Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 4 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i04.59849

PAPER

Development of Hybrid Feature Based Models for Dysarthric Speech Recognition

Faila Nadhifatul Aryza, Syahroni Hidayat (✉)

Universitas Negeri Semarang, Semarang, Indonesia

syahronihidayat@mail.unnes.ac.id

ABSTRACT

Speech recognition for individuals with dysarthria remains challenging due to unstable acoustic signals, high temporal variability, and frequent articulatory distortions, all of which hinder the ability of acoustic models to consistently capture phonetic patterns. This study aims to identify the most effective feature extraction strategy among three approaches, namely Wav2Vec 2.0, MFCC combined with Wav2Vec 2.0, and Wavelet-MFCC combined with Wav2Vec 2.0, evaluated using the UA-Speech dataset. All models were trained using the Wav2Vec 2.0 Base architecture with a CTC decoding mechanism to map audio signals to character sequences in an end-to-end manner. The experimental results demonstrate that the MFCC-Wav2Vec 2.0 combination yields the best performance, achieving a Word Error Rate (WER) of 0.2990. These findings indicate that combining traditional acoustic features with self-supervised representations yields a more robust speech recognition system for dysarthric speech.

KEYWORDS

dysarthria word recognition, Wav2Vec 2.0, mel-frequency cepstral coefficients (MFCC), wavelet transform, hybrid feature extraction

1 INTRODUCTION

Communication is a fundamental aspect of human life, serving as a primary means for conveying information, expressing thoughts, and facilitating social interaction [1]. One of the most dominant forms of communication is verbal communication through speech. With the rapid advancement of digital technologies, speech signal processing no longer serves merely as a medium for interpersonal communication but has also become an essential component in human-machine interaction systems, such as virtual assistants [2], voice-based search engines [3], and automatic transcription services [4]. This development marks a paradigm shift from human-human communication toward human-machine interaction that is increasingly natural and adaptive to user context.

Aryza, F. N., Hidayat, S. (2026). Development of Hybrid Feature Based Models for Dysarthric Speech Recognition. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(4), pp. 26–42. https://doi.org/10.3991/ijoe.v22i04.59849

Article submitted 2025-11-28. Revision uploaded 2026-01-03. Final acceptance 2026-01-05.

Speech recognition technology, or automatic speech recognition (ASR), has advanced rapidly with the development of machine learning and deep learning techniques [5]. Modern commercial ASR systems are capable of recognizing speech with high accuracy [6], [7], particularly for speakers with typical speech abilities and under controlled acoustic conditions [8]. However, such high performance is not consistently achieved when the systems are used by individuals with speech impairments. One of the most challenging user groups for ASR systems is individuals with dysarthria [9], [10].

Dysarthria is a neurological speech disorder caused by weakness or impaired coordination of the speech muscles, affecting phonation, articulation, and prosody [11]. This condition is commonly observed in individuals with stroke, Parkinson's disease, or Amyotrophic Lateral Sclerosis (ALS), with relatively high prevalence rates [12]–[14]. In Indonesia, the prevalence of stroke reaches 10.9 per 1,000 people, indicating a potentially large number of individuals living with dysarthria [15]. Reduced speech intelligibility in dysarthric speakers makes their speech difficult to understand, both by humans and machines. As a consequence, access to voice-based technologies in daily activities becomes increasingly limited [16]. These conditions highlight the need for ASR systems that are more inclusive and adaptive to the variability of pathological speech.

Various studies have been conducted to improve speech recognition accuracy for individuals with dysarthria. Common approaches include acoustic feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC) [17] and spectrogram-based representations [18], which are then used in machine learning [9] or deep learning models [10], [19]. MFCC features offer advantages in capturing spectral characteristics related to human phonetic perception and are compact for model training. However, as a handcrafted feature, MFCC primarily encodes low-level acoustic information and often fails to capture complex temporal variations, particularly in speakers with articulatory impairments [20], [21]. Consequently, MFCC-based systems tend to exhibit reduced performance when applied to dysarthric speech, which is highly variable in both prosody and articulatory patterns.

Recent advancements in ASR have been marked by the emergence of self-supervised learning approaches such as Wav2Vec 2.0 [22], which enable models to learn representations directly from raw waveforms without relying on manually engineered features [23]. These representations capture richer and more contextual phonetic information, even when trained on large amounts of unlabeled data. Several studies have shown that such representations can be more adaptive to speaker variability, including pathological speech such as dysarthria [21]. However, processing raw audio directly introduces substantial computational complexity [24] and does not always effectively capture fine spectral structures in the same way as traditional acoustic features. Therefore, although modern deep learning approaches provide advantages in end-to-end representation learning, their effectiveness for disordered speech remains an open research question.

Most previous studies have focused on a single approach, either MFCC or Wav2Vec 2.0, without exploring the potential synergy between the two [21], [25]. Integrating handcrafted features designed based on acoustic theory and human auditory perception [26], such as MFCC or Wavelet-MFCC with self-supervised representations, offers the possibility of creating hybrid systems that combine spectral precision with contextual learning flexibility. Therefore, this study aims to identify the most effective feature-extraction method among three approaches, namely Wav2Vec 2.0, MFCC combined with Wav2Vec 2.0 (MFCC-Wav2Vec 2.0), and Wavelet-MFCC combined with Wav2Vec 2.0 (Wavelet-MFCC-Wav2Vec 2.0), for the

task of dysarthric word recognition. The findings are expected to contribute to the development of more robust, inclusive, and adaptive speech recognition systems for individuals with speech impairments.

2 RELATED WORK

Research on automatic speech recognition for individuals with dysarthria has evolved in response to the limitations of conventional ASR systems in handling pathological speech characterized by high articulatory and temporal variability. Previous studies have generally focused on traditional acoustic features, self-supervised learning-based representations, and hybrid approaches to improve system robustness. Hernandez et al. [27] evaluated the effectiveness of cross-lingual self-supervised speech representations for dysarthric speech recognition using the UA-Speech, PC-GITA, and EasyCall datasets. Wav2Vec 2.0, HuBERT, and the multilingual XLSR model were compared with conventional filterbank features. The results showed that Wav2Vec 2.0 reduced the Word Error Rate (WER) on UA-Speech from 32.9% to 29.3%, while XLSR achieved the best performance with a WER of 26.1%. However, performance variations across datasets indicate that self-supervised representations are not yet fully consistent across all dysarthric speech scenarios.

Javanmardi et al. [21] investigated the use of pre-trained models for dysarthria detection and severity level classification using the UA-Speech and TORGO datasets. The study compared features extracted from Wav2Vec 2.0 and HuBERT with traditional acoustic features such as MFCC, openSMILE, and eGeMAPS using a Support Vector Machine (SVM) classifier. Experimental results showed that HuBERT improved absolute accuracy by up to 2.86% for binary detection and 10.46% for four-level severity classification on the UA-Speech dataset, while also indicating that traditional acoustic features still provide complementary information.

