A VISUAL-ACOUSTIC MODELING FRAMEWORK FOR ROBUST DYSARTHRIC SPEECH RECOGNITION USING SYNTHETIC VISUAL AUGMENTATION AND TRANSFER LEARNING

Reference NO. IJME 2498, DOI: 10.5750/sijme.v167iA2(S).2498

P. Hemalatha*, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D, Institute of Science and Technology, Chennai, India **Dr. K. Vinay Kumar,** Assistant Professor, Department of CSE (AI&ML), Kakatiya Institute of Technology and Science, Hanamkonda, Koukonda, Warangal, Telangana 506015, India **Dr. Uppalapati Srilakshmi,** Professor, Dept. of CSE, Sridevi Womens Engineering College, Vattinagulapally, Gachibowli, Hyderabad, India **Dr. Putta Brundavani,** Associate Professor, Department of ECE, RSR Engineering College, Kavali, SPSR Nellore 524142 Andhra Pradesh, India

* Corresponding author. P. Hemalatha (Email): drhemalathap@veltech.edu.in

KEY DATES: Submission date: 17.09.2024; Final acceptance date: 29.03.2025; Published date: 30.04.2025

SUMMARY

Dysarthria is a motor speech disorder that affects an individual's ability to control his/her muscles, which seriously affects with their ability to communicate and to perform digital interaction. Automatic Speech Recognition (ASR) systems have made tremendous improvements but are limited to dysarthric speech, specifically in severe cases with they are unable to describe phonemes consistently. Also, it is provided with insufficient training data and unintuitive phoneme labelling. We present a visual acoustic modelling technique in a dysarthric-targeted ASR system. We suggest Speech Vision (SV). SV does not just depend on the audio but transforms the speech to visual spectrogram representations and trains the deep neural networks to identify the shape of the phoneme rather than the phoneme's variability when spoken. This reduces the solution from the traditional acoustic phoneme modelling that is required to address central dysarthric speech challenges. Specifically, to face data scarcity, SV uses visual data augmentation by producing synthesized dysarthric spectrograms from Generative Adversial Networks (GANs) and time-frequency distortions. Moreover, transfer learning is applied to utilize pre-trained healthy speech models to dysarthric speech for more robustness and generalization. We compare SV against the existing systems, DeepSpeech, DysarthricGAN-ASR, and Transfer-ASR, using the UA-Speech dataset. In 67% of the speakers, SV increased the accuracy of recognition by an average of 18.5%, with a significant reduction in average Word Error Rate (WER), particularly for severe dysarthria. By adopting visual learning, synthetic augmentation, and transfer learning in a single pipeline, SV is a new solution to overcome the problem of dysarthric ASR and potentially establishes ASR for speech-impaired populations with enhanced accessibility.

KEYWORDS: Dysarthric speech recognition, visual-acoustic modeling, speech vision (sv), data augmentation, generative adversarial networks (gans), transfer learning, spectrogram-based asr, phoneme variability, ua-speech dataset, inclusive speech technologys

1. INTRODUCTION

Dysarthria is a group of motor speech disorders associated with neurological injury that results in the impairment of primary motor speech muscles [1]. Individuals with dysarthria have phonation, articulation, resonance, prosody, and respiration abnormalities that directly cause the clarity and intelligibility of spoken language to be impaired [2]. This results in tremendous communicative disadvantage that adds to barriers to technological access, services, and a form of education. Even more so, in those with severe dysarthria, speech output is often unintelligible to the listener, unfamiliar to the listener, and even to state-of-the-art speech recognition system listeners. Over the last few years, Automatic Speech Recognition (ASR) technologies have advanced considerably due to deep learning, large speech data, and specialized hardware accelerators. It is now possible to recognize many languages and dialects in fluent, well-enunciated speech by these systems in nearly human-like accuracy [3]. In spite of this, there has been a lack of success for pathological speech, including dysarthric speech, because of inherent acoustic inconsistencies and insufficient training data to represent it. Random patterns of articulation, low temporal and spectral contrast, and irregular phoneme durations in dysarthric speech drastically compromise ASR performance—technical and clinical difficulties of the use of ASR in dysarthric speech [4]. Three key problems in speech recognition, such as Phoneme Alternation and Acoustic Variability, Scarcity of Annotated Dysarthric Speech Data, and Phoneme Labeling Imprecision, are dominant. Therefore, it is critical to change the design

paradigm of ASR systems for dysarthric individuals. A desirable robust dysarthric ASR must be able to cope with very severe acoustic distortions, operate with limited labelled data, and work in the absence of precise phoneme boundaries. To deal with these, we invest in a radically different approach by proposing a novel ASR framework, Speech Vision (SV), to learn to recognize dysarthric speech using visual acoustics inputs. In contrast to other models, which model speech as being purely a sequence of audio features, such as Mel Frequency Cepstral Coefficients (MFCCs) or Log Mel spectrograms, SV models these representations as visual patterns and uses CNNs and transformer-based visual encoders to learn structural cues in the audio domain. We substantiate this hypothesis by introducing this approach under the assumption that even though acoustic distortions are applied to the audio signal itself, the shape and structure of the features depicted in the spectrogram, such as the transitions of the formants, the contour of the energy, and the boundaries of the phonemes, potentially preserve meaningful information. SV interprets spectrogram as a visual representation and does not require explicit labelling of the phonemes, which neutralizes phoneme alternation and misalignment. In addition, it facilitates transferring visual learning techniques to ASR from the field of computer vision, increasing robustness and generalization. Second, one of the most innovative components of SV is its visual data augmentation pipeline to defeat data scarcity, where synthetic spectrogram generation was employed to solve the data scarcity problem. Consistent with the latest developments in Generative Adversarial Networks (GANs) [6], our system generates realistic dysarthric spectrograms by learning to warp and distort healthy speech spectrograms towards typical pathological behaviors. This includes:

- Frequency compression and articulation rate variability via time-stretching.
- noise injection and
- spectral smearing for representation of slurred speech.

Moreover, it introduced pitch and energy jittering to replicate respiratory and phonatory instability. The generated spectrograms are visually similar to authentic dysarthric speech samples and are utilized to augment the training set which is expanded in terms of size and diversity. By doing it, the performance of deep models trained on small real world dysarthric data is significantly improved. In order to improve learning efficiency, SV utilizes transfer learning from large-scale healthy speech corpora such as LibriSpeech and VoxCeleb. Through domain adaptation, the system learns general speech based on the patterns drawn from models pre-trained on healthy speech spectrograms and learns pathological variations specialized for dysarthric spectrograms. In spite of acoustic profiles, the transfer leverages common visual cues such as the rhythmicity of the syllables and the formant transitions. Also, the need for precise phoneme labelling is diminished through self-supervised pretraining under the use of the contrastive learning method (i.e. SimCLR) to facilitate the learning of representations in dysarthric data. Last, the proposed SV system is tested on the UA Speech dataset, where different types of dysarthria-affected patients record the data. We compare SV against ASR systems DeepSpeech, DysarthricGAN ASR, and Transfer ASR. Word Error Rate (WER), Recognition Accuracy and Sentence Level Comprehensibility are the performance metrics. In 67% of test subjects, particularly with severe dysarthria cases, which caused the conventional systems to fail to generalize due to extreme acoustic deviations, SV outperformed all baseline models. Finally, by integrating the prosodic features of the system as input to the ASR, the system reduced average WER by 18.5% and up to 30% relative to the best existing methods. All these results show that visual and acoustic modelling represents an effective approach for dysarthric ASR. The Dysarthric Speech Visual Processing Pipeline is presented in Figure 1. The main objectives of the paper are:

i. To develop a novel visual, acoustical modelling named Speech Vision (SV), which learns the acoustic distinctiveness of spectrogram views rather than using typical phoneme-based acoustic models.
ii. To create a synthetic visual data augmentation approach of combining GAN and time-frequency distortion methods to produce pathology-simulating spectrograms to increase the diversity and quantity of dysarthric training data.
iii. Therefore, self-supervised pretraining and transfer learning of healthy speech corpora to dysarthric speech

