Convergence
A Journal for Young Researchers

ORIGINAL RESEARCH

Received: 3-11-26; Accepted: 5-12-26;
Managing Editor: Jonas Katona;
Action Editor: Jonas Katona

Speech Recognition for Assisting Patients with Speech Difficulties

Akshobh Karthik

Palos Verdes Peninsula High School, Rancho Palos Verdes, California, United States

Abstract

Automated speech recognition (ASR) systems based on large pretrained models achieve near-human accuracy for typical speakers, yet consistently fail for individuals with dysarthria—a motor speech disorder affecting articulation, phonation, and prosody. This paper presents an empirical evaluation of fine-tuning strategies for adapting OpenAI Whisper Large V3, a state-of-the-art sequence-to-sequence ASR model, to dysarthric speech drawn from two established public datasets: UASpeech and TORGO. The study also examines the applied contribution of voice-cloning and speech-synthesis-based data augmentation as a practical approach to the chronic low-resource challenge in dysarthric ASR. Fourteen systematic experiments were conducted, varying training data composition, augmentation strategy, and hyperparameter configuration. Speaker groups were strictly separated between training and test partitions to ensure uncontaminated evaluation. The fine-tuned Whisper Large V3 model augmented with approximately 4,724 voice-cloned samples achieved a Word Error Rate (WER) of 15.36% on multi-word TORGO utterances—representing reductions of approximately 70% for severe, 50% for moderate, and 54% for mild dysarthric speech compared to the unmodified pretrained baseline. A deployed ASR web application, ClearVoiceAI, demonstrates the practical utility of the approach. These findings contribute an empirical foundation for applying established fine-tuning and data augmentation techniques to the underserved domain of dysarthric speech recognition.

Keywords: automatic speech recognition, dysarthria, transfer learning, voice cloning, speech synthesis, assistive technology, word error rate

1. Introduction

1.1. Dysarthria and Automated Speech Recognition Systems

Dysarthria is a motor speech disorder resulting from neurological impairment that affects the coordination, strength, and control of the muscles involved in speech production (Tomik & Guiloff, 2010). Dysarthria is strongly associated with neurological conditions such as cerebral palsy (CP), affecting approximately 40% of individuals with CP, and amyotrophic lateral sclerosis (ALS), where prevalence can reach up to 80% (Shih et al., 2022). Additional associated conditions include stroke, Parkinson's disease, multiple sclerosis, and traumatic brain injury (Schölderle et al., 2020).

The acoustic manifestations of dysarthria include irregular articulation, slurred or imprecise phoneme production, atypical prosody, reduced vocal stability, and inconsistent speech patterns (Young & Mihailidis, 2010). These characteristics vary significantly across individuals and severity levels, creating high variability in speech signals. For affected individuals, these characteristics create significant communication barriers in daily life, impacting social participation, personal identity, and overall quality of life (Page & Yorkston, 2022). Individuals with dysarthria often experience reduced confidence in communication, social isolation, and stigma associated with impaired speech.

Automatic speech recognition (ASR) systems have achieved near-human performance for neurotypical speech through large-scale pretraining on diverse datasets. However, these systems exhibit severe performance degradation when applied to dysarthric speech, due to a fundamental domain mismatch between the training data and the target input conditions. Prior studies have consistently demonstrated that standard ASR systems fail to generalize effectively to dysarthric speech, particularly for moderate-to-severe cases (Young & Mihailidis, 2010; Schu et al., 2022). This limitation is particularly evident in technology-mediated contexts, such as voice assistants, telemedicine systems, and speech-to-text interfaces, where reliance on visual and contextual cues is reduced. These challenges highlight the need for targeted adaptation strategies to improve accessibility for individuals with speech impairments. An overview of the end-to-end ASR pipeline developed in this study is provided in the Appendix, Figure 1.

1.2. Adapting Large-Scale ASR Models to Dysarthric Speech: Fine-Tuning and Synthetic Data Strategies

Recent advances in large-scale ASR models, such as Whisper (Radford et al., 2022), have demonstrated robust performance across a wide range of acoustic conditions due to training on extensive multilingual and weakly supervised datasets. However, their effectiveness in low-resource and atypical speech domains, including dysarthria, remains limited without domain-specific adaptation.

Fine-tuning has emerged as a primary approach for adapting pretrained ASR models to specialized domains. By updating model parameters using domain-specific datasets, fine-tuning enables the model to learn acoustic and phonetic patterns unique to dysarthric speech. Prior work has shown that transfer learning can improve transcription accuracy for dysarthric speakers (Rathod et al., 2023). However, the success of fine-tuning is constrained by the limited availability of labeled dysarthric speech data, increasing the risk of overfitting (Liu et al., 2024).

To address this challenge, data augmentation techniques such as voice cloning and speech synthesis have been explored. Voice cloning generates synthetic samples that preserve speaker-specific acoustic characteristics, while speech synthesis

produces generalized dysarthric-like speech patterns without requiring per-speaker reference audio. These approaches provide scalable methods for expanding training datasets and improving model robustness in low-resource scenarios.

1.3. Scope and Aims

This study evaluates the effectiveness of fine-tuning and data augmentation strategies for adapting Whisper Large V3 to dysarthric speech. The primary objective is to assess how voice cloning and speech synthesis can mitigate data scarcity and improve transcription accuracy across varying levels of dysarthria severity.

This work contributes by providing a systematic empirical evaluation of state-of-the-art ASR adaptation techniques applied to dysarthric speech datasets. It offers new insights into the relative effectiveness of augmentation strategies, the impact of severity-specific acoustic variability, and the practical challenges of deploying ASR systems in real-world assistive contexts.

1.4. Research Hypothesis

It is hypothesized that fine-tuning Whisper Large V3 on dysarthric speech datasets, augmented with voice-cloned and speech-synthesized samples, will substantially reduce Word Error Rate (WER) compared to the unmodified pretrained model across all levels of dysarthria severity.

2. Literature Review

2.1. Dysarthria and Its Impact on ASR

The acoustic characteristics of dysarthric speech, including reduced articulatory precision, irregular vocal quality, atypical prosody, and co-articulation breakdown, create a substantial domain mismatch with the training distributions of standard ASR systems (Young & Mihailidis, 2010). Schu et al. (2022) demonstrated that standard ASR systems trained without dysarthric-specific adaptation produce consistently poor results on both the UASpeech and TORGO benchmarks, establishing these datasets as appropriate evaluation grounds for dysarthric ASR research.

2.2. Whisper as a Base Model

Radford et al. (2022) introduced Whisper, a sequence-to-sequence ASR model trained on 680,000 hours of multilingual weakly-supervised web audio, achieving robust performance across diverse acoustic conditions. The Whisper Large V3 variant supports 1.5 billion parameters with multilingual capability. Its architecture—an encoder-decoder transformer with a log-Mel spectrogram front-end—is well-suited to fine-tuning via the Hugging Face transformers library. Liu et al. (2024) systematically evaluated Whisper fine-tuning strategies for low-resource ASR scenarios and identified key hyperparameter sensitivities relevant to small-dataset fine-tuning, which informed the experimental design of this study.

2.3. Fine-Tuning for Dysarthric Speech

Rathod et al. (2023) demonstrated that transfer learning applied to Whisper improves word recognition accuracy for dysarthric speech, establishing fine-tuning as a viable adaptation strategy. The challenge in this domain is that dysarthric speech corpora are inherently small—collecting and annotating such data requires clinical access and significant manual

effort—making standard fine-tuning approaches susceptible to overfitting (Liu et al., 2024).

2.4. Voice Cloning and Speech Synthesis as Data Augmentation

To address data scarcity, this study employs F5-TTS (Chen et al., 2024), a flow-matching-based TTS model that can clone voices from short reference audio clips. F5-TTS generates synthetic dysarthric utterances that preserve the acoustic characteristics of specific impaired speakers—including their articulatory irregularities. An alternative approach, speech synthesis, fine-tunes the base F5-TTS model on the complete TORGO dataset to synthesize novel dysarthric-like utterances without per-utterance reference audio, avoiding hallucination artifacts common in voice cloning when reference audio quality is low.

2.5. Evaluation Metric: Word Error Rate (WER)

Word Error Rate (WER) is the standard evaluation metric for ASR systems. It is computed as WER = (S + I + D) / N, where S = substitutions, I = insertions, D = deletions, and N = number of reference words. Importantly, WER can exceed 100% when the number of insertion errors alone surpasses the total number of reference words—a phenomenon observed in this study for severely dysarthric single-word utterances, where the unmodified Whisper model generates extensive hallucinated output (see Section 4.1 and Discussion Section 5.2).

3. Methodology

3.1. Methodology Pipeline Overview

The study follows a five-stage analytical pipeline that organises all experimental work reported in this paper, summarised graphically in the Appendix, Figure 13. In the first stage, dysarthric speech is sourced from two established public corpora—UASpeech (multi-word, 1,053 utterances across 129 speakers) and TORGO (multi-word, 2,702 utterances)—and a strict speaker-wise split is applied so that training and test speakers do not overlap. In the second stage, every audio file is preprocessed through a common pipeline of 16 kHz resampling, silence trimming, amplitude normalisation, and text cleaning (lowercasing, punctuation, and special-character removal) to standardise inputs across speakers and recording conditions. The third stage performs synthetic data augmentation using two complementary techniques: voice cloning, which preserves speaker-specific acoustic characteristics through short reference clips, and speech synthesis, which generates novel dysarthric-like utterances after fine-tuning a TTS model on the full training corpus. The fourth stage fine-tunes OpenAI Whisper Large V3 on the augmented data using parameter-efficient adaptation, mixed-precision training, gradient accumulation, and a learning-rate schedule tuned for low-resource ASR. Finally, the fifth stage evaluates the fine-tuned models on held-out UASpeech and TORGO test sets using Word Error Rate (WER) as the primary metric, with a performance breakdown across severity levels (mild, moderate, severe). The remainder of Section 3 details each of these stages in turn.

3.2. Base Model Selection

Three candidate ASR models were evaluated: Kaldi, Meta Wav2Vec 2.0, and OpenAI Whisper (Seagraves, 2022). Based on comparative benchmarking, Whisper's overall WER was 45% lower than Wav2Vec 2.0 and 63% lower than Kaldi across

diverse acoustic conditions. OpenAI Whisper Large V3 (Radford et al., 2022) was selected as the base pretrained model for all fine-tuning experiments.

3.3. Datasets

UASpeech Dataset. The UASpeech dataset (Kim et al., 2023) contains recordings from 15 individuals with dysarthria caused by cerebral palsy and 13 neurotypical controls, comprising approximately 57,000 predominantly single-word utterances. Speakers were classified into severity groups based on speech intelligibility percentages following Farhadipour and Veisi (2023):

- Severe (0–40% intelligibility): Patients F03, M12 — 5,185 test files
- Moderate/Manageable (40–80% intelligibility): Patients M06, M11 — 2,847 test files
- Mild (>80% intelligibility): Patients M08, F05 — 10,711 test files

Training used 9 patients (~38,700 files); testing used 6 different patients (~18,700 files). Speaker groups were strictly non-overlapping between training and test partitions.

