INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

REAL-TIME DYSARTHRIC SPEECH CONVERSION AND RECOGNITION WITH COLLABORATIVE AI

[1]Anunandana M, [2]Devika Babu, [3]Twinkle Sani, [4]Ani Sunny, [5]Richu Shibu

Department of Computer Science and Engineering,

Mar Athanasius College of Engineering (Autonomous), Kothamangalam

Abstract : Accurate recognition of dysarthric speech remains a major challenge for conventional Automatic Speech Recognition (ASR) systems due to distorted articulation, high speaker variability, and the limited availability of pathological speech datasets. This work proposes a collaborative AI-based framework designed to improve both dysarthric speech recognition and pronunciation feedback for assistive communication and speech therapy applications. The proposed system integrates a multi-stage data augmentation pipeline consisting of synthetic speech generation using Tacotron2, speed perturbation for temporal variability, and an enhanced CycleGAN-based speech conversion model incorporating Inception-ResNet blocks and temporal masking for non-parallel speech normalization. The converted and augmented speech samples are then used to fine-tune a pre-trained Whisper-Tiny ASR model, enabling improved transcription accuracy for dysarthric speech. In addition to speech recognition, the framework introduces a pronunciation feedback module that performs phoneme- and word-level analysis by comparing user speech with healthy reference patterns. The system automatically identifies mispronounced sounds, estimates pronunciation clarity, and provides interpretable feedback to guide targeted speech practice. Experiments conducted on the TORGO and UA Speech datasets demonstrate stable model convergence, improved acoustic smoothness, and enhanced recognition robustness. The results highlight the potential of the proposed framework as an effective assistive technology for improving communication and supporting speech rehabilitation for individuals with dysarthria.

Index Terms - Dysarthric speech, Automatic Speech Recognition (ASR), CycleGAN, Whisper-Tiny, Data Augmentation, Pronunciation Feedback, Speech Rehabilitation, Tacotron2, HiFi-GAN.

I. INTRODUCTION

Speech is one of the most natural and fundamental modes of human communication. However, for individuals with dysarthria, effective verbal interaction remains a persistent challenge. Dysarthria is a motor speech disorder caused by damage to the central or peripheral nervous system, which disrupts the control of muscles involved in speech production. This impairment results in slow, slurred, imprecise, and inconsistent articulation, often accompanied by abnormal prosody and reduced intelligibility. Such communication difficulties significantly restrict social interaction, independence, and overall quality of life, highlighting the need for intelligent assistive systems capable of accurately recognizing and supporting impaired speech.

Although Automatic Speech Recognition (ASR) systems have achieved remarkable performance on typical speech, their effectiveness degrades substantially when applied to dysarthric speech. This degradation arises from several key challenges, including the limited availability of clinically annotated dysarthric speech datasets, large inter-speaker variability in severity and articulation patterns, and significant acoustic mismatch between dysarthric and healthy speech. Consequently, even state-of-the-art ASR models trained on large-scale speech corpora exhibit high word error rates when deployed in dysarthric speech scenarios, limiting their practical usability in assistive applications.

To address these challenges, this work proposes a collaborative AI framework for dysarthric speech conversion and recognition supported by a robust multi-stage data augmentation pipeline. The framework integrates three complementary augmentation strategies: static data generation using Tacotron2 to synthesize clean and linguistically consistent speech samples, speed perturbation to model natural variations in speaking rate, and an enhanced CycleGAN-based speech conversion model incorporating Inception-ResNet blocks and temporal masking. This non-parallel conversion mechanism transforms dysarthric speech into normalized acoustic representations while preserving linguistic content and speaker characteristics, thereby reducing acoustic mismatch and data scarcity.

The augmented dataset generated through this pipeline is used to fine-tune a pre-trained ASR model, Whisper-Tiny, which leverages large-scale self-supervised learning to capture rich phonetic and acoustic representations. By combining synthetic, perturbed, and converted speech samples, the proposed system improves generalization across diverse dysarthric speech patterns. Experimental observations demonstrate reductions in recognition errors and enhanced robustness compared to baseline ASR systems trained without specialized augmentation.

Beyond speech recognition, the proposed framework introduces a dedicated feedback mode designed to support speech rehabilitation. This module performs phoneme- and word-level pronunciation analysis by comparing user speech with healthy

reference patterns. The system automatically identifies mispronounced sounds, estimates their severity, and provides interpretable feedback using intuitive phoneme descriptions (e.g., /u:/ as in you). This enables users to better understand articulation errors and engage in targeted pronunciation practice. As a result, the framework functions both as an assistive communication system and as a supportive tool for speech therapy.

The proposed system is implemented and evaluated on benchmark dysarthric speech datasets, including TORGO and UA Speech. Experimental results demonstrate stable training convergence, effective preservation of linguistic information, and improvements in speech intelligibility. By combining advanced data augmentation, adversarial speech normalization, and therapy-oriented pronunciation feedback, this work presents a practical and user-centric approach for enhancing communication and rehabilitation outcomes for individuals with dysarthria.

II. RELATED WORK

Research on dysarthric speech recognition spans several interconnected areas, including pathological speech modeling, data augmentation, adversarial speech conversion, self-supervised learning–based ASR, and speech therapy support systems. Early investigations into dysarthric speech recognition highlighted the severe limitations of conventional ASR systems when applied to pathological speech. Studies by Rudzicz et al. and other clinical speech researchers demonstrated that traditional Hidden Markov Model (HMM)–based and Gaussian Mixture Model (GMM)–based ASR frameworks suffer from high error rates due to articulatory imprecision, abnormal prosody, and substantial inter-speaker variability [1], [2]. These findings established the need for specialized modeling strategies tailored to impaired speech.

A major bottleneck in dysarthric ASR research is the scarcity of large, diverse, and well-annotated clinical speech datasets. Widely used corpora such as UA-Speech and TORGO provide valuable dysarthric speech samples but remain limited in size and severity coverage. To address this issue, recent studies have increasingly focused on data augmentation techniques. Classical augmentation methods, including speed perturbation, pitch shifting, spectral masking, and noise injection, have been shown to improve robustness by introducing acoustic variability [3], [4]. More advanced approaches employ neural text-to-speech (TTS) models such as Tacotron2 and FastSpeech to generate high-quality synthetic speech [5], which, when combined with real dysarthric data, improves ASR generalization [6]. However, survey studies note that while augmentation increases data volume, it does not fully address the acoustic mismatch caused by pathological articulation.

Adversarial learning–based speech conversion has emerged as a powerful technique for bridging the gap between dysarthric and typical speech. CycleGAN-based voice conversion models enable non-parallel mapping between impaired and normalized speech representations, making them particularly suitable for clinical scenarios where paired data are unavailable [7]. Extensions of CycleGAN incorporating residual connections, attention mechanisms, and Inception-ResNet blocks have demonstrated improved preservation of phonetic content and speaker identity [6]. Existing literature on pathological speech conversion reports that such models reduce spectro-temporal distortions, stabilize formant trajectories, and enhance intelligibility, thereby benefiting downstream ASR performance [8]. Additional techniques such as temporal masking and spectrogram smoothing further mitigate conversion artifacts.