A large-scale approach to improving ASR generalization to disordered speech was proposed by Tobin et al. [28]. By leveraging large-scale data from Project Euphonia and a Universal Speech Model based on the Conformer-CTC architecture, the study achieved WER reductions of up to 33% for prompted speech and 26% for conversational speech through fine-tuning. Despite its effectiveness, this approach relies on extremely large models and substantial computational resources, making it less suitable for lightweight implementations.

Hybrid time-frequency feature approaches have also been explored in other speech disorder domains. Hidayat et al. [29] proposed a speech recognition system for individuals with cleft lip and palate by integrating Wavelet-MFCC features with an LSTM model. Using 5-fold cross-validation, the results demonstrated that the Coif1 wavelet achieved the highest accuracy and sensitivity with stable performance, confirming that multi-resolution analysis can enrich speech signal representations.

In the broader context of machine learning applications in healthcare, Gharaibeh et al. [30] proposed a Swin Transformer-based approach with multi-scale feature pyramid fusion for Alzheimer's disease detection using the ADNI dataset. The proposed method achieved a classification accuracy of 98%, with a sensitivity of 0.90 and a specificity of 0.93, outperforming conventional baseline methods. Although the study focused on medical image analysis rather than speech recognition, it highlights the effectiveness of multi-level feature fusion in improving model robustness, which is conceptually relevant to pathological speech processing.

Based on this review, existing studies tend to explore traditional acoustic features and self-supervised representations independently or rely on large-scale models to

achieve robustness. Systematic integration of time-frequency-based handcrafted features, such as MFCC and Wavelet-MFCC, with self-supervised representations within an end-to-end ASR framework for dysarthric speech remains limited. Therefore, this study evaluates Wav2Vec 2.0, MFCC-Wav2Vec 2.0, and Wavelet-MFCC-Wav2Vec 2.0 to identify the most effective model for dysarthric speech recognition.

3 MATERIALS AND METHODS

This section describes the materials and methods used in this study to evaluate three feature-extraction approaches, namely Wav2Vec 2.0, MFCC-Wav2Vec 2.0, and Wavelet-MFCC-Wav2Vec 2.0. The research workflow consists of four main stages: preprocessing, feature extraction, model training, and model evaluation. Each stage was systematically designed to ensure consistency and reproducibility throughout the entire pipeline, as illustrated in Figure 1.

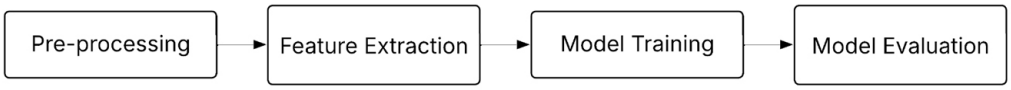

Fig. 1. Research workflow of the dysarthric word recognition system

3.1 Dataset

The dataset used in this study is the UA-Speech Corpus, a standard benchmark corpus for ASR research on dysarthric speech that includes speakers with varying severity levels as well as control speakers [31]. Two subsets were utilized, namely UA-Speech Normalized FM (dysarthric speakers) and UA-Speech Normalized Control (healthy speakers), both containing identical transcripts. Each subset consists of 30,855 audio files, and each speaker produces 765 isolated words spanning several categories, including digits, radio alphabet, computer commands, common words, and uncommon words. In the normalized version, amplitude variations have been equalized to reduce intensity differences across speakers. The duration of audio files in the FM subset ranges from 1 to 56 seconds, while the Control subset ranges from 1 to 13 seconds, reflecting differences in speaking rate, stability, and articulatory clarity between dysarthric and healthy speakers.

To ensure computational efficiency and maintain class balance, this study selected 7,500 samples from the FM subset and 7,500 samples from the Control subset, resulting in a total of 15,000 audio files. All files were stored in mono 16-bit PCM format with a sampling rate of 16 kHz. The labeling structure of the dataset is presented in Table 1, which provides examples of file name pairs and their corresponding word transcripts.

Table 1. Examples of labels and transcripts from the UA-speech normalized dataset

File Name Dysartria	File Name Control	Transcript
F02_B3_UW79_M6.wav	CF02_B3_UW79_M6.wav	powwow
F02_B2_C9_M3.wav	CF02_B2_C9_M3.wav	shift
F02_B1_C5_M6.wav	CF02_B1_C5_M6.wav	tab
F02_B2_C11_M6.wav	CF02_B2_C11_M6.wav	paragraph
F02_B2_LO_.wav	CF02_B2_LO_M6.wav	oscar

The File Name column in Table 1 contains information about the speaker ID, word category, and repetition index, while the transcript column provides the corresponding target word. This labeling structure serves as the basis for the feature extraction and model training processes in this study.

3.2 Pre-processing

The preprocessing stage aims to ensure that all audio data are standardized and ready for use in the word recognition system based on Automatic Speech Recognition (ASR) [32]. This procedure was applied consistently across all three feature-extraction approaches, namely Wav2Vec 2.0, MFCC-Wav2Vec 2.0, and Wavelet-MFCC-Wav2Vec 2.0. Each audio file was loaded using torchaudio.load() and converted to mono by averaging the channels (y = y.mean(dim = 0)). All signals were then resampled to 16 kHz using torchaudio.functional.resample(). This process ensured that every audio sample followed a uniform format, specifically mono 16-bit PCM with a 16 kHz sampling rate, consistent with the UA-Speech corpus specifications [31].

Amplitude normalization was applied to equalize intensity levels across recordings, ensuring that differences in loudness did not affect the feature-extraction process [33]. In this pipeline, normalization was handled by the Wav2Vec2FeatureExtractor through the parameter do_normalize = True, which normalizes the signal to zero-mean and unit-variance [22], before it enters the Wav2Vec 2.0 encoder. Since the normalized version of the UA-Speech corpus already applies initial RMS equalization, no additional manual normalization was performed outside this pipeline. Trimming, silence removal, and voice activity detection were also not applied in order to preserve uniformity in the duration of the isolated words.

3.3 Feature extraction

The feature-extraction stage aims to transform raw speech signals into meaningful numerical representations that can be effectively learned by the model [34]. This process was conducted using three approaches, namely Wav2Vec 2.0, MFCC-Wav2Vec 2.0, and Wavelet-MFCC-Wav2Vec 2.0. All extracted features were subsequently used as the primary inputs during model training to develop the Wav2Vec 2.0-based word recognition system.

Wav2Vec 2.0

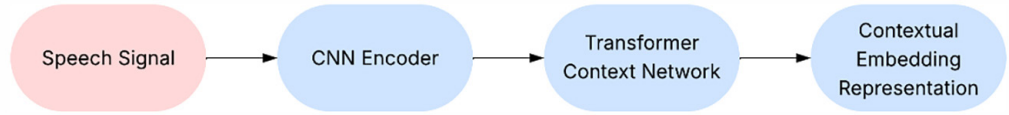

Fig. 2. Wav2Vec 2.0 feature extraction process

According to Figure 2, the feature extraction process in the Wav2Vec 2.0 approach begins with speech signals that have been normalized and resampled to 16 kHz. These signals are then passed into the Wav2Vec 2.0 Feature Encoder, a multi-layer convolutional module designed to extract low-level acoustic representations directly from the time domain [22]. In the present implementation, feature extraction is performed using the Wav2Vec2FeatureExtractor, which standardizes the raw waveform into a normalized input tensor before it is forwarded to the model's main encoder.