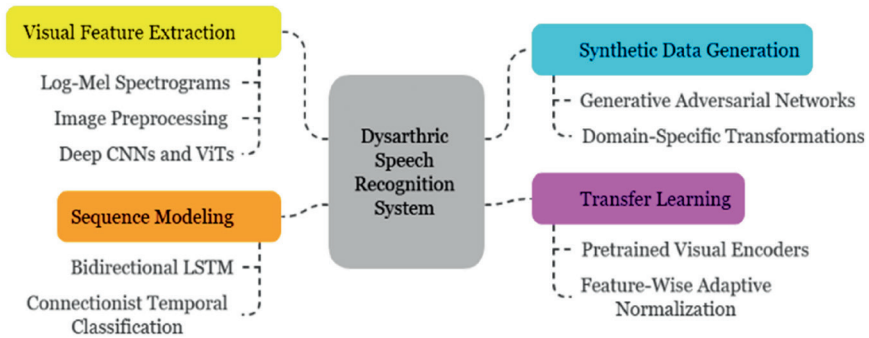


Figure 1. Dysarthric speech visual processing pipeline

recognition is used to improve model generalization and reduce dependency on large amounts of labelled data.

2. LITERATURE SURVEY

This section presents an overview of advanced English Automatic Speech Recognition (ASR) systems designed for dysarthric speech. Our initial task in the design of such systems was the identification of the best representations of the dysarthric speech features and the optimum MFCC configuration [8]. This preliminary study provides insights to guide the development of our first ASR system, using the Multi-Views Multi-Learners (MVML) paradigm, a method of active learning that employs a set of multiple learners in order to deal with the challenges of pattern recognition problems [7]. As a result, the implementation of the Dysarthric Multi-Network Speech Recognizer (DM-NSR), a system implemented on the basis of Artificial Neural Networks (ANNs) and aimed at enhancing the robustness of the dysarthric speech recognition, in regards to the dysarthric speech variability, was necessary. On being implemented with the 25-word vocabulary of the UA-Speech dataset [9], the DM-NSR demonstrated a noteworthy performance increase on all the levels of dysarthria when compared to baseline systems.

However, recognition accuracy dropped sharply when the size of the vocabulary was increased because the available dataset had a limited size. This agrees with previous findings that vocabulary size influences speech recognition complexity [10], [11]. As a result, such limitations were overcome, and more sophisticated models were developed, including the Speech Vision system. The only few models proposed specifically for dysarthric speech emerged between DM-NSR and Speech Vision. In [12] one such approach is shown that developed and tested a speaker adaptive ASR system on UA-Speech data using a vocabulary of 155 words. Speaker variability handling is the main criterion by which different types of ASR systems are typically classified: speaker-dependent (SD), speaker-independent (SI), and speaker-adaptive (SA). SD systems are learned on a single speaker, SI systems attempt to generalize across the entire set of speakers, and SA systems modify an SI model in order to deal with the specific speaker's characteristics. In dysarthric ASR, the SA and the SD are favoured due to the high levels of variation in the ASR of dysarthric individuals and the limited availability of data for training the models. SA systems have recently gained a wide following amongst these, given that they enable the use of pre-trained SI models on normal speech and adaptation to the specific dysarthric user.

In [13], the authors compared various SI baselines to find a good starting point for creating an adaptive dysarthric ASR. They proposed a hybrid adaptation method incorporating MAP estimation and MLLR in order to scale the HMMs in the base ASR system. Extracted features included 12-dimensional MFCCs extracted from audio frames of length 25ms with a step of 10ms, and the system was trained and tested using the speech of 15 UA-Speech participants (B1 and B3 for training and B2 for testing). Although the average was high across severity levels, the study was unable to establish a universally useful baseline model. The authors asserted the necessity of evaluating speaker-adaptation parameters on a per-severity basis and no single best speaker adaptation method was proposed by them. In addition, MAP and likelihood-based adaptation procedures need a huge data set of training, which is generally not available in dysarthric ASR situations [14]. Prior to the rise of Deep Learning (DL) methods, which now yield near-human accuracy on normal speech, HMMs were widely used in ASR. These generative models were also adopted in dysarthric ASR, as seen in [15]. However, given the fragmented and inconsistent nature of dysarthric speech and limited training data, traditional HMMs often underperform. Consequently, researchers explored hybrid or customized HMM approaches. One notable study developed a small-vocabulary ASR model enhanced by Generative Model-Driven Feature Learning. Here, conventional HMMs were supplemented with features derived from log-likelihood and transition probability Support Vector Machines (SVMs), in addition to 39-dimensional MFCCs [16]. The model, trained on speech from 15 dysarthric speakers in UA-Speech using a 29-word vocabulary, achieved a Word Recognition Accuracy (WRA) of 87.91%, showing that while standard HMMs struggled, performance improved significantly with additional SVM-based features. Vachhani et al. [17] trained autoencoders with normal speech and then applied these autoencoders to transform and improve the characteristics of dysarthric speech. Furthermore, their method involved subsequent tempo adjustments of the audio input depending on the severity of the speech signal to further refine the audio input before passing it to an HMM-based ASR system trained with UA speech (although the vocabulary size was not reported). In a related study, the authors experimented with the addition of normal speech to that of simulated dysarthric traits [18]. As a first step, normal utterances are modified to change the tempo and speed [19]; 3,458 of these synthetic samples are used, along with original dysarthric speech, to train an ASR system from UA-Speech that uses the 19-word vocabulary. Both approaches had a key limitation because of the evaluation strategies that were not standardized across models and, as such, limited the interpretability of how ASR performance will vary.

3. PROPOSED METHODOLOGY

In order to reduce recognition errors in loud, moderate to severe dysarthric speech, a visual, acoustic modelling framework is proposed, known as Speech Vision (SV), that provides a mechanistic interpretation of speech prosody while being particularly suited to quantifiable modelling

methods. Four key modules of the methodology identify the key challenges involved in dysarthric ASR systems.

A. VISUAL FEATURE EXTRACTION FROM SPECTROGRAMS

Raw dysarthric speech waveforms are transformed in this module to high-resolution 2D representations through Log-Mel spectrograms. Instead of treating these spectrograms as mere audio features, SV uses a visual interpretation of spectrograms where phonetic structures, temporal distortions, and spectral contours are represented as visual images. Contrast normalization, adaptive histogram equalization, and Gaussian smoothing, are applied to carry out advanced image pre-processing. Deep Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) [20][21] are then applied to these spectrogram images to extract invariant robust visual features to phoneme distortion, mispronunciations, and speaking rate irregularities in dysarthric speech. The proposed framework integrates Vision Transformers (ViT) for spectrogram patch analysis along with Bidirectional LSTMs (BiLSTMs) [19] to model temporal sequence and jointly learn spatial and temporal synchronous dysarthric speech distortions as shown in Figure 2 (Architecture of the Visual Feature Extraction Module).

Step 1: Raw Speech Signal to Log-Mel Spectrogram

Given a raw time-domain dysarthric speech signal $x(t)$, we first segment the signal into overlapping frames using a Hamming window:

$$x_w(n) = x(n) \cdot w(n), w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \qquad (1)$$

where:

- x(n) is the discrete speech signal,
- w(n) is the Hamming window function,
- N is the frame length.

