CLARIS: Clear and Intelligible Speech from Whispered and Dysarthric Voices

Neil Shah[*]
neilkumar.shah@tcs.com
TCS Research & CVIT, IIIT Hyderabad
Hyderabad, Telangana, India

Yash Sonkar[*]
yash.sonkar@research.iiit.ac.in
CVIT, IIIT Hyderabad
Hyderabad, Telangana, India

Shirish Karande
shirish.karande@tcs.com
TCS Research
Pune, Maharashtra, India

Vineet Gandhi
vgandhi@iiit.ac.in
CVIT, IIIT Hyderabad
Hyderabad, Telangana, India

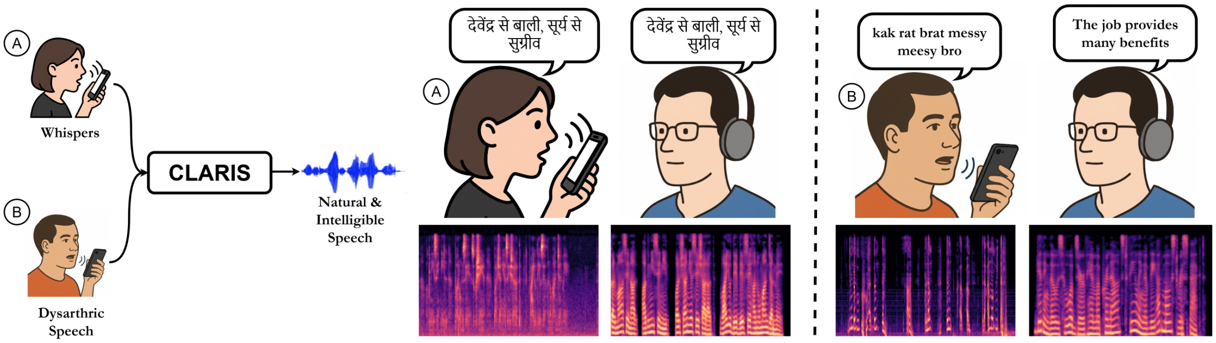
A
Whispers
B
Dysarthric Speech
CLARIS
Natural & Intelligible Speech
देवेंद्र से बाली, सूर्य से सुग्रीव
देवेंद्र से बाली, सूर्य से सुग्रीव
kak rat brat messy meesy bro
The job provides many benefits

Figure 1: *CLARIS* converts atypical forms of speech, such as (A) whispers and (B) dysarthria, into natural and intelligible voice. It is designed for everyday devices (e.g., mobile phones), so that atypical speech from the sender is transformed into clear, natural speech for the receiver. On the right, we show one example of each condition with corresponding spectrograms. In (A), the whispered input (left spectrogram) lacks low-energy formants due to the absence of vocal fold vibration, which are restored in the converted speech (right spectrogram). In (B), dysarthric speech (left spectrogram) shows fragmented and discontinuous acoustic structures, which are reconstructed into smooth and continuous trajectories in the converted speech (right spectrogram).

Abstract

Whispered and dysarthric speech hinder effective communication and undermine the reliability of voice-enabled systems. We present *CLARIS*, a compact speech-to-speech restoration system that turns such atypical input into clear, expressive speech. *CLARIS* requires no disorder-specific architectural tuning, generalizes across languages, and adapts quickly to new accents and speakers, enabling practical personalization. On whispered English, Hindi, and clinically challenging dysarthric speech, *CLARIS* delivers state-of-the-art intelligibility and naturalness, with listener studies confirming gains in quality, intelligibility, naturalness, and prosody. The system runs in real time, converting one second of input in about 30ms and enables inclusive, private, and personalized voice interaction. Audio samples are available at https://claris-w2s.github.io/CLARIS/

CCS Concepts

• **Human-centered computing** → **Accessibility technologies**; *Human computer interaction (HCI)*; • **Computing methodologies** → **Transfer learning**.

Keywords

speech interaction, whispered voice, dysarthric voice, voice conversion, silent speech, transfer learning, speech disorders

ACM Reference Format:
Neil Shah, Yash Sonkar, Shirish Karande, and Vineet Gandhi. 2026. CLARIS: Clear and Intelligible Speech from Whispered and Dysarthric Voices. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3772318.3791734

1 Introduction

Speech is the most natural and pervasive mode of human communication. It enables everything from casual conversations to professional collaboration, and is increasingly central to how people interact with digital systems. However, there are many contexts where vocal interaction becomes impractical. Speaking aloud in

[*] Both authors contributed equally to this research.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791734

public spaces, on a train, in an office, or during a conference call can compromise privacy, disturb others, or feel socially inappropriate. Moreover, for individuals with speech disorders such as dysarthria [66], Parkinson's disease [7], or stuttering [44], fluent and intelligible speech may be difficult or impossible to produce, often leading to misunderstandings, stigma, or exclusion from spoken interaction. User studies with people who stutter further show how mainstream speech technologies frequently fail to accommodate disfluencies, creating accessibility barriers in daily interactions [84]. These challenges highlight the broader opportunity to design voice interaction systems that are private, accessible, and inclusive across diverse speaking conditions.

Whispered speech offers one such alternative. It can be produced discreetly, including by individuals with impaired vocal folds [6], and recorded using standard microphones. Unlike voiced speech, driven by periodic vocal fold vibration, whispers emerge through a narrow glottal opening from turbulent airflow, producing noisy excitation. As a result, whispered speech carries lower energy and lacks a measurable fundamental frequency, making it difficult for humans and machines to interpret. These challenges motivate the development of *voice conversion systems*, which transform whispered input into natural, fluent speech, allowing speakers to whisper. At the same time, listeners perceive normal speech, supporting private and inclusive communication. A similar paradigm applies to pathological speech disorders such as dysarthria, where the source signal is slow, raspy, slurred, or distorted but can be converted into clear, intelligible output. In both cases, voice conversion bridges the gap between atypical speech production and natural-sounding perception, greatly enhancing communication opportunities for affected individuals.

Recent advances in whisper-to-speech conversion (e.g., WESPER [68], DistillW2N [79]) and dysarthric speech recognition [1, 29, 71] demonstrate encouraging progress using self-supervised and generative models. However, the performance of these systems remains limited. Whisper-to-speech systems trained in zero-shot or one-shot settings often break down on speakers with foreign accents or in new languages, exposing limits in generalization. Data augmentation pipelines still rely on hand-crafted simulations of whispers, which miss the variability of natural whispered speech. In dysarthria, systems typically succeed only in restricted-vocabulary settings and break down in severe cases, with error rates often exceeding 80% in open speech [66]. Studies on Parkinson's disease show how vocal impairments vary widely across individuals and progression stages, underscoring the need for speaker-specific solutions [7]. Compounding this, models trained on small speaker sets tend to overfit [75], reducing robustness. These findings point to a critical need for both *speaker-level* and *disorder-level* customization, dimensions that remain underexplored in existing research.

Meanwhile, on the modeling side, transformer architectures have become the backbone of speech generation and conversion, supported by self-supervised methods such as wav2vec [72] and HuBERT [30] that reduce dependence on labeled data. However, fundamental challenges remain. Non-autoregressive models, familiar in whisper-to-speech systems, require strictly aligned whisper–speech pairs, data that is infeasible to obtain in practice, as individuals with disordered speech cannot produce perfectly aligned normal counterparts. Autoregressive models relax alignment constraints but demand massive training corpora, which are scarce for low-resource or disordered speech. To compensate, researchers have turned to data augmentation, but current strategies fall short. Signal-processing pipelines (e.g., pitch shifting, filtering, or LPC-based pseudo-whispers [94]) are heavily engineered and produce speech that sounds mechanical and fails to generalize. Neural augmentation pipelines, such as Grad-TTS [46], GAN-based augmentation [35], and F5-TTS [9, 12], offer a more scalable path but have been evaluated mostly on normal speech, with little attention to multilingual or disordered settings. These limitations reveal a critical gap, as existing architectures and augmentation methods are not robust enough for cross-condition speech restoration, particularly in real-world deployments where systems must serve diverse speakers, accents, and disorders.

In this work, we present ***CLARIS*** (Clear and Accessible Restoration of Impaired Speech), illustrated in Figure 1, which converts atypical input such as whispers or dysarthric speech into fluent, intelligible speech and supports personalization across speakers and languages. *CLARIS* employs an autoregressive transformer to address alignment challenges, supports customization to unseen speakers and accents with minimal data, extends to clinically disordered speech, and integrates Text-to-Speech (TTS) augmentation with a gradient reversal mechanism for robust training on synthetic data. *CLARIS* advances the state-of-the-art in atypical-to-normal speech conversion through the following contributions:

- **Unified, alignment-free architecture:** We present *CLARIS*, a novel end-to-end speech-to-speech framework applicable to both whispered and dysarthric speech. *CLARIS* draws from recent efforts exploiting speech-unit–speech pipelines, while employing an autoregressive speech decoder that obviates the need for explicit alignment, augments disordered data through an integrated TTS module, and incorporates a gradient reversal discriminator to limit overfitting to synthetic artifacts. The framework operates using only disordered speech paired with text, without requiring any parallel clean recordings or alignment.
- **Cross-lingual, clinically relevant, and data-efficient personalization:** *CLARIS* is adaptable across languages, accents, and clinically disordered speech such as dysarthria. It supports practical personalization with 15–30 minutes of user-specific data, enabling accessible and inclusive voice restoration across diverse speaking conditions.
- **Strong empirical performance and deployability:** *CLARIS* delivers state-of-the-art results on whisper and dysarthric speech benchmarks, including 12% WER on wTIMIT and 31% WER on TORGO, with subjective evaluations confirming improvements in intelligibility, quality, and prosody. *CLARIS* is lightweight, with a 40.71M-parameter footprint at inference. It achieves runtimes of 32-ms on GPU and 170-ms on CPU per one-second utterance, making it suitable for real-world deployment.

The remainder of the paper is organized as follows. Section 2 reviews prior work on whisper-to-speech and dysarthric voice conversion. Section 3 introduces the *CLARIS* framework, covering augmentation, autoregressive modeling, the real–synthetic discriminator, and unit-to-speech rendering. Section 4 outlines the benchmark and generated datasets used in our evaluation. Section 5 reports systematic experiments across speakers, accents, languages, and disorder types. Section 6 presents ablation studies of key components. Finally, Section 7 discuss limitations, outline future directions, and conclude the paper.

2 Related Work

Our work builds on the long-standing field of Silent Speech Interfaces (SSI), which seek to generate intelligible speech when natural voicing is unavailable or infeasible. SSI systems have explored diverse modalities, including surface electromyography that translates muscle activity into speech [52, 93], ultrasound tongue imaging that maps tongue movements into audio [10, 42], lipreading from camera input [92], real-time MRI that records vocal tract movements and decodes them into speech [73], and non-audible murmurs captured by microphones placed behind the ears [59]. These studies demonstrate the feasibility of converting silent articulations into speech, but often at the cost of specialized hardware that limits portability [42, 59, 73], invasive electrodes [28], or restricted vocabulary coverage [16, 42]. More recent SSI designs emphasize non-invasive, accessible, and open-vocabulary decoding, such as StethoSpeech, which reconstructs non-audible murmurs captured with a stethoscope placed behind the ear [74], or headphone-integrated EMG systems that enable wireless and comfortable use [80].

In parallel, recent frameworks for speech restoration adopt speech to speech conversion pipelines [64], replacing specialized sensors with software-only solutions. These models use self-supervised encoders to derive linguistic units from atypical input and decode them into fluent speech. Work in this space includes unit-based pipelines for whisper-to-speech conversion [68, 79], encoder decoder architectures for pathological voices [5, 56], and lip-to-speech systems that translate visual articulations into audio [15, 65, 70]. Much of this progress is powered by self-supervised learning (SSL) advances, which leverage structure within raw audio to learn representations without manual labels. Early generative approaches reconstructed inputs or predicted future segments [14, 85], while contrastive methods such as wav2vec [3, 72] distinguished positive from negative samples. More recent predictive models like HuBERT [30], mHuBERT [45], WavLM [8], and Discrete BERT [2] combine unit discovery with auxiliary objectives to mitigate information loss during quantization. These models now form the backbone of many state-of-the-art speech-to-speech systems [8, 15, 64, 68]. Building on this line of research, our work uses only standard microphones to capture whispered input and converts it into fluent speech by predicting SSL-based target unit sequences. This software-only approach emphasizes accessibility and moves speech-to-speech conversion closer to inclusive everyday deployment.

