Neural Voice Reconstruction for Personalized Speech Rehabilitation in Patients With Neurological Disorders

Tabitha Warner
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
twarner@ucf.edu

Dariati Hidayanto
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
hidayanto1992@ku.edu

Mauro Bataglia
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
maurobataglia931@colostate.edu

Fredy Ziegert
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
fredyziegert@uab.edu

Abstract

Neurological disorders such as Parkinson disease, amyotrophic lateral sclerosis, stroke, and multiple sclerosis frequently impair speech production while leaving communicative intent, language competence, or residual vocal control partially intact. Existing text-to-speech and voice conversion systems can generate intelligible synthetic speech, but they are rarely designed for the clinical realities of progressive or heterogeneous dysarthria, speaker-specific identity preservation, residual speech learning, and rehabilitation-oriented feedback. This article develops a methodological framework for neural voice reconstruction that treats personalized speech rehabilitation as a constrained acoustic-articulatory inference problem rather than as a generic synthesis task. The framework combines a speech encoder, an articulatory-deficit estimator, a speaker-identity encoder, a controllable reconstruction decoder, and an uncertainty-aware adaptation layer. Reconstruction is formulated as a multi-objective learning problem that balances intelligibility, speaker similarity, articulatory error correction, calibrated confidence, and regularization. Personalization is handled through constrained speaker adaptation from limited residual speech and through a rehabilitation controller that selects practice stimuli according to estimated success probability and expected communicative gain. Evaluation logic is specified around clinically meaningful endpoints, including intelligibility to unfamiliar listeners, naturalness, identity preservation, listening effort, error profile, and communicative participation, with explicit separation between mechanistic hypotheses and observed evidence. The article also analyzes failure modes involving data scarcity, nonstationary disease progression, over-smoothing of pathological cues, privacy, and automation bias. The resulting framework provides a technically

substantive basis for aligning neural voice reconstruction with the goals of personalized speech rehabilitation in neurological care.

Keywords

neural speech reconstruction; personalized speech rehabilitation; dysarthria; voice conversion; articulatory deficit estimation; clinical speech technology

1. Introduction

Neurological disorders disrupt speech through multiple motor and sensorimotor pathways. Hypokinetic dysarthria in Parkinson disease typically reduces vocal intensity, narrows pitch and loudness variation, and impairs consonant precision; spastic and flaccid dysarthria after stroke or amyotrophic lateral sclerosis alter respiratory support, phonation, resonance, and articulation; ataxic dysarthria disturbs timing and coordination; and apraxia of speech interferes with the planning of articulatory gestures [14]. These conditions do not simply reduce acoustic quality. They alter the mapping between intended phonological units and observable acoustic evidence, producing substitutions, omissions, distortions, prosodic flattening, and variable intelligibility across speaking contexts. A voice reconstruction system that treats pathological speech merely as noisy speech risks learning a generic correction rule that erases speaker identity and clinically informative error structure.

Recent work in neural speech decoding and synthesis has shown that acoustic speech can be reconstructed from cortical activity [1], that sentence-level decoding can be achieved in a paralyzed person with anarthria [2], and that avatar-based speech output can combine decoding with expressive control [3]. Brain-to-text and electrocorticography-based synthesis studies have further demonstrated the feasibility of mapping neural activity to phone-like or acoustic representations [4,5]. These advances are important because they establish the technical plausibility of bypassing impaired articulatory pathways. However, the clinical problem addressed here is broader than direct neural decoding. Many patients retain residual speech that is partially intelligible, effortful, and variable. Others may have no reliable neural interface but can provide acoustic and articulatory samples. A rehabilitation-oriented system must therefore reconstruct a personalized voice from whatever residual evidence is available, while preserving the patient's identity and supporting gradual improvement rather than replacing residual ability outright.

Modern neural speech systems provide strong building blocks. Self-supervised speech encoders such as wav2vec 2.0 and HuBERT learn representations that transfer across low-resource speech tasks [6,7]. End-to-end text-to-speech models such as Tacotron and FastSpeech provide controllable acoustic generation and duration modeling [8,9]. Speaker-verification-based transfer learning and speaker embeddings enable multispeaker synthesis and identity conditioning [10,11]. Voice conversion methods can transform acoustic style while preserving linguistic content [12,13]. Clinical voice treatment, illustrated by intensive voice therapy for Parkinson disease, shows that structured practice with auditory and somatosensory feedback can improve vocal output [15]. Yet these strands are rarely integrated into a single framework that specifies how neural reconstruction should represent articulatory deficit, personalize to a patient's changing voice, calibrate uncertainty, and support rehabilitation decisions.