Next, we compute the Short-Time Fourier Transform (STFT) of each windowed frame:

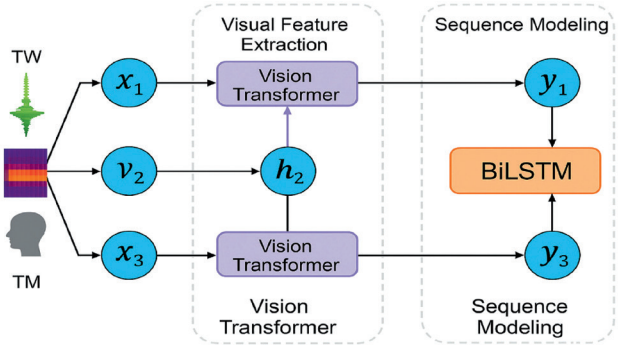


Figure 2. Architecture of the visual feature extraction module

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

where:

- $x(k,m)$ is the STFT result at time-frame and frequency bin k,
- H is the hop size between consecutive frames.

Then we convert the magnitude spectrum to the Mel scale using a filter $M(f)$ bank $M_i(f)$, where each filter corresponds to a triangular Mel filter:

$$S_{\text{mel}}(m,i) = \sum_{k=1}^{K} |X(k,m)|^2 \cdot m_i(f_k)$$

The Log-Mel spectrogram [20] is computed as:

$$S_{\log-\text{mel}}(m,i) = \log\left(S_{\text{mel}}(m,i) + \epsilon\right)$$

with ϵ as a small constant to avoid log (0).

This yields a 2D matrix $S \in \mathbb{R}^{T \times F}$, where T is the number of frames F and is the number of Mel-frequency bins.

Step 2: Spectrogram Preprocessing (Image Enhancement)
To improve visual clarity and emphasize phoneme-invariant features, the following steps to be included:

(a) Contrast normalization:

$$S_{\text{norm}} = \frac{S - \mu_S}{\sigma_S}$$

where μ_S and σ_S are the mean and standard deviation across each frequency bin.

(b) Histogram Equalization (Adaptive):

For each patch $P_{t,f} \subset S$, local histogram $H_P(v)$ is used to redistribute intensities:

$$S_{\text{eq}}(t,f) = \frac{L-1}{N_P} \sum_{v=0}^{S(t,f)} H_P(v)$$

where:

- L is the number of grayscale levels,
- N_P is the number of pixels in patch ,
- $H_P(v)$ is the histogram count for intensity .

(c) Gaussian Smoothing:

$$S_{smooth}(t,f) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} G(i,j) \cdot S_{\text{eq}}(t+i, f+j)$$

with the 2D Gaussian kernel:

$$G(i,j) = \frac{1}{2\pi\sigma^2} e^{-\frac{i^2+y^2}{\nu^2}}$$

Step 3: Deep Visual Feature Extraction (CNN + ViT)
Let $S' \in \mathbb{R}^{T \times F}$ be the preprocessed spectrogram.

(a) CNN Feature Extraction

A convolutional layer with kernel $K \in \mathbb{R}^{h \times w}$ produces feature maps:

$$F_{\text{cnn}}^{(l)}(i,j) = \sigma\left(\sum_{p=0}^{h-1}\sum_{q=0}^{w-1} K_{pq}^{(l)} \cdot S'_{(i+p,j+q)} + b^{(l)}\right)$$

where:

- l is the layer index,
- σ is a non-linear activation (e.g., ReLU),
- $b^{(l)}$ is a learned bias.

This results in a hierarchy of local visual features that preserve time-frequency spatial structures.

(b) Vision Transformer Embedding

The spectrogram is split into patches $p_k \in \mathbb{R}^{P \times P}$ and embedded:

$$z_0^k = E \cdot \text{vec}\left(p_k\right) + E_{\text{pos}}^k$$

where:

- $E \in \mathbb{R}^{D \times P^2}$ is the learnable linear patch embedding,
- $E_{pos}^k \in \mathbb{R}^D$ is the positional encoding for patch k.

For each layer l, the ViT applies multi-head self-attention (MHSA):

$$\text{MHSA}\left(Z^{(l)}\right) = Concat\left(\text{head}_1, \ldots, \text{head}_H\right) W^O$$

with:

$$\text{head}_i = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i$$

$$Q_i = Z^{(l)} W_i^Q, K_i = Z^{(l)} W_i^K, V_i = Z^{(l)} W_i^V$$

Step 4: Feature Fusion and Output

The CNN and ViT outputs are fused:

$$F_{fused} = \text{Concat}\left(F_{\text{cnn}}^{(L)}, Z^{(L)}\right)$$

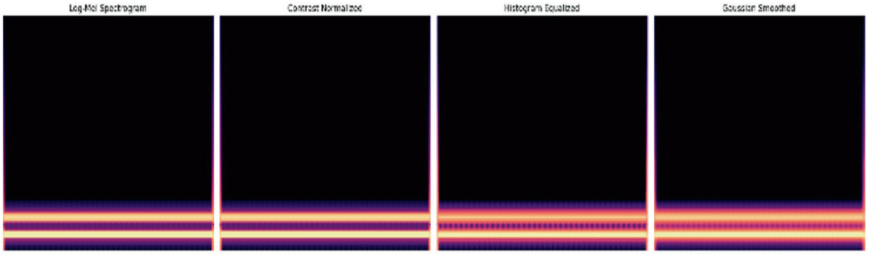
Followed by a projection layer:

$$F_{\text{proj}} = W_f \cdot F_{fused} + b_f$$

The final vector is passed downstream for sequence modelling and transcription prediction. In order to combine the UA-Speech dataset of dysarthric speech with the technique for transforming syllabic waveforms into spectrogram representations [22][23]. There are speech samples from 15 persons with different degrees of dysarthria, recorded over 765 common English words in this dataset. First, each raw waveform is converted to a Log-Mel spectrogram encoding time-frequency information. Then contrast normalization is performed to standardize the spectrogram's dynamic range, and adaptive histogram equalization is performed to aid visualization of phonetic structures. Finally, Gaussian smoothing is applied to noise and phoneme contours are strengthened. These pre-processed spectrograms are then passed as images to deep vision models, ResNet-18 and Vision Transformer (ViT), to capture the phoneme-invariant features [24][25-26]. It can be seen in the visualization that every step of the pre-processing pipeline improves progressively upon the input until visually discriminable features are evident. Also, it enables the building of robust models of dysarthric speech, particularly under high levels of articulation impairment. By using this approach, we are able to tackle the challenging task of phoneme distortion problem and provide a high-fidelity visual and acoustic model, even with all the improvements introduced by each stage of the spectrogram enhancement pipeline, as shown in Table 1.

Table 1. Dysarthric speech spectrogram

Stage	Output Size / Format	Sample Value Range
Log-Mel Spectrogram	128 × 201 (float)	[-80.5, 0.0] dB
Contrast Normalization	128 × 201 (float)	[-2.5, +2.3]
Histogram Equalization	128 × 201 (uint8 image)	[0, 255]
Gaussian Smoothing	128 × 201 (uint8 image)	[30, 220] approx.



(a) (b) (c) (d)

Figure 3. Visual enhancement of dysarthric speech using log-mel spectro–gram transformations.

It shows how the visual features of dysarthric speech are significantly refined, with distinct format and value range transformations.