In parallel, self-supervised learning (SSL) models have significantly advanced dysarthric speech recognition. Frameworks such as wav2vec 2.0, HuBERT, and Whisper learn robust acoustic representations from large-scale unlabeled speech corpora, making them well suited for transfer learning in low-resource clinical settings [9]. Multiple studies report that fine-tuning SSL-based ASR models on augmented or GAN-normalized dysarthric speech yields substantial reductions in word error rate across varying severity levels [10], [11]. Whisper-based models, in particular, exhibit strong robustness to atypical pronunciation patterns due to their large-scale multilingual and multitask pre-training [8].

Beyond transcription accuracy, recent research emphasizes the importance of intelligibility enhancement and speech therapy support. Several works explore phoneme-level error detection, forced alignment, and articulatory feature analysis to identify mispronounced sounds in dysarthric speech [12], [13]. Clinical speech therapy systems increasingly integrate ASR with phoneme-level feedback, visualization tools, and severity scoring to assist users in targeted pronunciation practice. Studies show that providing interpretable feedback—such as mapping phonemes to familiar word examples—helps users better understand articulation errors and supports rehabilitation progress [14]. These approaches highlight the potential of combining ASR with therapeutic feedback rather than focusing solely on transcription.

Summary: The existing literature indicates that effective dysarthric speech recognition benefits from a combination of advanced data augmentation, adversarial speech normalization, and fine-tuning of self-supervised ASR models. While significant progress has been made in improving recognition accuracy, relatively few systems integrate pronunciation-level feedback for therapeutic support. The proposed framework builds upon these advances by combining multi-stage data augmentation, CycleGAN-based speech normalization, SSL-based ASR fine-tuning, and a dedicated therapy mode for phoneme- and letter-level error identification. This integrated approach aims to support both assistive communication and speech rehabilitation for individuals with dysarthria.

III. METHODS

This section describes the proposed collaborative framework for dysarthric speech enhancement, recognition, and pronunciation assessment. The system integrates data augmentation, non-parallel speech conversion, high-fidelity audio reconstruction, automatic speech recognition, and a phoneme-level pronunciation feedback module. As illustrated in Fig. 1, dysarthric speech is first enhanced through an improved CycleGAN-based conversion model, followed by waveform reconstruction using a lightweight neural vocoder. The enhanced speech is then processed by a fine-tuned automatic speech recognition model to obtain stable textual output. Finally, a pronunciation feedback engine analyzes articulation quality at the phoneme level to support speech therapy applications.

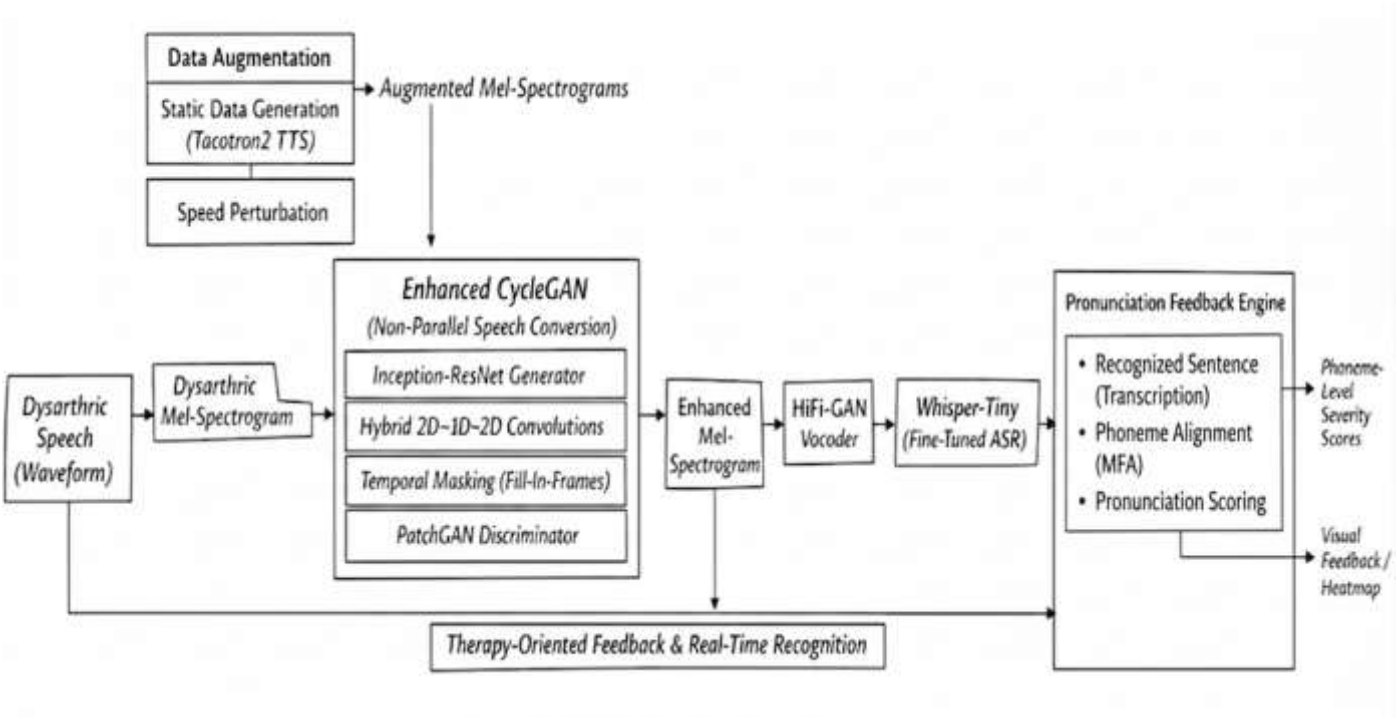

Fig. 1. Proposed collaborative framework for dysarthric speech processing and pronunciation feedback.

A. Data Augmentation

Dysarthric speech corpora are inherently limited in size and exhibit significant inter-speaker and intra-speaker variability, which negatively impacts model generalization. To address this challenge, a two-stage data augmentation strategy is adopted, consisting of static data generation and speed perturbation. This strategy increases linguistic coverage and simulates temporal variability characteristic of dysarthric speech.

1) Static Data Generation

Static data generation is performed using a neural text-to-speech (TTS) system to synthesize high-quality non-dysarthric speech from available text transcriptions. A Tacotron2-based model is used to generate mel-spectrograms that represent healthy speech pronunciations corresponding to the same linguistic content as the dysarthric utterances. This process expands the dataset without requiring additional speaker recordings and establishes a clean reference domain for subsequent speech conversion and pronunciation analysis.

2) Speed Perturbation

To simulate speaking rate variations commonly observed in dysarthric speech, speed perturbation is applied to both natural and synthesized audio samples. Speech signals are resampled at multiple speed factors (e.g., 0.9×, 1.0×, and 1.1×), introducing temporal diversity while preserving phonetic structure. This augmentation enhances robustness to articulation rate irregularities and temporal distortions during model training.

B. Enhanced CycleGAN-Based Speech Conversion

To reduce the acoustic mismatch between dysarthric and non-dysarthric speech, an Enhanced Cycle-Consistent Generative Adversarial Network (Enhanced CycleGAN) is employed for mel-spectrogram conversion. The model performs non-parallel bidirectional domain translation, eliminating the need for paired dysarthric–healthy speech recordings.