TORGO Dataset. The TORGO dataset (Rudzicz et al., 2012) contains recordings from 8 dysarthric speakers (with CP and ALS) and 7 controls. A subset of 5,600 files was used, aligned with the predefined TORGO severity scale. Files were partitioned into:

- One-word utterances: 4,888 files
- Severe: M01, M02, M04, F03 (2,476 files)
- Manageable: M05 (470 files)
- Mild: M03, F03, F04 (1,942 files)
- Imperative (multi-word, 3 words) utterances: 854 training files (patients M01, M02, F01, M03, F03) and 454 test files (patients M04, M05, F04)
- Severe test: M04 (145 files)
- Manageable test: M05 (140 files)
- Mild test: F05 (169 files)

No speaker overlap exists between the TORGO training and test partitions for multi-word experiments.

Voice-Cloned Augmentation Dataset. Using F5-TTS (Chen et al., 2024), 704 imperative sentences sourced from a public sentence corpus (Lettergram, 2019) were voice-cloned using training-set patient audio as reference. Five random samples per sentence were generated per patient. Following generation, 185 hallucinated or corrupted files were manually discarded (a one-week manual review process), yielding 4,724 valid voice-cloned samples. These were combined with the 854 real imperative sentences for a total training corpus of approximately 5,600 samples—a synthetic-to-real ratio of approximately 5.5:1.

Given the scarcity of real dysarthric speech data, a relatively high synthetic-to-real ratio was used. While this approach improves dataset size and diversity, it may introduce bias toward synthetic speech patterns, potentially affecting generalization to real-world speech.

Speech-Synthesized Augmentation Dataset. The base F5-TTS model was fine-tuned on the complete TORGO dataset over approximately 120,000 steps (~10 hours) to learn dysarthric vocal characteristics. This fine-tuned model was then used to synthesize novel imperative utterances, producing 5,124 synthesized files with no overt hallucinations detected during spot-checking and no manual preprocessing required. Combined with 854 real utterances, the synthetic-to-real ratio was approximately 6:1.

3.4. Audio Preprocessing

All audio files were preprocessed consistently across experiments:

- **Resampling**: All audio converted to 16 kHz mono to match Whisper's expected input format.
- **Silence/noise trimming**: Leading and trailing silence was trimmed using threshold-based voice activity detection. Background noise segments below a minimum energy threshold were removed.
- **Amplitude normalization**: Audio amplitude was normalized to a peak level of -3 dBFS to ensure consistent input levels across speakers and recording conditions.
- **Variable-length handling**: Whisper processes audio in fixed 30-second chunks. Audio files shorter than 30 seconds were zero-padded to the full window length. This is handled natively by Whisper's FeatureExtractor, which applies the log-Mel spectrogram transformation after padding.
- **Feature extraction**: Whisper's FeatureExtractor, Tokenizer, and Processor pipeline was used to convert audio to 80-channel log-Mel spectrograms before feeding them to the encoder.

3.5. Training Configuration and Hyperparameter Justification

Fine-tuning was implemented using the Hugging Face Seq2SeqTrainer following Gandhi (2022) and Liu et al. (2024). The training configuration parameters used for Seq2Seq fine-tuning are summarized in the Appendix, Figure 4. Key hyperparameters and their justifications:

- **Learning rate**: 1e-5. Selected based on Liu et al. (2024), who found that learning rates above 1e-4 destabilize Whisper fine-tuning on small datasets, and rates below 1e-6 produce negligible weight updates. 1e-5 with linear decay provided the best convergence behavior across pilot experiments.
- **Batch size**: 8 per device (effective batch 16 with gradient accumulation) — Constrained by GPU memory (A6000 NVIDIA, 48GB VRAM). Gradient accumulation steps of 2 were used to achieve an effective batch size of 16 without exceeding memory limits.
- **Warmup steps**: 500. Standard warmup for transformer fine-tuning; allows the optimizer to stabilize before the full learning rate is applied.
- **Max steps**: 1,000–5,000. Varied by experiment based on dataset size. For larger datasets (~5,600 samples), 5,000 steps allowed approximately 14–17 epochs. For smaller datasets (854 samples), 1,000 steps were sufficient before overfitting (one epoch ≈ 53 steps at an effective batch size of 16).
- **Evaluation steps**: Reduced from 1,000 (Experiment 5) to 100 (Experiments 6–14) after Experiment 5 revealed that overfitting occurred before the 1,000-step evaluation checkpoint.
- **Early stopping criterion**: WER stabilization. As documented in the training log for Experiment 11 (Appendix, Figure 5), the minimum validation loss occurred at approximately step 2,250, while WER continued to decrease until step

4,250. Using validation loss alone as the stopping criterion would have terminated training ~2,000 steps prematurely at a WER of ~17.05% rather than ~15.30%. WER-based stabilization is therefore used as the primary stopping criterion, consistent with recommendations in Liu et al. (2024).

- **Precision**: fp16. Mixed-precision training used to reduce memory footprint and accelerate training on NVIDIA hardware.

3.6. Experiment Summary

Table 1 summarizes all 14 experiments, including dataset partitions, speaker assignments, sample counts, and WER results.

Table 1. Summary of All 14 Experiments: Datasets, Speaker Partitions, and WER Results.

Exp	Description	Base Model	Train Dataset	Train Speakers	Train Samples	Test Dataset	Test Speakers	Test Samples	Key Hyperparameters	Overall WER (%)
1	Whisper Small baseline	Whisper Small	Hindi (Common Voice)	N/A	Standard	Hindi (Common Voice)	N/A	Standard	LR=1e-5, steps=4000	32.40
2	Pretrained Large V3 on UASpeech	Whisper Large V3	None (pretrained)	—	—	UASpeech (1-word)	F03, M12, M06, M11, M08, F05	18,743	No fine-tuning	127.57
3	Pretrained Large V3 on TORGO (1-word)	Whisper Large V3	None (pretrained)	—	—	TORGO (1-word)	M01, M02, M04, F03, M05, M03, F04	4,888	No fine-tuning	108.65
4	Pretrained Large V3 on TORGO (multi-word)	Whisper Large V3	None (pretrained)	—	—	TORGO (multi-word)	M04, M05, F05	454	No fine-tuning	43.17
5	Fine-tune on UASpeech (longer eval)	Whisper Large V3	UASpeech	9 patients	38,700	UASpeech	6 different patients	18,743	LR=1e-5, steps=5000, batch=16	— (overfit)
6	Fine-tune on UASpeech (shorter eval)	Whisper Large V3	UASpeech	9 patients	38,700	UASpeech	6 different patients	18,743	LR=1e-5, steps=1500, eval=100	28.41

7	Exp 6 model → UASpeech 1-word test	Exp 6 model	—	—	—	UASpeech (1-word)	F03, M12, M06, M11, M08, F05	18,743	Inference only	34.53 (avg)
8	Exp 6 model → TORGO 1-word test	Exp 6 model	—	—	—	TORGO (1-word)	M01, M02, M04, F03, M05, M03, F04	4,888	Inference only	65.64 (avg)
9	Fine-tune on TORGO multi-word	Whisper Large V3	TORGO (multi-word)	M01, M02, F01, M03, F03	854	TORGO (multi-word)	M04, M05, F04	454	LR=1e-5, steps=1000, eval=100	18.77
10	Exp 9 model → TORGO multi-word test	Exp 9 model	—	—	—	TORGO (multi-word)	M04, M05, F05	454	Inference only	17.69 (avg)
11	Fine-tune + TORGO + Voice Clone augment	Whisper Large V3	TORGO + Voice Clone	M01, M02, F01, M03, F03	5,578 (854+ 4,724)	TORGO (multi-word)	M04, M05, F04	454	LR=1e-5, steps=5000, eval=100	15.36
12	Exp 11 model → TORGO multi-word test	Exp 11 model	—	—	—	TORGO (multi-word)	M04, M05, F05	454	Inference only	15.53 (avg)
13	Fine-tune + TORGO + Speech Synthesis	Whisper Large V3	TORGO + Synthesis	M01, M02, F01, M03, F03	5,978 (854+ 5,124)	TORGO (multi-word)	M04, M05, F04	454	LR=1e-5, steps=5000, eval=100	16.84
14	Exp 13 model → TORGO multi-word test	Exp 13 model	—	—	—	TORGO (multi-word)	M04, M05, F05	454	Inference only	16.98 (avg)

Note: WER = Word Error Rate. Train/test speaker groups are non-overlapping in all experiments involving fine-tuning. Exp 5 is excluded from WER reporting due to confirmed overfitting before the first evaluation checkpoint.

4. Results

4.1. Why WER Exceeds 100%: An Explanation

Several experiments (2, 3) yield WER values exceeding 100%. This is mathematically possible because WER = (S + I + D) / N, where N is the number of words in the reference transcription. When a model generates extensive hallucinated output—inserting many spurious words not present in the reference—the count of insertions (I) alone can exceed N, pushing WER above 1.0 (100%). For severely dysarthric single-word utterances, the unmodified Whisper Large V3 model frequently produces multi-word hallucinated transcriptions in response to acoustically ambiguous or highly atypical input, resulting in the observed WER values of 122–163%.

4.2. Baseline: Pretrained Whisper Large V3 (Experiments 2–4)

The unmodified Whisper Large V3 model was evaluated on both datasets to establish baselines. Results are presented in Table 2.

Table 2. Baseline WER of Unmodified Whisper Large V3 by Dataset and Severity.

Experiment	Dataset	Utterance Type	Severe WER (%)	Manageable WER (%)	Mild WER (%)	Overall WER (%)
2	UASpeech	Single-word	142.57	163.77	110.68	127.57
3	TORGO	Single-word	122.95	114.71	86.24	108.65
4	TORGO	Multi-word	90.95	34.08	5.74	43.17

Note: WER = Word Error Rate. *Data from Rajamony & Karthik (2025), Patent FIG. 5 (Appendix, Figure 5).*

The particularly high WER for manageable speakers in UASpeech (163.77%) reflects hallucination behavior: the model's decoder, receiving acoustically ambiguous input characteristic of moderate dysarthria, produces spurious word sequences longer than the reference. The substantially lower WER for multi-word utterances (Experiment 4) compared to single-word utterances (Experiments 2–3) demonstrates that linguistic context enables the decoder to partially recover from acoustic ambiguity. The severity-stratified comparison between baseline and fine-tuned models is presented in the Appendix, Figure 10.