The first stage of feature extraction is performed by the Wav2Vec 2.0 CNN Encoder, which consists of seven temporal convolutional layers with progressively increasing kernel sizes and strides, yielding an overall temporal downsampling of approximately 20×. This module extracts latent speech representations in the form of 768-dimensional acoustic features that encode local energy patterns, phonetic transitions, and short-term temporal structure. These representations are domain-agnostic, meaning that they are not constrained by hand-crafted features such as MFCCs or filterbanks [23], [35].

Next, the latent representations are passed to the Transformer Context Network, which consists of 12 transformer blocks configured with a hidden size of 768, a feed-forward inner dimension of 3072, and a multi-head self-attention mechanism. This network models long-range dependencies in the speech signal, integrating both intra-word and cross-unit acoustic context so that each output vector captures information drawn from the entire input sequence [36]. Although the original Wav2Vec 2.0 pretraining procedure includes an internal temporal masking strategy, this study employs only the fine-tuning stage as implemented in the *facebook/wav2vec2-base* model. The output of this pipeline is a contextual embedding representation, an ordered sequence of 768-dimensional vectors that encodes higher-level phonetic and semantic structure. These embeddings are subsequently used as input to the CTC classification head during model training [37].

MFCC-Wav2Vec 2.0. The MFCC-Wav2Vec 2.0 approach aims to integrate MFCC with Wav2Vec 2.0 so that the model benefits from both spectrally grounded features and context-rich temporal representations [23], [38]. As illustrated in Figure 3, the feature extraction process begins with a mono speech signal normalized and resampled to 16 kHz. The signal is then propagated through two parallel processing paths.

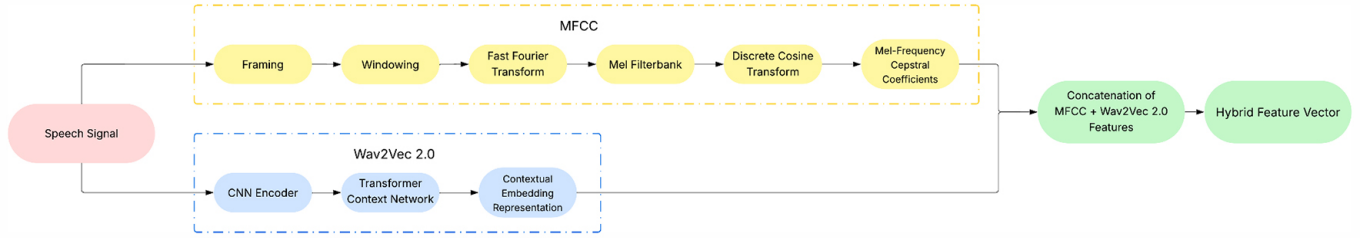

Fig. 3. MFCC-Wav2Vec 2.0 feature extraction process

In the MFCC branch, the signal is first segmented into short-time frames using a window length of 2048 samples (≈128 ms) and a hop length of 512 samples (≈32 ms) with a Hann window [39]. Each frame is then transformed into the frequency domain using a 2048-point Fast Fourier Transform (FFT). The resulting magnitude spectrum is passed through a Mel filterbank that models the perceptual frequency sensitivity of the human auditory system [40]. The energy in each filter is subsequently converted into the cepstral domain using the Discrete Cosine Transform (DCT), producing 13 MFCC coefficients. This procedure yields a compact spectral representation that captures formant structure, frequency-energy distribution, and articulatory cues that are particularly relevant for dysarthric speech. The MFCC formulation is presented in Eq. (1) [29]:

$$MFCC(k) = \sum_{n=1}^{N} log(E(n)).cos\left(\frac{k\pi}{N}(n-0.5)\right) \quad (1)$$

Concurrently, the Wav2Vec 2.0 branch processes the raw waveform using the Wav2Vec2FeatureExtractor, which applies zero-mean, unit-variance normalization. The standardized signal is then passed into the Wav2Vec 2.0 CNN Encoder, consisting of seven temporal convolutional layers with progressively increasing kernel sizes and strides, producing an overall temporal downsampling of approximately 20×. This module generates 768-dimensional latent speech representations for each time step. These representations are subsequently forwarded to the Transformer Context Network, which comprises 12 transformer layers with a hidden size of 768 and a feed-forward dimension of 3072. The network models long-range temporal dependencies and phonetic structure across frames, yielding 768-dimensional contextual embeddings at every timestep.

The two branches are subsequently integrated in the feature fusion stage. In this implementation, the MFCC features are first projected using a linear projection layer and standardized using Layer Norm to ensure that their scale is compatible with the Wav2Vec 2.0 hidden representations [41]. The MFCC sequence length is then aligned with the Wav2Vec 2.0 sequence through linear interpolation, a common technique used to match the temporal resolution of feature streams with differing sampling rates [42]. Once aligned, the MFCC features are directly added (additive fusion) to the contextual embeddings produced by Wav2Vec 2.0, forming a hybrid feature vector that combines fine-grained spectral cues from MFCC with deep contextual information from Wav2Vec 2.0. This fused feature vector is subsequently used as the input to the CTC classification head during training for the dysarthric word recognition task.

Wavelet-MFCC-Wav2Vec 2.0. In the Wavelet-MFCC-Wav2Vec 2.0 method, the raw speech signal is first processed as a mono waveform with a sampling rate of 16 kHz, after which it is simultaneously routed into two parallel feature-extraction branches.

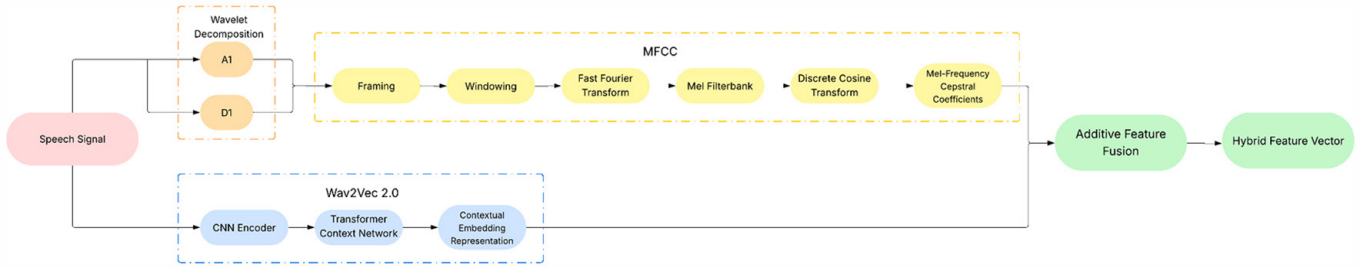

Fig. 4. Wavelet-MFCC-Wav2Vec 2.0 feature extraction process

As illustrated in Figure 4, the speech waveform is decomposed using a level-1 Discrete Wavelet Transform (DWT) with a Daubechies-1 mother wavelet, producing two sets of coefficients: the approximation component (A1), which captures low-frequency structural information, and the detail component (D1), which encodes high-frequency transient characteristics [43]. Each component is subsequently processed through an independent MFCC pipeline. This procedure includes Hann-window-based framing, a 2048-point FFT, perceptual spectral mapping using a Mel filterbank, and the computation of 13 cepstral coefficients via the DCT. To maintain temporal consistency between the two branches, the MFCC sequences obtained from A1 and D1 are aligned by trimming both to the shorter sequence length before concatenation along the feature axis. This results in a set of Wavelet Cepstral Coefficients (WCC) consisting of 26 features per frame, effectively integrating both low-frequency and high-frequency spectral characteristics into a unified representation.

