INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
*(Volume 11, Issue 2 - V11I2-1282)*
Available online at: https://www.ijariit.com

# Survey Paper on Advancements in Dysarthric Speech Recognition Systems

*Sushmita Chaudhari*
*sushmitachaudhari90966@gmail.com*
*Pune Vidhyarthi Grihas College of Engineering and Technology, Pune, Maharashtra*

*Mansi Chopkar*
*chopkarmansi@gmail.com*
*Pune Vidhyarthi Grihas College of Engineering and Technology, Pune, Maharashtra*

*Harshvardhan Gaikwad*
*21113028@pvgcoet.ac.in*
*Pune Vidhyarthi Grihas College of Engineering and Technology, Pune, Maharashtra*

*Anuj Raj*
*21023075@pvgcoet.ac.in*
*Pune Vidhyarthi Grihas College of Engineering and Technology, Pune, Maharashtra*

## ABSTRACT

***Dysarthria, a motor speech disorder resulting from neurological injuries, severely impairs intelligibility, making automatic speech recognition (ASR) a vital tool for enhancing communication. Over the years, significant research has explored computational approaches to improve ASR performance for dysarthric speech, from early rule-based models to deep learning architectures. This survey presents a comprehensive review of the evolution of ASR techniques tailored to dysarthric speech, categorizing methods by architecture type (HMM, DNN, CNN, LSTM, Transformers), learning paradigm (supervised, self-supervised, meta-learning), and input modality (audio-only, multimodal). The study examines the role of acoustic features like MFCC, PLP, and raw waveform-based learning, and compares key models including Wav2Vec2.0, TDNN, and UTran-DSR across datasets such as UA-Speech, TORGO, and CommonVoice. A critical evaluation of strategies like speaker adaptation, transfer learning, end-to-end pipelines, and contrastive learning is provided, along with their impact on accuracy and generalization. The paper highlights emerging trends such as emotion-aware ASR, multimodal fusion, and personalized adaptation, while addressing persistent challenges including data scarcity, speaker variability, and real-time deployment. This survey aims to provide a clear roadmap of the progress and ongoing efforts in dysarthric ASR, guiding future research toward more inclusive and intelligent speech interfaces.***

**Keywords:** *Dysarthric Speech Recognition, Automatic Speech Recognition (ASR), Model Adaptation, Speech Synthesis, Data Augmentation, Deep Learning for Speech Disorders*

## 1. INTRODUCTION

Automatic Speech Recognition (ASR) systems have witnessed remarkable advancements with deep learning and self-supervised architectures, yet their performance remains significantly challenged when applied to dysarthric speech—an impaired form of articulation resulting from neurological disorders. Unlike typical ASR tasks, dysarthric speech recognition demands specialized models due to high variability in speech patterns, limited annotated datasets, and speaker-dependent distortions. This survey presents a comprehensive analysis of developments in dysarthric ASR, encompassing the evolution of model architectures, enhancement strategies, dataset utilization, feature extraction, and comparative research trends. By highlighting existing challenges and recent breakthroughs, we aim to offer insights into the design of robust, adaptive ASR systems for impaired speech, paving the way for more inclusive and accessible speech technologies.

## 2. ARCHITECTURAL EVOLUTION

The field of automatic speech recognition (ASR) for dysarthric speech has progressed from early statistical methods to advanced deep learning and self-supervised architectures capable of modeling impaired and inconsistent speech patterns.

This evolution has been driven not only by architectural advancements—ranging from HMMs to Transformers—but also by enhancement strategies such as speaker adaptation, transfer learning, and contrastive learning. This section presents a year-wise overview of how ASR models and training strategies have co-evolved to improve recognition accuracy and generalization in dysarthric speech recognition.

## 2.1. Evolution Of Models In Dysarthric Speech Recognition

The progression of ASR architectures has been central to improving recognition performance on dysarthric speech. Early statistical models like Hidden Markov Models (HMMs) provided basic temporal modeling but were limited in handling variability. As deep learning emerged, architectures evolved from shallow ANNs to DNNs, CNNs, and LSTMs, offering better generalization. Recent advancements such as Transformer-based models and self-supervised frameworks like Wav2Vec2.0 have further enhanced accuracy while reducing dependence on large labeled datasets. This subsection outlines the year-wise development of these models and their growing effectiveness in dysarthric ASR.