1) Unpaired Bidirectional Mapping

The Enhanced CycleGAN learns two mappings: a generator G that converts dysarthric speech to non-dysarthric speech, and a generator F that maps non-dysarthric speech back to the dysarthric domain. Two discriminators distinguish real samples from generated samples in each domain. Adversarial training encourages the generated outputs to resemble real samples from the target domain, while cycle-consistency constraints ensure preservation of linguistic and phonetic content.

2) Inception-ResNet-Based Generator Architecture

To better capture the multi-scale distortions present in dysarthric speech, the generator architecture incorporates Inception-ResNet blocks. Each block employs parallel convolutional filters with varying receptive fields to capture both short-term phoneme distortions and longer temporal speech patterns. The outputs of these parallel filters are fused through residual connections, stabilizing training and preserving important acoustic information.

3) Hybrid 2D–1D–2D Convolution Strategy
The generator adopts a hybrid convolutional design consisting of 2D convolutional layers for downsampling and upsampling combined with 1D convolutional layers for temporal modeling. Two-dimensional convolutions preserve the spectro-temporal structure of mel-spectrograms, while one-dimensional convolutions effectively model temporal dependencies along the time axis. This hybrid design prevents structural degradation caused by repeated resampling operations.

4) Temporal Masking with Fill-In-Frames Strategy
To improve robustness to irregular timing patterns common in dysarthric speech, a temporal masking strategy is introduced during training. Consecutive time frames in the input mel-spectrogram are randomly masked, forcing the generator to reconstruct missing acoustic segments using surrounding contextual information. This strategy improves the model's ability to handle pauses, elongated phonemes, and incomplete articulatory cues.

5) Discriminator Design and Training Objective
For computational efficiency, PatchGAN discriminators are employed instead of full-scale discriminators. PatchGAN evaluates realism at the level of local spectrogram patches, reducing model complexity while maintaining sensitivity to fine-grained acoustic patterns. The Enhanced CycleGAN is trained by optimizing adversarial loss together with cycle-consistency constraints, enabling stable training and realistic speech conversion.

C. Audio Reconstruction Using HiFi-GAN
The Enhanced CycleGAN produces converted mel-spectrograms that must be transformed into time-domain waveforms for perceptual evaluation and recognition. HiFi-GAN, a lightweight neural vocoder, is used for this purpose. HiFi-GAN directly maps mel-spectrograms to waveform signals using a fully convolutional generator and multi-period discriminators, enabling high-fidelity speech synthesis with low computational latency. This design supports near real-time audio reconstruction while maintaining naturalness.

D. Automatic Speech Recognition Fine-Tuning
Automatic speech recognition is performed using the Whisper-Tiny model, a transformer-based encoder–decoder architecture pretrained on large-scale multilingual speech data. The model is fine-tuned using a combination of original dysarthric speech, CycleGAN-enhanced speech, and augmented samples. This training strategy enables the model to adapt to dysarthric speech characteristics while maintaining efficient inference performance.

E. Pronunciation Feedback Module
To support speech therapy functionality, a pronunciation feedback module is introduced to evaluate speech clarity at the word and phoneme levels. User speech is first transcribed using a Wav2Vec2-based automatic speech recognition system. Unlike conventional ASR systems that discard low-confidence outputs, Wav2Vec2 preserves approximate phonetic realizations of slurred or weakly articulated words, enabling more accurate reflection of the user's speech attempt.

The resulting transcript is aligned with the target sentence using a similarity-based matching algorithm that computes character-level similarity between spoken and expected words. A window-based alignment strategy is employed to accommodate skipped words, insertions, or reordered speech segments that frequently occur in dysarthric speech.

Phoneme-level comparison is performed using the eng-to-ipa library, which converts words into International Phonetic Alphabet (IPA) representations. The predicted phoneme sequence is compared with the expected phoneme sequence to measure pronunciation similarity.

During inference, feedback is displayed at the word level in the user interface, where each word is categorized as correctly pronounced, unclear, or missing based on predefined similarity thresholds. An overall pronunciation score is computed using a weighted combination of word-level matches. Additionally, recurring phonetic error patterns are analyzed to generate concise insights that highlight well-produced sounds and suggest articulatory targets for focused practice.

IV. RESULTS AND DISCUSSION

This section presents a comprehensive evaluation of the proposed collaborative AI framework. The results are organized to highlight the effectiveness of the Enhanced CycleGAN in speech normalization, improvements in acoustic smoothness through waveform analysis, and the convergence behavior of the fine-tuned Whisper-Tiny ASR model. All experiments were conducted using the TORGO dysarthric speech dataset along with augmented samples generated during training.

A. CycleGAN Training Stability
The first observation concerns the stability of the Enhanced CycleGAN training process. The cycle-consistency loss measures how accurately the model reconstructs the original speech signal after performing forward and reverse mappings between dysarthric and

non-dysarthric domains. As shown in Fig. 2, the loss decreases smoothly throughout the training epochs, indicating that the generators progressively learn an effective transformation between the two speech domains.

A steady reduction in cycle-consistency loss suggests that the model successfully preserves linguistic content during domain translation. Moreover, the absence of large oscillations in the curve confirms stable adversarial training and indicates that the model avoids mode collapse.

