (AFM) signal model combined with CNNs, attaining a superior accuracy of 97.8% across TORGO, UA-Speech, and

Parkinson datasets. This suggests that leveraging alternative signal processing techniques, such as AFM, could further enhance Hilbert Spectrum-based models. Similarly, (Mahum et al., 2024) introduced Tran-DSR, a hybrid model that integrates transformers with an ensemble of CNNs, achieving an exceptional accuracy of 99.18%. The significantly higher performance suggests that CNN-Hilbert models might benefit from transformer-based feature enhancement.

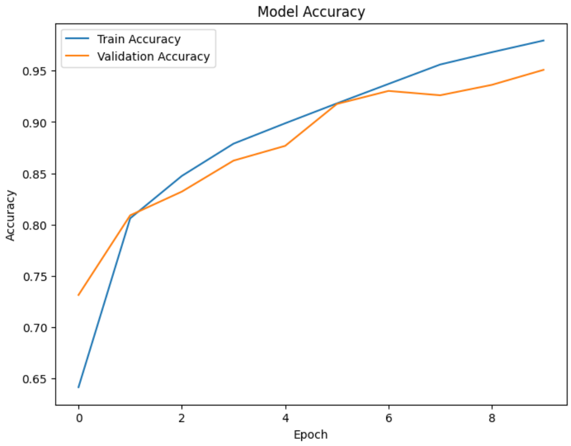


**Fig 13.** *CNN + Hilbert Accuracy*

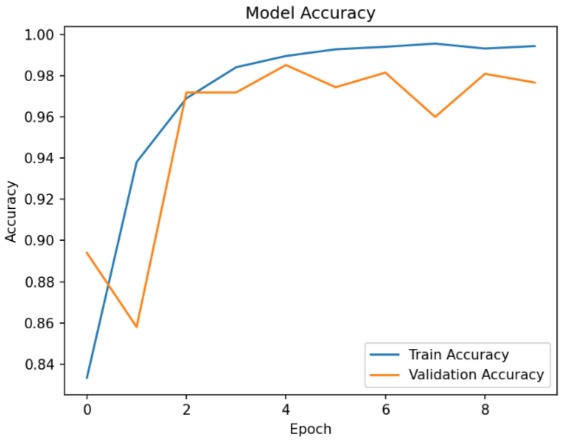


**Fig 14.** *CNN + Spectrogram Accuracy*

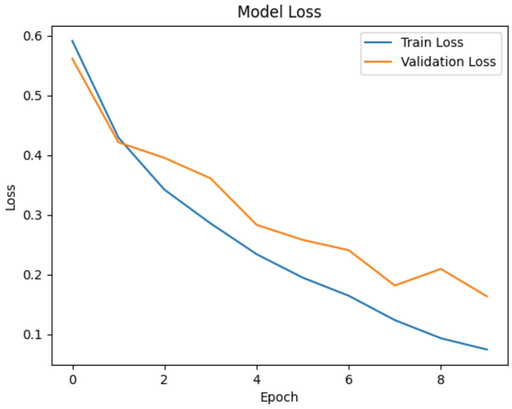


**Fig 15.** *CNN + Hilbert loss*

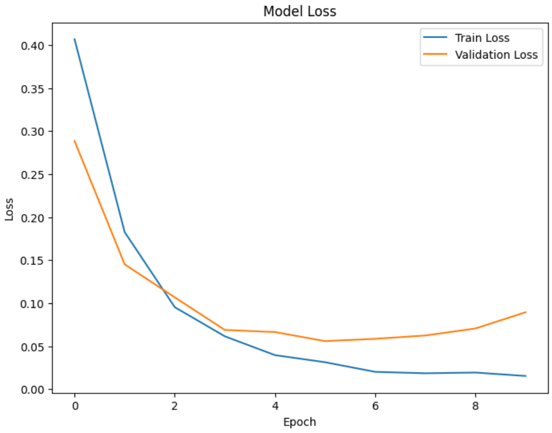


**Fig 16.** *CNN + Spectrogram loss*

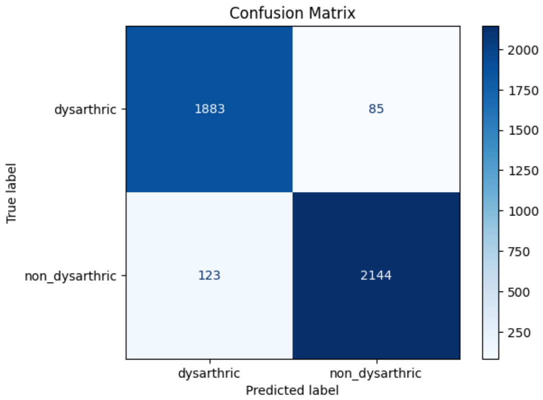


**Fig 17.** *CNN + Hilbert Confusion Matrix*

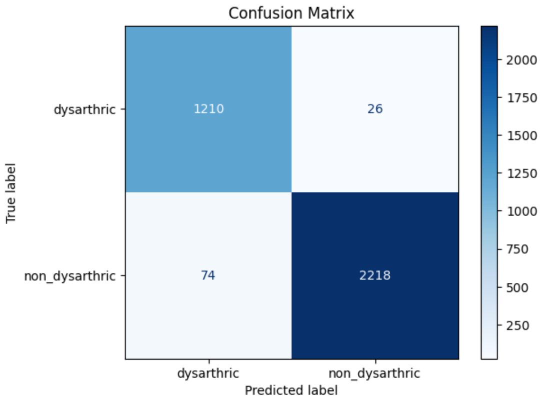


**Fig 18.** *CNN + Spectrogram Confusion Matrix*

The HuBERT model, as illustrated in **Fig 19**, achieved a validation accuracy of 97.27%, highlighting the effectiveness of self-supervised learning in dysarthric speech classification. The corresponding training and validation accuracy curves, along with the loss curves shown in **Fig 20**, indicate stable convergence with a final training loss of 0.091 and a validation loss of 0.131. The confusion matrix presented in **Fig 21** demonstrates balanced performance, with dysarthric speech being classified with a precision of 0.92 and a recall of 0.93, and non-dysarthric speech with a precision of 0.98 and a recall of 0.96.

(Javanmardi et al. 2024) also explored the effectiveness of HuBERT and other self-supervised models like Wav2Vec2, finding that HuBERT achieved superior classification performance of 95.50% accuracy. Our findings align with these prior works, as summarized in **Table 3**, reinforce the viability of HuBERT and the potential of Hilbert Spectrum features when paired with optimized architectures.