The central contribution of this article is a methodological framework for personalized neural voice reconstruction in neurological speech rehabilitation. The framework treats reconstruction as the inference of a corrected, identity-preserving acoustic trajectory from

residual speech and, where available, auxiliary articulatory or neural signals. It separates content, speaker identity, articulatory deficit, and expressive style, and it introduces a rehabilitation controller that uses calibrated reconstruction confidence to select practice material. The article does not report a clinical trial or a benchmark evaluation. Instead, it develops a technically explicit formulation, derives the consequences of design choices, relates them to established speech technology and clinical evidence, and specifies evaluation logic that can distinguish mechanistic reasoning from observed clinical benefit.

2. Neurological Speech Impairment and Reconstruction Requirements

2.1 Heterogeneous mechanisms and variable residual ability

Neurological speech disorders are not a single acoustic class. In hypokinetic dysarthria, reduced amplitude and impaired scaling of movement produce monopitch, monoloudness, imprecise consonants, and short rushes of speech. In spastic dysarthria, hypertonia and reduced range of movement produce strained-strangled voice, slow articulation, and prosodic limitation. Flaccid dysarthria reflects weakness and may produce hypernasality, audible inspiration, and consonant imprecision. Ataxic dysarthria involves incoordination and gives rise to irregular articulatory breakdowns and distorted vowel space. Mixed dysarthrias are common, particularly in amyotrophic lateral sclerosis and multiple sclerosis. Apraxia of speech adds a planning deficit in which the patient may produce inconsistent articulatory attempts despite adequate muscle strength. These differences imply that a single global degradation model is insufficient. A reconstruction system must estimate which dimensions of speech production are affected and must distinguish errors that are stable from errors that vary with fatigue, medication state, emotional arousal, or speaking rate.

Residual ability also changes the appropriate reconstruction target. A patient with mild dysarthria may benefit from subtle correction that preserves most of the original acoustic signal. A patient with severe anarthria may require synthesis from text, neural signals, or small residual vocalizations. Between these extremes lies a large clinical population for whom reconstruction should be partial, adaptive, and reversible. The system should not force all users toward a single normalized voice. It should expose controls for intelligibility gain, identity preservation, speaking rate, pause insertion, and expressive range, and it should document how those controls affect the patient's communicative participation.

2.2 Clinical and technical requirements

The first requirement is intelligibility. Reconstruction must reduce phoneme confusions and prosodic ambiguity without introducing new errors that listeners cannot resolve from context. The second is identity preservation. Patients often regard their voice as part of selfhood, and a synthetic voice that sounds like a different person can undermine acceptance and social participation. The third is naturalness, but naturalness should not be pursued by suppressing all pathological variation. Some cues, such as slow rate or reduced loudness, may be compensatory and meaningful. The fourth is controllability. Clinicians and patients need to adjust reconstruction strength, speaking rate, and output modality across settings. The fifth is latency and robustness. Conversational use requires short turn-taking delays and graceful degradation under noise, fatigue, or device mismatch. The sixth is privacy and security. Voice models encode biometric and health-related information and must be governed accordingly. The seventh is clinical interpretability. The system should report which acoustic dimensions were modified and whether confidence is sufficient for communicative use or for rehabilitation feedback.

Augmentative and alternative communication practice emphasizes that communication success depends on the fit among user, device, task, and partner [17]. A reconstruction system that improves word recognition in isolated tests but increases conversational repair in natural dialogue has not met the clinical goal. Similarly, studies of acquired dysarthria in conversation show that understandability problems arise from interactional as well as acoustic sources [18]. The reconstruction framework must therefore be evaluated in communicative context, not only through acoustic distance metrics.