2.1 Whisper-to-Speech Conversion

Research on whisper-to-speech conversion has moved from statistical feature mapping to neural encoder–decoder pipelines. WESPER [68] introduced a real-time, zero-shot, language-independent system that encodes whispered and normal speech into discrete self-supervised units, aligns them, and decodes natural-sounding speech. Our work draws inspiration from WESPER's use of HuBERT embeddings within an encoder–decoder framework. The key distinction is that, unlike WESPER, our approach neither pre-trains HuBERT nor requires the HuBERT encoder during inference. WESPER remains a strong baseline for whisper-to-speech conversion, though its ability to capture speaker-specific articulation patterns is limited, and its performance on severe speech disorders such as dysarthria remains unclear. DistillW2N [79] extends this line by distilling HuBERT-Soft units into a ConvNeXt encoder and SoundStream decoder, supporting one-shot conversion with reduced cost. Generative approaches such as AGAN-W2SC and MaskCycleGAN [20, 86] further improve perceptual quality, while earlier statistical methods mapped whispered MFCCs to normal MFCCs using Gaussian Mixture Models [89]. Although not initially designed for whisper-to-speech, general voice conversion systems have also been applied. FreeVC [48] disentangles content and speaker representations for zero-shot conversion, and QuickVC [26] builds on HuBERT-Soft features and a VITS backbone with inverse STFT decoding to achieve efficient any-to-many conversion. Across both neural and statistical approaches, a recurring limitation is their reliance on pseudo-whisper generation pipelines, which produce mechanical artifacts and bias models away from natural whispered speech [94].

Parallel efforts have explored hardware-based whisper sensing. Whisphone [18] leverages bone conduction microphone with noise cancellation for robust detection in noisy settings. ReHEarSSE [16] uses in-ear ultrasonic sensing of canal shape changes to classify silently spelled words. QuietSync [78] combines jaw-motion sensing with surface electromyography from head-worn electrodes, achieving high recognition accuracy with only a few calibration samples. Other systems adapt everyday devices, such as HPSpeech [95], which repurposes headphones as sonar arrays, or EchoWhisper [22], which uses smartphone speakers and microphones to track Doppler shifts from tongue and lip motion. Collectively, these works illustrate a rich design space of specialized wearables and consumer device adaptations. In contrast, our approach emphasizes software-only, non-intrusive conversion using a standard microphone to transform whispered or disordered speech into fluent output without requiring additional hardware.

2.2 Dysarthric-to-Speech Conversion

Most work on dysarthric speech has focused on recognition, mapping impaired speech into text and optionally resynthesizing it with TTS. Early datasets such as UASpeech [40], Nemours [57], and Whitaker [13] are limited to single-word utterances or nonsensical sentences. TORGO [69] has become a widely used benchmark, as it provides continuous speech from eight clinically diagnosed dysarthric speakers. Early approaches using Hidden Markov Models or Time-Delay Neural Networks yielded high error rates for moderate and severe speakers [29]. Later work introduced discriminative objectives such as LF-MMI, which improved performance but remained fragile under variability [29]. More recent studies on TORGO have explored adversarial augmentation [35], self-supervised encoders such as HuBERT and Wav2Vec2 [31, 47], and generative frameworks including GANs and flow matching [12, 34, 39], reporting measurable gains. However, most of these systems [29, 31, 35] operate under closed vocabularies tied to dataset

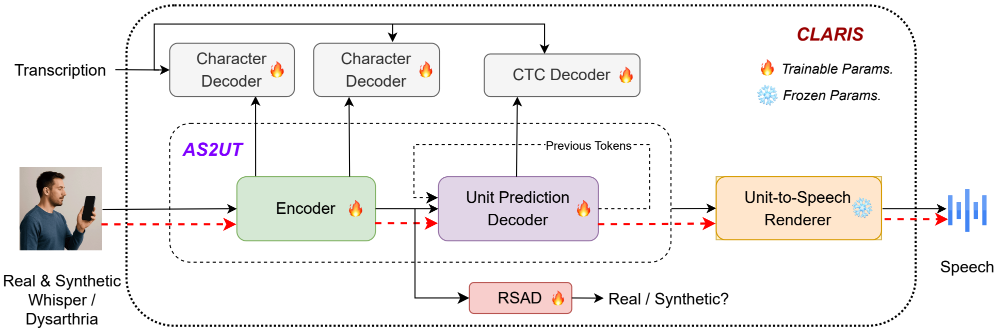

Figure 2: High-level overview of the *CLARIS* speech restoration framework. Whisper or dysarthria-like atypical speech inputs are collected from the user. During training, limited paired audio–text data is expanded through a TTS-based augmentation pipeline to generate synthetic inputs. Both real and synthetic audio are encoded by the AS2UT Encoder; embeddings are supervised by Character Decoders and aligned by the Real–Synthetic Alignment Discriminator RSAD. The unit prediction decoder then generates speech units, aided by a CTC Decoder, which are converted into natural speech by the Unit-to-Speech Renderer. The red dotted path illustrates inference: user audio is passed through the encoder, decoder, and renderer to restore intelligible speech.

transcripts and rely heavily on N-gram language models, raising concerns of overfitting given TORGO's small corpus size.

In contrast, open-vocabulary continuous speech recognition remains challenging, with error rates exceeding 80% for severe dysarthria [51]. Approaches that synthesize dysarthric data through TTS-based augmentation [88] or exploit raw phase and magnitude spectra have shown competitive results, but typically depend on extensive Automatic Speech Recognition (ASR) backbones such as Whisper Medium (769M parameters) or Large (1.55B parameters) [67]. Despite these advances, most work stops at predicting text representation, which requires an additional TTS engine for synthesis and adds substantial model complexity. Only limited efforts [12] have examined *end-to-end dysarthria-to-speech conversion*, where impaired input is mapped directly into fluent acoustic output. This is our direction: investigating whether a unified speech-to-speech architecture can support both whisper-to-speech and clinically disordered speech, such as dysarthria, without requiring disorder-specific tuning or specialized model modifications.

2.3 Personalization for Atypical Speech

A consistent finding across atypical speech technologies is that personalization is critical for intelligibility. In dysarthria-to-speech systems, studies such as Parrotron [5] and Project Euphonia [53] show that speaker-dependent fine-tuning can cut error rates dramatically. At the same time, more recent work explores adaptation through self-supervised features, adversarial learning, or personalized data augmentation [39, 58, 90, 91]. By contrast, whisper-to-speech systems have advanced rapidly but with little attention to speaker-specific adaptation, focusing instead on zero-shot or one-shot generalization across speakers [68, 79]. However, whispers vary widely in articulation, rhythm, and accent, making personalization equally important for intelligibility. Our study observes that when whisper accents shift across geographic regions, zero-shot systems often fail to produce intelligible speech. Building on these insights, our work introduces personalization for whispered and dysarthric speech, showing how modest user-specific data can substantially improve clarity and naturalness.

2.4 Positioning of this Work

Despite advances, many speech conversion architectures have structural limitations that hinder personalization and generalization. Non-autoregressive models such as WESPER offer efficient training, yet alignment remains challenging. WESPER uses a FastSpeech2-based decoder but omits the duration predictor and assumes a constant frame rate. This simplification can lead to misalignment under atypical rhythms or disfluencies and limit performance on severe disorders such as dysarthria. For context, a long dysarthric utterances from TORGO dataset, lasting up to 17-20 seconds, can be spoken by a typical speaker in approximately 4-5 seconds. Attention-based sequence-to-sequence models such as Tacotron2 [76] can learn alignments implicitly but often fail or produce artifacts when source and target prosody diverge or when encountering an unseen speaker [17, 23]. Many-to-one converters such as CycleGAN-VC [37, 38], parallel-data systems such as Parrotron [5], and adversarial approaches [24] have also been explored, but they often suffer from overfitting, data requirements, or training instability. Using an autoregressive decoder, synthetic augmentation, and a Real–Synthetic Alignment Discriminator (RSAD), our work stabilizes training for the autoregressive architecture and mitigates alignment concerns. It enables fine-tuning with only a few minutes of user-specific whispered or disordered speech, allowing the model to capture individual articulation. This mitigates the personalization gap in zero-shot systems without imposing fixed rhythm or duration, thereby addressing key limitations of prior architectures and improving practical applicability.

3 The *CLARIS* Speech Restoration Model

The design of *CLARIS* is guided by three practical challenges: (1) data scarcity, as atypical speech is complex to collect at scale; (2) data imbalance, since synthetic augmentation easily overwhelms limited real samples; and (3) alignment, because whispered and disordered speech cannot be paired frame-by-frame with natural speech. *CLARIS* addresses these through four components (Figure 2): a scalable TTS-based augmentation pipeline, the Atypical Speech-to-Unit Transformer (AS2UT), a Real–Synthetic Alignment Discriminator, and a unit-to-speech renderer.

The augmentation pipeline generates hundreds of hours of synthetic atypical audio (e.g., whispers, dysarthria) from as little as $15-30$ minutes of paired audio–text data. AS2UT employs an encoder–decoder architecture, extracting latent representations from atypical audio and autoregressively predicting discrete speech units, which the renderer converts into natural waveforms. RSAD mitigates real–synthetic imbalance by adversarially aligning encoder embeddings via a gradient reversal layer. Linguistic supervision is incorporated through auxiliary character decoders at intermediate encoder layers and a Connectionist Temporal Classification (CTC) decoder [25] conditioned on the AS2UT decoder.

3.1 Data Augmentation Strategy

Atypical-to-normal speech conversion faces a core challenge: the scarcity of atypical speech data. Collecting large corpora of whispers or dysarthric speech is impractical, yet transformer-based models [45] depend heavily on scale. Prior work has attempted to bridge this gap using signal-processing methods (e.g., pitch-shifting, LPC-based pseudo-whispering [94]), but such approaches produce unnatural artifacts and fail to generalize across disorder conditions. We address this gap by asking how atypical data can be synthesized in a way that is *scalable*, *faithful to real user speech*, and *generalizable across distinct speech disorders*.

Our preliminary experiments tested zero-shot prompting with the state-of-the-art NaturalSpeech3 [36] or F5-TTS [9], synthesizing whisper or dysarthria-like speech from short prompts. These models performed poorly because they had never been exposed to such styles during training. We therefore turned to VITS [41], a powerful TTS system that integrates variational inference, normalizing flows, and adversarial training. When trained from scratch on whispers paired with their corresponding text, VITS demonstrated that even with limited data, it could convincingly reproduce whisper-like and other atypical speech patterns, indicating robustness under low-data conditions.

To enable personalization, we modified the speaker-embedding layer of the multi-speaker trained VITS to represent only a single speaker. We then trained this adapted model using the original VITS training schedule, but with a reduced initial learning rate of 10^{-5}. For adaptation, we trained 5,000 steps for accent transfer and 10,000 steps for cross-language whispers or dysarthria. This process allowed us to generate hundreds of hours of synthetic atypical speech from only minutes of real data, scaling up training corpora while preserving speaker identity and disorder-specific characteristics. This augmentation strategy provides the volume required to train AS2UT effectively, while maintaining diversity across speakers and conditions. Combined with the RSAD module (Section 3.3), which explicitly aligns real and synthetic domains, our pipeline turns scarce and fragmented atypical recordings into a robust training signal for cross-disorder speech restoration.

3.2 AS2UT Encoder

Figure 3(A) shows the AS2UT encoder, which processes 80 dimensional mel-filterbank features extracted every 10 ms, normalized with cepstral mean and variance normalization, and augmented with SpecAugment [63] time and frequency masking. A convolutional subsampling layer reduces sequence length before passing the features into a 12-layer Transformer stack with multihead self-attention, feed-forward blocks, and residual connections.

We use mel-spectrograms as inputs, diverging from recent HuBERT based approaches. While HuBERT units excel in silent-speech tasks [73, 74], when directly applied to atypical speech (e.g., whispers or dysarthria), they discard the low-level acoustic detail that often carries the key to intelligibility. For whispered speech, WESPER effectively mitigates representation mismatch by jointly training HuBERT on whispered and normal speech. However, applying similarly content-focused HuBERT representations to atypical speech, such as dysarthria, remains challenging when the linguistic content is severely distorted and often unintelligible. Mel-spectrograms, in contrast, preserve fine-grained frequency cues and disorder-specific structures, enabling the encoder to learn representations that remain sensitive to subtle deviations in articulation and prosody.

To provide richer supervision during training, we attach auxiliary character decoders at the 8th and 10th encoder layers (Figure 3(B)). Each branch is a two-layer Transformer decoder that autoregressively predicts character sequences from a vocabulary automatically constructed from training transcripts. We use Label-Smoothed Cross Entropy (LSCE) loss for the same:

$$\mathcal{L}_{\text{LSCE}} = -\sum_{k=1}^{K} y_k^{\delta} \log(p_k), \tag{1}$$

where K is the vocabulary size, p_k is the predicted probability for class k, and $y_k^{\delta} = (1-\delta)y_k + \delta/K$ is the smoothed target distribution. When $\delta = 0$, this reduces to standard cross-entropy loss with one-hot targets, and for any other $\delta \in (0,1)$, it represents the loss with label smoothing. In all our experiments, we use $\delta = 0.2$.