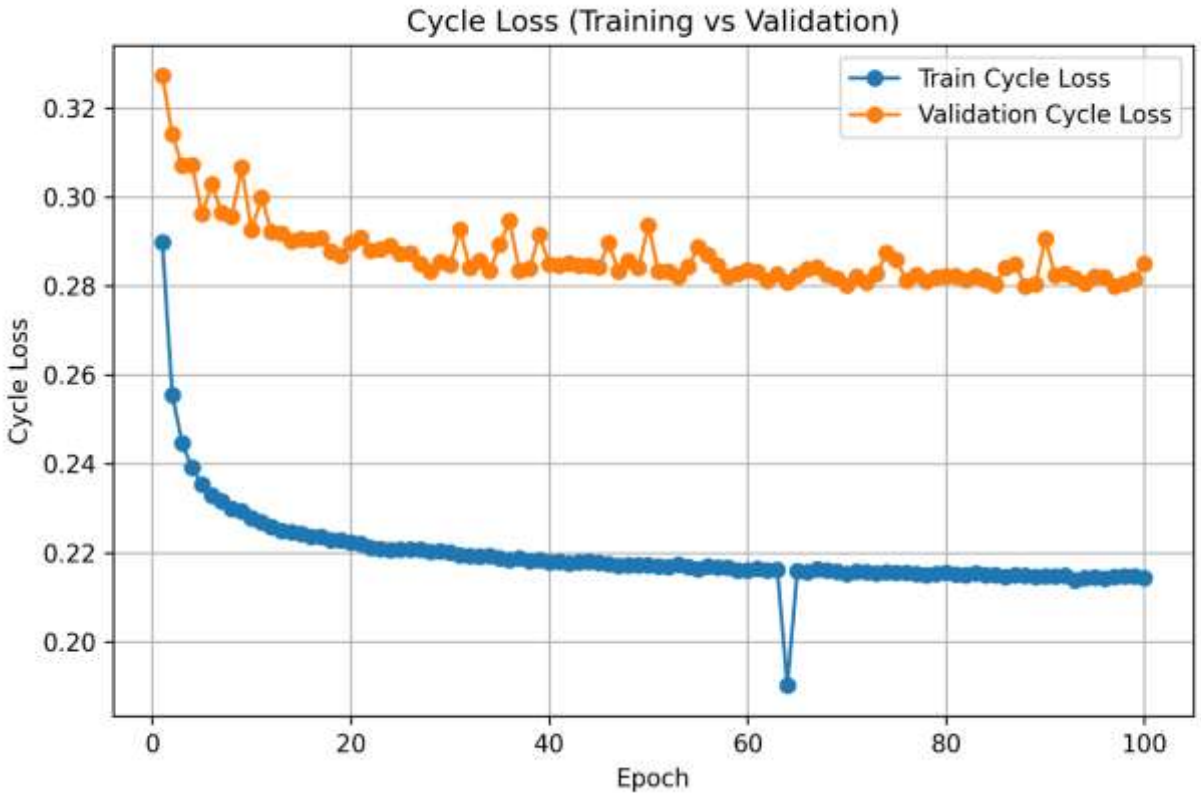

Fig. 2. Cycle-Consistency Loss Curve of the Enhanced CycleGAN Model. The smooth downward trend confirms stable bidirectional domain mapping.

B. Identity Preservation

In addition to reconstruction stability, it is important to ensure that the model preserves speaker-specific acoustic characteristics. The identity loss curve shown in Fig. 3 remains consistently low during the entire training process.

This behavior indicates that when speech from the target domain is used as input, the generator produces an output that closely matches the original signal. Maintaining a low identity loss ensures that the model avoids excessive normalization that could distort speaker identity. Preserving speaker characteristics is particularly important for downstream ASR systems, which rely on consistent acoustic features for accurate recognition.

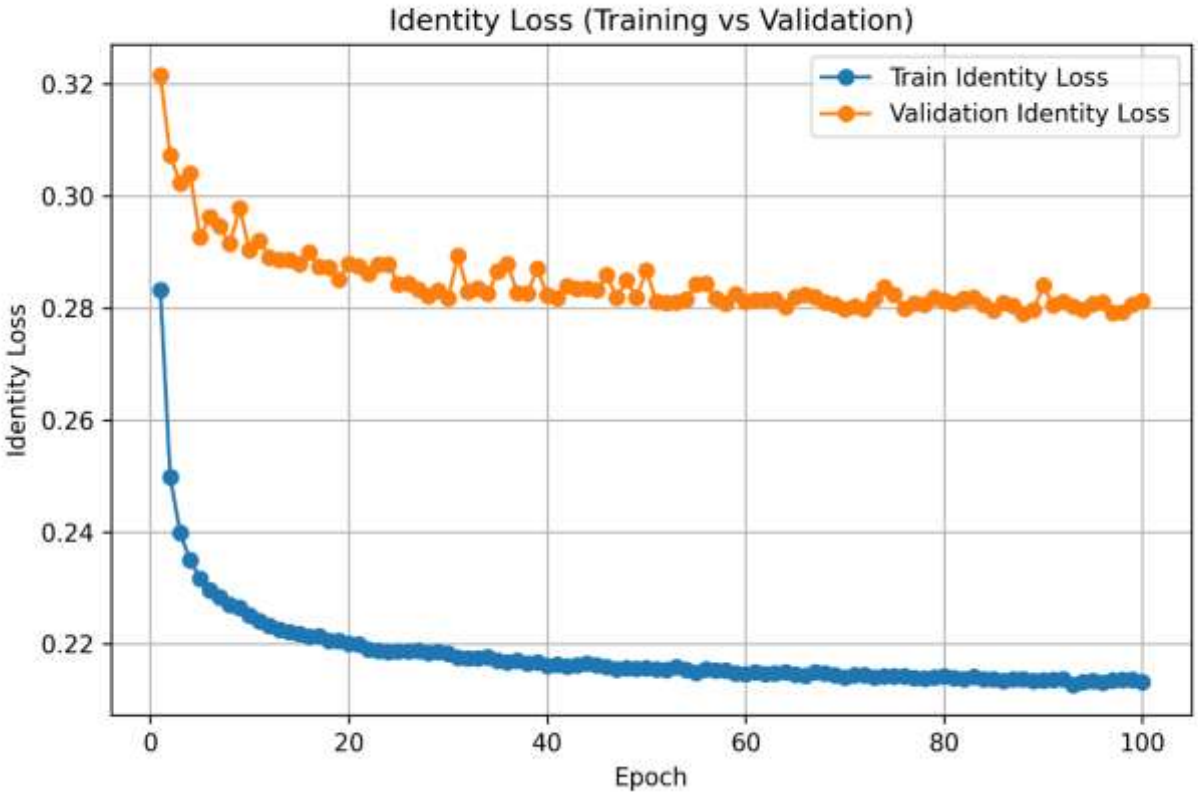

Fig. 3. Identity Loss Curve. The consistently low values demonstrate that the generator preserves speaker-specific characteristics without introducing distortion.

C. Waveform-Level Speech Quality Improvement

To evaluate perceptual improvements in speech quality, waveform-level comparisons were conducted between original dysarthric speech signals and the converted outputs produced by the Enhanced CycleGAN model.

As illustrated in Fig. 4, the original dysarthric waveform exhibits irregular amplitude patterns, abrupt discontinuities, and uneven energy distribution. These characteristics are typical symptoms of impaired articulation and reduced motor control. In contrast, the converted waveform displays smoother amplitude transitions and more consistent temporal structure.

These observations confirm that the Enhanced CycleGAN not only normalizes spectral properties but also reduces temporal irregularities in the speech signal. The resulting waveform is more stable and intelligible, which improves performance in downstream speech recognition tasks.

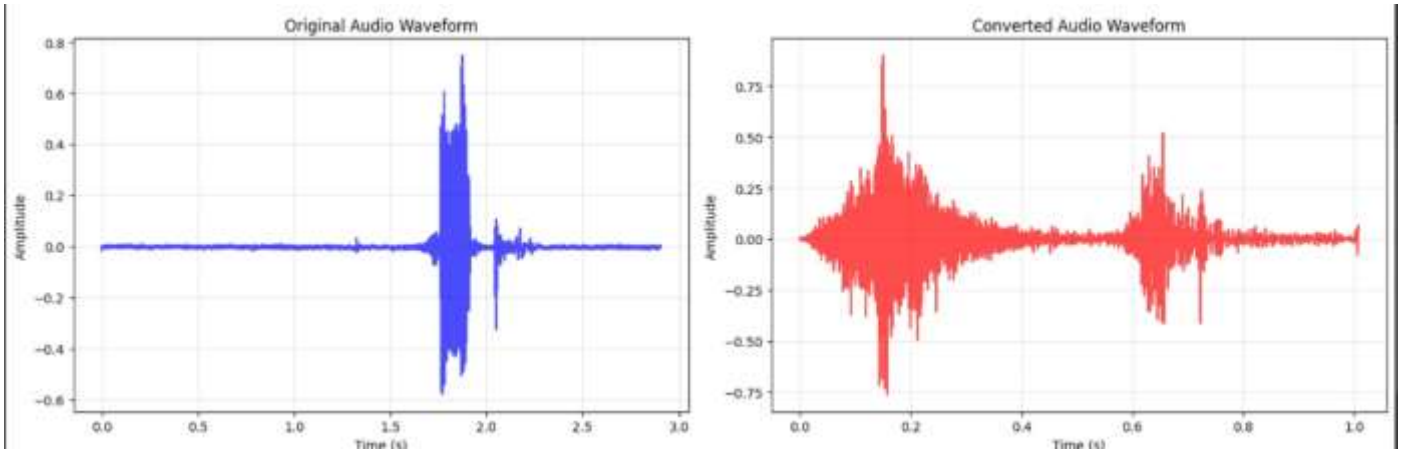

Fig. 4. Waveform Comparison Between Original Dysarthric Speech (top) and Converted Speech (bottom). The converted waveform shows smoother amplitude transitions and improved temporal regularity.