3. Methodological Framework

3.1 Problem formulation

The framework represents a patient's residual speech as an observation x, which may include the acoustic waveform, derived articulatory features such as formant trajectories or nasalance, and optional neural activity when an interface is available. The target y is a reconstructed acoustic representation intended for communication or rehabilitation practice. Speaker identity is represented by s, an articulatory state estimate by a, and context or control variables by c, including speaking rate, intended emphasis, and clinician-selected reconstruction strength. The reconstruction mapping is specified as follows.

$$\hat{y} = G(x, s, a, c; \theta)$$

Here G denotes the reconstruction network parameterized by θ, and ŷ is the reconstructed acoustic signal or acoustic feature sequence. The input x is not assumed to be clean or complete. It may contain substitutions, omissions, and temporal distortions. The identity variable s can be derived from a speaker encoder trained on the patient's residual speech and reference samples. The articulatory state a is an estimate of the patient's current production deficit, not a direct measurement of underlying physiology. The context c allows the same input to be rendered at different rates or correction strengths. This formulation makes explicit that reconstruction is conditional on patient identity and deficit state rather than being a single deterministic correction of noise.

The learning objective must reflect several partially competing goals. A reconstruction that maximizes intelligibility may over-normalize identity and prosody. A reconstruction that maximizes speaker similarity may preserve pathological cues that impair intelligibility. The framework therefore uses a composite objective with explicit weighting terms.

$$L(\theta) = Lrec + \lambda spk\, Lspk + \lambda art\, Lart + \lambda unc\, Lunc + \lambda reg\, \Omega(\theta)$$

In this expression, Lrec is the reconstruction loss between ŷ and a reference acoustic target when such a target is available, or a self-supervised consistency loss when it is not. Lspk measures speaker-identity preservation, Lart measures agreement between estimated and expected articulatory correction, Lunc penalizes miscalibrated predictive uncertainty, and Ω(θ) is a regularizer. The coefficients λspk, λart, λunc, and λreg set the relative influence of these terms. They are not universal constants. In a clinical deployment they should be treated as configuration parameters that reflect the patient's goals, the severity of dysarthria, and the intended use of the output. For example, a patient who uses reconstructed speech for public speaking may prefer higher intelligibility, whereas a patient who uses it mainly with familiar listeners may prioritize identity preservation.

3.2 Speaker identity preservation

Speaker identity is not a single acoustic attribute. It includes vocal tract resonances, pitch range, glottal source characteristics, temporal rhythm, and learned pronunciation habits. The

framework encodes identity through a speaker embedding and constrains reconstruction so that the generated voice remains close to the patient's reference identity in the embedding space. A cosine-based identity loss provides a simple and interpretable formulation.

$$Lspk = 1 - cos(ey, e\hat{y})$$

Here ey is the speaker embedding extracted from a reference voice sample, eŷ is the embedding extracted from the reconstructed voice, and cos denotes cosine similarity. The reference sample may come from before disease onset, from a stable period of residual speech, or from multiple samples aggregated to reduce day-to-day variability. The loss encourages similarity without requiring waveform-level identity, which is appropriate because exact acoustic matching is impossible when the patient's production system has changed. The limitation is that speaker embeddings are trained on largely healthy speech and may not represent pathological voice qualities faithfully. Identity preservation should therefore be assessed by both embedding-based measures and listener judgments, and the clinical team should decide whether the target identity is the patient's premorbid voice, current residual voice, or a clinician-guided compromise.

3.3 Articulatory deficit estimation and correction

A key difference between generic speech enhancement and rehabilitation-oriented reconstruction is the need to estimate which articulatory dimensions are impaired. The framework uses a deficit estimator D that maps acoustic and auxiliary features to a low-dimensional articulatory error state. For each time frame t, the estimator predicts deviations along interpretable dimensions such as vocal intensity, pitch range, consonant constriction, velopharyngeal closure, and articulatory timing.

$$Lart = (1/T) \sum t \, ||D(x_t) - dt||^2$$

In this equation, T is the number of frames, x_t is the input representation at frame t, D(x_t) is the predicted deficit vector, and dt is a target deficit vector derived from clinician annotation, acoustic reference data, or a structured perturbation model. The target dt should not be interpreted as a direct physiological measurement. It is a supervised or weakly supervised representation of the correction that the system is expected to apply. The loss enforces consistency between the estimated deficit and the correction that the decoder actually performs. If the decoder increases consonant constriction, for example, the deficit estimator should indicate a constriction-related deficit in the corresponding frame. This coupling makes the system more interpretable and provides a basis for clinical feedback.