4.3. Fine-Tuning on UASpeech (Experiments 5–8)

Experiment 5 revealed that fine-tuning with eval_steps=1000 allowed overfitting before the first evaluation point—training loss reached zero before a model checkpoint was saved with a valid WER. Experiment 6 reduced max_steps to 1,500 and eval_steps to 100, preventing premature convergence and achieving WER=28.41% on the UASpeech test set. The Experiment 6 model was subsequently tested on both UASpeech (Experiment 7) and TORGO single-word utterances (Experiment 8), showing improvements across all severity levels:

Table 3. WER of Fine-Tuned UASpeech Model (Experiments 7–8) vs. Baseline

Exp	Test Dataset	Severe WER (%)	Manageable WER (%)	Mild WER (%)	vs. Baseline Improvement
7	UASpeech (1-word)	81.13	18.34	5.56	43.1% / 88.8% / 95.0% better
8	TORGO (1-word)	76.53	70.53	49.87	37.8% / 38.5% / 42.2% better

Note: WER = *Word Error Rate.*

4.4. Fine-Tuning on TORGO Multi-Word Utterances (Experiments 9–10)

Experiment 9 fine-tuned Whisper Large V3 exclusively on TORGO imperative sentences (854 files, 5 training patients, 3 different test patients). Despite the small training set, WER reached 18.77%—a substantial reduction from the 43.17% baseline. Experiment 10 confirmed: WER of 33.04% (severe), 17.52% (manageable), and 2.50% (mild) compared to baseline values of 90.95%, 34.08%, and 5.74%, representing 63.7%, 48.5%, and 56.4% improvements respectively.

4.5. Voice-Cloning Augmentation (Experiments 11–12)

Experiment 11 augmented the 854-sample real training corpus with 4,724 voice-cloned samples (synthetic:real ratio ≈ 5.5:1), expanding training to ~5,578 samples. Training ran for 5,000 steps (14–17 epochs at effective batch size 16). WER on the validation set reached 15.36%—an improvement of 1.41 percentage points over Experiment 9. The augmentation did not substantially alter the severity balance of the training set, as voice cloning was applied uniformly across training-set patient groups. Experiment 12 tested this model on TORGO multi-word utterances:

Table 4. Voice-Clone-Augmented Model WER (Experiment 12) vs. Pre Trained Baseline

Severity	Baseline WER (%)	Fine-Tuned + Clone WER (%)	WER Reduction (%)
Severe	90.95	26.89	70.4%
Manageable	34.08	17.04	50.0%
Mild	5.74	2.66	53.7%

Note: WER = *Word Error Rate.*

4.6. Speech-Synthesis Augmentation (Experiments 13–14)

Experiment 13 replaced voice-cloned data with speech-synthesized samples (~5,124 files; synthetic:real ratio ≈ 6:1). WER reached 16.84% on the validation set—1.48 percentage points higher than the voice-clone model. Experiment 14 confirmed:

Speech synthesis produced slightly higher WER than voice cloning despite eliminating hallucination artifacts, suggesting that fine-tuning F5-TTS on the full TORGO corpus introduces averaging effects across speakers that reduce the acoustic individuality of the generated samples.

Table 5. Speech-Synthesis-Augmented Model WER (Experiment 14) vs. Voice-Clone Model (Exp 12)

Severity	Voice-Clone WER (%)	Speech-Synth WER (%)	Difference
Severe	26.89	29.44	+2.55 (worse)
Manageable	17.04	18.49	+1.45 (worse)
Mild	2.66	3.00	+0.34 (worse)

Note. WER = *Word Error Rate.*

4.7. Training Dynamics: WER vs. Validation Loss Divergence

The detailed training log for Experiment 11 (Appendix, Figure 5) illustrates a key finding: validation loss reached its minimum of 0.4121 at step 2,250, while WER continued improving to 15.36% at step 5,000. Had training been stopped at minimum validation loss, WER would have been approximately 17.05%—1.69 percentage points higher than the final model. This empirically justifies WER-based early stopping over validation-loss-based stopping, consistent with Liu et al. (2024). The training-loss convergence curve is shown in the Appendix, Figure 6; the linear-decay learning-rate schedule in the Appendix, Figure 7; the WER trajectory across training steps in the Appendix, Figure 8; and the non-monotonic evaluation-loss curve illustrating the WER vs. validation-loss discrepancy in the Appendix, Figure 9.

5. Results

5.1. Restatement of Contributions

This study demonstrates that fine-tuning large-scale ASR models can improve recognition accuracy for dysarthric speech, with voice-cloning-based augmentation outperforming speech synthesis across all severity levels.

More importantly, this work provides a systematic empirical evaluation of how state-of-the-art ASR models can be adapted to a highly underserved and clinically relevant domain. The study offers new insights into (1) the effectiveness of synthetic data augmentation strategies, (2) the relationship between dysarthria severity and ASR performance, and (3) the importance of WER-based training optimization in low-resource scenarios.

Together, these contributions extend existing knowledge by demonstrating not only that adaptation is possible, but how it can be practically achieved and deployed in real-world assistive applications.

5.2. Why Severe Dysarthria Produces Higher WER

The consistent severity-WER relationship across all experiments reflects the underlying acoustic properties of dysarthric speech. Severe dysarthria involves pronounced breakdown of articulatory precision: phonemes are produced with reduced acoustic distinctiveness, co-articulation patterns become irregular, and prosodic cues (stress, rhythm, intonation) that support ASR decoding are largely absent (Young & Mihailidis, 2010). The Whisper encoder, trained on typical speech spectrograms, produces degraded latent representations for severely atypical input, and the decoder—which relies on both acoustic evidence and language model priors—defaults to higher-probability word sequences that may not match the

reference (hallucinations). This explains the observed WER >100% for severe speakers under the unmodified baseline (Experiments 2–3).

After fine-tuning with dysarthric training data, the encoder learns to map atypical spectrograms to more appropriate latent representations, and the decoder learns dysarthric-specific acoustic-phonetic correspondences. However, the improvement is largest for severe speakers (70% WER reduction) precisely because the baseline is highest—leaving the most room for improvement—rather than because severe speech is inherently easier to model after fine-tuning. Despite these improvements, severe dysarthric speech likely remains more challenging for ASR systems to decode accurately than mild or moderate speech, due to its reduced acoustic distinctiveness and breakdown of articulatory precision.

5.3. Comparison to Prior Dysarthric ASR Systems

Schu et al. (2022) reported consistent WER improvement when using ASR systems specifically adapted for UASpeech and TORGO, establishing that domain-specific adaptation is necessary. Rathod et al. (2023) demonstrated that transfer learning with Whisper improves dysarthric transcription accuracy, consistent with the findings of this study. The WER values achieved here (15.36–18.77% for multi-word TORGO utterances with fine-tuning) compare favorably to the baseline values reported in Schu et al. (2022) for unadapted systems, though direct numerical comparison is constrained by differences in dataset partitioning and evaluation protocols.

5.4. Voice Cloning vs. Speech Synthesis

Voice cloning outperformed speech synthesis in all severity categories (Experiments 12 vs. 14). This is likely because voice cloning preserves the speaker-specific acoustic fingerprint of individual dysarthric patients, including their idiosyncratic articulatory patterns, whereas speech synthesis averages over the full training population, producing samples that are acoustically less representative of any specific speaker's impairment profile. The manual quality-control step (removing 185 hallucinated voice-cloned files) was essential to the quality of the augmented dataset.

5.5. Contextual Information and Single vs. Multi-Word Utterances

The dramatic WER difference between single-word (Experiments 2–3: WER >100%) and multi-word utterances (Experiment 4: WER=43.17% baseline; Experiment 11: WER=15.36% fine-tuned) reflects the language model's ability to leverage sequential context. In multi-word sequences, the Whisper decoder uses probability distributions over preceding tokens to constrain candidate words, partially compensating for degraded acoustic features. Single-word utterances provide no such context, forcing the decoder to rely entirely on a single degraded acoustic representation—leading to hallucination in severely impaired cases.

5.6. Practical Application: ClearVoiceAI

The fine-tuned model was deployed in ClearVoiceAI, a web application (clearvoiceai.com) enabling dysarthric users to record speech or upload audio files for real-time transcription. The application stack comprises a JavaScript/CSS frontend, a Python/FastAPI backend, and the fine-tuned Whisper model hosted on AWS. The deployed system architecture is shown in the Appendix, Figure 3; the underlying computing-device implementation, network architecture, system hierarchy, and detailed block diagram are depicted in the Appendix, Figure 12 and Figures 14–16. An extension of the framework into a

virtual-meeting AI agent for telemedicine use is illustrated in the Appendix, Figure 11. The voice-clone-augmented model (Experiment 11) serves as the default ASR engine. Practical deployment revealed that environmental noise and spontaneous (non-scripted) dysarthric speech present additional challenges beyond those captured in the controlled-environment TORGO and UASpeech recordings.

5.7. Limitations

Several limitations should be acknowledged when interpreting the findings of this study. Each experiment was conducted as a single training run due to computational cost constraints (each run required 1.5–6.5 hours on a dedicated GPU server), so no cross-run variance estimates are reported. Variance estimation through repeated runs or cross-validation over speaker partitions should be incorporated in future work to strengthen statistical claims. Evaluation was also restricted to the controlled-environment recordings in TORGO and UASpeech, leaving real-world acoustic conditions untested. The total training corpus was limited to approximately 5,600 samples; larger corpora would likely improve generalization. Finally, speaker-specific fine-tuning—personalizing models for individual patients—was not explored in the present work.

Limitations of WER as a Sole Metric. WER does not distinguish between substitution, insertion, and deletion error types, which carry different clinical implications. Substitution errors (incorrect word recognized) are typically less disruptive than deletion errors (words missed entirely) for communication-aid applications. Future work should report error-type breakdowns per severity level to provide a more clinically meaningful characterization of model behaviour.

5.8. Future Directions

Several promising directions extend the work presented here. Longitudinal WER monitoring could be developed as a clinical tool to track dysarthria progression over time, while speaker-adaptive fine-tuning would allow models to be personalized to individual patients in near real-time. Beyond dysarthria, the same fine-tuning and augmentation framework can be extended to other speech disorders such as dysphagia, apraxia, and aphasia, and cross-linguistic evaluation should investigate dysarthric speech in languages other than English. Robustness in deployment can be further improved through evaluation under real-world noise conditions and through edge deployment that enables on-device inference without cloud dependency, which is particularly important for users with limited or intermittent connectivity.

6. Conclusion

This study empirically evaluated fine-tuning strategies for adapting Whisper Large V3 to dysarthric speech, using the publicly available UASpeech and TORGO datasets with strictly non-overlapping speaker partitions between training and test sets. Across 14 systematic experiments, voice-cloning augmentation produced the best-performing model (WER=15.36% on TORGO multi-word utterances), with severity-specific reductions of approximately 70% (severe), 50% (moderate), and 54% (mild) relative to the unmodified pretrained baseline.

The primary contributions of this work lie in the systematic application and evaluation of state-of-the-art ASR adaptation techniques to the underserved domain of dysarthric speech recognition. Through 14 controlled experiments, the study provides new empirical insights into the effectiveness of fine-tuning strategies, the comparative impact of voice cloning versus speech synthesis, and the influence of severity-specific acoustic variability on model performance.