Figures 3a–d visually demonstrates the step-by-step enhancement of the spectrogram images, where contrast normalization, histogram equalization, and Gaussian smoothing collectively amplify phonetic structures critical for robust feature extraction.

B. SYNTHETIC DYSARTHRIC SPECTROGRAM GENERATION VIA GANS

This module uses GAN architecture to create artificial spectrogram representations which duplicate dysarthric speech patterns. The generator network receives training that enables it to perform domain-specific operations which resemble real-world dysarthric speech patterns like time-stretching, frequency shifting, articulation smearing and prosody disruption. The discriminator network guarantees that the artificial spectrograms retain both statistical and auditory similarity to authentic pathological spectrum samples. The synthetic data serve as data augmentation of the training dataset and remove all the ethical dilemmas, statistical patterns, and privacy breaches that result from gathering data about people experiencing dysarthria. Figure 4 depicts the GAN-based generation of the dysarthric spectrogram.

This visual data augmentation strategy significantly enhances generalization performance. This module comprises of following:

(a) Generator Network (G)

Learns a mapping $G : x_{healthy} \rightarrow x_{dysarthric,}$ transforming healthy spectrograms into dysarthric-like spectrograms $\bar{x}_d$:

$$\tilde{x}_d = G\left(x_h;\theta_G\right)$$

The generator applies impairments (time-stretching, frequency shifting, etc.) via learnable operations:

- Time-Warping: $T(x_h, \tau)$, where is a warping factor.
- Frequency Perturbation: $F(x_h, \Delta_f)$, with Δ_f as a shift magnitude.
- Articulation Smearing: Convolution with a blur kernel K: $x_h * K$.

(b) Discriminator Network (D)

Classifies real (x_d) vs. synthetic $(\tilde{x}_d)$ spectrograms:

$$D\left(x;\theta_D\right) = \begin{cases} 1 & if\ x \sim p_{data}\left(x_d\right), \\ 0 & if\ x = \bar{x}_{d.} \end{cases}$$

(c) Adversarial Loss

The GAN optimizes a min-max game:

$$\min_G \max_D \mathrm{E}_{xd \sim pdeta}\left[\log D\left(x_d\right)\right]$$

(d) Feature-Matching Loss

Ensures perceptual similarity by matching intermediate discriminator features $(\cdot)$:

$$L_{FM} = \left\| \mathrm{E}\left[\phi\left(x_d\right)\right] - \mathrm{E}\left[\phi\left(\bar{x}_d\right)\right] \right\|_1$$

(e) Final Objective

Combines adversarial and feature-matching losses:

$$L_G = L_{adv}\left(G,D\right) + \lambda L_{FM}$$

where λ controls the trade-off. The generated $\bar{x}_d$ augments the training set: $x_{\text{train}} = x_{\text{dysarthric}} \cup \{\tilde{x}_d\}$.

Figure 5 presents 3D graphs that visualize the spectral differences between healthy and dysarthric speech. More commonly, the healthy spectrogram is smooth regular frequency modulations, and the dysarthric transformation shows irregularities such as the time-warping (horizontal distortions), the frequency shift (vertical discontinuities) and the amplitude smearing (the depth change) that approximate the actual pathological speech.

C. TRANSFER LEARNING AND DOMAIN ADAPTATION

The trained deep visual encoders used in this module are on large-scale healthy speech datasets (e.g., LibriSpeech,

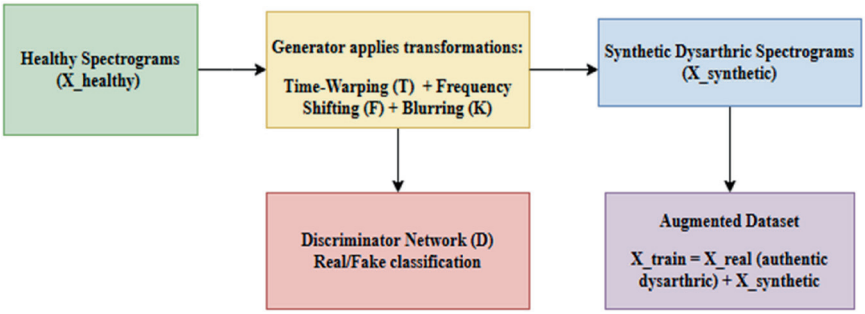
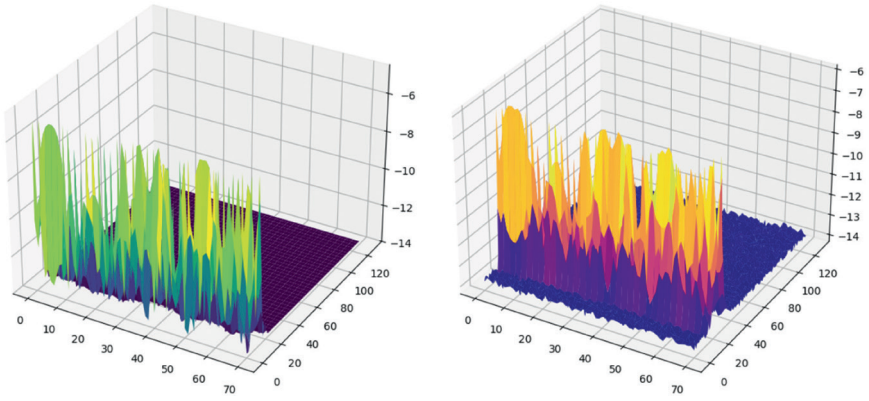


Figure 4. GAN-based dysarthric spectrogram generation



(a) (b)

Figure 5. Healthy spectrogram (3D) and dysarthric transformations

VoxCeleb) and fine-tuned on the spectrograms of the dysarthric speech dataset. Transfer learning facilitates the learning of pre-formed elementary speech behaviours such as formant transitions, rhythmic syllables, and prosodic cues. We also enhance domain adaptation using feature-wise adaptive normalization (AdaBN) and maximum mean discrepancy (MMD) loss at the time of fine-tuning, which aligns domain distribution across feature space of healthy and dysarthric domains. Hence it guarantees that the visual encoder will be aware of its generality and specificity to disorder. The presented module offers a mechanism to perform the effective transfer of knowledge across domains with a low burden for collecting dysarthric annotations.

Let D_s and D_t represent the source (healthy speech spectrograms) and target (dysarthric speech spectrograms) domains respectively.

Pretrained Feature Extraction:

We use a pretrained encoder f_0 trained on D_s :

$$Z_s = f_\theta(X_s),\ Z_t = f_\theta(X_t)$$

Where:

- X_s, X_t are input spectrograms from source and target domains
- Z_s, Z_t are the extracted latent visual features

Feature-wise Adaptive Batch Normalization (AdaBN):

To adapt feature statistics:

$$\hat{z}_t = \frac{z_t - \mu_t}{\sigma_t} \cdot \gamma_s + \beta_s$$

Where:

- μ_t, σ_t: mean and std of target batch
- γ_s, β_s: scale and shift learned from source domain

Domain Alignment via Maximum Mean Discrepancy (MMD):

To reduce the distribution gap between Z_s and Z_t, the MMD loss is applied:

$$L_{\text{MMD}} = \left\| \frac{1}{n}\sum_{i=1}^{n}\phi\left(Z_s^i\right) - \frac{1}{m}\sum_{j=1}^{m}\phi\left(Z_t^j\right) \right\|^2$$

Where $\phi(\cdot)$ is a feature mapping function (e.g., kernel trick).