### 2.1.1. Hidden Markov Models (1993)

The earliest attempts at automatic speech recognition (ASR) for dysarthric speech relied on Hidden Markov Models (HMMs), which were commonly applied to datasets like Whitaker and Nemours. These models were effective at capturing temporal dependencies in speech sequences and were widely used for isolated word recognition. However, HMMs struggled with the inconsistency and variability of dysarthric speech, especially in the presence of slurring and abnormal articulation patterns, which limited their real-world applicability.

### 2.1.2. Hybrid HMM-ANN Models (2000-2010)

In the early 2000s, researchers began combining HMMs with Artificial Neural Networks (ANNs) to improve recognition accuracy. These hybrid systems retained the temporal modeling capabilities of HMMs while leveraging ANNs for better pattern recognition. Despite showing moderate improvements, the neural networks used at this stage were relatively shallow, which restricted their ability to generalize across speakers and speech types. The models continued to rely on handcrafted acoustic features such as MFCCs, FFT, and LPC, and showed only incremental gains in performance.

### 2.1.3. Deep Neural Networks(2011)

The introduction of Deep Neural Networks (DNNs) marked a significant turning point in dysarthric ASR. These multi-layered architectures allowed models to extract higher-level abstract features from speech inputs. Studies like those by Joy and Umesh demonstrated that DNNs, when paired with speaker normalization techniques, could outperform traditional systems in speaker-dependent scenarios. Although still reliant on engineered features like MFCCs, DNNs laid the groundwork for deep learning in dysarthric speech processing.

### 2.1.4. Autoencoder-Enhanced DNNs(2014)

In 2014, autoencoders were integrated into ASR pipelines to enhance input features before they were processed by DNNs. Vachhani et al. proposed a system where MFCC features were pre-processed using autoencoders trained on normal speech, then passed into a DNN trained on dysarthric speech. This approach resulted in an absolute accuracy gain of 16% on the UA-Speech dataset, proving that unsupervised feature learning could significantly improve recognition performance for disordered speech.

### 2.1.5. Convolutional Neural Networks(2015)

CNNs began gaining attention in 2015 for their ability to model spatial patterns in spectrograms. Zaidi et al. applied CNNs to the Nemours database using PLP features as input. The networks captured localized time-frequency correlations within speech data, yielding noticeable improvements in word recognition. However, CNNs lacked the capability to model long-range temporal dynamics, limiting their standalone effectiveness for continuous dysarthric speech.

### 2.1.6. Long Short-Term Memory Networks(2016)

To better address the temporal variability of speech, researchers adopted Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) models. These networks were capable of learning time-based dependencies, making them suitable for slurred or slowly articulated speech. Xiong et al. improved recognition performance by combining acoustic features with articulatory movement data, demonstrating the benefit of integrating auxiliary inputs with LSTM-based architectures.

### 2.1.7. Time Delay Neural Networks(2017)

TDNNs offered a simpler, more efficient alternative to RNNs for capturing temporal context. These models operate over fixed time windows, enabling them to learn frame-level dependencies without requiring recurrent connections. TDNNs were especially appealing due to their faster training time and compatibility with feature extraction pipelines. Later studies incorporated TDNNs with transfer learning techniques and bottleneck autoencoders for improved speaker adaptation.

### 2.1.8. Transformer-Based ASR(2020)

Transformers revolutionized sequence modeling in natural language processing and found their way into ASR by 2020. These models used self-attention mechanisms to capture global dependencies in input sequences, allowing them to process entire speech utterances in parallel. In dysarthric ASR, transformers demonstrated strong performance, particularly in multilingual and cross-lingual setups. However, they required substantial training data and computational resources, limiting their accessibility in low-resource environments.

#### 2.1.9. Wav2Vec2.0 (2021)

The release of Wav2Vec2.0 by Facebook AI introduced self-supervised learning to speech recognition. This model learned rich audio representations directly from raw waveform inputs without the need for labeled data. Once pre-trained on large normal speech corpora, it could be fine-tuned on small dysarthric datasets like UA-Speech, achieving over 85% accuracy. Its ability to generalize across speakers and adapt with minimal supervision made it a landmark model in dysarthric ASR.

#### 2.1.10. DyPCL (2022)

In 2022, the DyPCL model brought contrastive learning into dysarthric speech recognition. It was designed to distinguish between subtle speech variations among different dysarthria severities and speakers. By leveraging contrastive pretraining on patient-level samples, DyPCL achieved a 22.1% relative reduction in word error rate, highlighting the effectiveness of patient-aware representations in disordered speech contexts.