In addition to these experimental findings, the deployment of the fine-tuned model in the ClearVoiceAI system demonstrates the practical feasibility of translating research advances into real-world assistive technology. These contributions collectively strengthen both the scientific understanding and applied capabilities of dysarthric speech recognition systems.

7. References

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., & Chen, X. (2024). *F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2410.06885

Farhadipour, A., & Veisi, H. (2023). *Gammatonegram representation for end-to-end dysarthric speech processing tasks: Speech recognition, speaker identification, and intelligibility assessment* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.03296

Gandhi, S. (2022, November 3). *Fine-tune Whisper for multilingual ASR with Transformers*. Hugging Face. https://huggingface.co/blog/fine-tune-whisper

Kim, H., Hasegawa-Johnson, M., Gunderson, J., Perlman, A., Huang, T., Watkin, K., Frame, S., Sharma, H. V., & Zhou, X. (2023). *UASpeech* [Data set]. IEEE DataPort. https://doi.org/10.21227/f9tc-ab45

Lettergram. (n.d.). *Sentence classification* [Code repository]. GitHub. Retrieved June 10, 2026, from https://github.com/lettergram/sentence-classification

Liu, Y., Yang, X., & Qu, D. (2024). Exploration of Whisper fine-tuning strategies for low-resource ASR. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024, Article 29. https://doi.org/10.1186/s13636-024-00349-3

Page, A. D., & Yorkston, K. M. (2022). Communicative participation in dysarthria: Perspectives for management. *Brain Sciences*, 12(4), Article 420. https://doi.org/10.3390/brainsci12040420

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2212.04356

Rajamony, K., & Karthik, A. (2025, December 2). *Speech recognition for assisting patients with speech difficulties* (U.S. Patent No. 12,488,786 B1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US12488786B1

Rathod, S., Charola, M., & Patil, H. A. (2023). Transfer learning using Whisper for dysarthric automatic speech recognition. In A. Karpov, K. Samudravijaya, K. T. Deepak, R. M. Hegde, S. S. Agrawal, & S. R. M. Prasanna (Eds.), *Speech and computer: SPECOM 2023* (Lecture Notes in Computer Science, Vol. 14338, pp. 579–589). Springer. https://doi.org/10.1007/978-3-031-48309-7_46

Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, 46(4), 523–541. https://doi.org/10.1007/s10579-011-9145-0

Schölderle, T., Haas, E., & Ziegler, W. (2021). Dysarthria syndromes in children with cerebral palsy. *Developmental Medicine & Child Neurology*, 63(4), 444–449. https://doi.org/10.1111/dmcn.14679

Schu, G., Janbakhshi, P., & Kodrasi, I. (2022). *On using the UA-Speech and TORGO databases to validate automatic dysarthric speech classification approaches* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2211.08833

Seagraves, A. (2025, May 30). *Benchmarking top open source speech recognition models: Whisper, Facebook wav2vec2, and Kaldi*. Deepgram. https://deepgram.com/learn/benchmarking-top-open-source-speech-models

Shih, D.-H., Liao, C.-H., Wu, T.-W., Xu, X.-Y., & Shih, M.-H. (2022). Dysarthria speech detection using convolutional neural networks with gated recurrent unit. *Healthcare*, 10(10), Article 1956. https://doi.org/10.3390/healthcare10101956

Tomik, B., & Guiloff, R. J. (2010). Dysarthria in amyotrophic lateral sclerosis: A review. *Amyotrophic Lateral Sclerosis*, 11(1–2), 4–15. https://doi.org/10.3109/17482960802379004

Young, V., & Mihailidis, A. (2010). Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, 22(2), 99–112. https://doi.org/10.1080/10400435.2010.483646

Code and Data Availability

The fine-tuned Whisper models are publicly available on Hugging Face. Training code is maintained on GitHub. Training metrics, loss curves, and WER logs are available on WandB.AI.

- **Code repository: GitHub** — Fine-tuning and inference scripts
- **Model repository**: Hugging Face — Fine-tuned Whisper Large V3 models
- **Training metrics**: WandB.AI — Training loss, validation loss, WER logs across all experiments
- **Application**: ClearVoiceAI — Deployed ASR web application

U.S. Patent Documents Cited by Examiner

The following prior-art patents and patent applications were cited by the U.S. Patent and Trademark Office examiner during the prosecution of U.S. Patent No. 12,488,786 B1. They are listed here for completeness and are formatted in APA 7th edition style.

Burkhardt, F. (2016, April 14). *Method for the interpretation of automatic speech recognition* (U.S. Patent Application Publication No. 2016/0104477 A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20160104477A1

Burns, T. (2020, September 3). *Voice cloning for hearing device* (U.S. Patent Application Publication No. 2020/0279549 A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20200279549A1

Chang, E. F., & Moses, D. A. (2022, September 22). *Method of contextual speech decoding from the brain* (U.S. Patent Application

Publication No. 2022/0301563 A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20220301563A1

Ingel, B. A., & Zass, R. (2025, January 2). *Generating and operating personalized artificial entities* (U.S. Patent Application Publication No. 2025/0006182 A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20250006182A1

Kanevsky, D., Nesbitt, P. A., Sainath, T. N., & Woodward, E. V. (2014, July 31). *System and method for improving voice communication over a network* (U.S. Patent Application Publication No. 2014/0214426 A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20140214426A1

Koul, A., Kasam, M., Johnston, M., Machanavajhala, S., & Halper, E. (2023, July 11). *Automated real time interpreter service* (U.S. Patent No. 11,699,360 B2). U.S. Patent and Trademark Office. https://patents.google.com/patent/US11699360B2

Krishna, G., Carnahan, M., Chandar, A., Shamapant, S., Tewfik, A., & del R. Millán, J. (2023, May 4). *EEG based speech prosthetic for stroke survivors* (U.S. Patent Application Publication No. 2023/0139394 A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20230139394A1

Li, X., Cheng, X., & Yang, X. (2024, October 17). *Systems and methods for enhanced speaker diarization* (U.S. Patent Application Publication No. 2024/0347064 A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20240347064A1

Lin, T.-J., Sung, C.-H., Pai, C.-C., & Yeh, C.-W. (2020, October 1). *System for improving dysarthria speech intelligibility and method thereof* (U.S. Patent Application Publication No. 2020/0312302 A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20200312302A1

Lin, T. J., Yeh, C. W., Yang, S. P., & Liao, C. Z. (2021, July 22). *Device and method for generating synchronous corpus* (U.S. Patent Application Publication No. 2021/0225384 A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20210225384A1

McNair, D. S. (2023, September 14). *Tool for assisting people with speech disorder* (U.S. Patent Application Publication No. 2023/0290353 A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20230290353A1

McNulty, S. F., & McNulty, M. S. (2024, October 31). *Systems, methods, and devices to curate and present content and physical elements based on personal biometric identifier information* (U.S. Patent Application Publication No. 2024/0361827 A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20240361827A1

Phillips, M. S., & Nguyen, J. N. (2011, March 3). *Sending a communications header with voice recording to send metadata for use in speech recognition and formatting in mobile dictation application* (U.S. Patent Application Publication No. 2011/0054896 A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20110054896A1

Sharma, M. K. (2025, March 27). *Intelligent system and method of providing speech assistance during a communication session* (U.S. Patent Application Publication No. 2025/0104689 A1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US20250104689A1

Wang, X., Zechner, K., & Hamill, C. (2025, March 11). *Targeted content feedback in spoken language learning and assessment* (U.S. Patent No. 12,249,324 B1). U.S. Patent and Trademark Office. https://patents.google.com/patent/US12249324B1

Author Note

Correspondence concerning this article should be addressed to Akshobh Karthik, Rancho Palos Verdes, California, United States.

Competing Interests: The authors declare that the methods and systems described in this article are the subject of U.S. Patent No. 12,488,786 B1, granted December 2, 2025, assigned to ARTIK LLC. Akshobh Karthik and Karthik Rajamony are named co-inventors.

Acknowledgements

I would like to thank my parents for their continuous encouragement, guidance, and support throughout this research project and my academic journey. I also thank my school for fostering an environment that encourages scientific inquiry, independent learning, and interdisciplinary research. Finally, I appreciate the broader STEM and research community, whose published work and open-source contributions supported the development of this project.

Author Biography

Akshobh Karthik is a high school researcher and student at Palos Verdes Peninsula High School in California. His academic interests span multiple STEM disciplines with a strong focus on biomedical science, artificial intelligence, and computational science. He has conducted work in biomedical AI and secure voice systems, including speech recognition technologies designed to assist individuals with speech impairments.

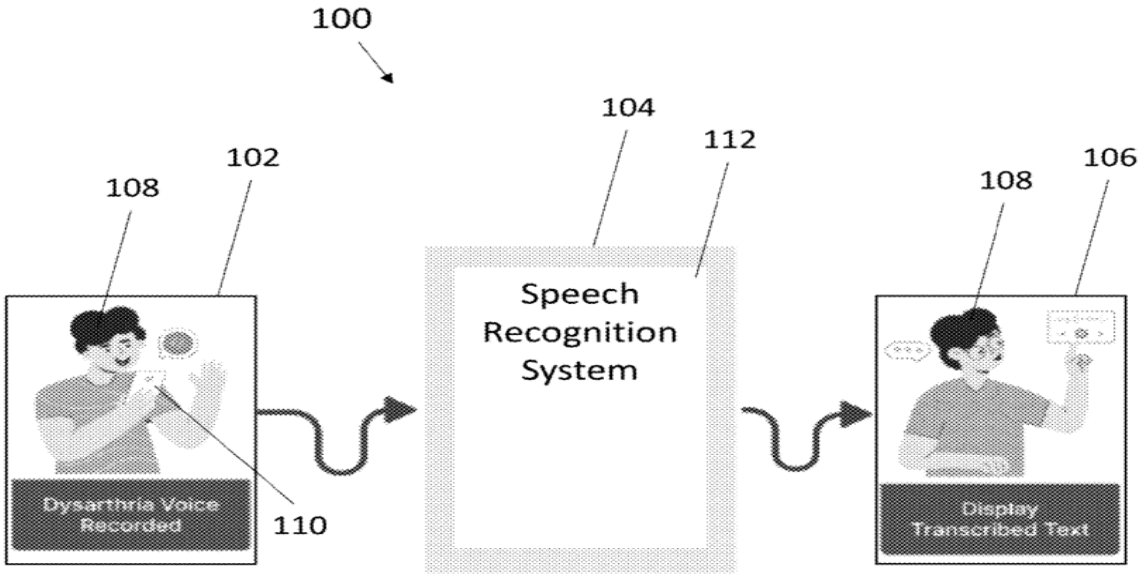
Mentor Contribution Statement

Mr. Greenberg served as Akshobh's science research mentor throughout the project. He provided guidance on research methodology, experiment planning, and the overall structure of the research process. He also supported the review of the presentation and the scientific communication of the project.