Mel features, SpecAugment augmentation, and transcript-level auxiliary supervision enable the AS2UT encoder to model both linguistic and paralinguistic aspects of atypical speech. These encoder representations are then shared with both the RSAD (Section 3.3) and the unit prediction decoder (Section 3.4). This design makes the encoder effective for whispered speech and severe dysarthric cases, where intelligibility depends on acoustic subtleties often lost in higher-level unit representations.

3.3 Real–Synthetic Alignment Discriminator

Atypical speech corpora are inherently small, but synthetic data can be produced at scale. While this seems attractive, naively mixing synthetic with real data creates a new problem: the encoder quickly learns to model artifacts of the TTS voices, yielding strong results on synthetic benchmarks but weak generalization to actual user recordings.

To address this, we introduce the Real–Synthetic Alignment Discriminator (RSAD), which explicitly regularizes the encoder to encourage domain-invariant embeddings as shown in Figure 3(F). RSAD takes inspiration from prior efforts in domain adversarial

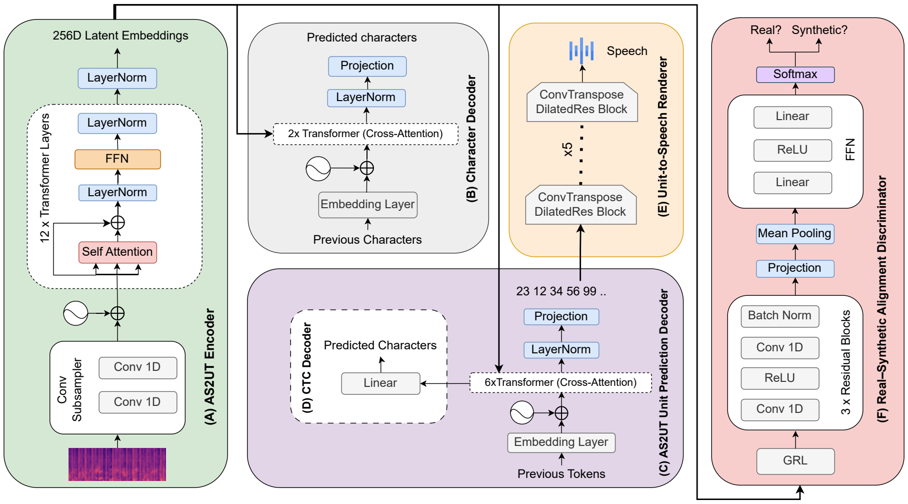

Figure 3: Detailed architecture of the *CLARIS* speech restoration pipeline. (A) The AS2UT Encoder processes mel-filterbank inputs into latent representations. (B) Character Decoders attached to the 8th and 10th encoder layers provide transcript-level supervision during training. (C) The AS2UT Unit-Prediction Decoder autoregressively generates discrete speech units. (D) The CTC Decoder attached to the 3rd layer of the AS2UT decoder provides character-level supervision during training. (E) The Unit-to-Speech Renderer generates speech from the rendered units. (F) The Real–Synthetic Alignment Discriminator aligns embeddings from real and synthetic inputs via adversarial training.

learning [19, 77, 83]. RSAD is implemented as a convolutional classifier with residual blocks, projection layers, global pooling, and a two-layer feedforward head. During training, its gradients are inverted through a gradient reversal layer [19], so instead of rewarding the encoder for distinguishing real from synthetic, the system aims to eliminate those differences. This adversarial alignment promotes consistency between features learned from synthetic data and those from scarce real inputs. No prior whisper-to-speech or silent-speech system has explicitly aligned synthetic and real domains in this way. RSAD employs a weighted cross-entropy loss:

$$\mathcal{L}_{CE} = -\sum_{n=1}^{N}\sum_{c=1}^{C} w_c\, y_{n,c}\, \log \frac{\exp(x_{n,c})}{\sum_{i=1}^{C}\exp(x_{n,i})}, \tag{2}$$

where N is the batch size, C is the number of classes (real vs. synthetic), $y_{n,c}$ is the one-hot label for sample n and class c, $x_{n,c}$ is the discriminator's logit for class c, and w_c is a class-specific weight to address the imbalance between real and synthetic samples. In practice, we set $w_c = 5$ for real audio and $w_c = 1$ for synthetic, more heavily penalizing misclassification of scarce real data. Overall, using RSAD facilitated better exploitation of synthetic data and yielded performance gains across all experimental settings.

3.4 AS2UT Unit Prediction Decoder

The AS2UT decoder predicts discrete units from encoded features. This is in contrast to mel-to-mel architectures such as AGAN-W2SC [21], which map mel inputs directly to mel outputs. Predicting quantized units reduces variance and simplifies training by prioritizing linguistic content over speaker identity.

The AS2UT decoder is autoregressive by design to address alignment concerns. Perfect alignment is unattainable since atypical and normative speech cannot be produced simultaneously. Prior works address this by separately recording natural speech and applying alignment techniques such as DTW to match atypical and natural speech [54, 87], or using MFA [55] to estimate phoneme durations from atypical speech–text alignments, which are then used to regenerate ground-truth speech. Such alignment based approaches are challenging for tasks such as dysarthric-to-normative speech mapping, where substantial temporal mismatch exists between the input and target speech sequences. Across the entire TORGO dataset, dysarthric utterances are on average three times longer than their normative counterparts, with extreme cases where a 17-second dysarthric utterance maps to only 4 seconds of normative speech. This large duration mismatch, compounded by substantial inter-speaker variability in duration patterns, makes the application of non-autoregressive models particularly difficult.

For supervision, we use dataset transcripts to synthesize fluent reference speech with the Google Cloud TTS engine[1], which, while not temporally aligned with the atypical inputs, provides high-quality examples of natural speech. From these reference speech samples, we extract mHuBERT embeddings and quantize them into 1,000 cluster centroids using k-means, following the textless speech-to-speech pipeline in [45]. These discrete units serve as the prediction targets for the decoder. As shown in Figure 3(C), the decoder is a six-layer Transformer with multihead self-attention, cross-attention to encoder outputs, and positional embeddings. A linear projection maps hidden states to the 1,000-unit vocabulary that interfaces with the unit-to-speech renderer. We attach a CTC

[1] https://cloud.google.com/text-to-speech/docs/reference/rest

branch (Figure 3(D)) to the third decoder layer to preserve linguistic information. This branch consists of convolutional layers, ReLU activations, and a linear projection over the character vocabulary derived from transcripts. It is trained with the standard CTC loss:

$$\mathcal{L}_{\text{CTC}} = -\log \sum_{\pi \in \mathcal{B}^{-1}(C)} \prod_{t=1}^{T} P(\pi_t \mid h_t), \tag{3}$$

where C is the target character sequence, π is a valid alignment in $\mathcal{B}^{-1}(C)$, h_t is the hidden state at time t, and $P(\pi_t \mid h_t)$ is the probability of symbol π_t given h_t. This branch is used only during training, ensuring unit predictions remain linguistically faithful without adding inference cost. For unit prediction, we use the label-smoothed cross-entropy loss (Eq. 1), comparing predicted unit distributions against ground-truth mHuBERT units.

3.5 Unit-to-Speech Renderer

The final step in our pipeline is converting predicted unit sequences into fluent speech. To achieve this, we adopt a unit-to-speech framework and train a unit-based HiFi-GAN vocoder with duration prediction. We employ the code-HiFiGAN variation [64] which is an adaption of HiFiGAN-v2 vocoder [43] for unit-to-speech prediction. Duration prediction provides stability because instead of regressing directly to waveforms, it explicitly models unit-to-frame alignment, improving naturalness and robustness across different utterance lengths. A key design choice was constraining the renderer to a single-speaker target voice. In preliminary experiments, multi-speaker targets caused inconsistencies in unit distributions, degrading intelligibility. We fix the target to one speaker and provide the decoder with a stable acoustic reference, leading to clearer outputs. Once trained, the renderer can synthesize fluent speech from any sequence of units predicted by AS2UT unit prediction decoder Figure 3(E).

3.6 Model Size, Training, and Compute

All components of the AS2UT model are trained jointly using the multi-objective losses described in Section 3. The encoder, decoder, RSAD module, auxiliary character decoders, and the CTC branch are optimized end-to-end with a weighted combination of the label-smoothed cross-entropy loss (×1), RSAD loss (×1), CTC loss (×4), and auxiliary character losses (×4). Training is conducted on a single NVIDIA A6000 GPU and typically converges within 1.5–2 days. The Adam optimizer is used with $(\beta_1, \beta_2) = (0.9, 0.98)$, gradient-norm clipping of 10, and a warm-up schedule that increases the learning rate from 10^{-7} to 1.5×10^{-4} over the first 15k steps. For a batch size of 64 samples, the peak GPU consumption is about 24 GB. The AS2UT decoder uses beam search with a beam size of 20.

During training, *CLARIS* comprises 48.70M parameters, spanning the AS2UT encoder (17.5M), decoder (9.7M), RSAD module (1.28M), two character decoders (3.17M each), the CTC decoder (0.4M), while the Unit-to-Speech renderer (13.48M) is trained separately for the target voice. The ground-truth units are pre-computed using a frozen HuBERT-base encoder.At inference, *CLARIS* uses only 40.71M parameters and consisting of the AS2UT encoder, decoder, and a Unit-to-Speech renderer. Its maximum memory footprint is approximately 500 MB (without any quantization). This makes the system considerably smaller than recent whisper-to-speech baselines such as FreeVC (354.79M), QuickVC (134.63M), and WESPER (142.91M), while DistillW2N (10.91M) is smaller but demonstrates substantially weaker performance. *CLARIS* also offers efficient runtime behavior. For inference, we convert our models to the framework-agnostic ONNX[2].

4 Dataset Configuration

To evaluate *CLARIS* fairly across disorder conditions, we consider three complementary settings: English whispers, Hindi whispers, and clinically disordered speech from people with dysarthria, which together test the model's ability to handle low volume inputs, cross-lingual contexts, and clinically impaired voices. To ensure consistency, we report intelligibility of all whispered and dysarthric samples using WER and Character Error Rate (CER). WER captures the proportion of words misrecognized by an ASR system, with lower scores indicating clearer output. CER complements this measure by offering a finer-grained view of errors, such as dropped syllables or partial word substitutions that still affect comprehension. Both metrics are computed using the OpenAI Whisper-small multilingual ASR model[3], unless otherwise noted.

4.1 English Whisper

4.1.1 ***US-accent Real Whisper.*** We began with the widely used wTIMIT (Whispered TIMIT) corpus [49], which contains whispered and normal speech from 48 speakers (28 American and 20 Singaporean), each reading 450 TIMIT sentences. The corpus provides about 29 hours of speech in both whispered and normal modes. When directly fed to ASR, the whispered speech yields a WER of 23.07% and a CER of 10.25%.

4.1.2 ***Synthetic Whisper.*** To expand coverage beyond wTIMIT, we generated synthetic English whispers (see Section 3.1) using approximately 200k text sentences sampled from three large-scale corpora: Emilia [27], LibriSpeech [61], and LJSpeech [32]. Emilia is a multilingual corpus with 49.5k hours of speech across diverse languages; LibriSpeech is a 1000-hour English audiobook dataset; and LJSpeech is a single-speaker dataset containing 24 hours of read speech. Together, these corpora provided diverse prompts for our augmentation pipeline, enabling synthesis of whisper-style speech with broader lexical coverage than wTIMIT alone.

4.1.3 ***Indian-accent Real Whisper.*** To assess cross-accent generalization, we collected new English whispers from 3 speakers in India (two male, one female), aged 20–32, with diverse native language backgrounds. Each participant read sentences drawn from LibriSpeech [61], with non-overlapping subsets assigned per speaker to avoid overlap with augmentation data. The recordings totaled 8.19 hours, with speaker01 contributing 5.66 hours, speaker02 2.30 hours, and speaker03 0.23 hours. These samples produced a baseline WER of 16.29% and a CER of 8.26%. Corresponding transcriptions were collected alongside the recordings, and normal speech was generated using the Google Cloud TTS engine (US-Standard-C voice). Together, wTIMIT, large-scale synthetic whispers, and new

[2] https://onnx.ai/ format, enabling direct deployment across diverse hardware architectures. On average, the inference for one second of input speech takes 0.032s on a single Nvidia RTX 4090 GPU and 0.17s on a consumer grade Intel i5-12500H CPU
[3] https://huggingface.co/openai/whisper-small

Indian-accent whisper data form a benchmark that spans diverse text distributions and varied accents, providing a robust basis for evaluating whisper-to-speech conversion in English.