The estimator also serves a safety function. If the predicted deficit is outside the range represented in the patient's calibration data, the system can lower reconstruction strength or request a clinician review. It can also distinguish stable deficits from transient fluctuations. A sudden increase in estimated deficit may reflect fatigue, medication fluctuation, or device error rather than disease progression, and the system should not silently treat all variation as a new voice target.

3.4 Uncertainty-aware reconstruction

Clinical use requires calibrated confidence. A reconstructed utterance that sounds fluent but is wrong can be more harmful than a clearly marked uncertain output. The framework therefore models reconstruction as a predictive distribution rather than a point estimate. For an acoustic target y, the decoder predicts a mean μ and a variance σ^2. A Gaussian negative log-likelihood provides a standard uncertainty-aware loss.

$$Lunc = (y - \mu)^2/(2\sigma^2) + 0.5\, log\, \sigma^2$$

Here μ is the predicted acoustic value, σ² is the predictive variance, and y is the reference target when available. The first term penalizes large errors more when the model claims low variance, and the second term discourages the model from assigning arbitrarily large variance to every input. In deployment, the variance can be summarized at the phoneme, word, or utterance level and presented as a confidence indicator. Calibration should be evaluated with reliability diagrams and proper scoring rules, not only with average error [20]. The system should also distinguish aleatoric uncertainty, reflecting irreducible ambiguity in the patient's residual signal, from epistemic uncertainty, reflecting unfamiliar inputs or insufficient speaker data. The former may be communicated to the listener, whereas the latter should trigger adaptation or clinician review.

4. Neural Architecture and Learning Strategy

4.1 Encoder-decoder organization

The reconstruction architecture separates four functions: content encoding, identity encoding, articulatory state estimation, and acoustic decoding. A self-supervised speech encoder provides robust content and phonetic representations from limited patient data [6,7]. The encoder should be trained or adapted with masking objectives that preserve phonological contrasts while reducing sensitivity to speaker-specific and microphone-specific nuisance factors. A speaker encoder supplies the identity embedding used in the loss described above. Speaker verification architectures such as ECAPA-TDNN offer strong identity representations and can be adapted to pathological voice samples [11]. The articulatory deficit estimator operates on intermediate representations and produces a low-dimensional state that conditions the decoder. The decoder generates acoustic features or a waveform through a neural vocoder, with control inputs for rate, reconstruction strength, and expressive range.

This organization follows the logic of disentangled voice conversion and multispeaker synthesis. Voice conversion methods have shown that content and speaker style can be partially separated and recombined [12,13]. Multispeaker text-to-speech systems have shown that speaker embeddings can condition a shared generator [10]. Style transfer in out-of-domain singing voice synthesis further illustrates that source style, target style, and speaker identity can be modeled as separate factors when training data are heterogeneous [16]. The clinical adaptation is that the "style" factor is not merely expressive; it includes pathological articulatory deviation that must be estimated and controlled. The framework therefore does not treat the patient's impaired speech as a style to be imitated. It treats the estimated deficit as a condition for correction, while preserving the patient's identity as a separate constraint.

4.2 Training data and supervision

The framework must operate under sparse and heterogeneous supervision. Some patients provide paired samples of impaired and corrected speech after therapy, but many do not. Some provide only residual speech and a small number of intelligible reference utterances. Others provide text and audio pairs, and a smaller number may provide neural signals. The learning strategy should therefore combine reconstruction loss on available paired data with self-supervised consistency, speaker-identity constraints, and clinically structured weak supervision. Clinician ratings of intelligibility, naturalness, and error type can provide labels at the utterance level, but they are costly and variable. Acoustic-phonetic measures such as vowel space area, consonant closure duration, and spectral moments can provide auxiliary targets, but they should not be treated as direct substitutes for communicative outcomes.

Data augmentation can improve robustness if it simulates plausible sources of variation rather than arbitrary noise. Speaking rate perturbation, channel filtering, and additive noise can improve generalization. However, augmenting pathological speech by randomly deleting phonemes or distorting formants may teach the model to correct errors that the patient does not have. A better approach is to use a structured perturbation model that generates synthetic deficits along the same interpretable dimensions used by the deficit estimator. The model can then learn to invert perturbations that are clinically plausible, while remaining cautious on inputs that fall outside the simulated range.