Appendix

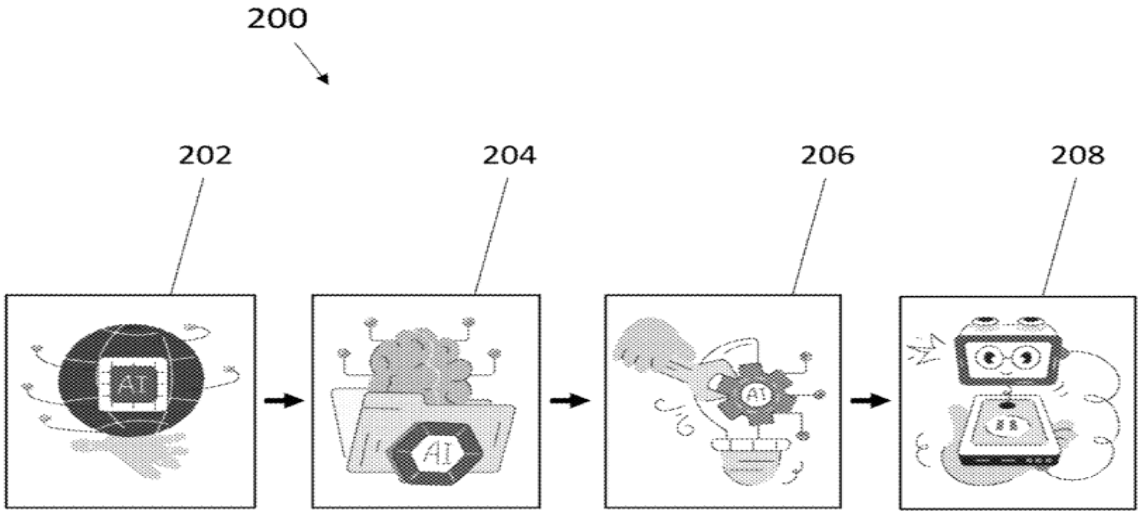
Patent Figures from U.S. Patent No. 12,488,786 B1 — Rajamony, K., & Karthik, A. (2025). Speech recognition for assisting patients with speech difficulties. ARTIK LLC.

Appendix, Figure 1. Overview of the speech recognition system for dysarthric speech, illustrating the end-to-end flow from voice recording to transcribed text display.



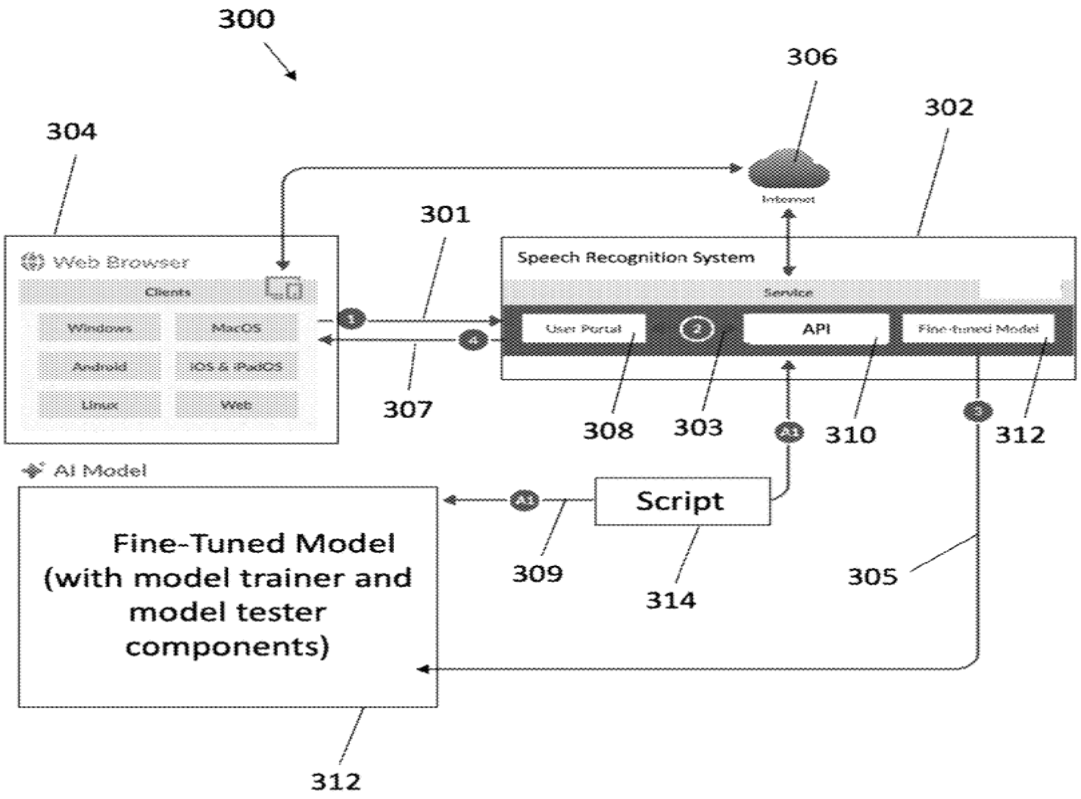
Source: Patent FIG. 1. U.S. Patent No. 12,488,786 B1. Cited in Section 1.1.

Appendix, Figure 2. Four-stage processing pipeline depicting AI-powered speech recognition: input acquisition, feature extraction, model inference, and output generation.



Source: Patent FIG. 2. U.S. Patent No. 12,488,786 B1. Provides a complementary view of the deployed inference pipeline.

Appendix, Figure 3. System architecture showing web browser clients, the Speech Recognition System service layer (User Portal, API, Fine-tuned Model), and the model training infrastructure with trainer and tester components.

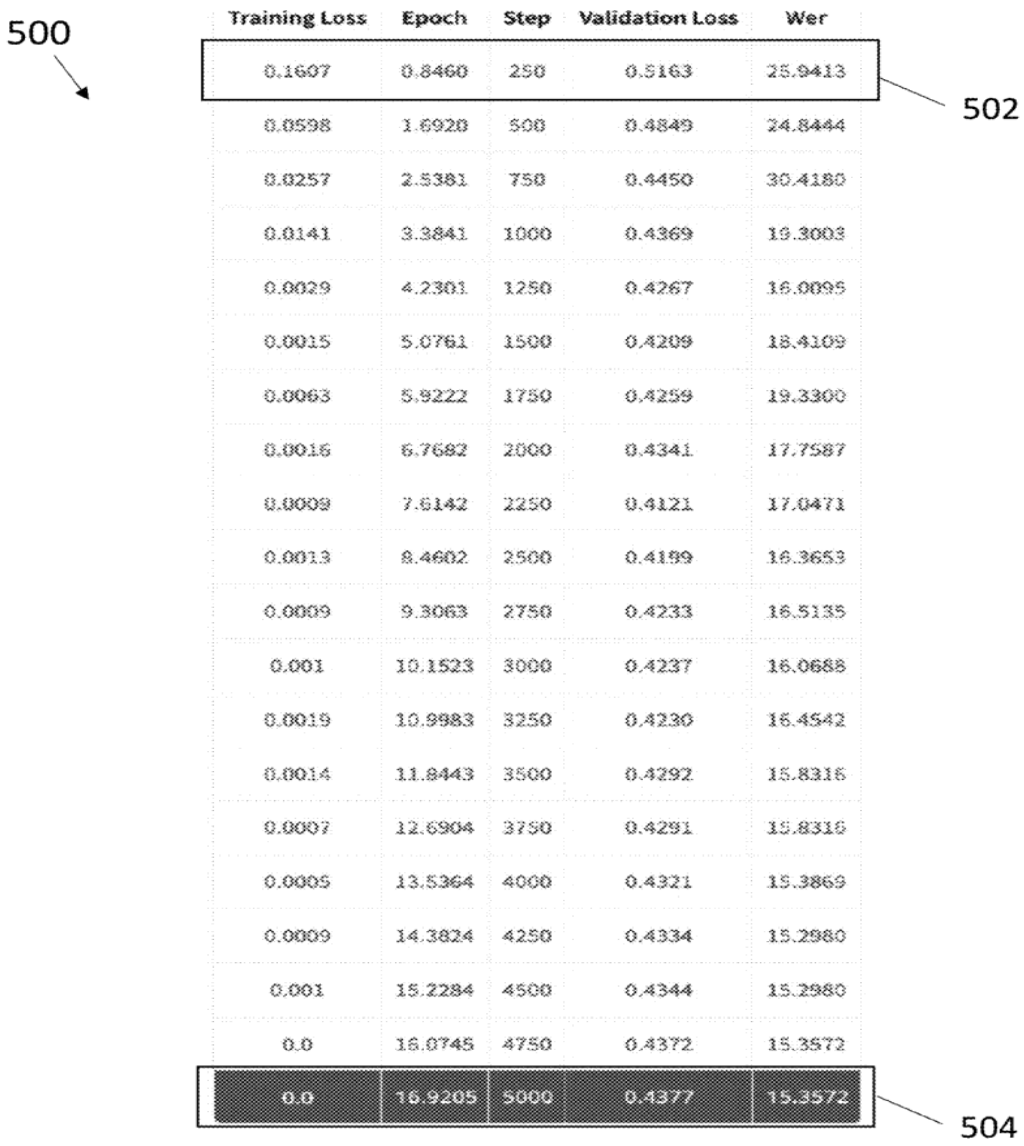


Source: Patent FIG. 3. U.S. Patent No. 12,488,786 B1. Cited in Section 5.6.

Appendix, Figure 4. Training configuration parameters for Seq2Seq model fine-tuning, including learning rate, batch size, gradient accumulation, warmup steps, and WER-based evaluation settings.

Source: Patent FIG. 4. U.S. Patent No. 12,488,786 B1. Cited in Section 3.5.