In parallel, the original waveform is processed by Wav2Vec 2.0, which comprises a hierarchical convolutional feature encoder that generates low-level acoustic representations, followed by a Transformer context network consisting of 12 self-attention layers that model long-range temporal dependencies and phonetic structures in a contextualized manner. This module produces a sequence of 768-dimensional embeddings for each time step. To enable integration with the wavelet branch, the WCC sequence is first temporally aligned with the Wav2Vec 2.0 output via linear interpolation. The interpolated WCC features are then projected into a 768-dimensional space through a linear layer followed by layer normalization, GELU activation, and dropout, ensuring that the WCC representation attains a scale and structural form compatible with the transformer embeddings.

After the alignment process, the two representations are combined through additive feature fusion, implemented as a residual summation between the Wav2Vec 2.0 embeddings and the projected WCC vectors at each time step. This additive fusion strategy is chosen because it preserves the native representational space of Wav2Vec 2.0 while enriching the embeddings with multi-resolution information derived from the Wavelet-MFCC branch, without introducing substantial additional model complexity [44]. The resulting output is a 768-dimensional hybrid feature vector that remains fully compatible with the standard Wav2Vec 2.0 output format, allowing it to be directly forwarded to the CTC classification head for end-to-end decoding.

3.4 Model training

The speech recognition model in this study employs the Wav2Vec 2.0 architecture using the Wav2Vec2ForCTC configuration from *facebook/wav2vec2-base*. As illustrated in Figure 5, the architecture comprises two primary components: a CNN feature encoder and a Transformer context network. The CNN feature encoder extracts low-level latent speech representations directly from the raw audio signal through seven hierarchical temporal convolutional layers [22]. These initial representations are subsequently processed by the Transformer network, which consists of 12 self-attention blocks with a hidden size of 768, a feed-forward inner dimension of 3072, and 12 attention heads. This configuration enables the model to capture long-range temporal dependencies, phonetic structures, and prosodic patterns characteristic of dysarthric speech [36].

In the final stage, the model employs a Connectionist Temporal Classification (CTC) decoding head, implemented as a linear projection layer that maps the Transformer output into the character-token space defined by the vocabulary constructed from the UA-Speech transcripts. The CTC mechanism enables end-to-end training without requiring explicit alignment between acoustic frames and character sequences [37], making it particularly suitable for speech recognition tasks characterized by substantial temporal variability. During inference, the model utilizes greedy CTC decoding, consistent with the decoding strategy applied during training.

Model training was conducted for 30 epochs using the AdamW optimizer with an initial learning rate of 1×10^{-4}, 500 warmup steps, and a batch size of 4, supported by mixed-precision training (fp16 = True) to accelerate GPU computation. AdamW was selected due to its demonstrated ability to provide more stable gradient behavior compared to the classical Adam optimizer [45]. Although this study incorporates three different feature-extraction paradigms, namely Wav2Vec 2.0, MFCC-Wav2Vec 2.0, and Wavelet-MFCC-Wav2Vec 2.0, all experiments employ the same Wav2Vec 2.0 architecture as the core encoder, with differences appearing only in the feature-fusion modules of the respective hybrid approaches.

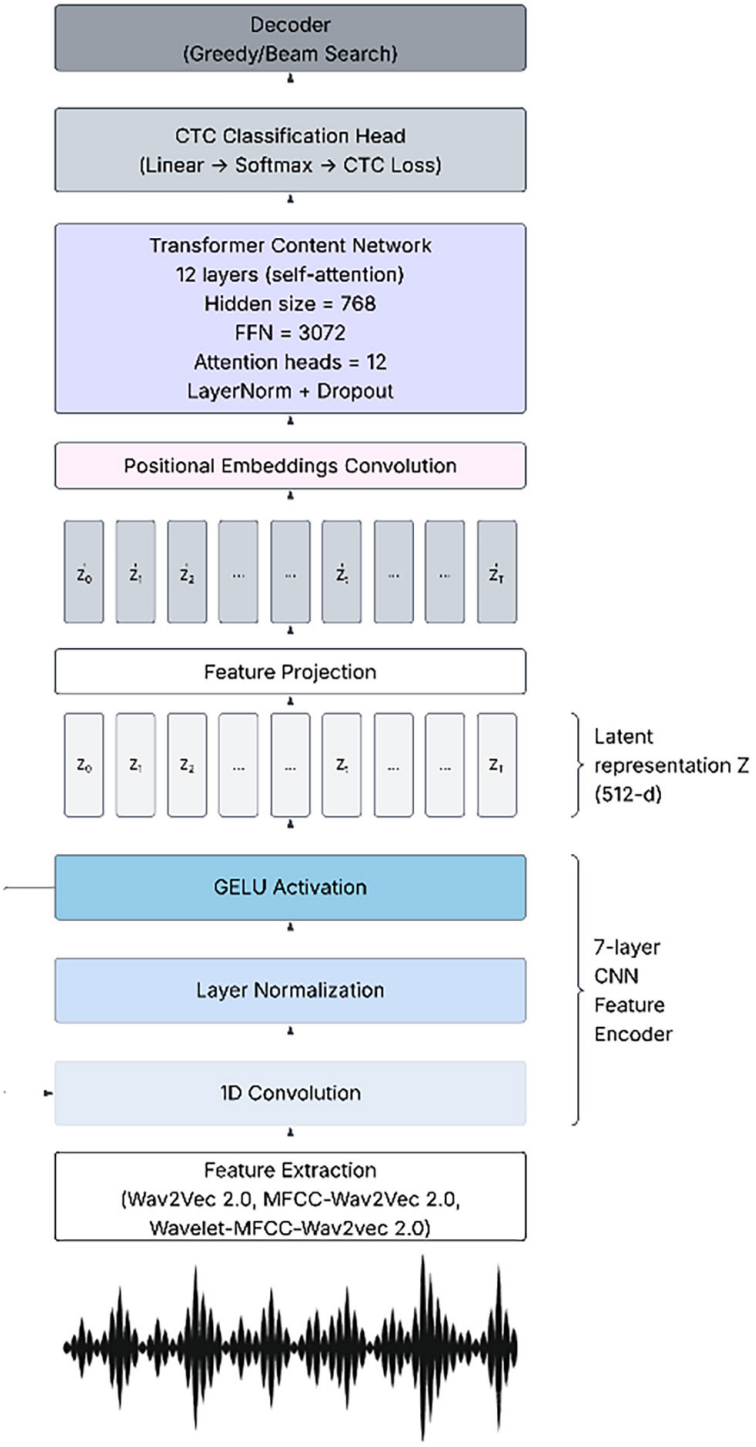

Fig. 5. Wav2Vec 2.0 architecture for word recognition in dysarthric speech

3.5 Model evaluation

The evaluation stage aims to assess the performance of the trained word-recognition system using the held-out test set [46]. This process measures how accurately the model is able to transcribe dysarthric speech relative to the available ground-truth annotations. The dataset was partitioned into two subsets, namely 80%

for training and 20% for testing, using a randomized split that preserves the speaker distribution, ensuring class balance and preventing bias during model assessment.