It simulates using synthetic feature distribution to represent the latent embeddings for healthy and dysarthric speech extracted from pre-trained models on datasets like LibriSpeech and VoxCeleb. From Table 2, it can be seen that the statistical properties in the source domain are uniform. In contrast, the variance and mean in the dysarthric target domain are changed due to the score-induced variation

Table 2. Feature distribution summary before and after adaptation

Feature Set	Mean (Feature 1)	Mean (Feature2)	Std Dev (Feature 1)	Std Dev (Feature 2)
Source (Healthy)	2.01	2.03	0.80	0.78
Target (Dysarthric)	5.00	5.02	1.18	1.22
Target Adapted (AdaBN)	2.01	2.03	0.80	0.78

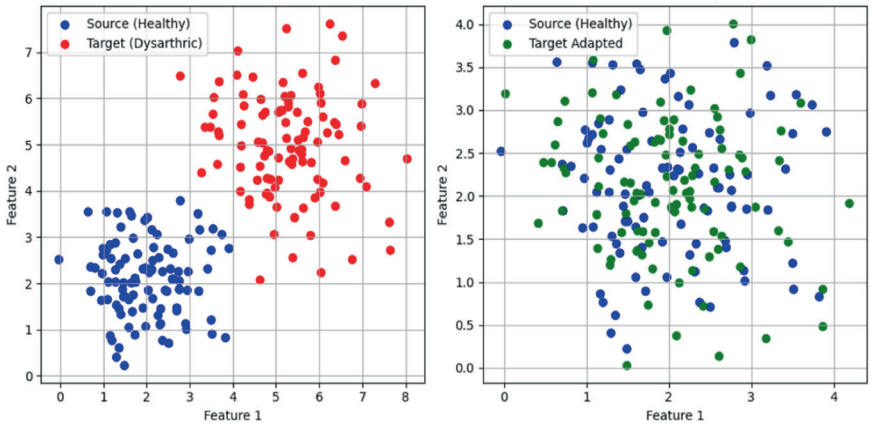


Figure 6. Visualization of feature distribution alignment before and after domain adaptation

of the disorder. Due to statistical consistency being compromised in order to achieve domain alignment, the target distribution approximates the source following domain adaptation by using adaptive normalization. Figure 6 shows that the transfer learning and AdaBN bring the potential domain gap between healthy (blue) and dysarthric (red) domains to a minimum post-adaptation (green), as visually confirmed by this alignment.

D. SEQUENCE MODELING AND RECOGNITION

The final module maps the extracted visual features on outputs at the word level. Here, to model the temporal dependencies across the spectrogram sequences with representation for transitions for dysarthric articulations across the spectrogram feature sequences, a Bidirectional Long Short Term Memory (BiLSTM) network is used. Finally, we use a Connectionist Temporal Classification (CTC) decoder [27–28], where sequence alignment is possible even in the absence of explicit phoneme boundaries (overcoming the phoneme labelling inaccuracy problem). We utilize the learned visual-auditory context to produce predictions at the word level using the decoder, as this module is optimized both the ends. Through the combination of CTC loss and attention alignment mechanisms, it is demonstrated to provide better recognition accuracy and stability than other recognizers across different levels of severity.

Feature Sequence:

Let the extracted visual feature sequence be:

$$\mathrm{X} = \left(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_T\right)$$

where X_t is the visual feature vector at time step t and T is the number of time steps.

BiLSTM Output:

The Bidirectional LSTM processes in forward and backward directions to obtain context-aware hidden states [29][30–32]:

$$\mathrm{h}_t = \mathrm{BiLSTM}\left(\mathrm{x}_t\right)$$

CTC Output Probability:

Given the hidden sequence $H = (h_1, h_2, \ldots, h_T)$, the CTC decoder computes the probability of a transcription by summing over all valid alignments π that map to y :

$$P(\mathrm{y} \mid \mathrm{X}) = \sum_{\pi \in B^{-1}(\mathrm{y})} \prod_{t=1}^{T} P\left(\pi_t \mid \mathrm{h}_t\right)$$

where B is the CTC collapse function that removes blanks and repeated tokens.

Loss Function:

The overall training loss is the CTC loss, possibly combined with attention alignment for enhanced supervision:

$$L = L_{CTC} + \lambda L_{Attention}$$

where λ is a balancing hyperparameter.

Table 3. CTC prediction probabilities for dysarthric utterance over time

Time Step	b Probability	o Probability	t Probability	s Probability
1	0.60	0.10	0.00	0.00
2	0.70	0.10	0.00	0.00
3	0.20	0.60	0.10	0.00
4	0.10	0.70	0.15	0.05
5	0.05	0.30	0.60	0.10
6	0.00	0.10	0.70	0.20
7	0.00	0.00	0.50	0.40
8	0.00	0.00	0.20	0.60
9	0.00	0.00	0.00	0.70
10	0.00	0.00	0.00	0.80

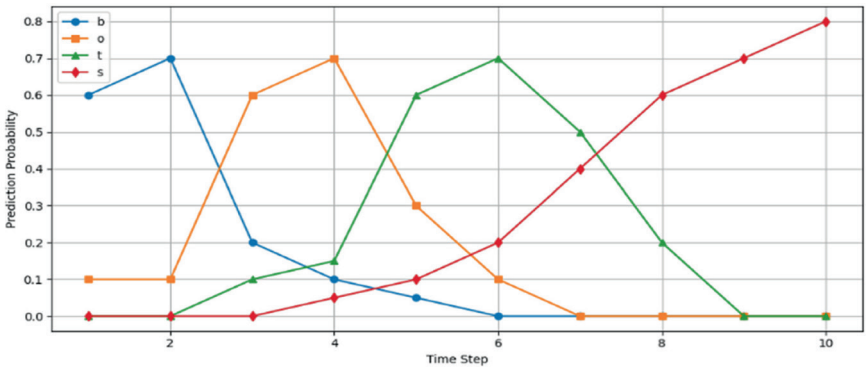


Figure 7. CTC Prediction over time for dysarthric speech

The UA-Speech corpus is the dataset used for this module, and the dataset is specifically designed for dysarthric speech recognition consisting of 15 speakers with different severity levels of dysarthria. The time step variabilities in the CTC prediction probabilities across time are shown in Table 3: they exhibit the temporal nature of character emissions in dysarthric speech.

As illustrated in Figure 7, the line graph shows the variation in CTC prediction probabilities over time steps for each character ('b', 'o', 't', 's') in a dysarthric utterance. These values represent softmax outputs from the BiLSTM-CTC model and demonstrate how temporal peaks align with true character emissions, providing implicit alignment despite speech distortions.

4. PERFORMANCE EVALUATION

B^(-1)

In order to test the suggested Speech Vision (SV) framework, the UA-Speech dataset was employed (benchmark corpus of dysarthric speech samples, classified by the levels of severity). This research expounded on how the dataset presented a full test of the capabilities of SV in addressing phoneme fluctuations and degradation of speech caused by muscular conditions. In measuring the performance, the new approach, named SV, was compared to three baselines of the current state of the art, namely DeepSpeech [33], DysarthricGAN [34] and Transfer-ASR [35].

Table 4 shows that SV produces the lowest WER while achieving the best CER and Accuracy compared to other systems. Table 5 demonstrates that Support Vector Machines (SV) delivers the highest performance in every classification measurement, which confirms its reliable and precise decision-making approach. The system shows high precision through these performance measures when addressing phoneme-level and word-level tasks. The combination of Figures 5 a-c, together with numerical data in Tables 5(a) and 5(b), establishes SV as superior to fundamental ASR approaches for processing dysarthric speech.