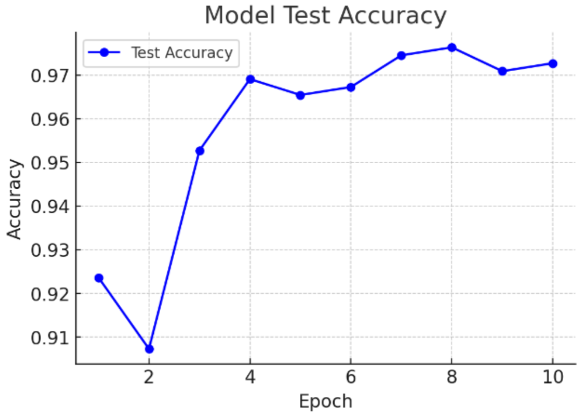


***Fig 19.*** *HuBERT Accuracy*

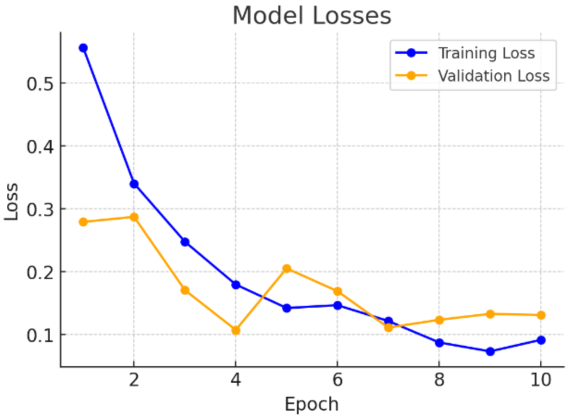


***Fig 20.*** *HuBERT loss*

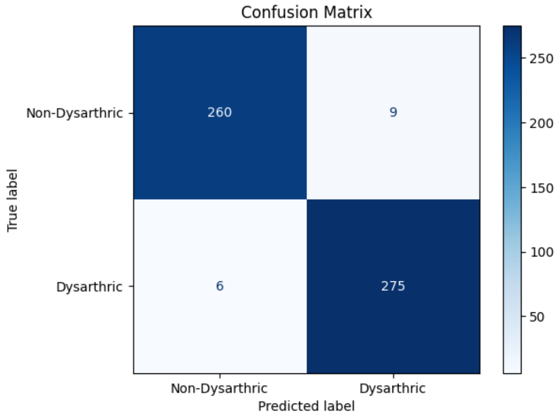


***Fig 21.*** *HUBERT Confusion Matrix*

***Table 3****. Comparative Evaluation of Dysarthric Speech Models Across Recent Studies*

| **Study / Model** | **Architecture / Approach** | **Dataset(s) Used** | **Accuracy (%)** | **Notable Features or Findings** |
|---|---|---|---|---|
| **Shabber & Sumesh, 2024 [41]** | CNN + Amplitude-Frequency Modulation (AFM) | TORGO, UA-Speech, Parkinson | 97.8 | AFM signal modeling enhanced CNN accuracy; |
| **Mahum et al., 2024 [42]** | Tran-DSR (Transformers + CNN ensemble) | Nemours database | 99.18 | Hybrid transformer-CNN model significantly outperforms others |

| | | | | |
|---|---|---|---|---|
| **Javanmardi et al., 2024 [43]** | HuBERT vs. Wav2Vec2 vs. Acoustic Features | UA-Speech, TORGO | 95.50 | HuBERT consistently outperformed Wav2Vec2 and conventional features in both detection and severity classification |
| **This Study** | **CNN + Hilbert Spectrum** | **TORGO** | **95** | **Slightly lower performance than spectrogram; competitive precision; viable alternative** |

Table 4 provides a comprehensive comparison of all models carried out in this work. The CNN-Spectrogram model leads with a 97% validation accuracy, while the CNN-Hilbert model follows closely with 95%. The slight reduction in performance for the Hilbert-based models, observed in both the ViT and CNN frameworks, is offset by their competitive precision and substantial correct classifications. This demonstrates that the Hilbert Spectrum representation, when paired with appropriate neural architectures, has significant potential for dysarthric speech classification. Although the ViT-Hilbert model exhibits lower recall compared to its spectrogram-based counterpart, its overall performance, combined with the robust results of the CNN-Hilbert model, indicates that the Hilbert Spectrum approach is a viable alternative. With further refinements in feature extraction and model architecture, the Hilbert Spectrum representation could serve as an effective and robust tool for dysarthric speech analysis.

***Table 4**. Comparative Evaluation of all models*

| Model Configuration | Validation Accuracy (%) | Precision (Dysarthric) | Recall (Dysarthric) | F1-Score (Dysarthric) | Precision (Non-Dysarthric) | Recall (Non-Dysarthric) | F1-Score (Non-Dysarthric) |
|---|---|---|---|---|---|---|---|
| **ViT with Spectrograms** | 90.53 | 0.95 | 0.77 | 0.85 | 0.89 | 0.98 | 0.93 |
| **ViT with Hilbert Spectrum** | 81 | 0.91 | 0.65 | 0.76 | 0.75 | 0.95 | 0.84 |
| **CNN with Spectrograms** | 97 | 0.96 | 0.97 | 0.97 | 0.99 | 0.98 | 0.98 |
| **CNN with Hilbert Spectrum** | 95 | 0.93 | 0.94 | 0.93 | 0.98 | 0.96 | 0.97 |
| **HuBERT** | 97.27 | 0.92 | 0.93 | 0.92 | 0.98 | 0.96 | 0.97 |

The fine-tuned Whisper model was evaluated for its transcription performance on dysarthric speech. The model achieved an overall Word Error Rate (WER) of approximately 22.40% on the test set.