During evaluation, the model generates text predictions for each audio file in the test set. These predictions are then compared to the reference transcripts to compute the recognition error using the WER metric, formulated in Eq. (2):

$$WER = \frac{S + D + I}{N} \tag{2}$$

where S denotes the number of substitutions (incorrect words), D the number of deletions (missing words), I the number of insertions (additional words), and N the total number of words [47]. WER ranges from 0 to 1, with lower values indicating better model performance [48]. In this study, WER is computed automatically using the *jiwer* library, which applies the Levenshtein distance algorithm to quantify discrepancies between the model's transcription and the reference text. WER is considered the most appropriate metric for this work, as it effectively captures both phonetic and semantic errors that arise from the high acoustic variability present in dysarthric speech [49], [50].

4 RESULTS

4.1 Training and validation

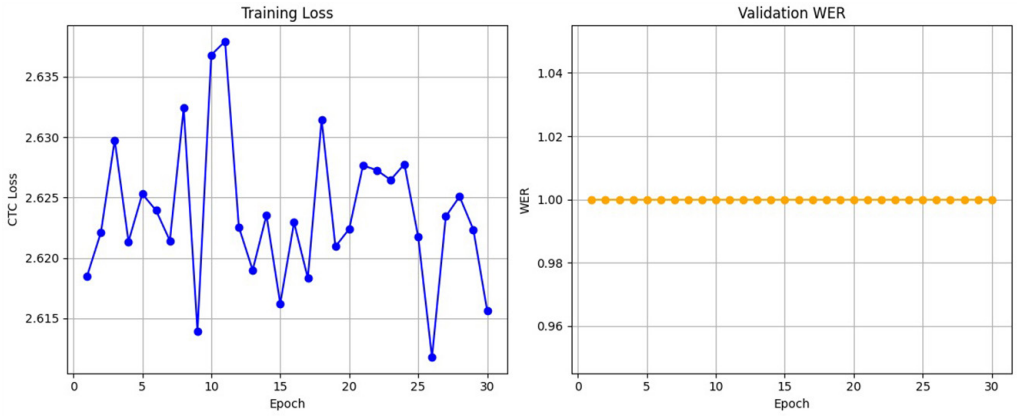

Fig. 6. Training loss and validation WER across methods Wav2Vec 2.0

Figure 6 illustrates the training loss pattern and validation WER obtained using the Wav2Vec 2.0 feature extraction method. Throughout the 30 training epochs, the training loss remained within the range of 2.61–2.64 and did not exhibit any clear downward trend. The fluctuations observed were largely random and did not indicate a stable optimization process. This stagnation suggests that the model did not improve its ability to map acoustic representations to character sequences during training. Meanwhile, the validation WER remained constant at 1.00 across all epochs, showing no variation or reduction. This outcome indicates that the model failed to generalize to the validation data and was unable to produce correct transcription predictions, even after the full training cycle. The stagnant pattern observed in both metrics indicates that the baseline Wav2Vec 2.0 training configuration in this method was not successful in capturing the acoustic characteristics of dysarthric speech effectively.

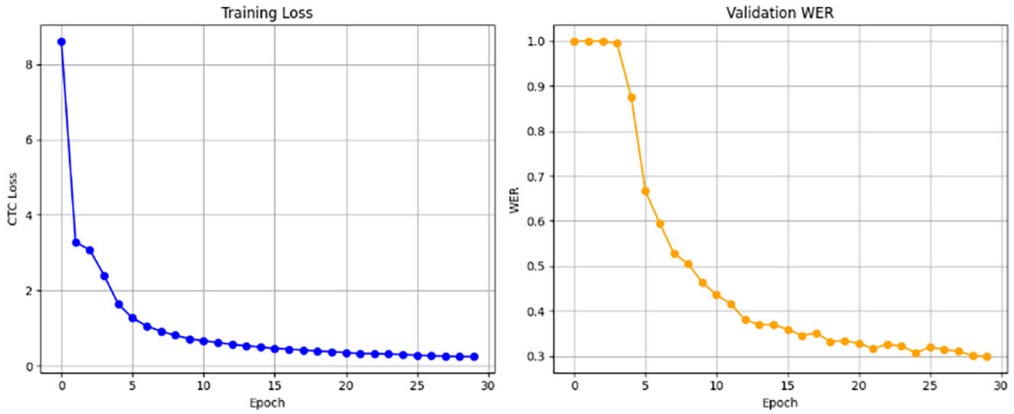

Fig. 7. Training loss and validation WER across methods MFCC-Wav2Vec 2.0

Subsequently, Figure 7 presents the training loss and validation WER dynamics for the MFCC-Wav2Vec 2.0 method. In contrast to the pure Wav2Vec 2.0 approach, this method shows a clear and stable downward trend in the training curve. The training loss decreases sharply from approximately 8.2 in the initial epoch to around 1.0 by epoch 10 and continues to decline gradually, reaching values close to 0.8 at epoch 30. Consistent with this behavior, the validation WER shows a steady reduction from 0.85 at the beginning of training to approximately 0.29 by epoch 30. No substantial fluctuations or large gaps between the training and validation curves are observed, indicating that the model achieved stable performance improvements without signs of overfitting. These results suggest that incorporating MFCC features prior to Wav2Vec 2.0 leads to a more effective convergence process compared to the baseline method.