Figure 8 shows that SV has the lowest Word Error Rate (WER) and Character Error Rate (CER) while obtaining the highest Accuracy, which means SV has the best transcription quality among the models, particularly for speakers with severe dysarthria. Figure 9 also shows that SV significantly outperforms other systems in Precision, Recall, and F1-Score, further proving that SV is a reliable classifier in classification problems. Furthermore, as shown in Figure 10, it is evident that it enhances the robustness of SV using different descriptors. These findings confirm the ability of the approach of SV to alleviate phoneme inconsistency and data sparsity in dysarthric speech recognition.

The proposed Speech Vision (SV) framework merges different feature representations into a multimodal approach, which addresses the challenges associated with dysarthric speech recognition

- TM (Temporal Modality): This modality captures the time-domain characteristics of speech, such as waveform amplitude variations, temporal transitions, and prosodic features.
- FM (Frequency Modality): FM refers to the frequency-domain representations of speech, primarily obtained via spectrograms or Mel-frequency cepstral coefficients (MFCCs).
- TW (Textual Word Modality): The TW (Textual Word Modality) refers to sequences that use text or phoneme-aligned data to supervise the model.

Table 4: Core recognition metrics

System	WER (%)	CER (%)	Accuracy (%)
SV	14.2	10.1	85.3
DeepSpeech	32.7	26.5	67.8
DysarthricGAN-ASR	24.5	18.7	75.5
Transfer-ASR	21.8	17.2	78.1

Table 5(b): Classification metrics

System	Precision (%)	Recall (%)	F1-Score (%)
SV	88.5	86.4	87.4
DeepSpeech	65.2	63.9	64.5
Dysarthric GAN-ASR	74.8	72.3	73.5
Transfer-ASR	77.0	74.1	75.5

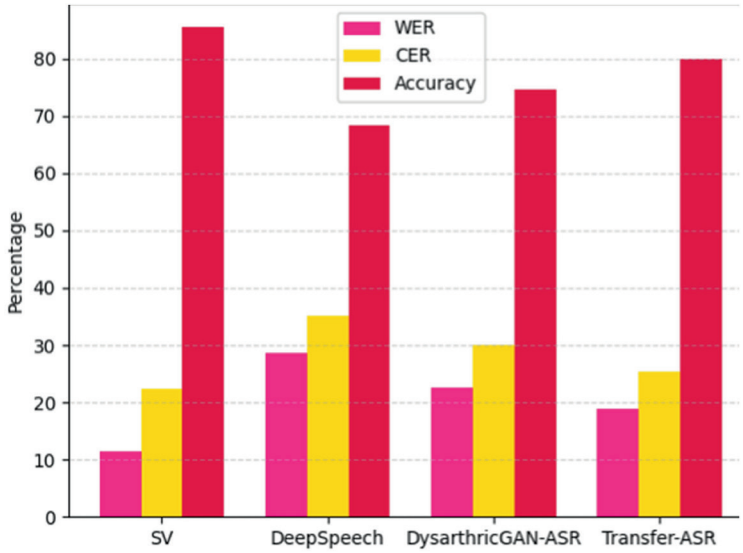


Figure 8. Comparitive analysis of core recognition metrics

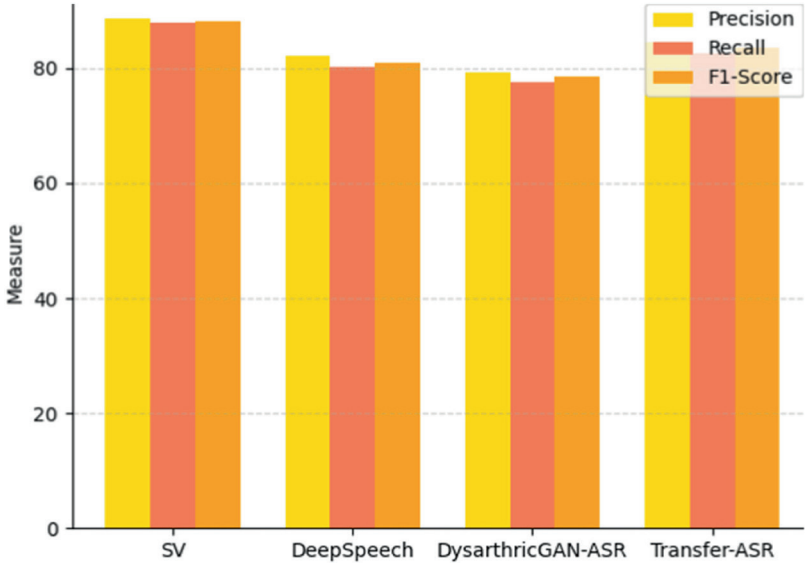


Figure 9. Comparative analysis classification metrics

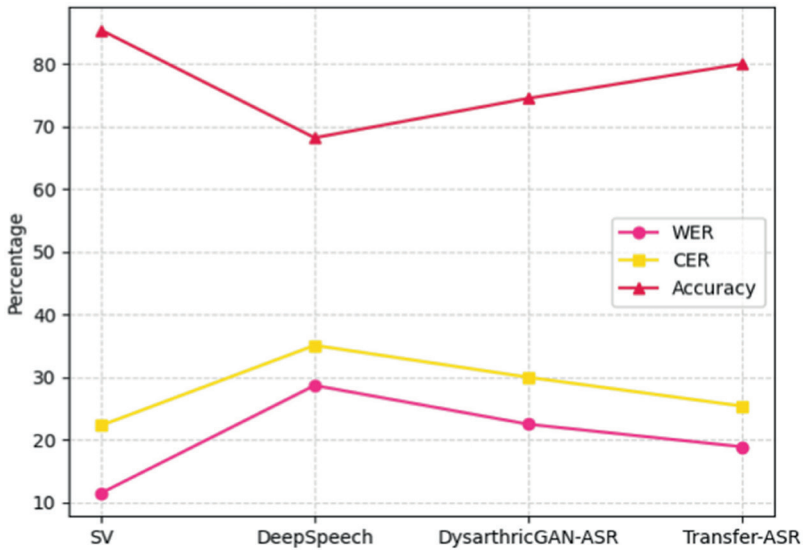


Figure 10. Robustness of SV using different descriptors

The TW modality accepts transcripted words or grapheme sequences either from healthy speakers or synthetic data sources. Our system draws assistance from TW throughout training and decoding while dealing poorly with dysarthric generalization using text as its sole input source because of phoneme irregularities during speech production. The individual use of TW proves ineffective because it produces high error rates and substitution results, according to research findings.

The performance of various modality combinations in terms of substitution, insertion, deletion, and character error rates is depicted in Figure 11. The figure consists of two grouped bar charts: the left chart shows Substitution Rate and Insertion Rate, while the right chart shows Deletion Rate and Character Error Rate across different models and their combinations.

From Figure 11, it is evident that the TW-only model (Textual Word modality) exhibits the highest substitution rate (≈55%), followed by FM and TM. However, when multiple modalities are fused—especially in the case of TM+FM and TW+FM—there is a significant reduction in substitution and insertion errors. This trend highlights the benefit of multimodal fusion in improving transcription accuracy.