#### 2.1.11. UTran-DSR (2023)

The UTran-DSR model extended standard transformer architectures by incorporating speaker-aware and emotion-aware layers. This hybrid architecture allowed the system to model not only phonetic variability but also affective cues in speech, which are often altered in dysarthric speakers. UTran-DSR achieved state-of-the-art results in emotion- and speaker-generalizable ASR systems for dysarthric speech.

#### 2.1.12. Multimodal Audio-Visual Fusion Models (2024)

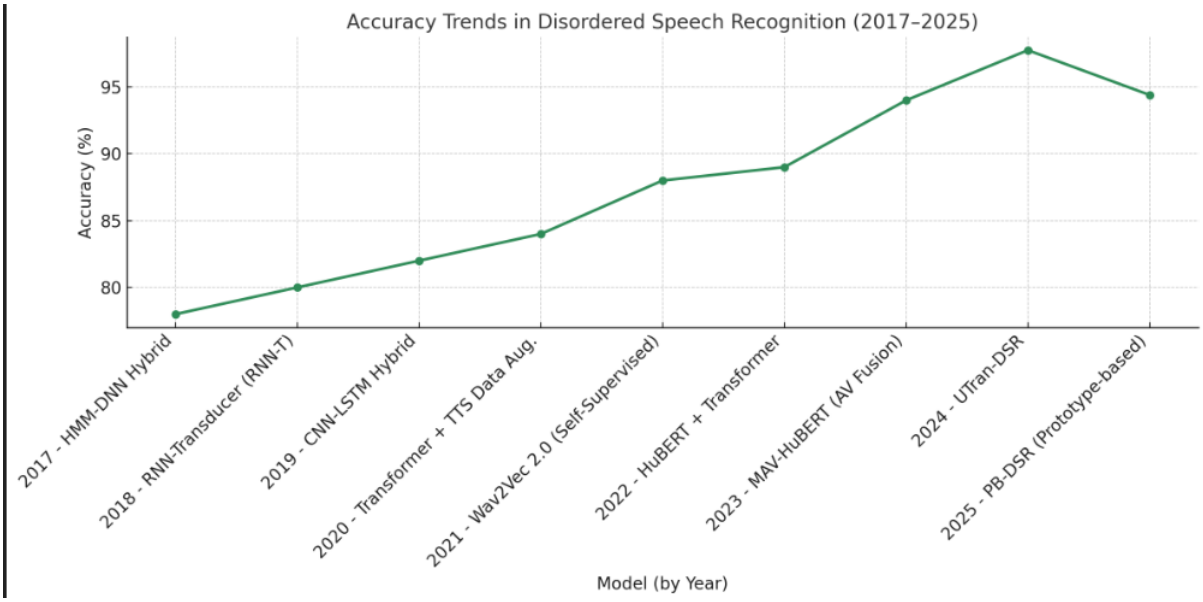
In 2024, multimodal systems combining audio with visual cues, such as lip movements, gained prominence. These models were particularly useful for severely dysarthric individuals whose speech was largely unintelligible through audio alone. Studies showed that integrating visual data significantly boosted recognition accuracy — in some cases achieving up to 67% accuracy in previously unrecognizable speech segments.

### 2.2. Strategies Applied To Models And Their Impact On Accuracy

Summary of Year-Wise Progression in Dysarthric ASR Models: Datasets, Strategies, and Accuracy Improvements