## Conclusion

This study presents a concise, end-to-end workflow for dysarthric speech analysis that spans from data preprocessing to final output generation. The pipeline begins with a detailed dataset description, followed by visualizations of audio waveforms and spectrograms, and employs advanced feature extraction techniques. Various models and architectures were then evaluated for classification and transcription. Transformer-based approaches, particularly HuBERT, achieved slightly higher accuracy than conventional CNN models, while the Hilbert Spectrum representation demonstrated competitive performance as an alternative to traditional spectrogram features. Additionally, the fine-tuned Whisper model achieved a Word Error Rate of approximately 22.40%, underscoring its efficacy in transcribing dysarthric speech. Collectively, these results validate the feasibility of a fully integrated system for dysarthric speech processing, paving the way for future refinements aimed at enhanced feature optimization and improved clinical and assistive applications.

## References:

[1] Zhang, Z., Yang, L., Wang, X., & Li, H. (2022). Automated Detection of Wilson's Disease Based on Improved Mel-frequency Cepstral Coefficients with Signal Decomposition. In *INTERSPEECH* (pp. 2143-2147).

[2] Lauraitis, A., Maskeliūnas, R., Damaševičius, R., & Krilavičius, T. (2020). Detection of speech impairments using cepstrum, auditory spectrogram and wavelet time scattering domain features. *IEEE Access*, *8*, 96162-96172.

[3] Igarashi, T., Umeda-Kameyama, Y., Kojima, T., Akishita, M., & Nihei, M. (2023). Questionnaires for the assessment of cognitive function secondary to intake interviews in in-hospital work and development and evaluation of a classification model using acoustic features. *Sensors*, *23*(11), 5346.

[4] Dahmani, H., Selouani, S. A., O'shaughnessy, D., Chetouani, M., & Doghmane, N. (2013). Assessment of dysarthric speech through rhythm metrics. *Journal of King Saud University-Computer and Information Sciences*, *25*(1), 43-49.

[5] Yunusova, Y., Weismer, G. G., & Lindstrom, M. J. (2011). Classifications of vocalic segments from articulatory kinematics: Healthy controls and speakers with dysarthria.

[6] Mendoza Ramos, V., Lowit, A., Van den Steen, L., Kairuz Hernandez-Diaz, H. A., Hernandez-Diaz Huici, M. E., De Bodt, M., & Van Nuffelen, G. (2021). Acoustic identification of sentence accent in speakers with dysarthria: cross-population validation and severity related patterns. *Brain Sciences*, *11*(10), 1344.

[7] Fonville, S., Van Der Worp, H. B., Maat, P., Aldenhoven, M., Algra, A., & Van Gijn, J. (2008). Accuracy and inter-observer variation in the classification of dysarthria from speech recordings. *Journal of Neurology*, *255*, 1545-1548.

[8] T. Bhattacharjee, C. V. Thirumala Kumar, Y. Belur, A. Nalini, R. Yadav and P. K. Ghosh, "Static and Dynamic Source and Filter Cues for Classification of Amyotrophic Lateral Sclerosis Patients and Healthy Subjects," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10094959.

[9] Pernon, M., Assal, F., Kodrasi, I., & Laganaro, M. (2022). Perceptual classification of motor speech disorders: The role of severity, speech task, and listener's expertise. *Journal of Speech, Language, and Hearing Research*, *65*(8), 2727-2747.