4.3 Personalization with limited data

Personalization should not require retraining the entire network for each patient. The framework uses a shared backbone with patient-specific adapters or embeddings. A base model is trained across speakers and disorder types to learn general reconstruction and identity-preservation priors. A small patient-specific adapter is then fit to the patient's residual speech, reference identity samples, and clinician-defined correction targets. The adaptation objective regularizes the adapter toward the shared parameters to reduce overfitting.

$$\theta s = argmin\theta\ (1/|Ss|) \sum i \in Ss\ L(\theta;\ xi, yi) + \lambda ad\ ||\theta - \theta 0||^2$$

Here θs denotes the adapted parameters for patient s, Ss is the patient's adaptation set, L is the composite reconstruction loss, θ0 denotes the shared base parameters, and λad controls regularization strength. The adaptation set may contain only a few utterances, so the regularization term is essential. The formulation also allows the clinical team to freeze certain parameters, such as those controlling identity, while adapting those controlling articulatory correction. This separation supports safer personalization: the system can learn patient-specific correction patterns without drifting into a different speaker identity.

Temporal drift is a central challenge. Neurological conditions may progress, fluctuate with medication, or improve during rehabilitation. The adapter should therefore be updated with change detection rather than continuously. A practical rule is to monitor reconstruction loss, identity similarity, and calibration error on a held-out patient sample. When these indicators exceed a clinically reviewed threshold, the system prompts recalibration. The threshold should be selected per patient and should not be inferred from a single global default.

5. Personalized Adaptation and Rehabilitation Control

5.1 Reconstruction as a rehabilitation instrument

Speech rehabilitation is not merely the provision of an output device. It involves practice, feedback, and the gradual reorganization of motor planning. Neural voice reconstruction can support rehabilitation in two ways. First, it can serve as a communication prosthesis that converts residual speech into more intelligible output in real time. Second, it can serve as a practice instrument that renders an idealized or partially corrected version of the patient's intended utterance, allowing the patient to compare their production with the reconstructed target. The first use emphasizes robustness and latency; the second emphasizes interpretability and controlled difficulty. The same model can support both, but the interface and evaluation criteria differ.

For practice, the system should not always present the easiest reconstruction. Motor learning benefits from challenge that is attainable but not overwhelming. The rehabilitation

controller therefore selects practice stimuli and reconstruction conditions based on the patient's current success probability and the expected communicative gain. Let P be a set of candidate stimuli or practice conditions, such as syllable sequences, phrase length, speaking rate, or reconstruction strength. The controller selects the condition p_t at practice step t according to the following rule.

$$p_t = argmax p \in P\, E[\Delta Q(p)]\ subject\ to\ Pr(success \mid p, \theta s) \geq \rho$$

In this expression, E[ΔQ(p)] is the expected improvement in a clinically meaningful performance measure Q under condition p, Pr(success | p, θs) is the estimated probability that the patient can complete the task successfully given the adapted model, and ρ is a minimum success probability set by the clinician. The measure Q may combine intelligibility, phoneme accuracy, vocal intensity, or listener-confirmed comprehensibility depending on the treatment goal. The constraint prevents the controller from selecting tasks that are so difficult that they produce failure and fatigue. The expected improvement term prevents it from selecting tasks that are too easy to produce learning. The controller is thus a decision aid, not an autonomous therapist. It should present options and rationale to the clinician and patient.

5.2 Feedback design and communicative participation

Feedback should be multimodal and cautiously delivered. Visual feedback can display pitch range, loudness, rate, and articulatory error estimates. Auditory feedback can present the patient's original production, the reconstructed production, or a blended output. Blended output is often preferable in early rehabilitation because it preserves the patient's voice while enhancing intelligibility. As performance improves, the blend can be adjusted toward the patient's unassisted speech or toward a more natural reconstruction. The system should avoid implying that the reconstructed voice is the "correct" voice. The goal is communicative effectiveness and patient-defined participation, not conformity to a healthy acoustic norm.