| Year | Model | Dataset Used | Strategy Applied | Accuracy (Before) | Accuracy (After) | Milestone / Highlight |
|---|---|---|---|---|---|---|
| 1993 | HMM | Whitaker, Nemours | Speaker-dependent/adaptive HMM | ~65% | ~68% | First attempts at dysarthric ASR |
| 2000–2010 | HMM + ANN | Nemours, Synthetic | Feature engineering (MFCC, LPC), Speaker Norm. | ~68% | ~70% | Hybrid ASR systems with shallow learning |
| 2011 | DNN | TORGO | Speaker Normalization (CMN, VTLN) | ~72% | ~75% | First deep learning implementation |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2014 | DNN + AE | UA-Speech | Autoencoder feature enhancement | ~72% | ~88% | +16% gain with AE pre-processing |
| 2015 | CNN | Nemours | Transfer Learning with PLP | ~73% | ~82% | Spectrogram-based feature modeling |
| 2016 | LSTM | TORGO | Acoustic + Articulatory Fusion | ~74% | ~78% | Temporal modeling of slurred speech |
| 2017 | TDNN | UA-Speech | Bottleneck AE + LHUC speaker adaptation | ~68% | 69.4% | Lightweight ASR with moderate gains |
| 2020 | Transformer | LibriSpeech + UA-Speech | Transfer learning, Multilingual Fine-tuning | ~77% | ~84% | Attention-based ASR with E2E support |
| 2021 | Wav2Vec2.0 | LibriSpeech + UA-Speech | Self-supervised learning + Fine-tuning | ~75% | 85%+ | SOTA with minimal labeled data |
| 2022 | DyPCL | UA-Speech | Contrastive learning (Patient-level) | ~85% (baseline) | ↓WER by 22.1% | Improved severity-aware recognition |
| 2023 | UTran-DSR | TORGO + IEMOCAP | Emotion-aware + Speaker-aware Transformer | ~82% | ~85% | Emotion-sensitive ASR innovation |
| 2024 | Multimodal AV Fusion | UA-Speech + Custom Visual | Audio-Visual Fusion + Contrastive Learning | ~65% (severe cases) | 67% | Major leap in severe speech recognition |




## 3. COMPARATIVE ANALYSIS OF DYSARTHRIC ASR RESEARCH

| Paper/Study | Model Used | Dataset(s) | Accuracy / WER | Limitation | Future Work |
|---|---|---|---|---|---|
| Depthwise CNNs for Dysarthria [7] | Depthwise Separable CNN | UA-Speech, TORGO | High accuracy (real-time capable) | Difficulty adapting to different dysarthria severities | Integrate self-supervised learning for robustness |
| PB-DSR for Unseen Speakers [8] | Prototype-based adaptation with Transformers | Mozilla CommonVoice, Europarl, etc. | WER: 5.6% | Struggles with diverse speech disorders | Real-time adaptation and multi-disorder support |
| Self-Supervised Pretraining for ASR [4] | Transformer (Wav2Vec2.0, HuBERT) | CommonVoice, LibriSpeech | 70–99.7% accuracy | Weak multilingual & noise robustness | Real-time and noise-resilient ASR |
| Cross-lingual Self-Supervised ASR [3] | Fbank + [7]Wav2Vec 2.0 + XLSR | CommonVoice, UA-Speech | ~90% (healthy), improved on disordered | Low dysarthric coverage across languages | Expand corpora, improve low-resource accuracy |
| MAV-HuBERT AVSR [1] | Multimodal AV HuBERT (audio + video) | LRS3, UA-Speech | WER: 6.05% (mild), 30.77% (severe) | Sensitive to facial/head movement | Robust AVSR in real-time & head-movement settings |
| WFST Cascade [2] | Weighted Finite-State Transducers | UA-Speech | 70% recognition accuracy | Manual alignments, low scalability | Neural-WFST hybrid, smoothing strategies |
| UTran-DSR [10] | U-shaped Transformer + Feature Enhancer | UA-Speech (spec.) | 97.75% | Lacks multimodal integration | Add audio-visual + real-time support |
| Synthetic Speech Augmentation [6] | DNN-based Speech Generator | Paired healthy-dysarthric speech | >95% classification accuracy | Realism of synthetic speech questionable | Real-time generation, telemedicine use |
| Comprehensive LSTM Survey [9] | LSTM / RNN | UA-Speech, TORG [7]O | 85–95% | Less effective in real-time / noisy input | Cross-lingual and low-latency deployment |
| Multi-Dataset Transformer Models[5] | FastSpeech 2 + Transformer | UA-Speech, TORGO + synthetic | Moderate-high accuracy | Real vs. synthetic mismatch | Improve synthesis realism and augmentation |

## 4. CHALLENGES AND LIMITATIONS IN DYSARTHRIC SPEECH RECOGNITION

Despite significant advancements in dysarthric ASR, several challenges continue to limit its robustness, generalizability, and real-world applicability. One of the most pressing issues is the wide variability in speech patterns caused by differing neurological

conditions and severity levels. This diversity makes it difficult for ASR systems to generalize across speakers, with many models requiring fine-tuning for each individual—a process that is neither scalable nor practical in assistive applications.