D. Spectrogram-Level Acoustic Enhancement

In addition to waveform-level analysis, spectrogram comparisons were conducted to evaluate how effectively the Enhanced CycleGAN model normalizes dysarthric speech in the spectral domain. Spectrograms provide a time–frequency representation of the speech signal, revealing harmonic structures and energy distributions across frequency bands.

Figure 5 compares the original dysarthric spectrogram with the converted spectrogram produced by the Enhanced CycleGAN model. The original spectrogram exhibits irregular frequency bands and blurred harmonic structures, which are typical indicators of impaired articulation and inconsistent phoneme production. In contrast, the converted spectrogram shows clearer harmonic patterns and more stable frequency distributions. These improvements indicate that the Enhanced CycleGAN successfully reduces spectral distortions while preserving important phonetic information.

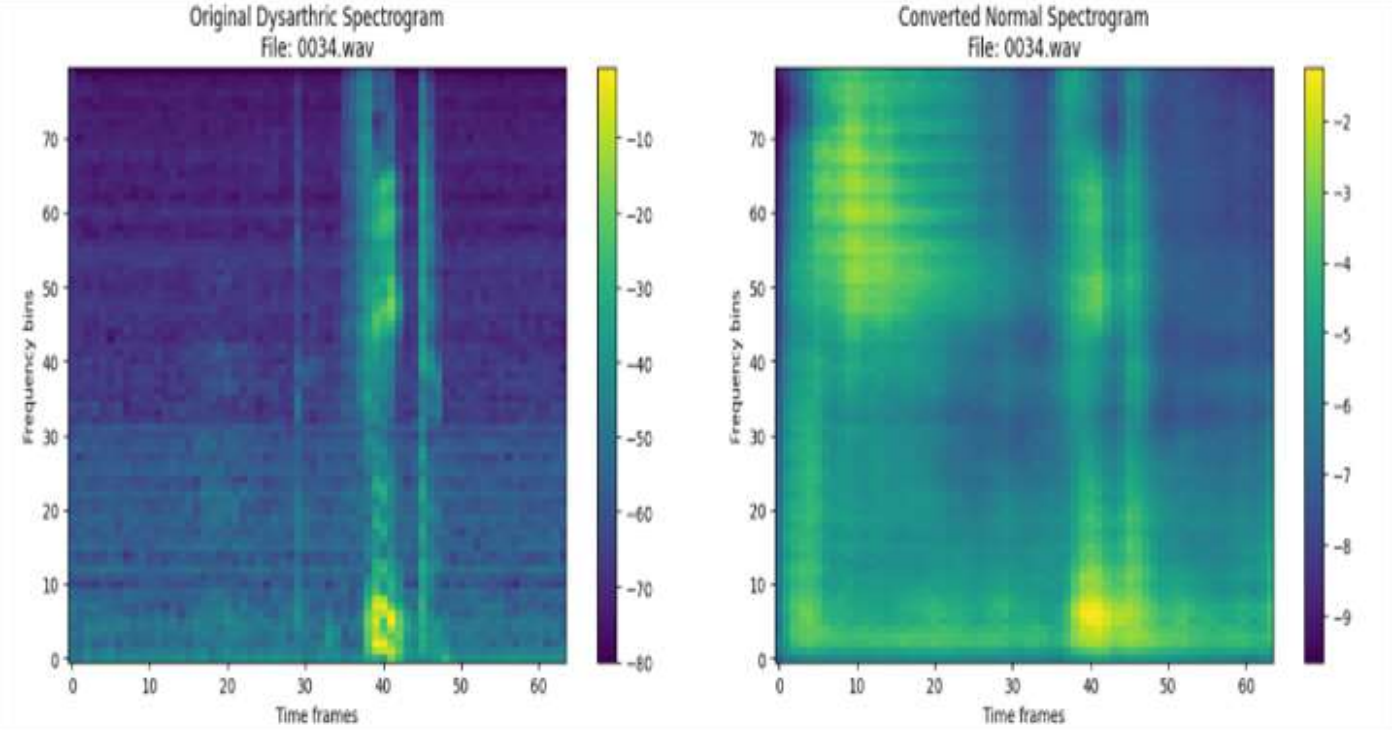

Fig. 5. Mel-spectrogram comparison between original dysarthric speech (left) and CycleGAN-converted speech (right).

E. ASR Model Training Behavior

Following data augmentation and speech normalization, the Whisper-Tiny ASR model was fine-tuned on the enhanced dataset. The training loss curve shown in Fig. 6 demonstrates a consistent downward trend, indicating effective optimization and stable gradient behavior during training. The absence of sudden oscillations suggests that the dataset provides sufficient diversity to support robust learning without causing instability.