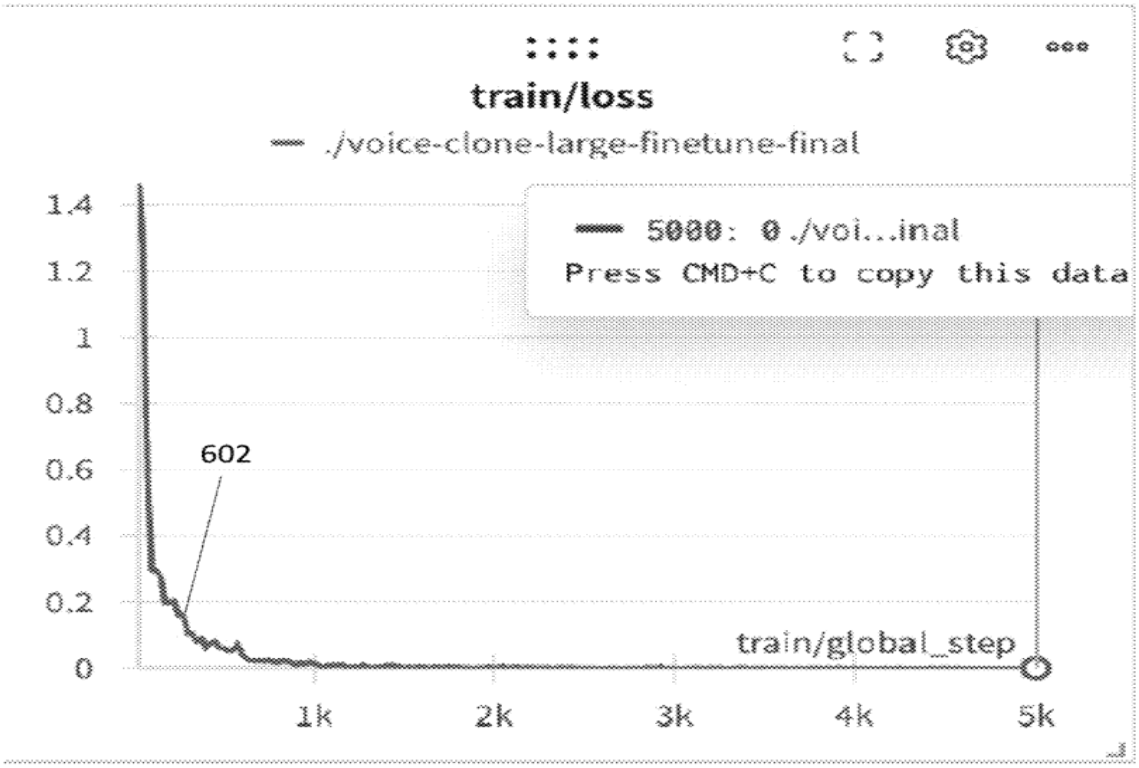
Appendix, Figure 5. Full training metrics log across 5,000 optimization steps, showing training loss, epoch, validation loss, and Word Error Rate (WER) at each checkpoint.



Training Loss	Epoch	Step	Validation Loss	Wer
0.1607	0.8460	250	0.5163	25.9413
0.0598	1.6920	500	0.4849	24.8444
0.0257	2.5381	750	0.4450	30.4180
0.0141	3.3841	1000	0.4369	19.3003
0.0029	4.2301	1250	0.4267	16.0095
0.0015	5.0761	1500	0.4209	18.4109
0.0063	5.9222	1750	0.4259	19.3300
0.0016	6.7682	2000	0.4341	17.7587
0.0009	7.6142	2250	0.4121	17.0471
0.0013	8.4602	2500	0.4199	16.3653
0.0009	9.3063	2750	0.4233	16.5135
0.001	10.1523	3000	0.4237	16.0688
0.0019	10.9983	3250	0.4230	16.4542
0.0014	11.8443	3500	0.4292	15.8316
0.0007	12.6904	3750	0.4291	15.8316
0.0005	13.5364	4000	0.4321	15.3869
0.0009	14.3824	4250	0.4334	15.2980
0.001	15.2284	4500	0.4344	15.2980
0.0	16.0745	4750	0.4372	15.3572
0.0	16.9205	5000	0.4377	15.3572

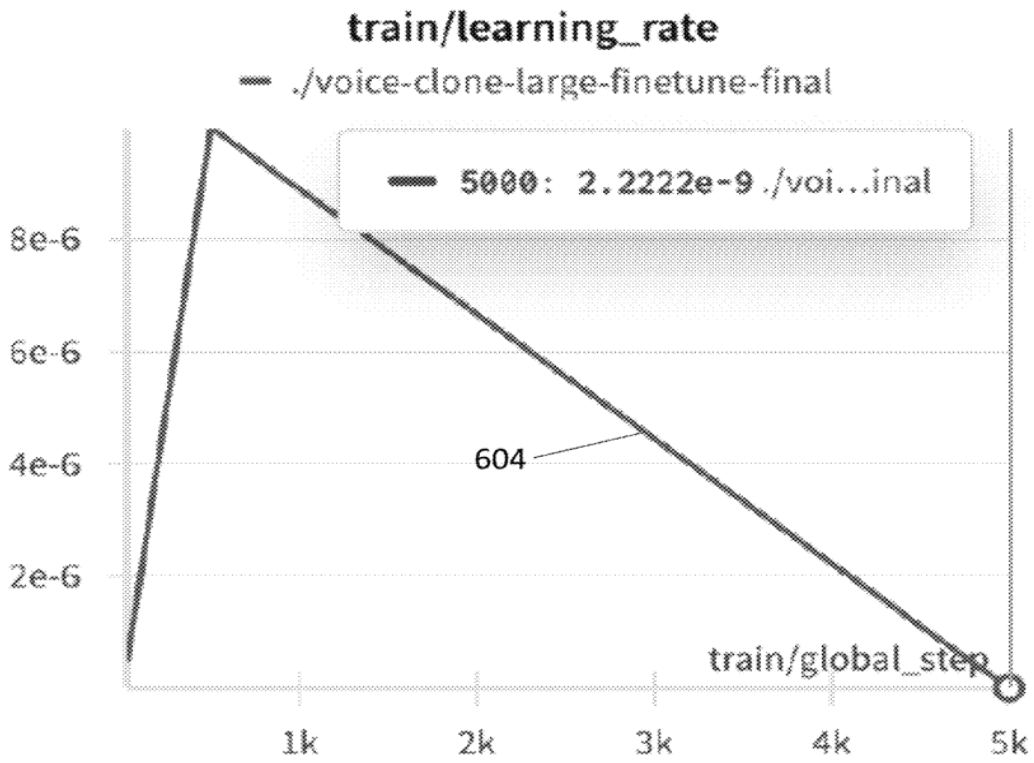
Source: Patent FIG. 5. U.S. Patent No. 12,488,786 B1. Cited in Sections 3.5, 4.2, and 4.7.

Appendix, Figure 6. Training loss convergence curve across global training steps, demonstrating a rapid decrease from ~1.4 to near zero by step 5,000.



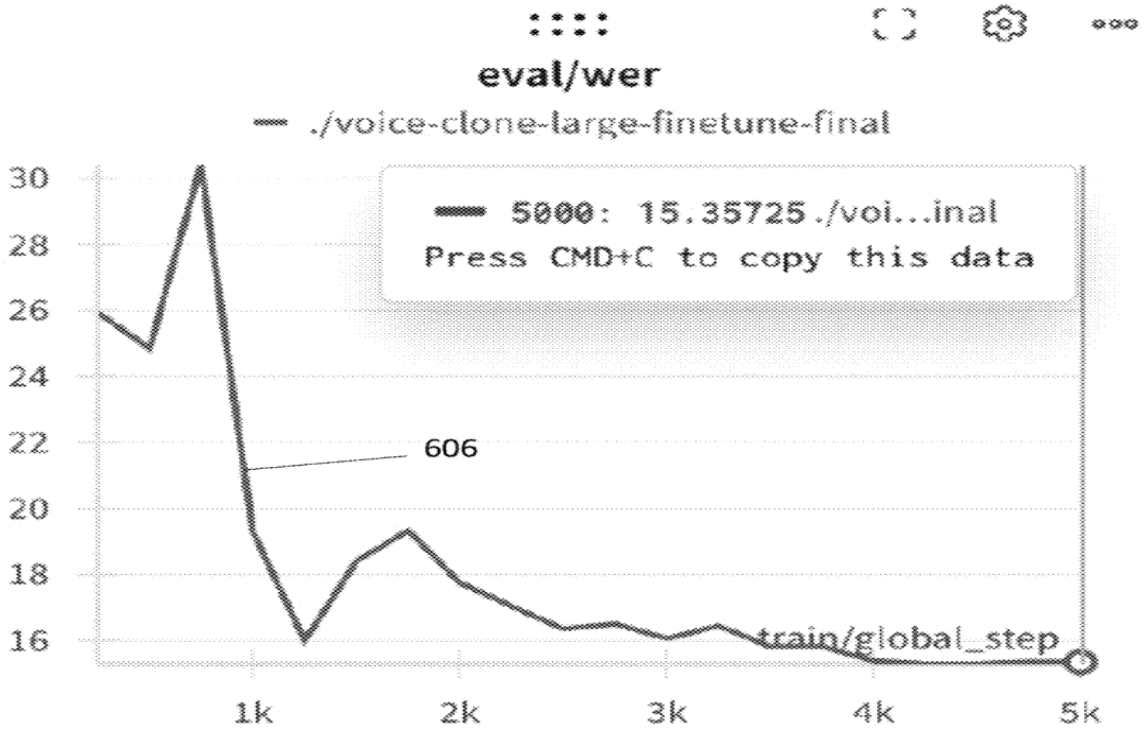
Source: Patent FIG. 6A. U.S. Patent No. 12,488,786 B1. Cited in Section 4.7.

Appendix, Figure 7. Learning rate schedule over the course of fine-tuning, showing a linear warmup phase followed by linear decay from ~1e-5 to near zero.



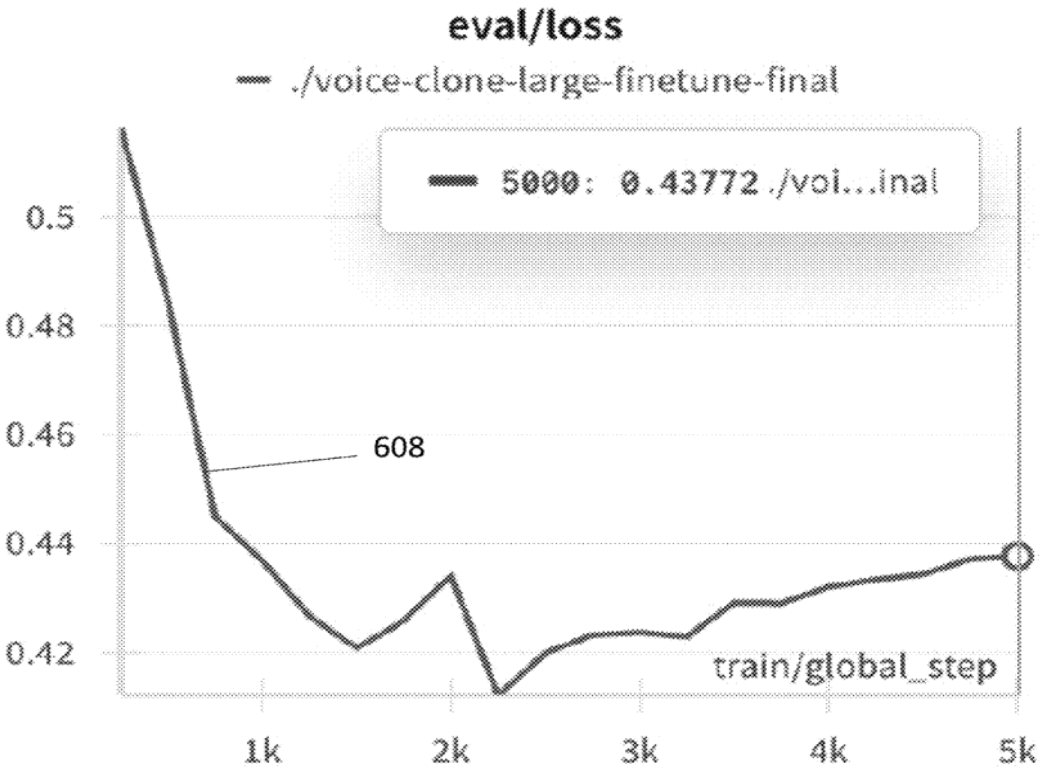
Source: Patent FIG. 6B. U.S. Patent No. 12,488,786 B1. Cited in Section 4.7.