4.2 Hindi Whisper

To test cross-lingual adaptation, we collected whispered speech in Hindi, a language linguistically distant from English and widely spoken by millions of users. Since no public Hindi whisper corpus exists, we recruited 10 speakers (seven male, three female), aged 18–32, representing diverse regional accents and dialects. Participants whispered Hindi sentences drawn from Project Vaani [81], Gram Vaani [4], and Kathbath [33]. For a fair evaluation of the generalization ability of the models evaluated, each participant was assigned a distinct set of texts, and the sentences used for augmentation were drawn from a disjoint set. In total, the dataset comprises about 8.9 hours of whispered audio, with individual contributions ranging from a few minutes to several hours: speaker01 4.34h, speaker02 0.75h, speaker03 0.43h, speaker04 0.80h, speaker05 0.58h, speaker06 0.18h, speaker07 0.54h, speaker08 0.24h, speaker09 0.52h, speaker10 0.60h. Baseline intelligibility was low, with a Hindi fine-tuned Whisper-small ASR[4] yielding a WER of 43.36% and a CER of 20.87%. These results highlight how challenging whisper recognition becomes in Hindi, motivating the need for cross-lingual whisper-to-speech conversion methods that can generalize beyond English benchmarks.

4.3 Dysarthric Speech

We further extend our evaluation to converting dysarthric speech into normal speech. Dysarthric voices present a different challenge because whispers lack vocal fold vibration. In contrast, dysarthric speech reflects articulatory impairment and often results in shorter, fragmented utterances (example spectrogram is presented in Figure 1(B)). For this analysis, we used the TORGO dataset [69], which contains recordings from eight speakers with dysarthria (five male, three female). The dataset includes single-word and multi-word utterances, capturing the disrupted articulation patterns and prolonged, effortful phrasing characteristic of dysarthric speech. Recognition accuracy on the raw dysarthric speech ranged from 5.68% to 250.00% WER, illustrating how quickly intelligibility breaks down and underscoring the need for restoration models that can cope with diverse levels of impairment. For augmentation, we followed the same strategy as in English whispers (Section 4.1). We sampled approximately 200k English text sentences and generated corresponding normal speech using the Google Cloud TTS engine, ensuring consistency in synthetic references and enabling direct comparison of whisper and dysarthric conditions under a unified pipeline.

4.4 Ethics and Recruitment

All data collection and evaluation followed institutional ethical guidelines. Participants provided informed consent and could withdraw at any time. English (Indian accent) and Hindi whisper speakers were recruited via local outreach, primarily among university students, representing diverse linguistic and regional backgrounds. Age, gender, and language background were recorded, and all data were anonymized. None of the participants reported hearing or speech pathologies. Recordings were conducted in daily lab environments using built-in laptop microphones (Apple MacBook Pro) without headphones, reflecting realistic deployment conditions without specialized hardware. The walking data were collected using a hand-held smartphone (iPhone 16E) from an English speaker (speaker01). For dysarthric speech, we used the publicly available TORGO dataset [69], which had been collected under prior ethical approval. Subjective evaluations were conducted with university students, who were compensated for their time.

5 Evaluation

5.1 Setup

We evaluate *CLARIS* using a combination of objective metrics that capture both intelligibility and linguistic fidelity. Beyond error rates, we assess sentence-level fidelity with BLEU [62] and ROUGE-L [50], two n-gram overlap metrics widely used in translation and summarization. BLEU is precision-oriented and rewards outputs that closely match the ground-truth text sequence. At the same time, ROUGE-L emphasizes recall through the longest common subsequence and captures how much of the reference content is preserved. These metrics provide a comprehensive view in which WER and CER quantify intelligibility from the perspective of automated recognition, and BLEU and ROUGE-L evaluate whether the restored speech content faithfully conveys the intended message. For baseline comparisons, we employ the publicly released checkpoints of QuickVC, FreeVC, and DistillW2N. While we attempted to retrain these models on wTIMIT, the results did not yield improvements. For WESPER, as the training code is not publicly available, we relied on the officially distributed model[5] for evaluation.

One straightforward approach to whisper-to-speech is to pass whispered input through an ASR system, obtain the transcribed text, and then use a TTS model to synthesize speech. This intuitive and straightforward pipeline often delivers strong results and, in most cases, surpasses traditional baselines. Its main drawback is that the generated speech does not capture the source speaker's voice identity. However, this limitation is shared by *CLARIS* as well as by all other baselines we are aware of. Generating speech in the original speaker's voice, whether in the context of whispers or disordered speech, is only feasible when pre-morbid recordings of their natural speech are available, as shown by personalized TTS approaches such as [58]. A key distinction of our approach is model size. *CLARIS* has 40.71M parameters, whereas a representative ASR–TTS pipeline combines Whisper-small (244M) and Coqui XTTS[6] (466M) for a total of about 710M parameters. This makes *CLARIS* practical for deployment in resource-constrained settings. We additionally assess the recorded whispered inputs directly through ASR (baseline termed Whispers in all Tables and Figures), to determine the extent to which the various systems yield improvements in output speech intelligibility.

To complement objective metrics, we conducted a user study with twenty participants (aged 21–35, balanced for gender). All were fluent in English and Hindi and reported no hearing or speech impairments. Each participant was presented with the ground-truth

[4] https://huggingface.co/bohraanuj23/whisper-small-hindi

[5] https://github.com/rkmt/wesper-demo

[6] https://huggingface.co/coqui/XTTS-v2

text as a reference, followed by a sequence of audio clips including the original whispered/dysarthric speech and outputs from all systems (baselines and *CLARIS*). The order of utterances was independently randomized for every participant to avoid ordering or learning effects. Participants rated each clip on a 5-point MOS scale for overall quality and intelligibility. Randomizing the presentation order of audio clips for each participant minimized preference bias, ensuring that ratings reflected the perceptual quality of each system rather than being influenced by the sequence in which clips were heard. In addition, listeners indicated whether a voice sounded whispered or normal, providing a direct perceptual check of voiced speech conversion. Finally, they rated prosody consistency in terms of rhythm, stress, and intonation, as unstable prosody can reduce perceived naturalness even when intelligibility is high .

In the following subsections, we present results across diverse evaluation settings. Section 5.2 examines performance on the wTIMIT corpus, considering seen speakers, unseen speakers, and cross-accent generalization to Indian-accent English. Section 5.3 evaluates personalization, showing how *CLARIS* adapts to Indian-accent speakers with as little as 30 minutes of data. Section 5.4 extends the analysis to Hindi, reporting performance across seen and unseen speakers as well as personalization on a held-out speaker. Finally, Section 5.5 benchmarks *CLARIS* on the TORGO dysarthric dataset and analyzes its ability to generalize and personalize to an unseen speaker.

5.2 wTIMIT: Seen, Unseen, and Cross-Accent Evaluation

We consider three evaluation scenarios: (a) training and testing on 44 wTIMIT speakers, (b) training on 44 speakers train-set and testing on four unseen/held-out wTIMIT speakers, and (c) testing the wTIMIT trained model on Indian-accent English speakers. These settings assess performance on seen speakers, unseen speakers, and generalization across varying accents.

5.2.1 ***Training and testing on 44 speakers.*** We first evaluate *CLARIS* on the 44 wTIMIT speakers set. For each speaker, 95% of the data is used for training and 5% for testing. We also ensured that the same sentences were not seen in train-test splits (the initial 95% of sentences were used for training). A full 48 speaker trained model for WESPER and DistillW2N were used for comparisons. For training *CLARIS*, we first augment the real wTIMIT whispers with synthetic ones, generated using the data augmentation strategy described in Section 3.1. Approximately 200k sentences were distributed across the 44 speakers to ensure balanced coverage, synthesized in the corresponding whisper-style voices. Each synthetic whisper was paired (though not temporally aligned) with normal speech generated using Google Cloud TTS, yielding about 367 hours of whispered audio with corresponding normal speech. Combined with the original wTIMIT recordings, this provided a robust training set for *CLARIS*.

Results in Table 1 highlight four main observations. (1) The ASR–TTS pipeline preserved intelligibility better than all baselines (24.38% WER vs. 23.40% for original whispers). (2) *CLARIS* delivered the strongest performance by a large margin (12.22% WER, with higher BLEU and ROUGE-L), significantly surpassing all prior methods. (3) DistillW2N, trained solely on synthetic whispers, exhibited severe domain mismatch and produced the highest error rates (55.08% WER). (4) FreeVC and QuickVC, though yielding lower WERs (51.18% and 44.85%) than WESPER (45.18%), did so by generating outputs perceptually close to the original whispers, rather than natural speech, thereby offering slight improvement in whisper-to-natural speech conversion. Overall, while all baselines degraded performance relative to the input whispers, *CLARIS* consistently advanced recognition quality.

To assess the impact of model scaling on the wTIMIT test set, we reduced the AS2UT encoder size from 17.5M to 9.3M parameters by decreasing the number of AS2UT encoder layers from 12 to 6, resulting in an overall inference model size of 32.51M parameters. This downsizing leads to a WER degradation from 12.22% to 14.28%. Despite the decline, the lightweight model remains competitive, showing minimal performance loss.

Using a larger ASR model improves performance in the ASR–TTS pipeline. Switching from OpenAI's Whisper-small to Whisper-medium reduces the WER from 24.38% to 15.66% (36% improvement). While this increase in accuracy is substantial, it comes with a corresponding rise in model size, from 244M parameters (small) to 769M (medium). Replacing Whisper-small to Whisper-medium during evaluation (transcribing) also leads to gains. ASR-TTS improves from 15.66% to 15.4%, while the WER of *CLARIS* reduces from 12.22% to 8.16%.

The perceptual evaluations (Figure 4) further corroborate the objective findings. *CLARIS* consistently received the highest ratings across all dimensions: quality (4.5), intelligibility (4.38), natural voicing (5.00), and prosody (4.93), demonstrating clear superiority. Paired t-tests confirmed significantly higher quality compared to ASR–TTS ($t(19) = 11.73$, $p < 0.001$, $d = 2.62$) and higher prosody preservation ($t(19) = 5.91$, $p < 0.001$, $d = 1.32$), while intelligibility was comparable ($t(19) = 1.19$, $p = 0.25$, $d = 0.27$). The ASR–TTS pipeline also performed strongly, particularly in intelligibility (4.25) and voicing (5.00), reinforcing its robustness relative to baseline conversion models. In contrast, FreeVC and QuickVC, despite moderately better WERs than WESPER, were perceived as whisper-like (voicing: 2.25 and 2.75). DistillW2N (2.00 quality, 3.88 intelligibility) and WESPER (2.25 quality, 4.13 intelligibility) also underperformed, showing objective and perceptual weaknesses. The perceptual evidence indicates that *CLARIS* provides more natural, intelligible, and prosodically consistent speech, with statistically reliable improvements over the alternatives.

5.2.2 ***Unseen wTIMIT speakers.*** Table 2 reports results on the four held-out wTIMIT speakers (s127, s129, s130, s131), with training conducted on the remaining 44 speakers. For WESPER and DistillW2N, the publicly released models trained on all 48 speakers were used, while the other baselines are not dependent on wTIMIT. As expected, baseline performance remains unchanged, except for ASR–TTS, which shows a modest improvement, underscoring its sensitivity to the quality of the whispered input. In contrast, *CLARIS* demonstrates strong generalization, sustaining an error rate of approximately 12%, comparable to the seen-speaker setting when accent characteristics are similar.

5.2.3 ***Indian-accent English whispers.*** To test its generalization towards varying accents, we directly evaluated the *CLARIS*

Table 1: Recognition performance on the held-out 5% split of 44 wTIMIT speakers. *CLARIS* achieves the lowest WER and CER while also improving linguistic fidelity (BLEU, ROUGE-L) compared to whispers, ASR–TTS, and other baseline systems.

Metric	Whispers	ASR-TTS	FreeVC	QuickVC	DistillW2N	WESPER	*CLARIS*
WER ↓	23.40	24.38	51.18	44.85	55.08	45.18	**12.22**
CER ↓	10.28	11.07	27.43	23.08	30.37	24.44	**4.80**
BLEU ↑	76.27	65.20	54.19	43.32	35.06	65.89	**79.62**
ROUGE-L ↑	79.12	81.37	70.12	68.26	53.19	69.28	**89.45**

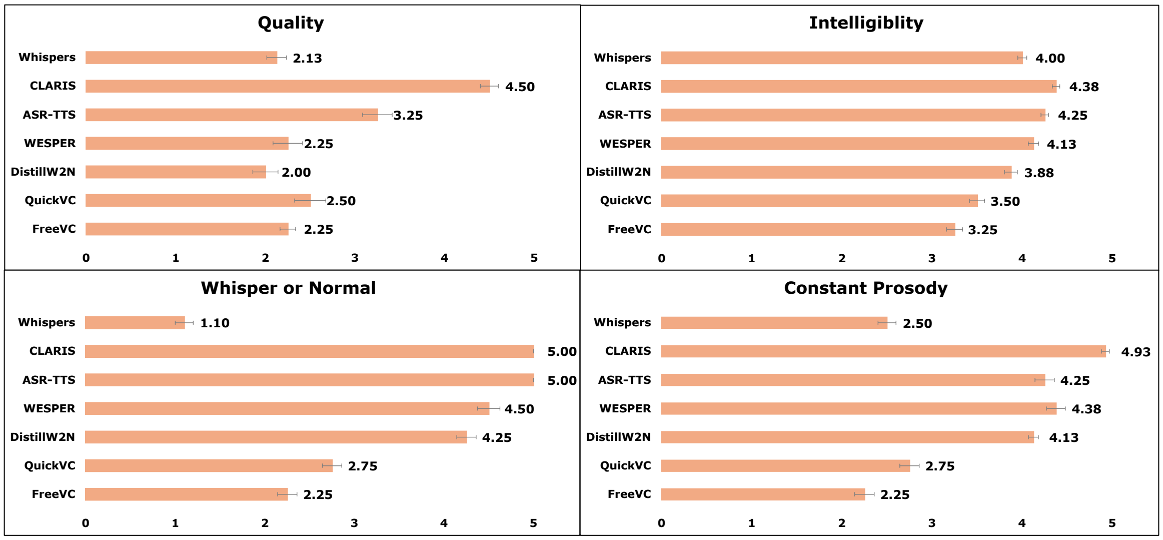

Figure 4: Subjective evaluation on the 5% split of 44 wTIMIT speakers. Listener ratings across all dimensions show that *CLARIS* consistently outperforms baseline systems, producing clearer and more natural speech.