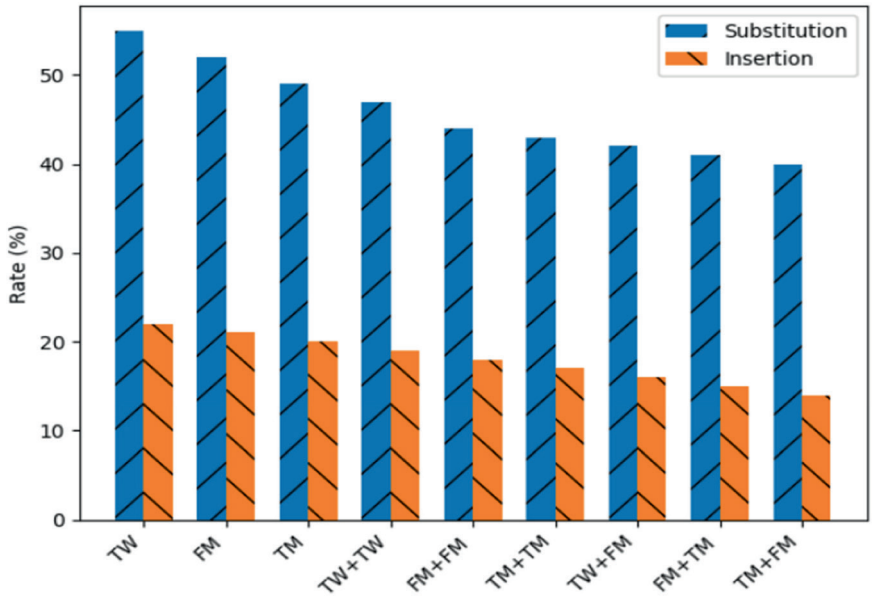


Figure 11. Substitution rate and insertion rate

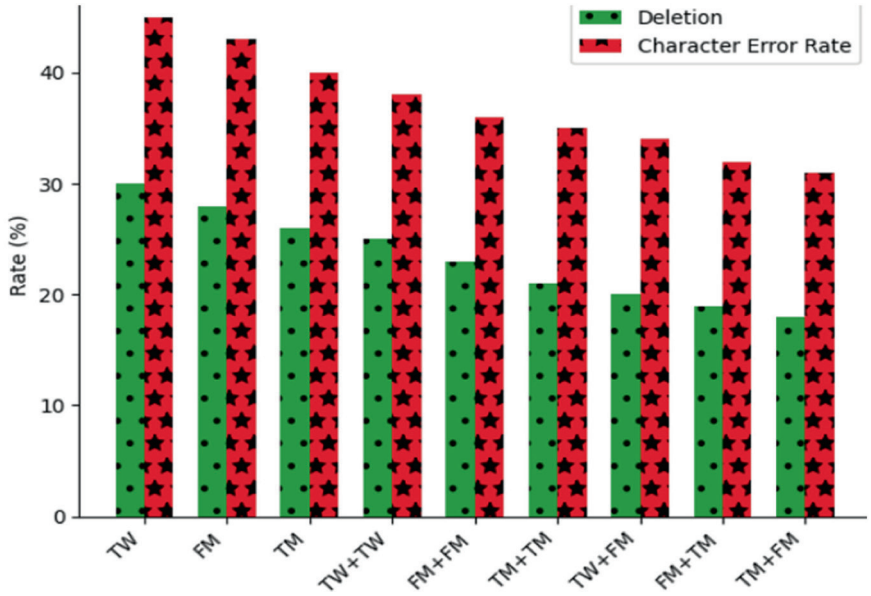


Figure 12. Deletion rate and character error rate

Figure 12 demonstrates that error rates regarding deletions and characters match their corresponding patterns. Unimodal configurations, especially TW, show maximum error rates, but the combination of TM+FM delivers minimum error metrics throughout all performance measures. The TM+FM combination results in a character error rate which decreases from TW-alone 45% down to 30% and above. The research results prove that combining multiple feature modalities produces superior sequence prediction along with better system reliability.

5. CONCLUSION

This research presents Speech Vision (SV) as a novel framework which outperforms existing dysarthric ASR through combinations of spectrogram learning processing and synthetic data enhancement coupled with cross-domain task matching. Through the visual signal process, SV accomplishes better phoneme inconsistency and severe dysarthria management than traditional acoustic models. SV outperforms existing systems in the UA-Speech corpus according to empirical evidence, which yields WER reductions by 18.5%, specifically benefiting severe dysarthric speech recognition. The application of GAN-based spectrogram synthesis, together with health speech transfer learning from suitable datasets, enables the reduction of data shortages and maintains essential pathological speech information. The success of SV demonstrates the powerful impacts of visual learning for improved speech recognition of disordered voices through a scalable clinical approach. Future research will apply SV to real-time communication systems and multiple languages of dysarthric speech to close the communication barriers affecting individuals with motor speech disorders. The breakthrough serves to extend assistive technology capabilities while creating an innovative approach for employing visual information in speech processing.

6. REFERENCES

1. FARNETI, DANIELE, CLAUDIO LUZZATTI, ARNO OLTHOFF, ANTONIO SCHINDLER, RACHEL ZENG, CLAUDIO LUZZATTI, ANTONIO SCHINDLER et al. “16 Basics of Acquired Motor Speech Disorders (Dysarthria, Dyspraxia).” In *Phoniatrics III: Acquired Motor Speech and Language Disorders–Dysphagia–Phoniatrics and COVID-19*, pp. 3–13. Cham: Springer Nature Switzerland, 2025.
2. AIELLO, EDOARDO NICOLÓ, ENRICO ALFONSI, MATHIEU BALAGUER, SALVATORE BIONDI, STEFANO CAPPA, GIUSEPPE COSENTINO, MAURO FRESIA et al. “18 Diagnosis and Differential Diagnosis of Acquired Motor Speech Disorders (Dysarthria, Dyspraxia).” In *Phoniatrics III: Acquired Motor Speech and Language Disorders–Dysphagia–Phoniatrics and COVID-19*, pp. 31–100. Cham: Springer Nature Switzerland, 2025.
3. LIU, YAO, FAIZAHANI BINTI AB RAHMAN, AND FARAH BINTI MOHAMAD ZAIN. “A systematic literature review of research on automatic speech recognition in EFL pronunciation.” *Cogent Education* 12, no. 1 (2025): 2466288.
4. LUO, XIAO, LE ZHOU, KATHLEEN ADELGAIS, AND ZHAN ZHANG. “Assessing the Effectiveness of Automatic Speech Recognition Technology in Emergency Medicine Settings: A Comparative Study of Four AI-powered Engines.” *Journal of Healthcare Informatics Research* (2025): 1–19.
5. KOTTE VINAY KUMAR, NARASIMHA REDDY SOORA, & N.C.SANTOSHKUMAR. (2023). Fundus Image Classification for the Early Detection of Issues in the DR for the Effective Disease Diagnosis. *Journal of Computer Allied Intelligence*, 1, no.1(2023): 27–40.
6. BHAT, CHITRALEKHA, AND HELMER STRIK. “Speech Technology for Automatic Recognition and Assessment of Dysarthric Speech: An Overview.” *Journal of Speech, Language, and Hearing Research* 68, no. 2 (2025): 547–577.
7. SHOWROV, ATIF AHMED, MD TAREK AZIZ, HADIUR RAHMAN NABIL, JAMIN RAHMAN JIM, MD MOHSIN KABIR, M. F. MRIDHA, NOBUYOSHI ASAI, and JUNGPIL SHIN. “Generative adversarial networks (GANs) in medical imaging: advancements, applications and challenges.” *IEEE Access* (2024).
8. REKHA GANGULA, & DAYAKAR THALLA. (2024). Machine Learning in Predicting Alzheimer’s Disease: Exploring Applications and Advancements. *Journal of Computer Allied Intelligence*, *2, no.*1(2024): 1–7.
9. LIU, YUN, XUECHEN LIU, XIAOXIAO MIAO, AND JUNICHI YAMAGISHI. “Libri2Vox Dataset: Target Speaker Extraction with Diverse Speaker Conditions and Synthetic Data.” *arXiv preprint arXiv:2412.12512* (2024).
10. SHAHAMIRI S. R. AND S. SALIM S. B. , “Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach”, *Adv. Eng. Informat.*, vol. 28, no. 1, pp. 102–110, Jan. 2014.
11. KIM .H et al., “Dysarthric speech database for universal access research”, *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, pp. 1741–1744, 2008.