[10] Mucha, J., Galaz, Z., Mekyska, J., Kiska, T., Zvoncak, V., Smekal, Z., ... & Alonso-Hernandez, J. B. (2017, July). Identification of hypokinetic dysarthria using acoustic analysis

of poem recitation. In *2017 40th International Conference on Telecommunications and Signal Processing (TSP)* (pp. 739-742). IEEE.

[11] Doyle, P. C., Leeper, H. A., Kotler, A. L., Thomas-Stonell, N., O'Neill, C., Dylke, M. C., & Rolls, K. (1997). Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. *Journal of rehabilitation research and development*, *34*, 309-316.

[12] Calvo, I., Tropea, P., Viganò, M., Scialla, M., Cavalcante, A. B., Grajzer, M., ... & Corbo, M. (2021). Evaluation of an automatic speech recognition platform for dysarthric speech. *Folia Phoniatrica et Logopaedica*, *73*(5), 432-441.

[13]Mengistu, K. T., & Rudzicz, F. (2011). Comparing humans and automatic speech recognition systems in recognizing dysarthric speech. In *Advances in Artificial Intelligence: 24th Canadian Conference on Artificial Intelligence, Canadian AI 2011, St. John's, Canada, May 25-27, 2011. Proceedings 24* (pp. 291-300). Springer Berlin Heidelberg.

[14]Thoppil, M. G., Kumar, C. S., Kumar, A., & Amose, J. (2017). Speech signal analysis and pattern recognition in diagnosis of dysarthria. *Annals of Indian Academy of Neurology*, *20*(4), 352-357.

[15]Sun, M., Gao, M., Kang, X., Wang, S., Du, J., Yao, D., & Wang, S. J. (2023). CDSD: Chinese Dysarthria Speech Database. *arXiv preprint arXiv:2310.15930*.

[16]Agarwal, P., & Kumar, S. (2022). Imagined word pairs recognition from non-invasive brain signals using Hilbert transform. *International Journal of System Assurance Engineering and Management*, *13*(1), 385-394.

[17]Liu, Z., Xu, Y., Ding, Z., & Chen, Q. (2020, December). Time-frequency analysis based on hilbert-huang transform for depression recognition in speech. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1072-1076). IEEE.

[18]Deng, S., Srinivasan, R., Lappas, T., & D'Zmura, M. (2010). EEG classification of imagined syllable rhythm using Hilbert spectrum methods. *Journal of neural engineering*, *7*(4), 046006.

[19]Guo, Y., Sun, Y., Feng, Y., Zhang, Y., & Yin, S. (2017). The relative weight of temporal envelope cues in different frequency regions for Mandarin sentence recognition. *Neural Plasticity*, *2017*(1), 7416727.

[20]De Boer, J. N., Voppel, A. E., Brederoo, S. G., Schnack, H. G., Truong, K. P., Wijnen, F. N. K., & Sommer, I. E. C. (2023). Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-recognition tool. *Psychological medicine*, *53*(4), 1302-1312.

[21]Wang, X., Zhu, M., Cui, H., Yang, Z., Wang, X., Zhang, H., ... & Li, G. (2020, July). The effects of channel number on classification performance for sEMG-based speech recognition. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 3102-3105). IEEE.

[22]Soleymanpour, M., Johnson, M. T., Soleymanpour, R., & Berry, J. (2024). Accurate synthesis of dysarthric speech for ASR data augmentation. *Speech Communication*, *164*, 103112.

[23]Lin, Y., Wang, L., Li, S., Dang, J., & Ding, C. (2020, October). Staged Knowledge Distillation for End-to-End Dysarthric Speech Recognition and Speech Attribute Transcription. In *INTERSPEECH* (pp. 4791-4795).

[24]S. R. Shahamiri, "Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 852-861, 2021, doi: 10.1109/TNSRE.2021.3076778.