Table 2: Recognition performance on four unseen wTIMIT speakers (s127, s129, s130, s131). *CLARIS* generalizes robustly to a new speaker's whispers without adaptation.

Metric	Whispers	ASR-TTS	FreeVC	QuickVC	DistillW2N	WESPER	*CLARIS*
WER ↓	24.05	15.78	52.10	43.67	54.21	46.12	**12.04**
CER ↓	10.95	6.13	28.02	22.41	29.58	25.06	**4.12**
BLEU ↑	75.21	77.01	55.03	44.28	36.41	64.37	**80.31**
ROUGE-L ↑	78.04	88.85	71.45	67.12	54.08	70.51	**90.19**

trained on 44 wTIMIT speakers with US and Singaporean accents (Section 5.2.1) on Indian-accented English whispers from the three speakers introduced in Section 4.1.

Recognition results for Indian-accented English whispers are summarized in Table 3. We observe that whispered inputs are already well intelligible (17.54% WER) and the ASR–TTS pipeline achieves the best whisper-to-speech conversion results (17.63% WER). The robust generalization of the default ASR model is expected, given its exposure to approximately 680K hours of multilingual and multi-accent training data. The performance degradation in large-scale pretrained voice conversion models such as QuickVC and FreeVC is comparatively moderate, likely reflecting their training on generic speech conversion tasks. In contrast, the drops observed in WESPER and *CLARIS* are substantially larger, primarily due to their exclusive reliance on the wTIMIT training data. Overall, Tables 1, 2, and 3 indicate that generic models, though relatively stable across accents, suffer from inherently poor base performance, making them unsuitable for practical use. The ASR–TTS baseline delivers strong generalized results but is constrained by deployment concerns like large model size. Focused models such as *CLARIS* excel on familiar accents yet exhibit substantial degradation under accent variation. A promising solution lies in finetuning of focused models on new distributions with minimal data, enabling effective personalization and customization.

Table 3: Recognition performance on Indian-accented English whispers, averaged across three speakers. ASR–TTS benefits from large-scale pretraining, whereas existing conversion systems, including *CLARIS*, fail to generalize.

Metric	Whispers	ASR-TTS	FreeVC	QuickVC	DistillW2N	WESPER	*CLARIS*
WER ↓	17.54	**17.63**	44.10	57.40	65.29	111.14	79.43
CER ↓	8.92	**8.54**	26.20	34.75	40.35	72.38	50.52
BLEU ↑	78.11	**72.70**	45.28	32.54	23.03	9.14	9.06
ROUGE-L ↑	81.37	**86.48**	61.82	53.57	44.85	25.88	33.63

5.3 Cross-Accent Personalization

We next evaluate personalization of *CLARIS* for Indian-accent speakers, a setting where the base wTIMIT model trained on 44 speakers failed to generalize. Personalization follows the same two-stage pipeline used in training. In the first stage, the augmentation model (pre-trained on wTIMIT) is fine-tuned with 5 – 30 minutes of whispered input from a single speaker, enabling it to generate customized synthetic whispers. In the second stage, *CLARIS* is (Section 5.2.1) is fine-tuned on the few minutes of recorded audio and the corresponding augmented data.

Table 4 reports results when the model is fine-tuned on one of the speakers. With only 5 minutes of fine-tuning data, WER drops from 76.78% in the zero-shot setting to 54.06%. With 15 minutes, WER falls sharply to 26.86%, and with 30 minutes, it reaches 12.63% WER and 6.06% CER, surpassing the ASR–TTS pipeline (17.01% WER, 8.99% CER). Notably, training *CLARIS* from scratch using the same 30 minutes of data achieves competitive results (15.14% WER), highlighting its efficiency in learning effectively from minimal data.

We repeated the 30-minute fine-tuning on two additional Indian-accent speakers to assess generality. *CLARIS* achieved 12.60% and 15.2% WER, respectively, confirming that improvements are consistent. Overall, *CLARIS* conclusively demonstrates scalable personalization to an unseen accent, adapting to new speakers with only 30 minutes of whispered data, surpassing large ASR–TTS pipelines and enabling accessible, speaker-specific whisper-to-voice conversion.

Subjective evaluations (Figure 5) reveal several key insights. First, fine-tuning leads to substantial gains in intelligibility (4.71 vs. 3.25; $t(19) = 10.39$, $p < 0.001$, $d = 2.32$), indicating that even limited accent-specific adaptation can markedly enhance listener-perceived clarity. Second, the divergence between WERs and the perceived intelligibility of WESPER outputs underscores the limitations of relying solely on ASR-based metrics for evaluation; despite relatively high WERs, listener ratings remained favorable. Finally, the ASR–TTS pipeline consistently demonstrated strong performance across subjective and objective measures while all other baselines reported only minor drops.

Noise augmented training and testing: To assess *CLARIS*'s robustness under challenging acoustic conditions, we incorporate noise both during training and evaluation. We augment 50% of the training data using diverse environmental noise samples (e.g., bus, kitchen, living room, driving) at an SNR of 0dB, following the procedure described in the DEMAND corpus [82]. For each training sample, we randomly select a noise sample from the DEMAND corpus and subsequently extract a random temporal segment for noise augmentation. As shown in Table 4, noise-augmented training further enhances performance, reducing the error rate to 8.09%. We construct two noisy test sets: one by synthetically augmenting noise using the DEMAND corpus, and another by collecting 10 minutes of whispered speech recorded on a mobile phone while walking outdoors (same speaker as in Table 4). The clean-trained model (30min FT) performs poorly on both sets (WER's reaching more than 40%), highlighting the challenges of in-the-wild evaluations. In contrast, the noise-augmented model (30min FT-noise) demonstrates robust performance, achieving WERs of 12.62% and 15.03% on the synthetic and walking test sets, respectively. Notably, the model generalizes to the walking test set despite being trained solely on synthetic noise.

5.4 Language adaptation: The Case of Hindi

To examine cross-lingual adaptation, we focus on Hindi, a language with relatively less mature ASR technology and scarce whisper-to-speech resources. The Hindi Whisper dataset was partitioned using speaker-wise 95–5% train–test splits, and training leveraged both collected data and simulated whispers from the training split. A Hindi-specific augmentation model was trained from scratch, generating 660 hours of synthetic whispered speech across ten voices, paired with normal speech synthesized via Google Cloud TTS (hi-IN-Standard-B voice). Using this augmented corpus, we trained *CLARIS* from scratch, thereby enabling Hindi whisper-to-speech conversion in a cross-language setting. We include WESPER, DistillW2N, and an ASR–TTS baseline for comparison. WESPER demonstrates promising cross-lingual generalization in its original evaluations (e.g., Japanese and German), while DistillW2N supports one-shot adaptation. However, neither WESPER nor DistillW2N was fine-tuned for Hindi, and both may not have been exposed to Hindi during training. Accordingly, the results should be regarded as indicative rather than directly comparable.

Table 5 reports averaged results on the 5% test split for all ten speakers. Directly feeding raw whispers to ASR yields a 43.36% WER. The ASR–TTS pipeline causes slight degradation across all three metrics. WESPER achieves 91.11% WER without Hindi-specific fine-tuning. In contrast, DistillW2N exhibits significant degradation relative to its performance on English datasets. *CLARIS* attains a WER of 29.21%, substantially improving the best performing ASR-TTS baseline. The results demonstrate the efficacy of our approach in training a Hindi whisper-to-speech system from scratch with limited data, highlighting its promise for broader cross-lingual adaptation in low-resource settings.

Personalization: We evaluated personalization via two leave-one-speaker-out trials with ten Hindi speakers. In the first, speaker01 was held out; in the second, speaker02. The model achieved WERs of 38.26% and 38.53% on held-in speakers, whereas performance on

Table 4: Cross-accent personalization on a single Indian-accented English whisper speaker. "No FT" = *CLARIS* trained on 44 wTIMIT speakers; "5/15/30 min FT" = *CLARIS* fine-tuned with the given amount of data; "30 min Scratch" = *CLARIS* trained from scratch with 30 minutes plus augmentation; "30 min FT-noise" = *CLARIS* fine-tuned with 30 minutes of noise-augmented whispers from speaker01.

Metric	Whispers	ASR-TTS	No FT	5 min FT	15 min FT	30 min FT	30 min Scratch	30 min FT-noise
WER ↓	16.98	17.01	76.78	54.06	26.86	12.63	15.14	**8.09**
CER ↓	8.90	8.99	45.12	38.21	15.32	6.06	7.42	**3.19**
BLEU ↑	77.94	76.70	12.13	38.89	57.30	77.92	73.97	**81.23**
ROUGE-L ↑	91.45	89.48	35.98	64.34	77.05	88.10	85.92	**89.19**

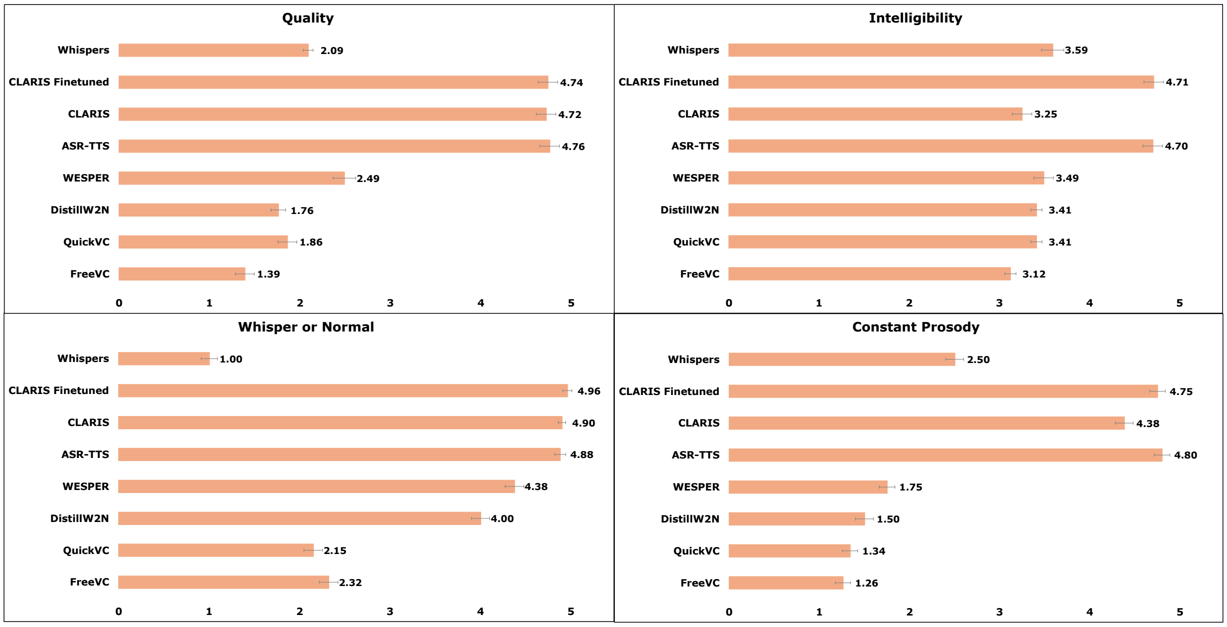

Figure 5: Subjective ratings for Indian-accent whispered speech, based on a mixed set of samples from all three recorded speakers. Ratings compare baselines with *CLARIS*, highlighting the gains from accent-specific fine-tuning over the pretrained model.

Table 5: Recognition performance of original whispers, baseline systems, and *CLARIS* on Hindi whispered speech; *CLARIS* was trained and evaluated on all ten speakers.