Appendix, Figure 8. Word Error Rate (WER) trajectory across training steps, illustrating non-monotonic convergence with a final WER of approximately 15.36% at step 5,000.



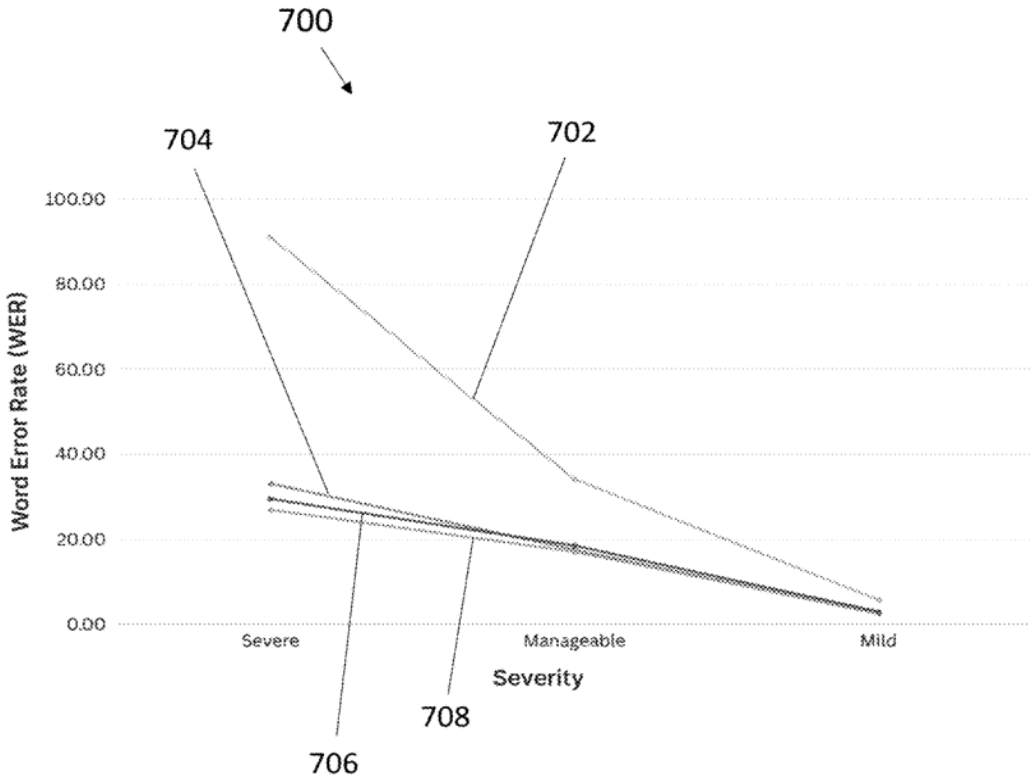
Source: Patent FIG. 6C. U.S. Patent No. 12,488,786 B1. Cited in Section 4.7.

Appendix, Figure 9. Evaluation loss across training steps, reaching its minimum of ~0.4121 at step 2,250 and diverging slightly thereafter, illustrating the WER vs. validation loss discrepancy discussed in Section 4.7.



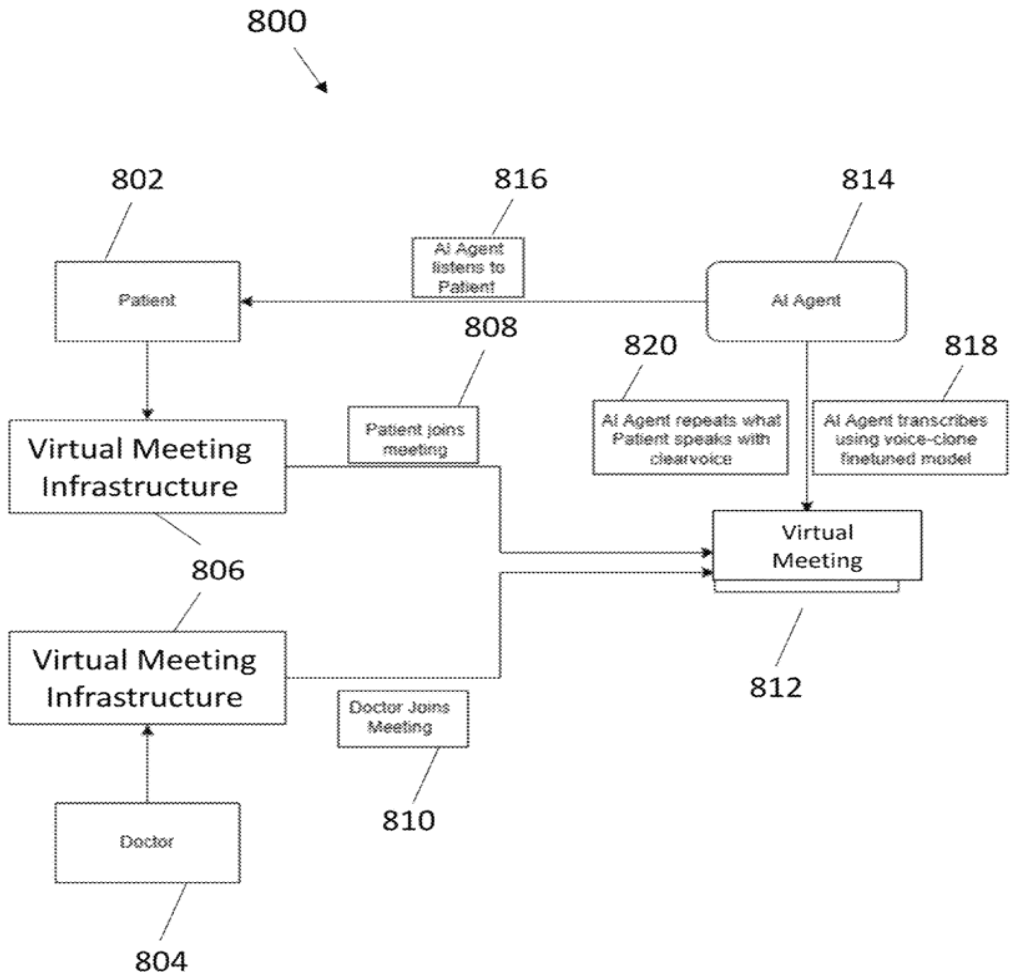
Source: Patent FIG. 6D. U.S. Patent No. 12,488,786 B1. Cited in Section 4.7.

Appendix, Figure 10. WER comparison across speech impairment severity levels (Severe, Manageable, Mild) for baseline versus fine-tuned ASR models, demonstrating the greatest improvement for severely impaired speakers.



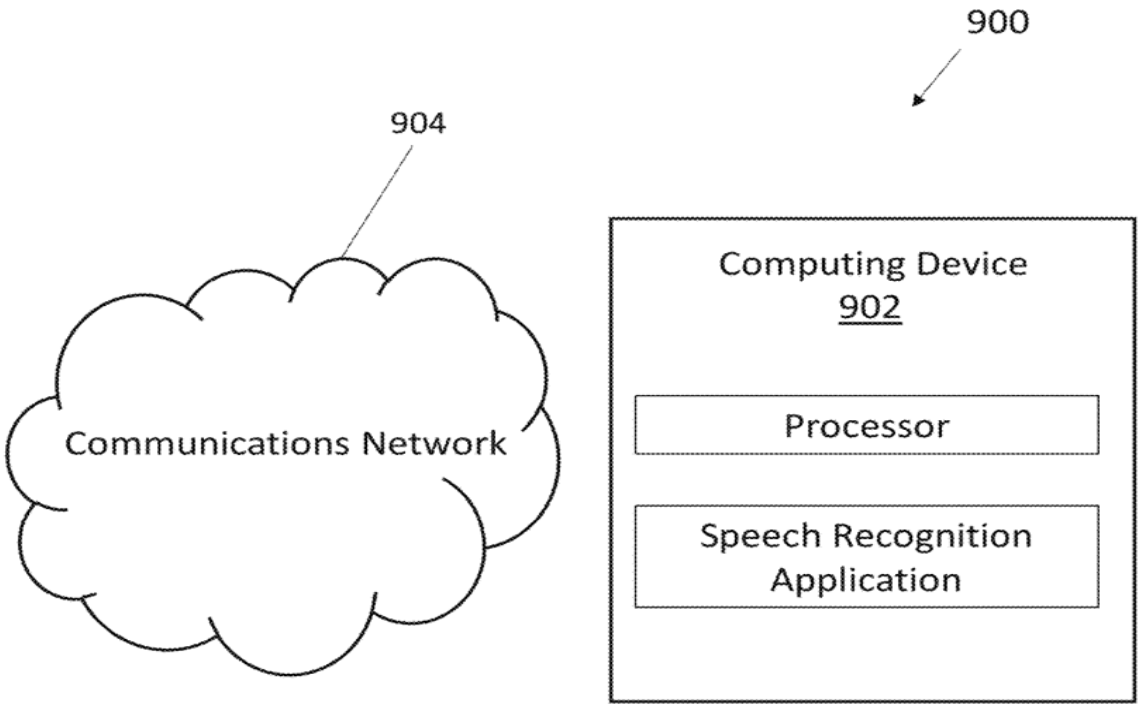
Source: Patent FIG. 7. U.S. Patent No. 12,488,786 B1. Cited in Section 4.2.

Appendix, Figure 11. Virtual meeting AI agent architecture enabling dysarthric patients to participate in telemedicine and virtual consultations via AI-powered transcription and clear-voice re-vocalization.



Source: Patent FIG. 8. U.S. Patent No. 12,488,786 B1. Cited in Section 5.6.

Appendix, Figure 12. Computing system network architecture for speech recognition and transcription, showing the computing device connected via communications network.



Source: Patent FIG. 9. U.S. Patent No. 12,488,786 B1. Cited in Section 5.6.