[25]Li, Y., Pi, S., Xiao, N. (2019). Speech Recognition Method Based on Spectrogram. In: Deng, K., Yu, Z., Patnaik, S., Wang, J. (eds) Recent Developments in Mechatronics and Intelligent Robotics. ICMIR 2018. Advances in Intelligent Systems and Computing, vol 856. Springer, Cham. https://doi.org/10.1007/978-3-030-00214-5_110

[26]Kumar, A., Solanki, S. S., & Chandra, M. (2022). Hilbert Spectrum based features for speech/music classification. *Serbian Journal of Electrical Engineering*, *19*(2), 239-259.

[27]Vazhenina, D., & Markov, K. (2020). End-to-end noisy speech recognition using Fourier and Hilbert spectrum features. *Electronics*, *9*(7), 1157.

[28]Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, *29*, 3451-3460.

[29]Gong, Y., Lai, C. I., Chung, Y. A., & Glass, J. (2022, June). Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 10, pp. 10699-10709).

[30]Baade, A., Peng, P., & Harwath, D. (2022). Mae-ast: Masked autoencoding audio spectrogram transformer. *arXiv preprint arXiv:2203.16691*.

[31]Atito, S., Awais, M., Wang, W., Plumbley, M. D., & Kittler, J. (2024). Asit: Local-global audio spectrogram vision transformer for event classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[32]Marma, Ž., Punys, M., & Lipnickas, A. (2023, September). Speech Emotion Recognition Using Combined Mel Spectrograms with 2D CNN Models. In *2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)* (Vol. 1, pp. 194-198). IEEE.

[33]Musaev, M., Khujayorov, I., & Ochilov, M. (2019, September). Image approach to speech recognition on CNN. In *Proceedings of the 2019 3rd international symposium on computer science and intelligent control* (pp. 1-6).

[34]Dayal, A., Yeduri, S. R., Koduru, B. H., Jaiswal, R. K., Soumya, J., Srinivas, M. B., ... & Cenkeramaddi, L. R. (2022). Lightweight deep convolutional neural network for background sound classification in speech signals. *The Journal of the Acoustical Society of America*, *151*(4), 2773-2786.

[35]Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, *29*, 3451-3460.

[36]Dimitriadis, A., Pan, S., Sethu, V., & Ahmed, B. (2023). Spatial HuBERT: Self-supervised Spatial Speech Representation Learning for a Single Talker from Multi-channel Audio. *arXiv preprint arXiv:2310.10922*.

[37]Yadav, H., Sitaram, S., & Shah, R. R. (2024). Ms-hubert: Mitigating pre-training and inference mismatch in masked language modelling methods for learning speech representations. *arXiv preprint arXiv:2406.05661*.

[38]Vinotha, R., Hepsiba, D., & Anand, L. V. (2024, July). Leveraging OpenAI Whisper Model to Improve Speech Recognition for Dysarthric Individuals. In *2024 Asia Pacific Conference on Innovation in Technology (APCIT)* (pp. 1-5). IEEE.

[39]Leung, W. Z., Cross, M., Ragni, A., & Goetze, S. (2024). Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis. *arXiv preprint arXiv:2406.08568*.

[40]L. P. Sahu and G. Pradhan, "Significance of Filterbank Structure for Capturing Dysarthric Information through Cepstral Coefficients," *2022 IEEE International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India, 2022, pp. 1-5, doi: 10.1109/SPCOM55316.2022.9840837.

[41]Shabber, S., & Sumesh, E. (2024). AFM signal model for dysarthric speech classification using speech biomarkers. *Frontiers in Human Neuroscience*, 18. https://doi.org/10.3389/fnhum.2024.1346297.

[42]Mahum, R., El-Sherbeeny, A., Alkhaledi, K., & Hassan, H. (2024). Tran-DSR: A hybrid model for dysarthric speech recognition using transformer encoder and ensemble learning. *Applied Acoustics*. https://doi.org/10.1016/j.apacoust.2024.110019.

[43]Javanmardi, F., Kadiri, S., & Alku, P. (2024). Pre-trained models for detection and severity level classification of dysarthria from speech. *Speech Commun.*, 158, 103047. https://doi.org/10.1016/j.specom.2024.103047.