Metric	Whispers	ASR-TTS	WESPER	DistillW2N	*CLARIS*
WER ↓	43.95	51.50	91.11	105.81	**29.21**
CER ↓	19.85	26.76	72.36	74.90	**12.35**
BLEU ↑	30.16	25.90	4.13	2.79	**47.11**

the unseen speakers dropped to 59.35% and 60.71%, respectively. Under the same conditions, the ASR–TTS baseline produced WERs of 48.12% and 49.02%. However, subsequent fine-tuning of the nine-speaker *CLARIS* with only 30 minutes of whispered speech from the held-out speaker reduced WER to 37.01% for speaker01 and 37.98% for speaker02.

Subjective Evaluations: Figure 6 presents listener ratings for the Hindi evaluation, comparing *CLARIS* with baseline systems. Among baselines, WESPER achieved the highest quality and intelligibility (3.24/3.23), outperforming DistillW2N (1.49/2.44) and whispered input (2.80/2.76). ASR–TTS performed more strongly (4.49/3.77), exceeding WESPER and whispers. Paired comparisons confirmed that *CLARIS* scored significantly higher than ASR–TTS in intelligibility (4.46 *vs.* 3.77; $t(19) = 6.81$, $p < 0.001$, $d = 1.52$) and prosody ($t(19) = 5.15$, $p < 0.001$, $d = 1.15$). For quality, the difference was marginal (4.64 *vs.* 4.49; $t(19) = 2.09$, $p = 0.050$, $d = 0.47$), while for the whisper–normal judgment, *CLARIS* and ASR–TTS did not differ significantly (4.88 *vs.* 4.63; $t(19) = 1.49$, $p = 0.15$, $d = 0.33$). Overall, *CLARIS* delivers robust gains in intelligibility and prosody, maintains competitive quality, and offers an effective, inclusive whisper-to-speech solution for Hindi.

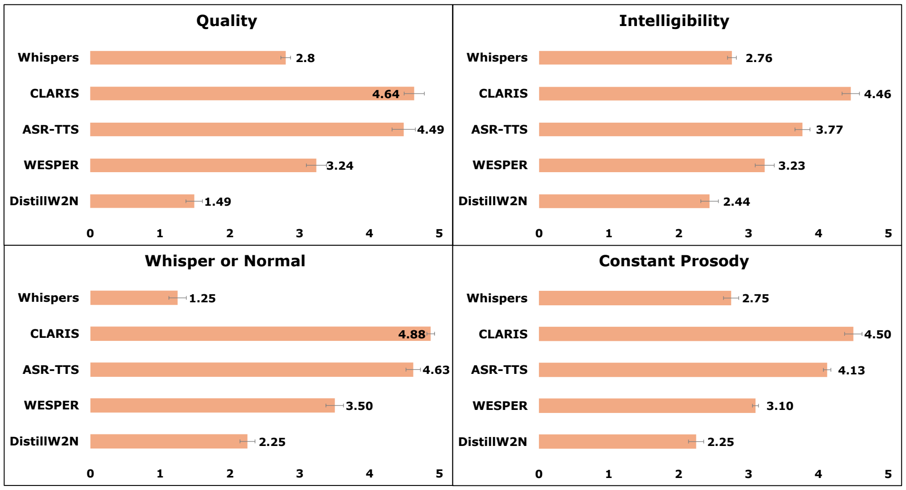

Figure 6: Subjective ratings for Hindi whispered speech converted by baseline systems and by *CLARIS*.

5.5 *CLARIS* for Dysarthric Speech

Dysarthria poses a greater challenge than whispering because it directly impacts articulation, rhythm, and loudness. In severe cases, speech may be weak or heavily slurred, making it difficult for human listeners and ASR systems to interpret. To examine robustness under these conditions, we evaluate *CLARIS* on the TORGO dataset, which contains clinical recordings from eight dysarthric speakers. Our analysis proceeds in three steps: first, benchmarking *CLARIS* against ASR–TTS and WESPER across all speakers; second, examining listener judgments on quality, intelligibility, and prosody; and finally, testing the applicability for personalization on dysarthric speech.

5.5.1 ***Training and Testing on Eight Speakers.*** We trained *CLARIS* on the 95% split of all eight speakers and tested on the held-out 5%, comparing against ASR–TTS and WESPER using their released checkpoints. Table 6 summarizes recognition outcomes at the individual-speaker level. Raw dysarthric speech is highly variable across speakers, with an average WER of 69.17% and a range from 5.68% for M03 to 250% for M04. ASR–TTS provides only limited improvements. It performs best when the input speech is already moderately intelligible, with the converted speech reaching 14.77% WER for M03 and 20.19% for F04. However, it fails on severe cases where, for example, M04 rises to 412.86% WER. WESPER performs consistently poorly with an average WER above 226%.

In contrast, *CLARIS* achieves a substantially lower average WER of 31.43%, improving over raw dysarthric recordings and all baselines. It restores intelligibility for the most challenging speakers, with M04 improving from 250% to 31.43%, M02 from 100% to 31.03%, and F01 from 77.78% to 11.11%. Even when the original speech was moderately clear, such as M03 and F04, *CLARIS* produced competitive results without degrading performance. The Torgo dataset contains many single-word or very short utterances, even phonetically similar output deviations (e.g., "sigh" vs. "psi", "ate" vs. "8" or "meat" vs. "meet") result in a WER of 100%, despite being interpretable by listeners. To better understand contextual usability, we evaluated only samples with five or more words, where contextual cues aid comprehension. On this subset, the WER dropped to 14.50%, indicating that most errors stem from short, context-free utterances and that the model is considerably more intelligible in natural conversational settings. Overall, we establish a strong benchmark on the TORGO dataset for dysarthric-to-speech conversion, evidencing that *CLARIS* can restore intelligibility and naturalness even under severe articulatory impairment. The study underscores substantial potential and offers real hope for enabling natural conversations for individuals affected by dysarthria.

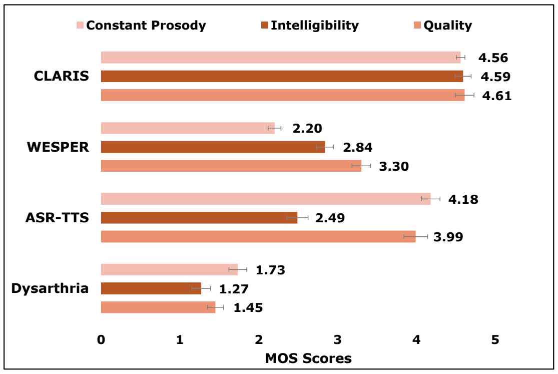

Figure 7: Subjective evaluation on the TORGO dysarthric dataset (5% held-out split from the 8-speaker training setup).

Subjective Evaluations: Figure 7 reports listener ratings on quality, intelligibility, and prosody for dysarthric speech conversion. Evaluation was conducted on the 5% held-out split, randomly sampled across all speakers. *CLARIS* achieved much higher quality (4.61) and intelligibility (4.59) than original dysarthric speech (1.45/1.27; Quality: $t(19) = 47.05$, $p < 0.001$, $d = 10.52$; Intelligibility: $t(19) = 43.51$, $p < 0.001$, $d = 9.73$), underscoring the

Table 6: Recognition performance on the 5% test split of the TORGO dysarthric dataset, reported overall and by speaker. We compare original dysarthric speech against ASR–TTS, WESPER, and *CLARIS* under multi-speaker training.

	Original				ASR–TTS				WESPER				*CLARIS*			
Speakers	**WER↓**	**CER↓**	**BLEU↑**	**R-L↑**	**WER↓**	**CER↓**	**BLEU↑**	**R-L↑**	**WER↓**	**CER↓**	**BLEU↑**	**R-L↑**	**WER↓**	**CER↓**	**BLEU↑**	**R-L↑**
Overall	69.17	40.29	28.67	47.81	91.58	52.00	24.58	42.77	226.47	132.37	9.22	33.91	**31.43**	**20.38**	**53.64**	**44.29**
F01	77.78	54.26	2.56	18.27	77.78	54.26	2.57	18.27	103.70	77.52	0.88	13.39	**11.11**	**3.88**	**62.84**	**86.00**
F03	37.40	24.46	39.43	57.43	47.15	31.57	39.28	**46.48**	63.41	42.31	22.58	34.64	**35.77**	**22.98**	**50.97**	39.76
F04	15.38	6.28	66.44	83.20	**20.19**	**11.17**	**65.70**	**76.77**	124.04	50.76	29.51	67.16	22.12	14.02	62.74	49.60
M01	79.59	53.23	20.38	21.86	81.63	56.35	20.16	19.15	328.57	166.37	0.58	13.18	**36.73**	**24.28**	**55.68**	**30.37**
M02	100.00	58.35	10.25	29.68	103.45	60.87	20.16	19.15	218.39	157.67	0.39	11.52	**31.03**	**17.85**	**51.50**	**48.98**
M03	5.68	4.19	62.95	87.50	**14.77**	**10.70**	**61.54**	**80.00**	17.05	10.70	59.23	71.47	25.00	14.65	52.04	58.51
M04	250.00	132.92	1.35	11.96	412.86	204.92	0.83	11.93	982.86	478.46	0.10	5.62	**31.43**	**24.31**	**52.77**	**34.13**
M05	47.06	30.90	35.42	31.90	54.41	38.21	33.98	26.22	82.35	192.69	21.47	27.52	**47.06**	**35.22**	**37.29**	**34.56**

difficulty of human perception. Compared to ASR–TTS, which obtained relatively high quality (3.99) and prosody (4.18) but low intelligibility (2.49), *CLARIS* scored significantly higher on intelligibility ($t(19) = 35.90$, $p < 0.001$, $d = 8.03$). WESPER reached slightly higher intelligibility (2.84) than ASR–TTS, but its quality (3.30) and prosody (2.20) lagged. *CLARIS* outperformed WESPER across all metrics (Quality: $t(19) = 16.57$, $p < 0.001$, $d = 3.71$; Intelligibility: $t(19) = 19.09$, $p < 0.001$, $d = 4.27$; Prosody: $t(19) = 25.69$, $p < 0.001$, $d = 5.74$). Overall, listener ratings for *CLARIS* were highest across dimensions, with statistically significant differences that reflected substantial effects. Ratings approached those of natural speech, indicating that the system restored lexical intelligibility and preserved expressive cues. We next examine whether *CLARIS* can generalize to an unseen speaker and adapt through fine-tuning with dysarthric speech from a held-out speaker.

Personalization: To examine the role of personalization, we trained *CLARIS* on seven dysarthric speakers while holding out F03. On the 5% held-out test split from the training set of seven speakers, we observed an average WER of 38.19%, compared to 31.43% when all eight speakers were used for training. In the zero-shot setting on F03, performance degraded, with a WER of 50.41% compared to 35.77% when F03 was included in the eight-speaker model. After fine-tuning the seven-speaker model on F03's data, the WER dropped to 38.21%, alongside consistent gains in CER and ROUGE-L. This case study demonstrates that even brief personalization can meaningfully improve intelligibility for dysarthria, underscoring the promise of speaker-specific adaptation in practical deployment.

6 Ablation Studies

To understand the contribution of key components in *CLARIS*, we performed ablation experiments where modules were selectively removed and their effect on recognition measured. Unless otherwise noted, results are based on the base model trained on 44 wTIMIT English whisper speakers (Section 5.2.1).

6.1 Effect of RSAD

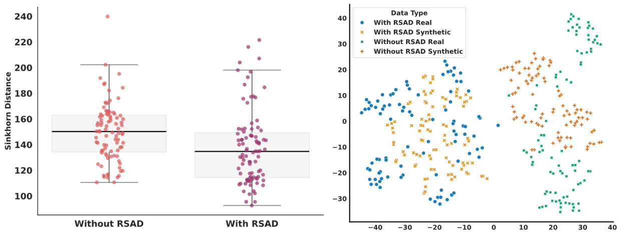

Figure 8: Effect of RSAD on AS2UT encoder embeddings for real *vs.* synthetic speech.

RSAD enables large-scale use of synthetic data without causing the autoregressive model to overfit to synthetic distributions. Table 7 shows that RSAD consistently improves recognition accuracy. On the wTIMIT test split, WER decreased from 13.28% to 12.22%, and on four unseen wTIMIT speakers from 14.38% to 12.04%. The benefits were most pronounced in low-resource personalization. Fine-tuning with only 30 minutes of Indian accented English reduced WER from 42.14% to 12.63%, and adapting to an unseen Hindi speaker improved performance from 58.32% to 39.29%. These results indicate that RSAD stabilizes training and strengthens generalization across accents and languages.