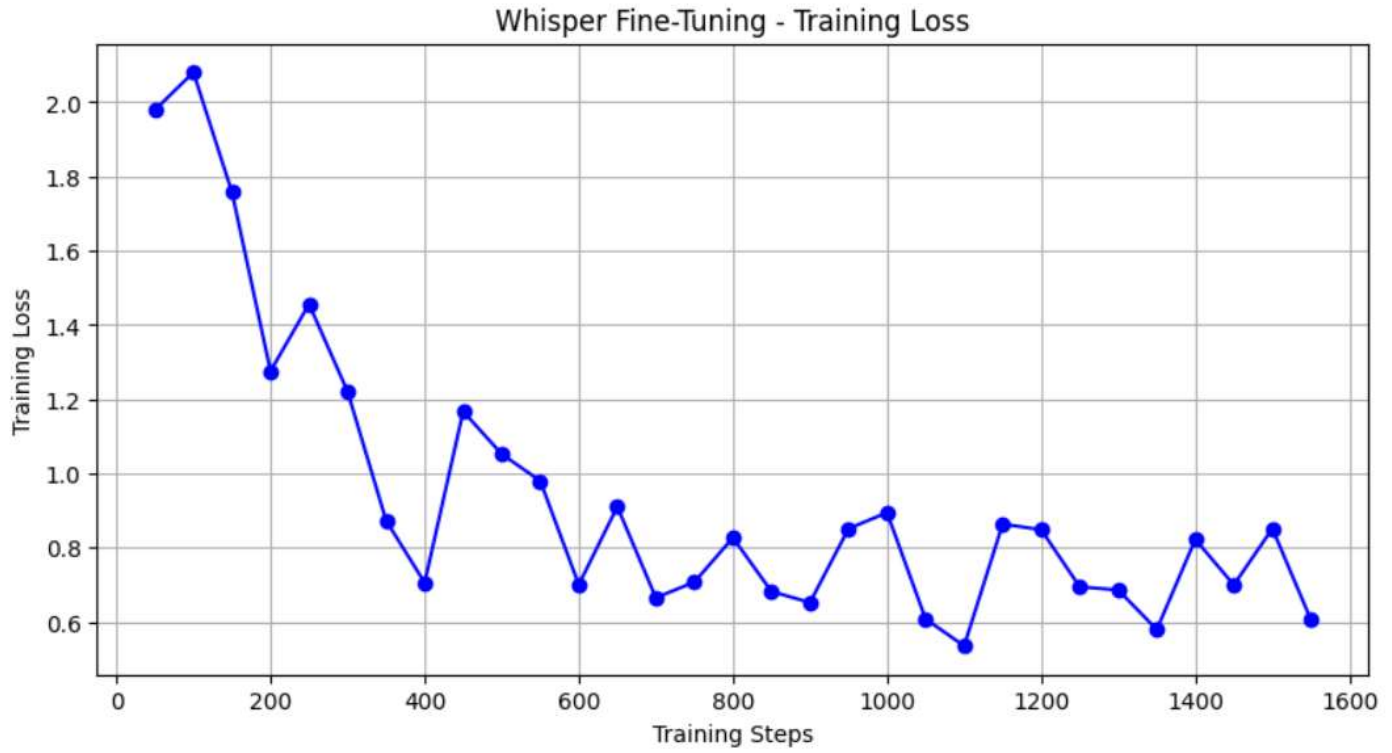

Fig. 6. Training Loss Curve of the Fine-Tuned Whisper-Tiny ASR Model.

F. ASR Model Generalization

The validation loss curve shown in Fig. 7 further supports the effectiveness of the proposed training strategy. As training progresses, the validation loss decreases steadily alongside the training loss. The narrowing gap between the training and validation curves suggests that the model generalizes well to unseen dysarthric speech samples.

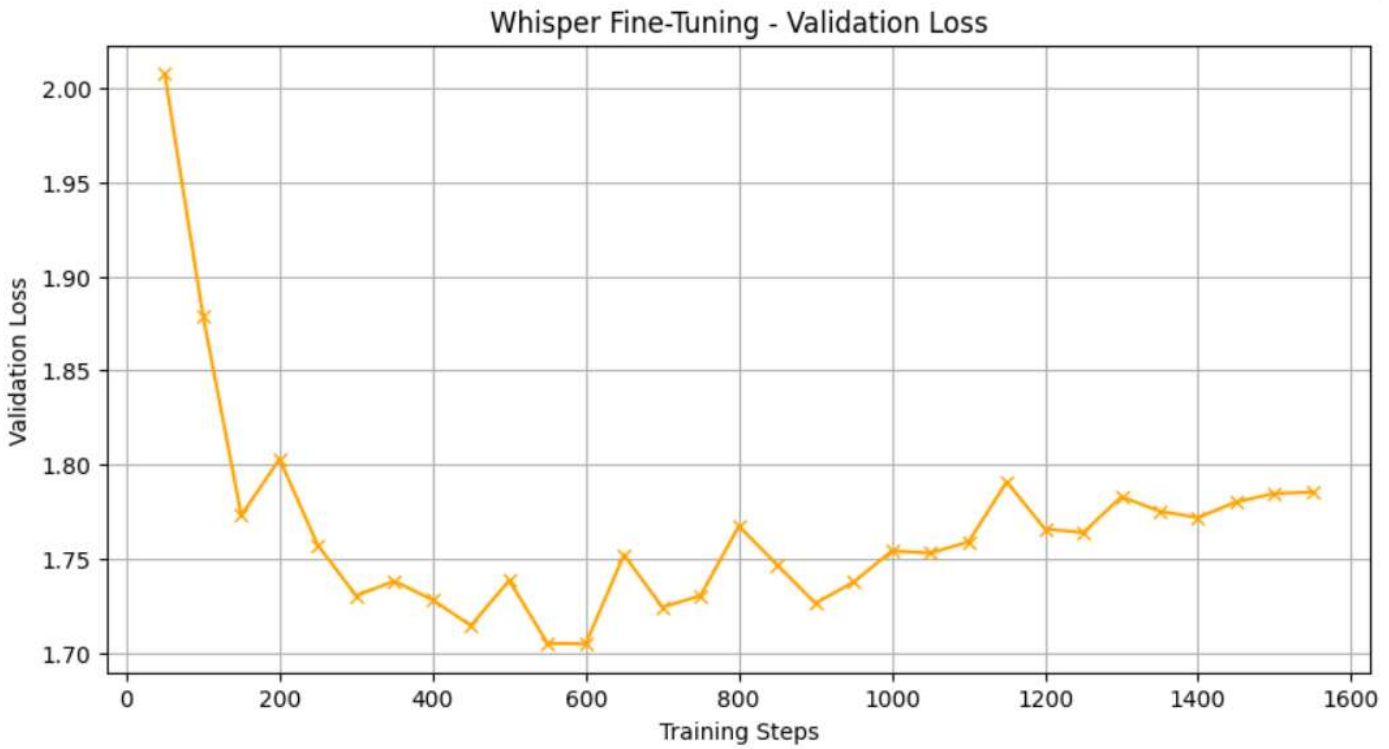

Fig. 7. Validation Loss Curve for Whisper-Tiny demonstrating improved generalization.

G. Acoustic and Recognition Analysis of Dysarthric Speech

To further analyze the characteristics of dysarthric speech and the behavior of the recognition system, an acoustic and recognition comparison was conducted between normal and dysarthric speech samples. The figures below illustrate waveform patterns, spectrogram representations, pitch contour variations, and word recognition accuracy. The waveform comparison highlights that normal speech exhibits more consistent amplitude patterns, whereas dysarthric speech shows irregular fluctuations caused by impaired motor control.

Spectrogram analysis further reveals that dysarthric speech contains distorted frequency patterns and less stable harmonic structures. Pitch contour analysis indicates that dysarthric speech often exhibits unstable fundamental frequency trajectories compared to normal speech.

The word recognition accuracy comparison demonstrates the practical impact of these acoustic distortions. While the normal speech sample achieves near-perfect recognition accuracy, the dysarthric speech sample produces lower recognition accuracy and a higher proportion of unclear words. This highlights the impact on ASR performance in real-world speech communication scenarios.