Personalization also extends to listener and setting. A patient may use a high-intelligibility mode with unfamiliar listeners and a more identity-preserving mode with family. The system should store these profiles and make them easy to switch. It should also log when reconstruction is active, because listeners may otherwise attribute synthetic fluency to the patient. Transparency is especially important in clinical and legal settings. Augmentative and alternative communication research emphasizes that device acceptance depends on social and interactional factors, not only on technical performance [17]. The reconstruction interface should therefore be designed with familiar communication partners and should support repair strategies such as repetition, spelling, and gesture.

6. Evaluation Logic and Clinical Translation

6.1 Endpoints and estimands

Evaluation must distinguish technical reconstruction quality from clinical benefit. Technical endpoints include phoneme error rate, word error rate from an automatic recognizer, spectral distortion, speaker-embedding similarity, and latency. These are useful for engineering iteration but insufficient for clinical claims. Clinical endpoints should include intelligibility to unfamiliar listeners, naturalness, listener effort, identity preservation, communicative participation, fatigue, and patient-reported satisfaction. Each endpoint should be defined as an estimand with a specified population, comparison, time horizon, and handling of missing data. For example, the effect of reconstruction on intelligibility should be estimated for patients with a specified disorder and severity range, comparing reconstructed

speech with unassisted speech and with a generic synthetic voice, over a defined practice period.

The evaluation should separate mechanistic hypotheses from observed outcomes. A mechanistic hypothesis might state that articulatory deficit estimation reduces phoneme substitutions by increasing consonant constriction in impaired segments. An observed outcome would be the change in listener identification accuracy for those segments. The framework predicts that calibrated uncertainty and identity preservation will improve acceptance, but that prediction must be tested rather than assumed. Similarly, the framework predicts that adaptive difficulty will improve practice engagement, but engagement is influenced by many factors outside the model. Clinical claims should be tied to the measured endpoint rather than to the model's internal representations.

6.2 Study design and confounding

A rigorous evaluation would use a within-participant design with counterbalanced conditions and blinded listener ratings. Speech samples should be elicited across tasks, including isolated words, sentences, conversation, and emotional or urgent communication. Listeners should include familiar and unfamiliar partners, because familiar listeners may use contextual repair strategies that mask intelligibility gains. Order effects, fatigue, and practice effects should be controlled. Because neurological conditions fluctuate, repeated baseline measures are necessary. Medication state, time of day, and recent therapy should be recorded. The analysis should account for the nested structure of utterances within speakers and listeners within rating sessions.

Confounding is a major risk. Patients who adapt well to the system may differ in motivation, cognitive status, or social support from those who do not. A comparison between reconstructed and unassisted speech within the same session reduces some confounding but cannot establish long-term rehabilitation effects. A randomized or stepped-wedge design may be appropriate for clinical evaluation, but it must be ethically sensitive to the possibility that reconstruction provides immediate communicative benefit. The evaluation should therefore separate immediate communication outcomes from rehabilitation outcomes such as improved unassisted speech. The former can be studied with short-term crossover designs; the latter requires longitudinal designs with prespecified retention measures.

6.3 Uncertainty, error analysis, and safety monitoring

Uncertainty should be evaluated at the level at which decisions are made. If the system flags low confidence, does the listener actually benefit from the flag? If the clinician overrides a low-confidence output, is the override correct? Calibration metrics alone do not answer these questions. The evaluation should include decision-curve analysis for communicative actions, such as whether to send an utterance, request repetition, or switch to a text-based modality. Error analysis should classify failures into acoustic, linguistic, speaker-identity, and interactional categories. For example, a reconstruction may be acoustically intelligible but pragmatically inappropriate because it removes a polite pause or alters emphasis. Such errors are invisible to word error rate but consequential in conversation.

Safety monitoring should include off-target synthesis, identity drift, and inappropriate correction of clinically meaningful cues. In Parkinson disease, reduced loudness and prosodic variation are both symptoms and potential communicative signals [19]. A system that normalizes all prosodic variation may improve intelligibility in one context but reduce the patient's ability to convey emotion or urgency in another. The clinician should be able to

constrain the reconstruction so that it does not alter selected dimensions. The patient should also be able to audit and delete stored voice models. Privacy protection should include on-device processing where feasible, encryption of voice templates, and clear consent for secondary use.