Figure 8 illustrates how RSAD reshapes the encoder's representational space. We observe quantitative and qualitative differences by comparing two models trained with and without RSAD. The boxplot on the left shows that the Sinkhorn distance [11] between real and synthetic embeddings is about eight percent lower and less variable with RSAD, which indicates more substantial alignment between the two distributions. The t-SNE projection on the right confirms this effect. Without RSAD, most real embeddings appear as outliers, which hinders reliable conversion. With RSAD, the clusters overlap, suggesting a shared embedding manifold. We hypothesize that RSAD achieves this through two mechanisms. First, it acts as a distributional regularizer that reduces representational drift between synthetic and real inputs, preventing the encoder from learning shortcuts tied to synthetic artifacts. Second, it encourages the encoder to allocate more capacity to phonetic and articulatory features that transfer across conditions rather than superficial cues such as spectral smoothness or quantization noise. This explains why RSAD provides significant gains in cross-accent and cross-lingual settings.

6.2 Effect of CTC Decoders

Atypical speech weakens clear linguistic cues, so auxiliary character and CTC decoders help the model recover missing information

Table 7: Recognition performance across evaluation scenarios, with and without RSAD. Each scenario reports metrics for RSAD = No / Yes. 'Accent mismatch' denotes Indian-accent English (30 min fine-tuning); 'Hindi' reports adaptation to an unseen sixth speaker.

	Seen wTIMIT		Unseen wTIMIT		Accent mismatch		Hindi	
Metrics	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
WER↓	13.28	**12.22**	14.38	**12.04**	42.14	**12.63**	58.32	**39.29**
CER↓	5.84	**4.80**	6.94	**4.12**	13.26	**6.06**	24.94	**15.28**
BLEU↑	78.15	**79.62**	75.26	**80.31**	49.42	**77.92**	48.45	**37.82**
ROUGE-L↑	88.86	**89.45**	83.27	**90.19**	43.22	**88.10**	38.43	**42.10**

Table 8: Effect of Character and CTC decoders on model performance. Using decoders on both encoder and decoder yields the best accuracy and fastest convergence.

Character+CTC Decoders	WER ↓	CER ↓
All (encoder + decoder)	**12.22%**	**4.83%**
Only on decoder	14.85%	7.53%
Only on encoder	13.65%	7.16%

by predicting text directly from hidden embeddings. These auxiliary tasks reinforce the model's ability to represent linguistic units. Table 8 shows that removing these decoders consistently degraded performance. With both encoder and decoder supervision, the model achieved 12.22% WER and 4.83% CER. Restricting decoders to only the AS2UT encoder or only the AS2UT decoder produced higher error rates of 13.65% and 14.85% WER, respectively. This demonstrates that joint supervision across the encoder and decoder provides stronger guidance and stabilizes optimization.

7 Discussion and Conclusion

Cross-condition Gains and Generalization. ASR-TTS outperformed evaluated methods for whispered speech when evaluated on unseen speakers. For severe dysarthria speakers such as M01, M02, M04 both ASR-TTS and *CLARIS* yielded limited improvements. However, when *CLARIS* is fine-tuned with a limited amount of speaker-specific data (F03), it consistently improved intelligibility, linguistic fidelity, and perceptual naturalness. The model generalized from whispered English to Hindi, a linguistically distant language, and to clinically challenging dysarthric voices. These findings show that autoregressive alignment-free modeling, paired with large-scale augmentation and adversarial regularization, can serve as a disorder-agnostic foundation for atypical speech restoration.

Personalization as a Core Design Requirement. A central challenge in speech technologies is that individuals, particularly those with speech disorders, exhibit distinct vocal behaviors that a one-size-fits-all model cannot capture. Our results show that even brief adaptation with user data can reliably transform unusable zero-shot outputs into intelligible, natural-sounding speech without degrading quality or prosody. For example, *CLARIS* trained on English whispers from one accent group failed to generalize across accents, yet thirty minutes of user-specific adaptation restored intelligibility to usable levels. In Hindi, a model trained on nine speakers performed poorly on an unseen tenth, but adaptation substantially reduced errors. The effect was most pronounced for dysarthric speech, where patient-specific variability makes personalization indispensable. These findings suggest that personalization is a performance boost and a design pathway for inclusive, adaptive, and user-centered speech technologies.

Limitations, Design Implications, and Future Directions. While results are promising, several limitations remain. Subjective evaluations were conducted primarily with non-impaired listeners; involving participants with similar conditions in longitudinal settings will be important to assess communicative success, usability, and comfort. The autoregressive design of *CLARIS* alleviates alignment challenges, but it inherently limits inference speed compared to non-autoregressive alternatives. Although it gives competitive runtime performance (32ms on GPU and 170ms on CPU per second of utterance), their remains a scope for further optimizations using libraries like TensorRT [60], and is left for future work.

Our study also focused on English dysarthric speech; extending to other disorders, such as Parkinson's, and building multilingual datasets will help evaluate broader applicability. Finally, personalization is used mainly to fine-tune the model for enhanced intelligibility and quality in a single normalized target voice, and is not intended to model expressive voice styles of the target speaker.

We will release the recorded datasets, code, and pre-trained checkpoints used in our evaluations. Beyond technical improvements, the work raises design questions: how should systems signal successful conversion, provide confidence feedback, or foster user trust? Personalization proved central to equity, as calibration allowed otherwise excluded users to achieve intelligible output. Embedding speech restoration into applications such as captioning, discreet mobile interfaces, and accessible assistants will allow studies on how restored voices shape identity, participation, and social interaction.

Conclusion. We introduced *CLARIS*, a lightweight autoregressive framework for atypical-to-normal speech conversion that restores intelligibility, naturalness, and prosody across whispers and dysarthric voices. With minimal adaptation data, it generalizes to new speakers and accents, and fine-tuning further enables personalized restoration for clinical speech. More broadly, our findings position speech conversion not as a niche task but as a core capability for accessible voice interaction, where personalization is essential for equity. We see *CLARIS* as a step toward everyday systems that restore voice, supporting more inclusive, expressive, and trustworthy interaction.

References

[1] Ahmed Aboeitta, Ahmed Sharshar, Youssef Nafea, and Shady Shehata. 2025. Bridging ASR and LLMs for Dysarthric Speech Recognition: Benchmarking Self-Supervised and Generative Approaches . In *Interspeech 2025*. 2123–2127. doi:10.21437/Interspeech.2025-1994

[2] Alexei Baevski and Abdelrahman Mohamed. 2020. Effectiveness of Self-Supervised Pre-Training for ASR. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7694–7698. doi:10.1109/ICASSP40776.2020.9054224

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates Inc. doi:10.48550/arXiv.2006.11477

[4] Anish Bhanushali, Grant Bridgman, Deekshitha G, Prasanta Ghosh, Pratik Kumar, Saurabh Kumar, Adithya Raj Kolladath, Nithya Ravi, Aaditeshwar Seth, Ashish Seth, Abhayjeet Singh, Vrunda Sukhadia, Umesh S, Sathvik Udupa, and Lodagala V. S. V. Durga Prasad. 2022. Gram Vaani ASR Challenge on spontaneous telephone speech recordings in regional variations of Hindi. In *Interspeech 2022*. 3548–3552. doi:10.21437/Interspeech.2022-11371

[5] Fadi Biadsy, Ron J. Weiss, Pedro J. Moreno, Dimitri Kanvesky, and Ye Jia. 2019. Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation. In *Interspeech 2019*. 4115–4119. doi:10.21437/Interspeech.2019-1789

[6] Bin Cao, Mimi Kim, Ted Mau, and Jun Wang. 2016. Recognizing Whispered Speech Produced by an Individual with Surgically Reconstructed Larynx Using Articulatory Movement Data. In *Proceedings of the 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*. 80–86. doi:10.21437/SLPAT.2016-14

[7] Fangyuan Cao, Adam P. Vogel, Puya Gharahkhani, and Miguel E. Renteria. 2025. Speech and Language Biomarkers for Parkinson's Disease Prediction, Early Diagnosis and Progression. *npj Parkinson's Disease* 11, 1 (2025), 57. doi:10.1038/s41531-025-00913-4

[8] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518. doi:10.1109/JSTSP.2022.3188113

[9] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025. F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 6255–6271. doi:10.18653/v1/2025.acl-long.313

[10] Tamás Gábor Csapó, Tamás Grósz, Gábor Gosztolya, László Tóth, and Alexandra Markó. 2017. DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface. In *Interspeech 2017*. 3672–3676. doi:10.21437/Interspeech.2017-939

[11] Marco Cuturi. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems 26 (NeurIPS 2013)*. Curran Associates Inc., 2292–2300. https://papers.nips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html

[12] Shoutrik Das, Nishant Singh, Arjun Gangwar, and S Umesh. 2025. Improved Intelligibility of Dysarthric Speech using Conditional Flow Matching. In *Interspeech 2025*. 2118–2122. doi:10.21437/Interspeech.2025-2617

[13] J. R. Jr. Deller, M. S. Liu, L. J. Ferrier, and P. Robichaud. 1993. The Whitaker database of dysarthric (cerebral palsy) speech. *The Journal of the Acoustical Society of America* 93, 6 (1993), 3516–3518. doi:10.1121/1.405684

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186. doi:10.18653/V1/N19-1423

[15] Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Haithem Boussaid, Ebtessam Almazrouei, and Merouane Debbah. 2023. Lip2Vec: Efficient and Robust Visual Speech Recognition via Latent-to-Latent Visual to Audio Representation Mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13790–13801. https://openaccess.thecvf.com/content/ICCV2023/papers/Djilali_Lip2Vec_Efficient_and_Robust_Visual_Speech_Recognition_via_Latent-to-Latent_Visual_ICCV_2023_paper.pdf

[16] Xuefu Dong, Yifei Chen, Yuuki Nishiyama, Kaoru Sezaki, Yuntao Wang, Ken Christofferson, and Alex Mariakakis. 2024. ReHEarSSE: Recognizing Hidden-in-the-Ear Silently Spelled Expressions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–16. doi:10.1145/3613904.3642095

[17] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, R.J. Skerry-Ryan, and Yonghui Wu. 2021. Parallel Tacotron 2: A Non-Autoregressive Neural TTS Model with Differentiable Duration Modeling. In *Interspeech 2021*. 141–145. doi:10.21437/Interspeech.2021-1461

[18] Masaaki Fukumoto. 2025. Whisphone: Whispering Input Earbuds. *arXiv preprint* (2025). doi:10.48550/arXiv.2501.01636

[19] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research* 17, 59 (2016), 1–35.

[20] Teng Gao, Qing Pan, Jian Zhou, Huabin Wang, Liang Tao, and Hon Keung Kwan. 2023. A novel attention-guided generative adversarial network for whisper-to-normal speech conversion. *Cognitive Computation* 15, 2 (2023), 778–792. doi:10.1007/s12559-023-10108-9

[21] Teng Gao, Jian Zhou, Huabin Wang, Liang Tao, and Hon Keung Kwan. 2021. Attention-guided generative adversarial network for whisper to normal speech conversion. *arXiv preprint arXiv:2111.01342* (2021).

[22] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-based Silent Speech Interface for Smartphone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 80 (Sept. 2020), 27 pages. doi:10.1145/3411830

[23] Efthymios Georgiou, Kosmas Kritsis, Georgios Paraskevopoulos, Athanasios Katsamanis, Vassilis Katsouros, and Alexandros Potamianos. 2023. Regotron: Regularizing the Tacotron2 Architecture Via Monotonic Alignment Loss. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. 977–983. doi:10.1109/SLT54892.2023.10023268

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Vol. 27. 2672–2680. https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[25] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh, Pennsylvania, USA) *(ICML '06)*. Association for Computing Machinery, New York, NY, USA, 369–376. doi:10.1145/1143844.1143891

[26] Houjian Guo, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. 2023. QuickVC: A Lightweight VITS-Based Any-to-Many Voice Conversion Model Using iSTFT for Faster Conversion. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 1–7. doi:10.1109/ASRU57964.2023.10389621

[27] Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2024. Emilia: An Extensive, Multilingual, and Diverse Speech Dataset for Large-Scale Speech Generation. In *Proceedings of the 2024 IEEE Spoken Language Technology Workshop (SLT)*. doi:10.1109/SLT61566.2024.10832365

[28] Dominic Heger, Christian Herff, Adriana de Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. 2015. Continuous speech recognition from ECoG. In *Interspeech 2015*. 1131–1135. doi:10.21437/Interspeech.2015-296

[29] Enno Hermann and Mathew Magimai-Doss. 2020. Dysarthric Speech Recognition with Lattice-Free MMI. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6109–6113. doi:10.1109/ICASSP40776.2020.9053549

[30] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. In *Proceedings of Interspeech 2021*. ISCA, 2513–2517. doi:10.21437/Interspeech.2021-1473

[31] Shujie Hu, Xurong Xie, Mengzhe Geng, Zengrui Jin, Jiajun Deng, Guinan Li, Yi Wang, Mingyu Cui, Tianzi Wang, Helen Meng, et al. 2024. Self-supervised asr models and features for dysarthric and elderly speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), 3561–3575.

[32] Keith Ito and Linda Johnson. 2017. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/.