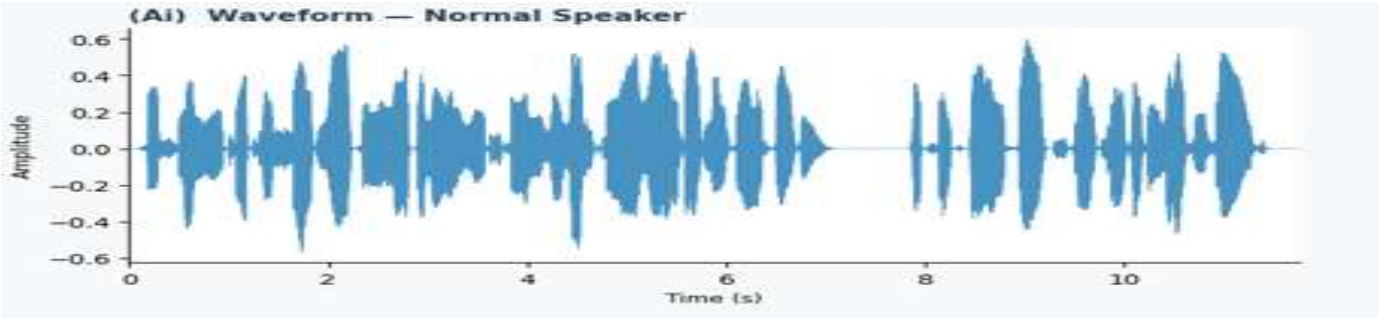

Fig. 8. Waveform of normal speech showing consistent amplitude patterns.

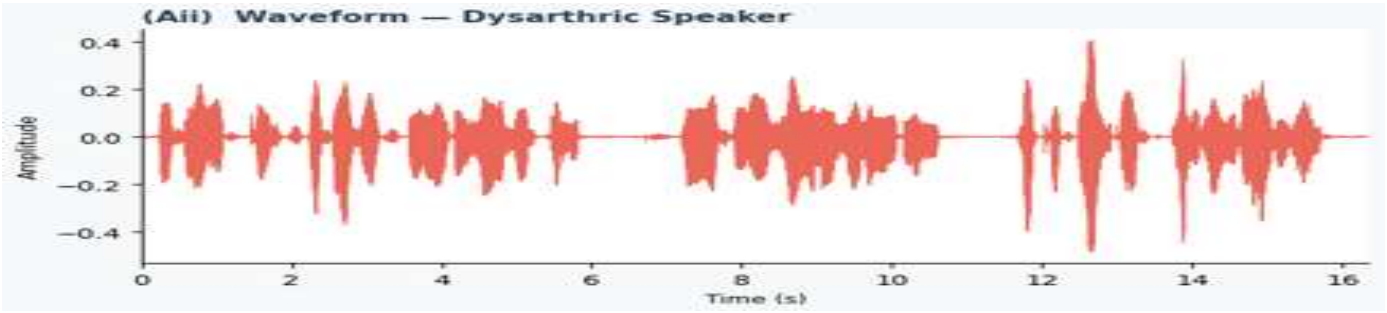

Fig. 9. Waveform of dysarthric speech showing irregular amplitude fluctuations.

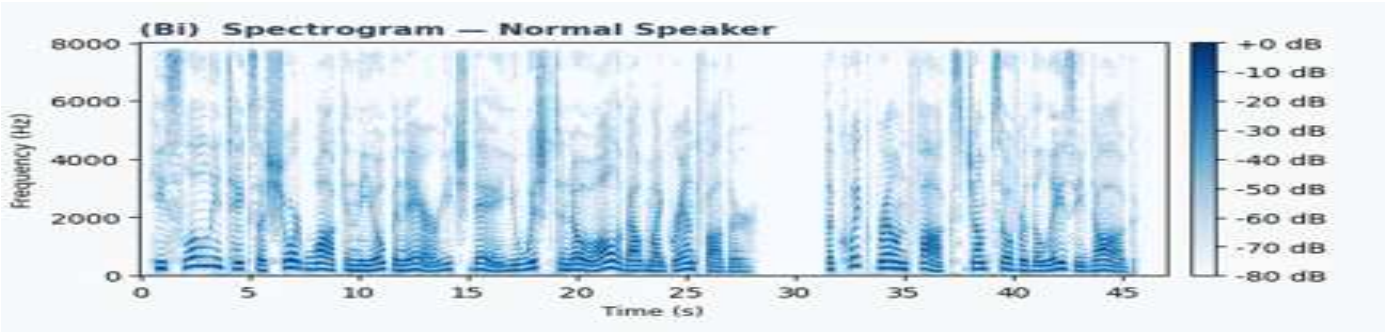

Fig. 10. Spectrogram of normal speech with clear harmonic structures.

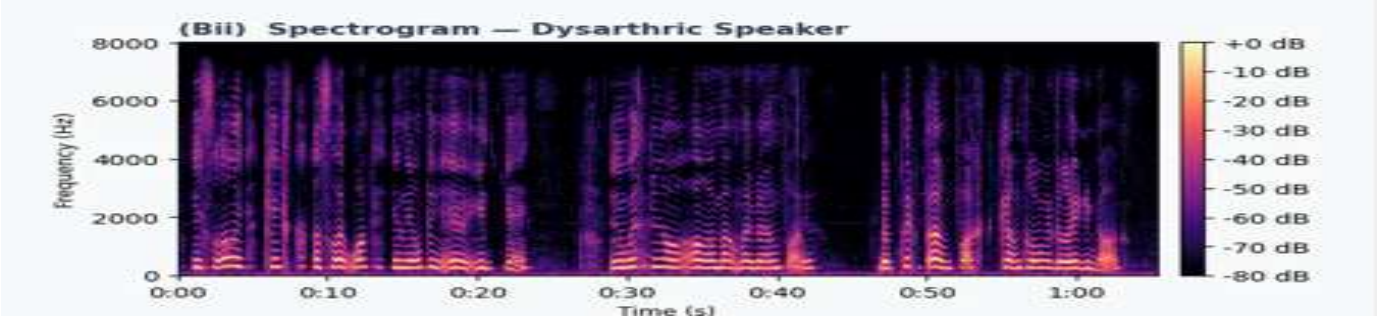

Fig. 11. Spectrogram of dysarthric speech showing distorted frequency patterns.

Overall, waveform and spectrogram analyses confirm that dysarthric speech exhibits significant temporal and spectral irregularities, impacting speech intelligibility and posing challenges for accurate automatic speech recognition. Additionally, pitch contour analysis reveals unstable fundamental frequency variations, while word recognition results indicate reduced accuracy and increased ambiguity compared to normal speech.