Appendix, Figure 13. Overall analytical methodology pipeline. The study follows a systematic five-stage pipeline: (1) collection of dysarthric speech datasets (UASpeech and TORGO) with strict speaker-wise splits; (2) audio preprocessing to standardize inputs;

(3) synthetic data augmentation using voice cloning and speech synthesis; (4) fine-tuning of Whisper Large V3 using parameter-efficient adaptation; and (5) evaluation on unseen test sets using Word Error Rate (WER) across dysarthria severity levels.

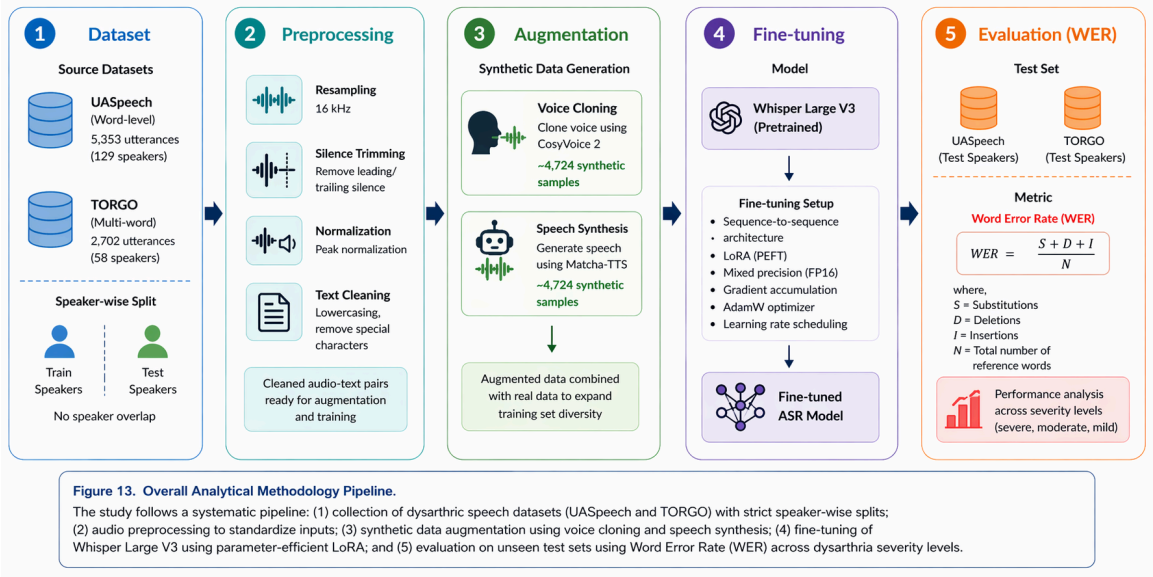
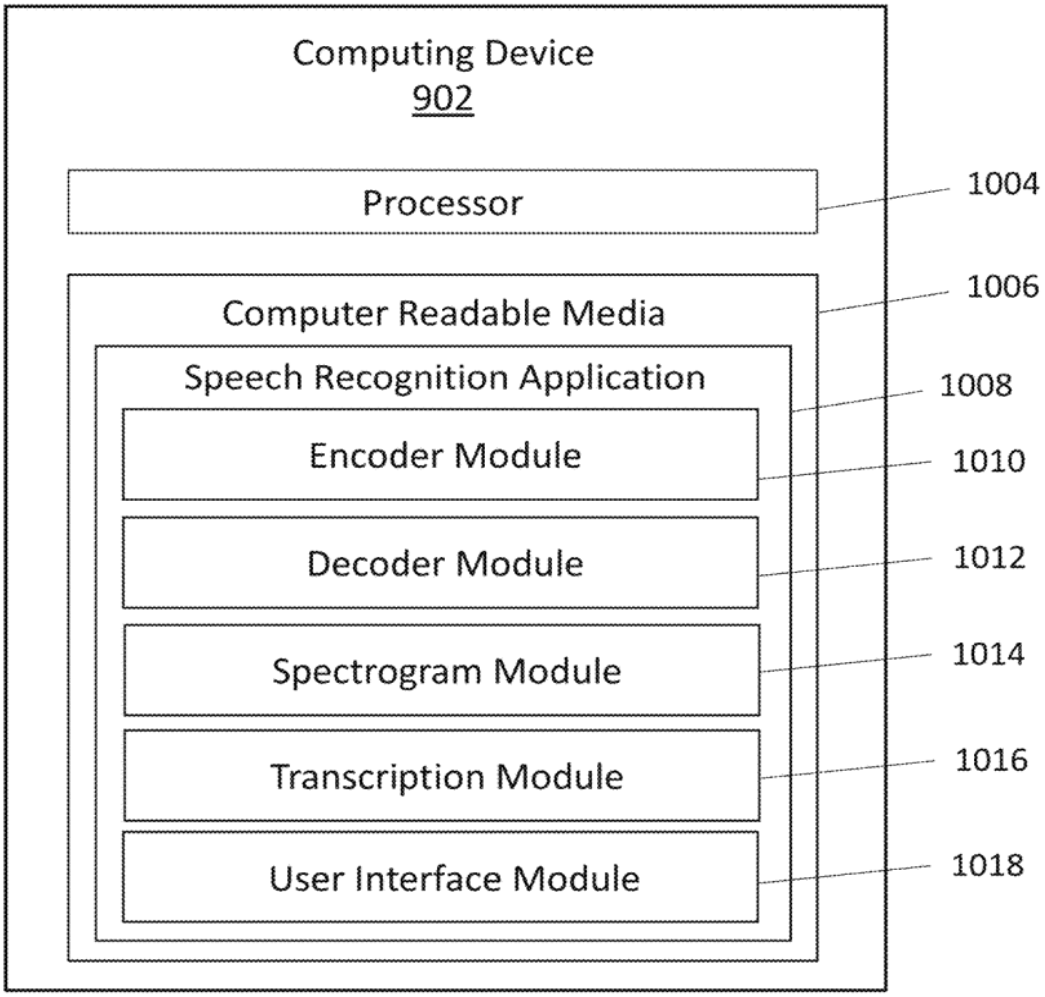


Figure 13. Overall Analytical Methodology Pipeline.
The study follows a systematic pipeline: (1) collection of dysarthric speech datasets (UASpeech and TORGO) with strict speaker-wise splits; (2) audio preprocessing to standardize inputs; (3) synthetic data augmentation using voice cloning and speech synthesis; (4) fine-tuning of Whisper Large V3 using parameter-efficient LoRA; and (5) evaluation on unseen test sets using Word Error Rate (WER) across dysarthria severity levels.

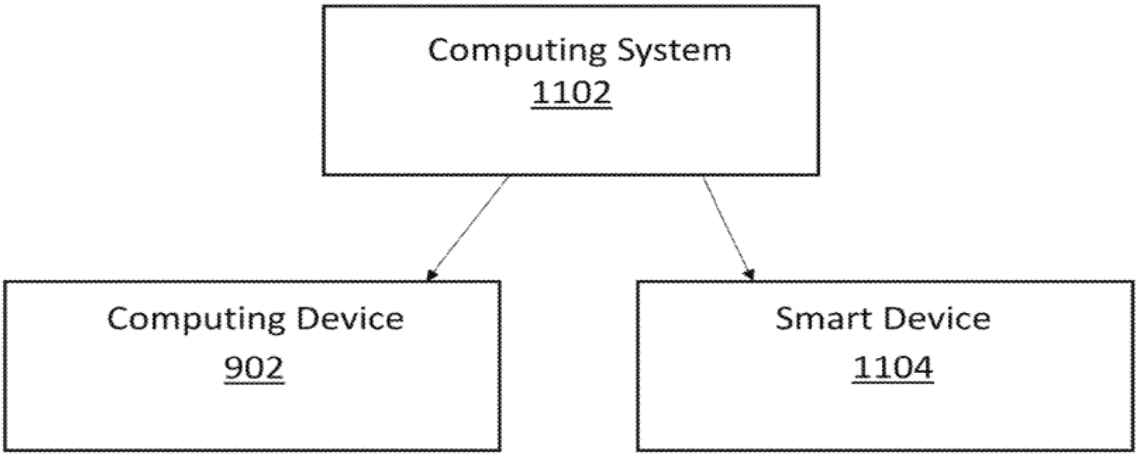
Source: Author-prepared by the authors. Cited in Section 3.1.

Appendix, Figure 14. Computing device (902) architecture with processor and speech recognition application comprising Encoder Module, Decoder Module, Spectrogram Module, Transcription Module, and User Interface Module.



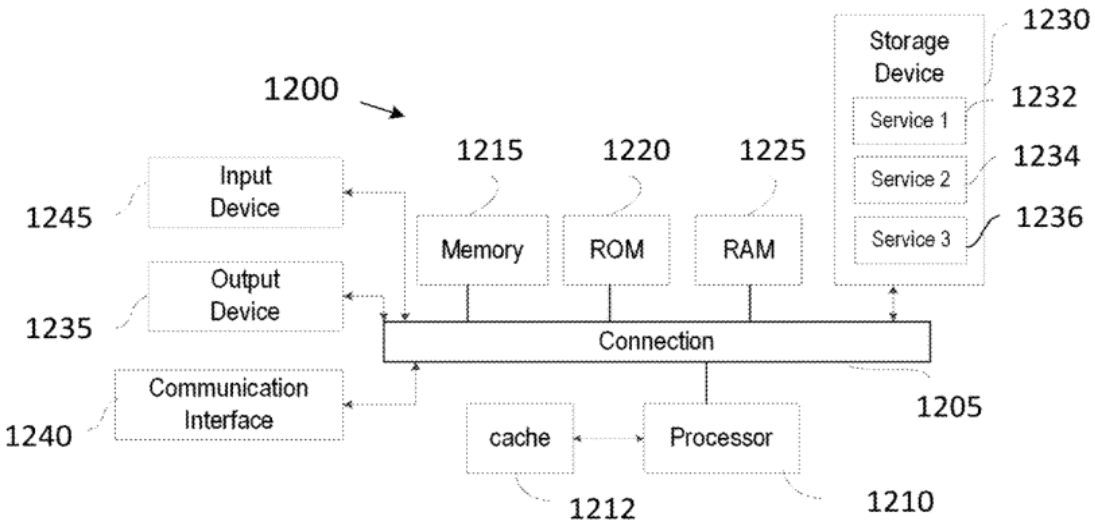
Source: Patent FIG. 10. U.S. Patent No. 12,488,786 B1. Cited in Section 5.6.

Appendix, Figure 15. System hierarchy showing the computing system (1102) integrating a Computing Device (902) and a Smart Device (1104) for voice-controlled IoT applications.



Source: Patent FIG. 11. U.S. Patent No. 12,488,786 B1. Cited in Section 5.6.

Appendix, Figure 16. Detailed computing system block diagram illustrating processor, memory hierarchy (Memory, ROM, RAM), storage services, I/O devices, and communication interface components.



Source: Patent FIG. 12. U.S. Patent No. 12,488,786 B1. Cited in Section 5.6.