[33] Tahir Javed, Kaushal Santosh Bhogale, Abhigyan Raman, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. IndicSUPERB: A Speech Processing Universal Performance Benchmark for Indian languages. doi:10.48550/ARXIV.2208.11761

[34] Yejin Jeon, Solee Im, Youngjae Kim, and Gary Geunbae Lee. 2025. Facilitating Personalized TTS for Dysarthric Speakers Using Knowledge Anchoring and Curriculum Learning . In *Interspeech 2025*. 2108–2112. doi:10.21437/Interspeech.2025-596

[35] Zengrui Jin, Mengzhe Geng, Jiajun Deng, Tianzi Wang, Shujie Hu, Guinan Li, and Xunying Liu. 2023. Personalized adversarial data augmentation for dysarthric and elderly speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2023), 413–429.

[36] Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. 2024. NaturalSpeech 3: zero-shot speech synthesis with factorized codec and

diffusion models. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 909, 19 pages. doi:10.5555/3692070.3692979

[37] Takuhiro Kaneko and Hirokazu Kameoka. 2018. CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*. 2100–2104. doi:10.23919/EUSIPCO.2018.8553236

[38] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2019. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6820–6824.

[39] Karl El Hajal and Enno Hermann and Sevada Hovsepyan and Mathew Magimai Doss. 2025. Unsupervised Rhythm and Voice Conversion to Improve ASR on Dysarthric Speech. In *Interspeech 2025*. 2760–2764. doi:10.21437/Interspeech.2025-2069

[40] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S. Huang, Kenneth Watkin, and Simone Frame. 2008. Dysarthric speech database for universal access research. In *Interspeech 2008*. 1741–1744. doi:10.21437/Interspeech.2008-480

[41] Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*. PMLR, 5530–5540.

[42] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3290605.3300376

[43] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems* 33 (2020), 17022–17033.

[44] Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P Bigham, and Leah Findlater. 2023. From User Perceptions to Technical Improvement: Enabling People Who Stutter to Better Use Speech Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 361, 16 pages. doi:10.1145/3544548.3581224

[45] Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2022. Textless Speech-to-Speech Translation on Real Data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 860–872. doi:10.18653/v1/2022.naacl-main.63

[46] Wing-Zin Leung, Mattias Cross, Anton Ragni, and Stefan Goetze. 2024. Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis. *arXiv preprint arXiv:2406.08568* (2024).

[47] Wing-Zin Leung, Mattias Cross, Anton Ragni, and Stefan Goetze. 2024. Training Data Augmentation for Dysarthric Automatic Speech Recognition by Text-to-Dysarthric-Speech Synthesis. In *Interspeech 2024*. 2494–2498. doi:10.21437/Interspeech.2024-1645

[48] Jingyi Li, Weiping Tu, and Li Xiao. 2023. Freevc: Towards High-Quality Text-Free One-Shot Voice Conversion. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. doi:10.1109/ICASSP49357.2023.10095191

[49] Boon Pang Lim. 2011. *Computational differences between whispered and non-whispered speech*. University of Illinois at Urbana-Champaign.

[50] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013/

[51] Anuprabha M, Krishna Gurugubelli, and Anil Kumar Vuppala. 2025. Fairness in Dysarthric Speech Synthesis: Understanding Intrinsic Bias in Dysarthric Speech Cloning using F5-TTS. In *Interspeech 2025*. 2750–2754. doi:10.21437/Interspeech.2025-1536

[52] Siyuan Ma, Dantong Jin, Ming Zhang, Bixuan Zhang, You Wang, Guang Li, and Meng Yang. 2019. Silent Speech Recognition Based on Surface Electromyography. In *2019 Chinese Automation Congress (CAC)*. 4497–4501. doi:10.1109/CAC48633.2019.8996289

[53] Robert L. MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A. Ladewig, Jimmy Tobin, Michael P. Brenner, Philip C. Nelson, Jordan R. Green, and Katrin Tomanek. 2021. Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia. In *Interspeech 2021*. 4833–4837. doi:10.21437/Interspeech.2021-697

[54] Harshit Malaviya, Jui Shah, Maitreya Patel, Jalansh Munshi, and Hemant A. Patil. 2020. Mspec-Net : Multi-Domain Speech Conversion Network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7764–7768. doi:10.1109/ICASSP40776.2020.9052966

[55] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017*. 498–502. doi:10.21437/Interspeech.2017-1386

[56] Hadil Mehrez, Mounira Chaiani, and Sid Ahmed Selouani. 2024. Using StarGANv2 voice conversion to enhance the quality of dysarthric speech. In *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 738–744.

[57] X. Menendez-Pidal, J.B. Polikoff, S.M. Peters, J.E. Leonzio, and H.T. Bunnell. 1996. The Nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, Vol. 3. 1962–1965 vol.3. doi:10.1109/ICSLP.1996.608020

[58] Péter Mihajlik, Éva Székely, Piroska Barta, Máté Soma Kádár, Gergely Dobsinszki, and László Tóth. 2025. Improved Dysarthric Speech to Text Conversion via TTS Personalization. *arXiv preprint arXiv:2508.06391* (2025).

[59] Yoshitaka Nakajima, Hideki Kashioka, Kiyohiro Shikano, and Nick Campbell. 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, Vol. 5. V–708. doi:10.1109/ICASSP.2003.1200069

[60] NVIDIA Corporation. 2024. NVIDIA TensorRT: High Performance Deep Learning Inference. https://developer.nvidia.com/tensorrt. Accessed: 2025-12-02.

[61] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 5206–5210.

[62] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Pierre Isabelle, Eugene Charniak, and Dekang Lin (Eds.). Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. doi:10.3115/1073083.1073135

[63] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proceedings of Interspeech*. 2613–2617. doi:10.21437/Interspeech.2019-2680

[64] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Interspeech 2021*. 3615–3619. doi:10.21437/Interspeech.2021-475

[65] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_CVPR_2020/papers/Prajwal_Learning_Individual_Speaking_Styles_for_Accurate_Lip_to_Speech_Synthesis_CVPR_2020_paper.pdf

[66] Zhaopeng Qian, Kejing Xiao, and Chongchong Yu. 2023. A Survey of Technologies for Automatic Dysarthric Speech Recognition. *EURASIP Journal on Audio, Speech, and Music Processing* 2023, 1 (2023), 48. doi:10.1186/s13636-023-00318-2

[67] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 1182, 27 pages.

[68] Jun Rekimoto. 2023. WESPER: Zero-shot and realtime whisper to normal voice conversion for whisper-based speech interactions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–12.

[69] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. 2012. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language resources and evaluation* 46, 4 (2012), 523–541.

[70] Neha Sahipjohn, Neil Shah, Vishal Tambrahalli, and Vineet Gandhi. 2023. RobustL2S: Speaker-Specific Lip-to-Speech Synthesis exploiting Self-Supervised Representations. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1492–1499. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10317357

[71] Paban Sapkota, Hemant Kumar Kathania, Sudarsana Reddy Kadiri, and Shrikanth Narayanan. 2024. Improving end-to-end speech recognition for dysarthric speech through in-domain data augmentation. In *2024 58th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 345–349.

[72] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862* (2019).

[73] Neil Shah, Ayan Kashyap, Shirish Karande, and Vineet Gandhi. 2025. MRI2Speech: Speech Synthesis from Articulatory Movements Recorded by Real-time MRI. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[74] Neil Shah, Neha Sahipjohn, Vishal Tambrahalli, Ramanathan Subramanian, and Vineet Gandhi. 2024. StethoSpeech: Speech Generation Through a Clinical Stethoscope Attached to the Skin. 8, 3, Article 123 (Sept. 2024), 21 pages. doi:10.1145/3678515

[75] Shakeel A Sheikh, Md Sahidullah, and Ina Kodrasi. 2025. Deep learning for pathological speech: A survey. *arXiv preprint arXiv:2501.03536* (2025). doi:10.48550/arXiv.2501.03536

[76] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783.

[77] Yusuke Shinohara. 2016. Adversarial multi-task learning of deep neural networks for robust speech recognition.. In *Interspeech*. San Francisco, CA, USA, 2369–2372.

[78] Tanmay Srivastava, R. Michael Winters, Thomas Gable, Yu Te Wang, Teresa LaScala, and Ivan J. Tashev. 2024. Whispering Wearables: Multimodal Approach to Silent Speech Recognition with Head-Worn Devices. In *Proceedings of the 26th International Conference on Multimodal Interaction* (San Jose, Costa Rica) *(ICMI '24)*. Association for Computing Machinery, New York, NY, USA, 214–223. doi:10.1145/3678957.3685720

[79] Tianyi Tan, Haoxin Ruan, Xinan Chen, Kai Chen, Zhibin Lin, and Jing Lu. 2025. DistillW2N: A Lightweight One-Shot Whisper to Normal Voice Conversion Model Using Distillation of Self-Supervised Features. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[80] Chenyu Tang, Josée Mallah, Dominika Kazieczko, Wentian Yi, Tharun Reddy Kandukuri, Edoardo Occhipinti, Bhaskar Mishra, Sunita Mehta, and Luigi G Occhipinti. 2025. Wireless Silent Speech Interface Using Multi-Channel Textile EMG Sensors Integrated into Headphones. *IEEE Transactions on Instrumentation and Measurement* (2025).

[81] VAANI Team. 2025. VAANI: Capturing the Language Landscape for an Inclusive Digital India (Phase 1). https://vaani.iisc.ac.in/.

[82] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. 2013. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics*, Vol. 19. Acoustical Society of America, 035081.

[83] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*. 4068–4076.

[84] Ashwin Vaidya, Ben Williams, Stoyan Stoyanov, Shari Trewin, and Chaitanya Reddy. 2023. Enabling People Who Stutter to Use Voice Assistants: Voice Accessibility as a Design Imperative. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM. doi:10.1145/3544548.3581224

[85] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6309–6318. https://dl.acm.org/doi/10.5555/3295222.3295378

[86] Dominik Wagner, Ilja Baumann, and Tobias Bocklet. 2023. Vocoder-Free Non-Parallel Conversion of Whispered Speech With Masked Cycle-Consistent Generative Adversarial Networks. *arXiv preprint arXiv:2306.06514* (2023). doi:10.48550/arXiv.2306.06514

[87] Dominik Wagner, Ilja Baumann, and Tobias Bocklet. 2024. Generative adversarial networks for whispered to voiced speech conversion: a comparative study. *International Journal of Speech Technology* 27, 4 (2024), 1093–1110. doi:10.1007/s10772-024-10161-1

[88] Huimeng Wang, Zengrui Jin, Mengzhe Geng, Shujie Hu, Guinan Li, Tianzi Wang, Haoning Xu, and Xunying Liu. 2024. Enhancing Pre-Trained ASR System Fine-Tuning for Dysarthric Speech Recognition Using Adversarial Data Augmentation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 12311–12315. doi:10.1109/ICASSP48485.2024.10447702

[89] Rui Wang and Askar Hamdulla. 2022. Fusion of MFCC and IMFCC for Whispered Speech Recognition. In *2022 3rd International Conference on Pattern Recognition and Machine Learning (PRML)*. 285–289. doi:10.1109/PRML56267.2022.9882209

[90] Yuejiao Wang, Xixin Wu, Disong Wang, Lingwei Meng, and Helen Meng. 2024. UNIT-DSR: Dysarthric Speech Reconstruction System Using Speech Unit Normalization. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 12306–12310. doi:10.1109/ICASSP48485.2024.10446921

[91] Seung Hee Yang and Minhwa Chung. 2020. Improving dysarthric speech intelligibility using cycle-consistent adversarial training. *arXiv preprint arXiv:2001.04260* (2020). doi:10.5220/0009163003080313

[92] Jeong Hun Yeo, Chae Won Kim, Hyunjun Kim, Hyeongseop Rha, Seunghee Han, Wen-Huang Cheng, and Yong Man Ro. 2025. Personalized lip reading: adapting to your unique lip movements with vision and language. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'25/IAAI'25/EAAI'25)*. AAAI Press, Article 1053, 9 pages. doi:10.1609/aaai.v39i9.33026

[93] Xianzhang Zeng, Beicheng Zhu, Yang Liu, and Longhan Xie. 2025. A Cross-Subject sEMG-to-Speech Conversion System Using Content Features and Model Calibration. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 33 (2025), 2215–2224. doi:10.1109/TNSRE.2025.3576860

[94] Zeta-Chicken. 2017. toWhisper. https://github.com/zeta-chicken/toWhisper. Accessed: 2025-08-24.

[95] Ruidong Zhang, Hao Chen, Devansh Agarwal, Richard Jin, Ke Li, François Guimbretière, and Cheng Zhang. 2023. HPSpeech: Silent Speech Interface for Commodity Headphones. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers* (Cancun, Quintana Roo, Mexico) *(ISWC '23)*. Association for Computing Machinery, New York, NY, USA, 60–65. doi:10.1145/3594738.3611365