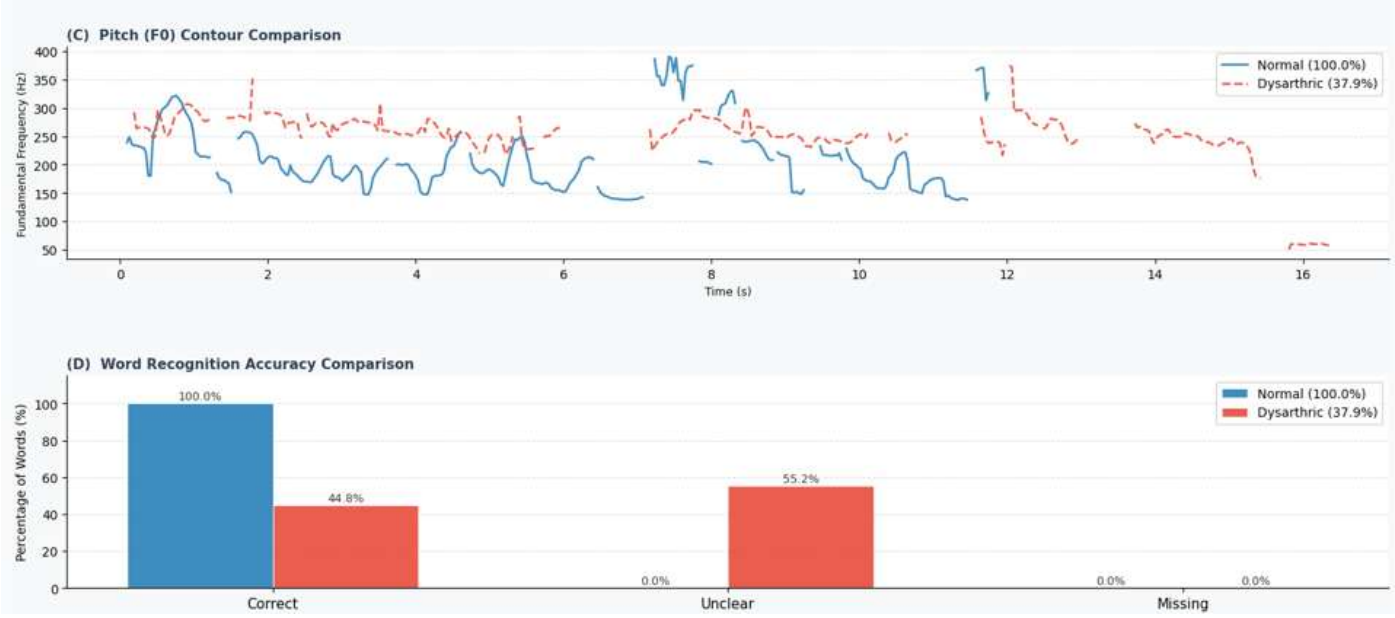

Fig. 12. Pitch contour and word recognition accuracy comparison between normal and dysarthric speech.

V. FUTURE WORK

Although the proposed collaborative AI framework demonstrates promising improvements in dysarthric speech normalization and recognition, several directions can further enhance its performance and practical applicability.

One important direction is the integration of adaptive learning mechanisms that allow the ASR system to continuously improve based on user interaction. Incorporating reinforcement learning or feedback-driven correction loops could enable the model to adapt to individual articulation patterns. Such personalization is particularly valuable for individuals with progressive neurological disorders, where speech characteristics may change over time.

Another area of future research involves developing lightweight and resource-efficient versions of the CycleGAN and Whisper models to enable deployment on edge devices such as smartphones, tablets, or assistive communication devices. Techniques including model pruning, quantization, and parameter-efficient fine-tuning could reduce computational requirements while maintaining high recognition accuracy, thereby supporting real-time inference in low-power environments.

Expanding the dataset to include multilingual and multi-accent dysarthric speech is also an important objective. Current systems are largely trained on English dysarthric speech datasets. Extending the framework to support additional languages and dialects would improve accessibility and broaden the system's usability in global contexts. Furthermore, incorporating samples from more severe dysarthria conditions and diverse neurological disorders could improve model robustness and generalization.

Future studies may also investigate advanced speech enhancement methods to further improve acoustic clarity prior to recognition. Emerging techniques such as diffusion-based speech restoration, transformer-based vocoders, and perceptually guided enhancement models could provide additional improvements in spectral quality and intelligibility.

Finally, comprehensive user-centered evaluations will be essential to assess the real-world impact of the proposed system. Collaborations with clinical experts, speech-language pathologists, and individuals with dysarthria will enable evaluation of usability, communication effectiveness, and long-term benefits in real assistive communication scenarios.

REFERENCES

[1] S. Liu et al., "Recent progress in the CUHK dysarthric speech recognition system," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 2267–2281, 2021.
[2] L. Moro-Velazquez et al., "Study of the performance of automatic speech recognition systems in speakers with Parkinson's disease," in Proc. Interspeech, Sep. 2019, pp. 3875–3879.
[3] M. Geng et al., "Investigation of data augmentation techniques for disordered speech recognition," arXiv preprint arXiv:2201.05562, 2022.
[4] Z. Jin et al., "Adversarial data augmentation for disordered speech recognition," arXiv preprint arXiv:2108.00899, 2021.
[5] J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in Proc. IEEE ICASSP, Apr. 2018, pp. 4779–4783.
[6] W.-Z. Zheng, J.-Y. Han, C.-Y. Chen, Y.-J. Chang, and Y.-H. Lai, "Improving the efficiency of dysarthria voice conversion system based on data augmentation," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 4613–4625, 2023.
[7] W.-C. Huang, B. M. Halpern, L. P. Violeta, O. Scharenborg, and T. Toda, "Towards identity preserving normal to dysarthric voice conversion," in Proc. IEEE ICASSP, May 2022, pp. 6672–6676.
[8] Y. He, K. P. Seng, and L.-M. Ang, "Collaborative AI dysarthric speech recognition system with data augmentation using generative adversarial neural network," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 33, 2025.
[9] S. Hu et al., "Self-supervised ASR models and features for dysarthric and elderly speech recognition," The Chinese University of Hong Kong and Institute of Software, Chinese Academy of Sciences, 2024.
[10] S. R. Shahamiri, V. Lal, and D. Shah, "Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 3407–3417, 2023.
[11] M. Geng et al., "Homogeneous speaker features for on-the-fly dysarthric and elderly speaker adaptation and speech recognition," The Chinese University of Hong Kong and Institute of Software, Chinese Academy of Sciences, 2025.
[12] M. Anuprabha, K. Gurugubelli, and A. K. Vuppala, "Dysarthric speech intelligibility assessment by custom keyword spotting," International Institute of Information Technology-Hyderabad and Samsung R&D Institute Bengaluru, 2025.
[13] J. Tobin et al., "Automatic speech recognition of conversational speech in individuals with disordered speech," Journal of Speech, Language, and Hearing Research, vol. 67, no. 11, pp. 4176–4185, Nov. 2024.
[14] I.-T. Hsieh and C.-H. Wu, "Hierarchical curriculum learning for dysarthric speech recognition via multi-level knowledge distillation," National Cheng Kung University, 2025.
[15] Z. Zhong et al., "Convolution-augmented transformers for enhanced speaker-independent dysarthric speech recognition," University of Auckland and University of Illinois at Urbana-Champaign, 2025.
[16] J. Zhao, Q. Huang, S. Wang, and S. Sun, "VB-Adapter: Variational Bayesian adapter for cross-domain speech representation learning," East China Normal University and Shanghai Jiao Tong University, 2025.
[17] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification: A study on acoustic features and deep learning techniques," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30, pp. 1147–1158, 2022.
[18] S. R. Shahamiri, "Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 852–861, 2021.

[19] D. Mulfari, L. Carnevale, and M. Villari, "Sequence-to-sequence models in Italian atypical speech recognition," in Proc. IEEE Symposium on Computers and Communications (ISCC), Jun. 2024, pp. 1–6.
[20] D. Mulfari, L. Carnevale, and M. Villari, "Toward a lightweight ASR solution for atypical speech on the edge," Future Generation Computer Systems, vol. 149, pp. 455–463, Dec. 2023.