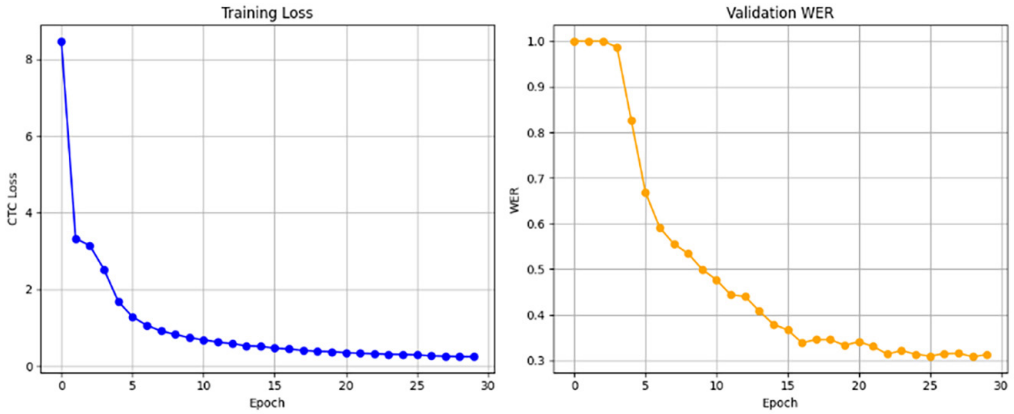

Fig. 8. Training loss and validation WER across methods Wavelet-MFCC-Wav2Vec 2.0

Figure 8 illustrates the performance of the Wavelet-MFCC-Wav2Vec 2.0 method, which exhibits a training pattern similar to that of the MFCC-Wav2Vec 2.0 approach. The training loss decreases gradually from an initial value of approximately 8.0 to 0.7 by epoch 30. The validation WER also shows a stable decline from 0.90 to 0.31 toward the end of training, with no major fluctuations. Although the final WER is slightly higher than that of the MFCC-Wav2Vec 2.0 method, the consistently smooth training curves indicate that this method is likewise capable of achieving stable and convergent learning behavior.

4.2 WER performance comparison

Table 2 presents the WER comparison for the three proposed feature-extraction methods. The standard Wav2Vec 2.0 approach yields the highest WER, namely 1.0000, indicating that the model failed to produce any correct word predictions on the test set. This result aligns with the flat WER curve observed in Figure 6, which shows no improvement throughout training. The MFCC-Wav2Vec 2.0 method achieves the lowest WER at 0.2990, making it the best-performing method among the three. Meanwhile, the Wavelet-MFCC-Wav2Vec 2.0 method attains a WER of 0.3126, differing by only 0.0136 from the MFCC-Wav2Vec 2.0 approach. These findings indicate that both hybrid methods outperform the baseline Wav2Vec 2.0 model by a substantial margin.

Table 2. WER performance of the three feature extraction methods

Method	WER
Wav2Vec 2.0	1.0000
MFCC-Wav2Vec 2.0	0.2990
Wavelet-MFCC-Wav2Vec 2.0	0.3126

Overall, the superior performance of the hybrid methods indicates that handcrafted features such as MFCC and Wavelet coefficients help stabilize the acoustic representations, particularly for dysarthric speech, which often exhibits irregular frequency patterns and temporal distortions. MFCCs are especially effective in capturing dominant spectral structures and reducing articulatory variability, while Wavelet-based features contribute multi-resolution information that enhances the representation of transient characteristics in the signal. These findings are consistent with prior studies showing that MFCCs improve representation stability in pathological speech [9] and that Wavelet transforms offer advantages in analyzing non-stationary signals [51].

5 DISCUSSION

The performance differences among the three methods can be explained by how each system extracts and represents dysarthric speech signals. The WER of 1.0000 obtained by the standalone Wav2Vec 2.0 model indicates a complete failure to map the input signal to the correct character sequence. This occurs because the Wav2Vec 2.0 pipeline processes the raw waveform directly through a convolutional feature encoder that was pre-trained on speech from typical speakers. When exposed to dysarthric speech, which is characterized by unstable articulation, irregular phonetic transitions, and distorted prosodic patterns, the CNN encoder is unable to construct meaningful latent speech representations. The resulting 768-dimensional embeddings become non-informative and fail to capture the underlying phonetic structure. As a result, the transformer context network propagates only temporal noise that cannot be learned by the CTC head, preventing fine-tuning from converging and causing the WER to remain at 1.00 throughout training. This finding is consistent with reports by Hernandez et al. [27] showing that waveform-based self-supervised models such as Wav2Vec perform poorly in pathological speech domains when no auxiliary acoustic features are provided.

In contrast, the MFCC-Wav2Vec 2.0 method achieves a significantly lower WER of 0.2990 because the MFCC branch provides a more stable acoustic structure before

the signal enters the encoder. The MFCC pipeline extracts formant patterns and frequency energy distributions through framing, FFT, Mel filterbank processing, and DCT, a technique that has long been recognized as effective for capturing pathological speech characteristics [52]. For dysarthric speakers who exhibit substantial temporal irregularities, MFCC serves as a frequency-domain stabilizer that reduces extreme variability present in the raw waveform. When the projected and normalized MFCC features are added to the Wav2Vec 2.0 contextual embeddings, this additive fusion produces a richer multimodal representation. Wav2Vec contributes long-range temporal context, while MFCC provides more structured spectral patterns. This combination results in a more stable reduction of loss and WER, reaching a value of 0.2990, thereby making MFCC-Wav2Vec 2.0 the best-performing model in this study. This performance improvement aligns with findings from previous studies that combine classical features such as MFCC with self-supervised representations to mitigate domain mismatch in pathological speech tasks [53].

Furthermore, the Wavelet-MFCC-Wav2Vec 2.0 method produces a WER of 0.3126, which is slightly higher than that of MFCC-Wav2Vec 2.0. This pipeline decomposes the signal using DWT so that the approximation (A1) and detail (D1) components can be captured separately, which is highly relevant because articulatory breakdown in dysarthria often appears in high-frequency transients. MFCC is then applied to both A1 and D1, and the resulting features are combined to form WCC consisting of 26 features per frame. This creates a multi-resolution representation that is more sensitive to local variations. However, this increased sensitivity also amplifies spectral variability that is not always relevant for word recognition, thereby introducing a small amount of structural noise into the fusion process. After the WCC features are projected into a 768-dimensional space and added to the Wav2Vec 2.0 embeddings, the model remains capable of learning, although it requires a greater degree of generalization. As a result, its WER is slightly higher, with a difference of 0.0136 compared to MFCC-Wav2Vec 2.0. This pattern is consistent with findings from other wavelet-based dysarthric ASR studies, in which increased multi-resolution sensitivity often leads to a trade-off between local precision and global representational stability [54].

6 CONCLUSION

This study demonstrates that the quality of acoustic representations plays a crucial role in determining the success of word recognition for speakers with dysarthria. The pure Wav2Vec 2.0 model was unable to adapt to articulatory and prosodic distortions present in pathological speech, resulting in a WER of 1.0000. The hybrid methods consistently produced significant performance improvements, with MFCC-Wav2Vec 2.0 achieving the best result with a WER of 0.2990 due to the ability of MFCC to stabilize irregular frequency structures. The Wavelet-MFCC-Wav2Vec 2.0 approach also improved accuracy with a WER of 0.3126, although the multi-resolution sensitivity of wavelet decomposition introduced additional spectral variability that slightly reduced generalization. Overall, the findings indicate that combining classical handcrafted features with self-supervised representations provides greater robustness against the temporal variability characteristic of dysarthric speech. For future research, several feature extraction parameters may be further optimized, such as evaluating MFCC window sizes of 512 or 1024 samples, exploring wavelet types other than Daubechies-1, and examining approximation channels within the wavelet representation. Enhancements to the feature extraction pipeline

have the potential to improve representational stability and model performance across a broader range of dysarthric speech variations.