A major technical bottleneck is the scarcity and imbalance of available datasets. Widely used corpora such as UA-Speech and TORGO feature a limited number of speakers and often lack representation across severity levels, languages, and spontaneous speech types. This not only increases the risk of overfitting but also restricts the applicability of models to real-world conditions. Additionally, the lack of multilingual dysarthric datasets further prevents the development of ASR systems that can serve diverse linguistic populations.

Deployment challenges also persist, particularly in achieving real-time, low-latency performance on resource-constrained devices. Advanced models like Transformers and Wav2Vec2.0, while accurate, often require high computational power, making them difficult to implement in portable or embedded systems. Another key issue is the inability of many ASR models to preserve emotional tone and prosody—an important aspect of natural and expressive communication that is especially critical in assistive contexts.

The absence of standardized benchmarking protocols creates further fragmentation, with studies using different datasets, evaluation metrics, and preprocessing techniques, making comparisons and reproducibility difficult. Even with recent innovations, recognition of severely dysarthric speech remains a challenge, with state-of-the-art models still underperforming in such cases. While multimodal approaches that integrate visual cues offer promise, they introduce new complications related to video collection, privacy, and system complexity.

Lastly, ethical concerns such as user consent, data privacy, and the responsible use of speech data in sensitive domains like healthcare are often overlooked in technical research. Addressing these limitations is essential to ensure that future ASR systems for dysarthric speech are not only accurate and efficient but also inclusive, ethical, and deployable in real-world scenarios.

## 5. CONCLUSION

Recent advancements in ASR have significantly improved the recognition of dysarthric speech through deep learning, self-supervised models, and strategies like speaker adaptation, transfer learning, and multimodal fusion. While systems trained on datasets like UA-Speech and TORGO now offer improved accuracy and robustness, major challenges remain, including data scarcity, speaker variability, and limited cross-lingual generalization. Future efforts must focus on developing lightweight, adaptive, and emotionally expressive ASR models that perform reliably in real-time scenarios. More importantly, building inclusive speech technologies will require interdisciplinary collaboration to ensure these systems are not only accurate but also accessible, ethical, and empowering for individuals with speech impairments.

## REFERENCES

[1] [1] Dysarthric Speech Recognition Using Depthwise Separable CNNs Author(s), "Dysarthric Speech Recognition Using Depthwise Separable CNNs,"
[2] Enhancing Dysarthric Speech Recognition for Unseen Speakers via Prototype-Based Adaptation (PB-DSR) S. Wang, S. Zhao, J. Zhou, A. Kong, and Y. Qin, "Enhancing Dysarthric Speech Recognition for Unseen Speakers via Prototype-Based Adaptation (PB-DSR)," arXiv preprint arXiv:2407.18461, 2024.
[3] Effectiveness of Self-Supervised Pre-Training for Speech Recognition A. Baevski, M. Auli, and A. Mohamed, "Effectiveness of Self-Supervised Pre-Training for Speech Recognition," arXiv preprint arXiv:1911.03912, 2019.
[4] Cross-Lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition 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," arXiv preprint arXiv:2204.01670, 2022.
[5] Multi-Stage Audio-Visual Fusion for Dysarthric Speech Recognition With Pre-Trained Models (MAV-HuBERT) Chongchong Yu , Xiaosu Su , and Zhaopeng Qian IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, VOL. 31, 2023 Application of Weighted Finite-State Transducers to Improve Recognition Accuracy for Dysarthric Speech Omar Caballero Morales and Stephen Cox
[6] UTran-DSR: A Novel Transformer-Based Model Using Feature Enhancement for Dysarthric Speech Recognition EURASIP Journal on Audio, Speech, and Music Processing, vol. 2024, Article 54, 2024.
[7] Improving Dysarthric Speech Segmentation With Emulated and Synthetic AugmentationSaeid Alavi Naeini, Leif Simmatis, Deniz Jafari, Yana Yunusova, Babak Taati, IEEE Senior Member A Comprehensive Survey of ASR Systems for Dysarthric Speech Using LSTM Networks

[8] Shailaja Yadav1 , Dinkar Manik Yadav2 , Kamalakar Ravindra Desai3, 2023 Few-Shot Dysarthric Speech Recognition Using Text-to-Speech Augmented TrainingEnno Hermann and Mathew Magimai.-Doss INTERSPEECH 2023 20-24 August 2023, Dublin, Ireland