7. Technical Discussion

The proposed framework differs from a standard cascaded automatic speech recognition and text-to-speech pipeline. A cascaded system first transcribes residual speech and then synthesizes a generic voice. It can achieve high intelligibility when recognition is accurate, but it discards speaker identity, prosody, and paralinguistic information, and it propagates recognition errors into synthesis. The reconstruction framework instead operates on acoustic and articulatory representations that preserve identity and timing. This is advantageous for rehabilitation because the patient's residual speech is not treated as an error to be erased but as evidence about the patient's current motor state. The trade-off is complexity: the reconstruction model requires more patient-specific calibration and more careful validation than a generic text-to-speech system.

The framework also differs from direct neural decoding. Neural decoding can bypass impaired articulation and may be transformative for severe anarthria [1–3]. However, it requires surgical access, stable electrode interfaces, and extensive calibration. Most patients in speech rehabilitation do not have such interfaces. The reconstruction framework is designed to use neural signals when available but to remain functional with acoustic and articulatory inputs alone. This modularity is a strength, but it also means that the system cannot claim the same neural specificity as an implanted decoder. The appropriate comparison depends on the patient population and the stage of disease.

Identity preservation presents a fundamental tension. A voice reconstruction system can preserve identity by staying close to the patient's residual voice, but residual voice quality may be the source of unintelligibility. It can instead move toward a premorbid voice, but premorbid recordings may be unavailable or may no longer match the patient's current anatomy and motor control. The framework handles this tension by making identity a controllable constraint rather than a fixed target. The clinician and patient choose the operating point. This is clinically realistic, but it complicates evaluation because there is no single ground-truth identity. Research should therefore report identity preservation across multiple reference conditions and should include patient-defined acceptability.

Data scarcity is another limitation. Neurological speech corpora are small, heterogeneous, and often collected under clinical constraints. Self-supervised encoders reduce dependence on labeled data [6,7], and adapter-based personalization reduces the need for full retraining. Nevertheless, the model may fail on rare disorder combinations, severe dysarthria, or atypical accents. Uncertainty estimation is intended to detect such failures, but uncertainty estimates can themselves be miscalibrated on out-of-distribution patients [20]. A safe deployment should include clinician review of low-confidence cases and periodic recalibration with new patient data.

Finally, the framework's rehabilitation controller depends on estimates of success probability and expected gain. These estimates are necessarily model-based and should not be treated as clinical prescriptions. The controller may propose a practice sequence, but the clinician should retain authority over goals, intensity, and stopping rules. The system should also monitor fatigue and frustration, because a model of motor learning that ignores affective

state can produce harmful practice schedules. The most defensible claim is therefore conditional: if articulation-aware reconstruction, calibrated uncertainty, and constrained practice selection are implemented with appropriate clinical oversight, they provide a coherent path toward personalized rehabilitation. Whether that path improves long-term outcomes is an empirical question that requires the evaluation designs described above.

8. Conclusion

This article has developed a methodological framework for neural voice reconstruction in personalized speech rehabilitation for patients with neurological disorders. The framework formulates reconstruction as a constrained inference problem over residual speech, speaker identity, articulatory deficit, and context. It specifies a multi-objective learning criterion that balances reconstruction accuracy, identity preservation, articulatory correction, uncertainty calibration, and regularization. It describes an encoder-decoder architecture with a self-supervised speech backbone, a speaker encoder, an articulatory deficit estimator, and a controllable decoder, and it explains how limited patient data can be handled through regularized adapter-based personalization. It further introduces a rehabilitation controller that selects practice conditions according to expected communicative gain and a minimum success probability, and it outlines evaluation logic centered on clinically meaningful endpoints, confounding control, error analysis, and safety monitoring.

The framework's main technical insight is that pathological speech should not be treated as generic noise. It contains information about the patient's current motor state, identity, and compensatory strategies. A reconstruction system that estimates and controls articulatory deficit while preserving identity can support both communication and rehabilitation more coherently than a generic speech-to-text-to-speech pipeline. The main clinical insight is that reconstruction strength and practice difficulty are not fixed parameters; they are treatment decisions that should be shared among patient, clinician, and system. The main evidentiary limitation is that these claims are derived from mechanism and design reasoning, not from a completed clinical trial. Future evaluation should test whether calibrated, identity-preserving reconstruction improves intelligibility and participation without increasing fatigue, identity disruption, or automation bias. The framework provides a technically explicit and clinically grounded basis for that evaluation.