7 DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT-5.1 by OpenAI in order to assist with English translation. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

8 REFERENCES

[1] S. Durga and V. Mehrotra, "Communication and its vital role in human life," *Int. J. Health Sci. (Qassim)*, vol. 6, no. S5, pp. 5940–5948, 2022. https://doi.org/10.53730/ijhs.v6nS5.10005

[2] A. Skoczylas, W. Koperska, M. Stachowiak, and N. Duda-Mróz, "Speech recognition and enhancement in underground mines for the use of smart voice assistants," *Procedia Comput. Sci.*, vol. 225, pp. 1964–1973, 2023. https://doi.org/10.1016/j.procs.2023.10.187

[3] R. Joshi and V. Kannan, "Attention-based end-to-end speech recognition for voice search in Hindi and English," in *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE '21)*, 2021, pp. 107–113. https://doi.org/10.1145/3503162.3503173

[4] A. Olev and T. Alumäe, "Estonian speech recognition and transcription editing service," *Balt. J. Mod. Comput.*, vol. 10, no. 3, pp. 409–421, 2022. https://doi.org/10.22364/bjmc.2022.10.3.14

[5] H. Kheddar, M. Hemis, and Y. Himeur, "Automatic speech recognition using advanced deep learning approaches: A survey," *Inf. Fusion*, vol. 109, p. 102422, 2024. https://doi.org/10.1016/j.inffus.2024.102422

[6] M. McGuire and J. Larson-Hall, "Assessing Whisper automatic speech recognition and WER scoring for elicited imitation: Steps toward automation," *Res. Methods Appl. Linguist.*, vol. 4, p. 100197, 2025. https://doi.org/10.1016/j.rmal.2025.100197

[7] M. Wang, H. Ma, Y. Wang, and X. Sun, "Design of smart home system speech emotion recognition model based on ensemble deep learning and feature fusion," *Appl. Acoust.*, vol. 218, p. 109886, 2024. https://doi.org/10.1016/j.apacoust.2024.109886

[8] S. Dutta, D. Irvin, and J. H. L. Hansen, "Exploring discrete speech units for privacy-preserving and efficient speech recognition for school-aged and preschool children," *Int. J. Hum. Comput. Stud.*, vol. 199, p. 103460, 2025. https://doi.org/10.1016/j.ijhcs.2025.103460

[9] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification: A study on acoustic features and deep learning techniques," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1147–1157, 2022. https://doi.org/10.1109/TNSRE.2022.3169814

[10] W. Ye, Z. Jiang, Q. Li, Y. Liu, and Z. Mou, "A hybrid model for pathological voice recognition of post-stroke dysarthria by using 1DCNN and double-LSTM networks," *Appl. Acoust.*, vol. 197, p. 108934, 2022. https://doi.org/10.1016/j.apacoust.2022.108934

[11] J. R. Duffy, *Motor Speech Disorders: Substrtes, Differential Diagnosis, and Management*. 4th ed. St. Louis, Missouri: Elsevier, 2019.

[12] C. Mitchell *et al.*, "Prevalence of aphasia and dysarthria among inpatient stroke survivors: Describing the population, therapy provision and outcomes on discharge," *Aphasiology*, vol. 35, no. 7, pp. 950–960, 2021. https://doi.org/10.1080/02687038.2020.1759772

[13] N. Do, S. Mitchell, L. Sturgill, P. Khemani, and M. K. Sin, "Speech and swallowing problems in Parkinson's disease," *J. Nurse Pract.*, vol. 18, pp. 848–851, 2022. https://doi.org/10.1016/j.nurpra.2022.05.019

[14] R. Dubbioso *et al.*, "Precision medicine in ALS: Identification of new acoustic markers for dysarthria severity assessment," *Biomed. Signal Process. Control*, vol. 89, p. 105706, 2024. https://doi.org/10.1016/j.bspc.2023.105706

[15] Kemenkes RI, "Laporan nasional riskesdas 2018," Lembaga Penerbit Badan Penelitian dan Pengembangan Kesehatan, 2019. [Online]. Available: https://repository.badankebijakan.kemkes.go.id/id/eprint/3514/1/Laporan Riskesdas 2018 Nasional.pdf

[16] J. Mills, O. Duffy, K. Pedlow, and G. Kernohan, "Exploring the perceptions of voice-assisted technology as a tool for speech and voice difficulties: Focus group study among people with Parkinson disease and their carers," *JMIR Rehabil. Assist. Technol.*, vol. 12, p. e75316, 2025. https://doi.org/10.2196/75316

[17] B. A. Al-Qatab and M. B. Mustafa, "Classification of dysarthric speech according to the severity of impairment: An analysis of acoustic features," *IEEE Access*, vol. 9, pp. 18183–18194, 2021. https://doi.org/10.1109/ACCESS.2021.3053335

[18] S. Sajiha, K. Radha, D. Venkata Rao, N. Sneha, S. Gunnam, and D. P. Bavirisetti, "Automatic dysarthria detection and severity level assessment using CWT-layered CNN model," *Eurasip. J. Audio, Speech, Music Process.*, vol. 2024, 2024. https://doi.org/10.1186/s13636-024-00357-3

[19] J. C. Prabhala, R. Ragoju, V. Kuppili, and C. Chesneau, "Enhanced early detection of dysarthric speech disabilities using stacking ensemble deep learning model," *Mach. Learn. with Appl.*, vol. 21, p. 100721, 2025. https://doi.org/10.1016/j.mlwa.2025.100721

[20] L. C. Chang and J. W. Hung, "A preliminary study of robust speech feature extraction based on maximizing the probability of states in deep acoustic models," *Appl. Syst. Innov.*, vol. 5, 2022. https://doi.org/10.3390/asi5040071

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

[22] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 12449–12460, 2020.

[23] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Conf. Int. Speech Commun. Assoc.*, 2019, pp. 3465–3469. https://doi.org/10.21437/Interspeech.2019-1873

[24] M. Kunešová, Z. Zajíc, L. Šmídl, and M. Karafiát, "Comparison of wav2vec 2.0 models on three speech processing tasks," *Int. J. Speech Technol.*, vol. 27, pp. 847–859, 2024. https://doi.org/10.1007/s10772-024-10140-6

[25] A. S. Al-Ali, R. M. Haris, Y. Akbari, M. Saleh, S. Al-Maadeed, and M. R. Kumar, "Integrating binary classification and clustering for multi-class dysarthria severity level classification: A two-stage approach," *Cluster Comput.*, vol. 28, 2025. https://doi.org/10.1007/s10586-024-04748-1

[26] F. G. Eriş and E. Akbal, "Enhancing speech emotion recognition through deep learning and handcrafted feature fusion," *Appl. Acoust.*, vol. 222, p. 110070, 2024. https://doi.org/10.1016/j.apacoust.2024.110070

[27] A. Hernandez, P. A. Pérez-Toro, E. Nöth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. Interspeech*, 2022, pp. 51–55. https://doi.org/10.21437/Interspeech.2022-10674