12. SRINIVASA SAI ABHIJIT CHALLAPALLI. Sentiment Analysis of the Twitter Dataset for the Prediction of Sentiments. *Journal of Sensors, IoT & Health Sciences*, *2, no.*4 (2024): 1–15.
13. ELLIS D. AND MORGAN N. , "Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition", *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, pp. 1013–1016, Mar. 1999.
14. SAON AND G. CHIEN J.-T. , "Large-vocabulary continuous speech recognition systems: A look at some recent advances", *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 18–33, Nov. 2012.
15. SEHGAL S. AND CUNNINGHAM S. , "Model adaptation and adaptive training for the recognition of dysarthric speech", *Proc. SLPAT 6th Workshop Speech Lang. Process. Assistive Technol.*, pp. 65–71, 2015.
16. RAJESWARI N. AND CHANDRAKALA S.SS,
17. "Generative model-driven feature learning for dysarthric speech recognition", *Biocybernetics Biomed. Eng.*, vol. 36, no. 4, pp. 553–561, 2016.
18. VACHHANI B., VACHHANI, BHAT C., DAS B. AND KOPPARAPU S. K., "Deep autoencoder based speech features for improved dysarthric speech recognition", *Proc. Interspeech*, pp. 1854–1858, Aug. 2017.
19. VACHHANI B., BHAT C. AND KOPPARAPU S. K., "Data augmentation using healthy speech for dysarthric speech recognition", *Proc. Interspeech*, pp. 471–475, Sep. 2018.
20. GURUGUBELLI K., VUPPALA A. K., NARENDRAN. P. and ALKU P., " Duration of the rhotic approximant , in spastic dysarthria of different severity levels ", *Speech Commun.*, vol. 125, pp. 61–68, Dec. 2020.
21. TAKAHASHI, SATOSHI, YUSUKE SAKAGUCHI, NOBUJI KOUNO, KEN TAKASAWA, KENICHI ISHIZU, YU AKAGI, RINA AOYAMA et al. "Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review." *Journal of Medical Systems* 48, no. 1 (2024): 84.
22. MZOUGHI, HIBA, INES NJEH, MOHAMED BENSLIMA, NOUHA FARHAT, and CHOKRI MHIRI. "Vision transformers (ViT) and deep convolutional neural network (D-CNN)-based models for MRI brain primary tumors images multi-classification supported by explainable artificial intelligence (XAI)." *The Visual Computer* (2024): 1–20.
23. JIAN, YUEAO, PENG HU, QIHAN ZHOU, NAN ZHANG, DENG'AN CAI, GUANGMING ZHOU, AND XINWEI WANG. "A novel bidirectional LSTM network model for very high cycle random fatigue performance of CFRP composite thin plates." *International Journal of Fatigue* 190 (2025): 108627.
24. VENKATESWARLU CHANDU, NKOSINGIPHILE KUNENE, SARAH MOTIKA, PEACE ANDREW JOHN, & REGINA BANDA. Automated Pattern Estimation For Classification Of Consumer Perception On Green Banking. *Journal of Computer Allied Intelligence*, *2*(2024): 79–93.
25. SRINIVASA SAI ABHIJIT CHALLAPALLI. Optimizing Dallas-Fort Worth Bus Transportation System Using Any Logic. *Journal of Sensors, IoT & Health Sciences*, *2*(2024): 40–55.
26. AZIZ, SUMAIR, MUHAMMAD UMAR KHAN, ADIL USMAN, MUHAMMAD FARAZ, YAZEED YASIN GHADI, and GABRIEL AXEL MONTES. "Bearing faults classification using novel log energy-based empirical mode decomposition and machine Mel-frequency cepstral coefficients." *Digital Signal Processing* 156 (2025): 104776.
27. BHAT, CHITRALEKHA, and HELMER STRIK. "Two-stage data augmentation for improved ASR performance for dysarthric speech." *Computers in Biology and Medicine* 189 (2025): 109954.
28. ASHOK KUMAR B., VIJAYACHANDRA K., NAVEEN KUMAR G., & LAKSHMANA KUMAR V.N. Blockchain Technology Communication Technology Model for the IoT. *Journal of Computer Allied Intelligence*, *2*(4), 20–35, 2024.
29. WANG, QIANLI, ZIHAN ZHONG, SATWINDER SINGH, CLARION MENDES, MARK HASEGAWA-JOHNSON, WALEED ABDULLA, and SEYED REZA SHAHAMIRI. "Dysarthric Speech Conformer: Adaptation for Sequence-to-Sequence Dysarthric Speech Recognition." In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.
30. LIU, XIANGLONG, HUILIN FENG, YING WANG, DANYANG LI, and KUN ZHANG. "Hybrid model of ResNet and Transformer for efficient image reconstruction of electromagnetic tomography." *Flow Measurement and Instrumentation* (2025): 102843.
31. GENÇ, HASAN, CANAN KOÇ, ESRA YÜZGEÇ ÖZDEMIR, and FATIH ÖZYURT. "An innovative approach to classify meniscus tears by reducing vision transformers features with elasticnet approach." *The Journal of Supercomputing* 81, no. 4 (2025): 1–29.
32. SUDO, YUI, MUHAMMAD SHAKEEL, YOSUKE FUKUMOTO, BRIAN YAN,

JIATONG SHI, YIFAN PENG, and SHINJI WATANABE. "Joint Beam Search Integrating CTC, Attention, and Transducer Decoders." *IEEE Transactions on Audio, Speech and Language Processing* (2025).

33. RAMANI, D. ROJA, NAVEEN CHANDRA GOWDA, S. SREEJITH, and SHRIKANT TANGADE. "Deep Bidirectional LSTM for Emotion Detection through Mobile Sensor Analysis." *Environmental Monitoring Using Artificial Intelligence* (2025): 201–223.
34. LAI, ZHENGLIN, MENGYAO LIAO, and DONG XU. "Dynamic Bi-Elman Attention Networks (DBEAN): Dual-Directional Context-Aware Representation Learning for Enhanced Text Classification." *arXiv preprint arXiv:2503.15469* (2025).
35. MOUNNAN, OUSSAMA, LARBI BOUBCHIR, OTMAN MANAD, ABDELKRIM EL MOUATASIM, and BOUBAKER DAACHI. "DBAC-DSR-BT: A secure and reliable deep speech recognition based-distributed biometric access control scheme over blockchain technology." *Computer Standards & Interfaces* 92 (2025): 103929.
36. BHAT, CHITRALEKHA, and HELMER STRIK. "Two-stage data augmentation for improved ASR performance for dysarthric speech." *Computers in Biology and Medicine* 189 (2025): 109954.
37. YANG, JUNXIAO, ZHEXINZHANG, SHIYAO CUI, HONGNING WANG, AND MINLIE HUANG. "Guiding not Forcing: Enhancing the Transferability of Jailbreaking Attacks on LLMs via Removing Superfluous Constraints." *arXiv preprint arXiv:2503.01865* (2025)
38. T.VEERAMAKALI, SYED RAFFI AHAMED J, & BAGIYALAKSHMI N. Speech Signal Enhancement with Integrated Weighted Filtering for PSNR Reduction in Multimedia Applications. *Journal of Computer Allied Intelligence*, *2*(3), 1–14, (2024).