References

1. Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. Nature, 568(7753), 493–498. https://doi.org/10.1038/s41586-019-1119-1

2. Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., Chartier, J., Dougherty, M. E., Liu, P. M., Abrams, G. M., Tu-Chan, A., Ganguly, K., & Chang, E. F. (2021). Neuroprosthesis for decoding speech in a paralyzed person with anarthria. New England Journal of Medicine, 385(3), 217–227. https://doi.org/10.1056/NEJMoa2027540

3. Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., Dougherty, M. E., Liu, J. R., Wu, P., Berger, M. A., Zhuravleva, I., Tu-Chan, A., Ganguly, K., Anumanchipalli, G. K., & Chang, E. F. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. Nature, 620(7976), 1037–1046. https://doi.org/10.1038/s41586-023-06443-4

4. Herff, C., Heger, D., de Pesters, A., Tsiouris, P., Brunner, P., Schalk, G., & Schultz, T. (2015). Brain-to-text: Decoding spoken phrases from phone representations in the brain. Frontiers in Neuroscience, 9, 217. https://doi.org/10.3389/fnins.2015.00217

5. Angrick, M., Herff, C., Mugler, E., Krusienski, D. J., Schalk, G., & Schultz, T. (2019). Speech synthesis from ECoG using densely connected 3D convolutional neural networks. Journal of Neural Engineering, 16(3), 036019. https://doi.org/10.1088/1741-2552/ab0c59

6. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems (Vol. 33, pp. 12449–12460). Curran Associates.

7. Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

8. Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Ying, Z., Senior, A., Vinyals, O., Norouzi, M., & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. In Proceedings of Interspeech 2017 (pp. 4006–4010). https://doi.org/10.21437/Interspeech.2017-1452

9. Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). FastSpeech: Fast, robust and controllable text to speech. In Advances in Neural Information Processing Systems (Vol. 32, pp. 3165–3174). Curran Associates.

10. Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Lopez Moreno, I., & Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In Advances in Neural Information Processing Systems (Vol. 31, pp. 4480–4490). Curran Associates.

11. Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In Proceedings of Interspeech 2020 (pp. 3830–3834). https://doi.org/10.21437/Interspeech.2020-2650

12. Qian, K., Zhang, Y., Chang, S., Yang, X., & Hasegawa-Johnson, M. (2019). AutoVC: Zero-shot voice style transfer with only autoencoder loss. In Proceedings of the 36th International Conference on Machine Learning (Vol. 97, pp. 5210–5219). PMLR.

13. Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., & Ling, Z. (2018). The Voice Conversion Challenge 2018: Promoting development of parallel and nonparallel methods. In Proceedings of the Odyssey 2018 Workshop (pp. 195–202). https://doi.org/10.21437/Odyssey.2018-28

14. Duffy, J. R. (2019). Motor speech disorders: Substrates, differential diagnosis, and management (4th ed.). Elsevier.

15. Ramig, L. O., Sapir, S., Countryman, S., Pawlas, A. A., O'Brien, C., Hoehn, M., & Thompson, L. L. (2001). Intensive voice treatment (LSVT) for patients with Parkinson's disease: A 2 year follow up. Journal of Neurology, Neurosurgery, and Psychiatry, 71(4), 493–498. https://doi.org/10.1136/jnnp.71.4.493

16. Zhang, Y., Huang, R., Li, R., He, J., Xia, Y., Chen, F., ... & Zhao, Z. (2024, March). Stylesinger: Style transfer for out-of-domain singing voice synthesis. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 17, pp. 19597-19605).

17. Beukelman, D. R., & Mirenda, P. (2013). Augmentative and alternative communication: Supporting children and adults with complex communication needs (4th ed.). Paul H. Brookes.

18. Bloch, S., & Wilkinson, R. (2009). Acquired dysarthria in conversation: Identifying sources of understandability problems. International Journal of Language & Communication Disorders, 44(5), 769–783. https://doi.org/10.1080/13682820802546969

19. Sapir, S., Ramig, L. O., & Fox, C. M. (2008). Speech and swallowing disorders in Parkinson disease. Otolaryngologic Clinics of North America, 41(2), 379–395. https://doi.org/10.1016/j.otc.2008.01.001

20. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321–1330). PMLR.