[28] J. Tobin, K. Tomanek, and S. Venugopalan, "Towards a single ASR model that generalizes to disordered speech," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. – Proc.*, 2025. https://doi.org/10.1109/ICASSP49660.2025.10888895

[29] S. Hidayat, T. Andrasto, F. P. Rochim, M. Khaira, M. H. Herdiansyah, and F. N. Aryza, "Wavelet-MFCC and LSTM-based speech recognition for cleft lip and palate," in *2025 International Electronics Symposium (IES)*, 2025, pp. 795–801. https://doi.org/10.1109/IES67184.2025.11162014

[30] N. Gharaibeh, A. A. Abu-Ein, O. M. Al-hazaimeh, K. M. O. Nahar, W. A. Abu-Ain, and M. M. Al-Nawashi, "Swin transformer-based segmentation and multi-scale feature pyramid fusion module for alzheimer's disease with machine learning," *Int. J. online Biomed. Eng.*, vol. 19, no. 4, pp. 22–50, 2023. https://doi.org/10.3991/ijoe.v19i04.37677

[31] H. Kim *et al.*, "Dysarthric speech database for universal access research," in *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech*, 2008, pp. 1741–1744. https://doi.org/10.21437/Interspeech.2008-480

[32] D. Geneva and G. Shopov, "Towards accurate text verbalization for ASR based on audio alignment," in *Proceedings of the Student Research Workshop Associated with RANLP-2019*, 2019, pp. 39–47. https://doi.org/10.26615/issn.2603-2821.2019_007

[33] A. Jeannerot, N. de Koeijer, P. Martínez-Nuevo, M. B. Møller, J. Dyreby, and P. Prandoni, "Increasing loudness in audio signals: A perceptually motivated approach to preserve audio quality," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. – Proc.*, 2022, pp. 1001–1005. https://doi.org/10.1109/ICASSP43922.2022.9747589

[34] I. Azam and S. Ali Khan, "Feature extraction trends for intelligent facial expression recognition: A survey," *Inform. Lithuanian Acad. Sci.*, vol. 42, no. 4, pp. 507–514, 2018. https://doi.org/10.31449/inf.v42i4.2037

[35] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv Preprint arXiv:1807.03748*, 2019. [Online]. Available: http://arxiv.org/abs/1807.03748

[36] A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017. https://doi.org/10.1109/2943.974352

[37] R. Fan, W. Chu, P. Chang, and A. Alwan, "A CTC alignment-based non-autoregressive transformer for end-to-end automatic speech recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 1436–1448, 2023. https://doi.org/10.1109/TASLP.2023.3263789

[38] Y. El Kheir, A. Das, E. E. Erdogan, F. Ritter-Guttierez, T. Polzehl, and S. Möller, "Two views, one truth: Spectral and self-supervised features fusion for robust speech deepfake detection," in *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2025. https://doi.org/10.1109/WASPAA66052.2025.11230938

[39] F. Jiao, J. Song, X. Zhao, P. Zhao, and R. wang, "A spoken English teaching system based on speech recognition and machine learning," *Int. J. Emerg. Technol. Learn.*, vol. 16, no. 14, pp. 68–82, 2021. https://doi.org/10.3991/ijet.v16i14.24049

[40] R. Darni, Y. Harisman, and I. N. A. F. Setiawan, "The implementation and empirical analysis of adaptive virtual mentor: Mobile technology empowers introverts' Business Communication Skills," *Int. J. Interact. Mob. Technol.*, vol. 19, no. 8, pp. 159–173, 2025. https://doi.org/10.3991/ijim.v19i08.53887

[41] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4580–4584. https://doi.org/10.1109/ICASSP.2015.7178838

[42] Y. Gong, Y. A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech*, 2021, pp. 571–575. https://doi.org/10.21437/Interspeech.2021-698

[43] D. Campo, O. L. Quintero, and M. Bastidas, "Multiresolution analysis (discrete wavelet transform) through Daubechies family for emotion recognition in speech," *J. Phys. Conf. Ser.*, vol. 705, no. 1, p. 012034, 2016. https://doi.org/10.1088/1742-6596/705/1/012034

[44] K. He, X. Zhang, S. Ren, and J. Su, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. https://doi.org/10.1246/cl.2003.428

[45] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019*, 2019. https://doi.org/10.48550/arXiv.1711.05101

[46] E. Kumalija and Y. Nakamoto, "Performance evaluation of automatic speech recognition systems on integrated noise-network distorted speech," *Front. signal Process.*, vol. 2, 2022. https://doi.org/10.3389/frsip.2022.999457

[47] S. Ouzerrout, "Universal-WER: Enhancing WER with segmentation and weighted substitution for varied linguistic contexts," in *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, 2024, pp. 29–35.

[48] R. Zhao, "Application of mobile interactive applications in college English teaching from the perspective of intelligent educational technology," *Int. J. Interact. Mob. Technol.*, vol. 19, no. 24, pp. 18–32, 2025. https://doi.org/10.3991/ijim.v19i24.59477

[49] M. Kim, B. Cao, K. An, and J. Wang, "Dysarthric speech recognition using convolutional LSTM neural network," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. Interspeech*, 2018, pp. 2948–2952. https://doi.org/10.21437/Interspeech.2018-2250

[50] S. W. Yang *et al.*, "SUPERB: Speech processing universal performance benchmark," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. Interspeech*, 2021. https://doi.org/10.21437/Interspeech.2021-1775

[51] G. Gidaye, J. Nirmal, K. Ezzine, and M. Frikha, "Wavelet sub-band features for voice disorder detection and classification," *Multimed. Tools Appl.*, vol. 79, nos. 39–40, pp. 28499–28523, 2020. https://doi.org/10.1007/s11042-020-09424-1

[52] Z. Qian and K. Xiao, "A survey of automatic speech recognition for dysarthric speech," *Electronics*, vol. 12, no. 20, pp. 1–23, 2023. https://doi.org/10.3390/electronics12204278

[53] S. Hu *et al.*, "Exploring self-supervised pre-trained ASR models for dysarthric and elderly speech recognition," in *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10097275

[54] Z. Qian, K. Xiao, and C. Yu, "A survey of technologies for automatic dysarthric speech recognition," *EURASIP J. Audio, Speech, Music Process.*, vol. 1, no. 48, 2023. https://doi.org/10.1186/s13636-023-00318-2

9 AUTHORS

Faila Nadhifatul Aryza is an undergraduate student in the Department of Informatics and Computer Engineering Education, Universitas Negeri Semarang (UNNES), Indonesia. Her research interests include speech recognition and image processing (E-mail: failaarz10@students.unnes.ac.id).

Syahroni Hidayat is a Lecturer in the Department of Electrical Engineering, Universitas Negeri Semarang (UNNES), Indonesia. He completed his Master of Engineering in Electrical Engineering in 2016. His research interests include automatic speech recognition, speaker recognition, speech and audio signal processing, and machine learning for speech technology (E-mail: syahronihidayat@mail.unnes.ac.id).