Computer Modeling in Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.072790

ARTICLE

Mordukhovich Subdifferential Optimization Framework for Multi-Criteria Voice Cloning of Pathological Speech

Rytis Maskeliūnas[1], Robertas Damaševičius[1,*], Audrius Kulikajevas[1], Kipras Pribuišis[2], Nora Ulozaitė-Stanienė[2] and Virgilijus Uloza[2]

[1]Center of Real Time Computer Systems, Kaunas University of Technology, Kaunas, 51423, Lithuania
[2]Department of Otorhinolaryngology, Academy of Medicine, Lithuanian University of Health Sciences, Kaunas, 50161, Lithuania

*Corresponding Author: Robertas Damaševičius. Email: robertas.damasevicius@ktu.lt

Received: 03 September 2025; Accepted: 05 November 2025; Published: 23 December 2025

ABSTRACT: This study introduces a novel voice cloning framework driven by Mordukhovich Subdifferential Optimization (MSO) to address the complex multi-objective challenges of pathological speech synthesis in under-resourced Lithuanian language with unique phonemes not present in most pre-trained models. Unlike existing voice synthesis models that often optimize for a single objective or are restricted to major languages, our approach explicitly balances four competing criteria: speech naturalness, speaker similarity, computational efficiency, and adaptability to pathological voice patterns. We evaluate four model configurations combining Lithuanian and English encoders, synthesizers, and vocoders. The hybrid model (English encoder, Lithuanian synthesizer, English vocoder), optimized via MSO, achieved the highest Mean Opinion Score (MOS) of 4.3 and demonstrated superior intelligibility and speaker fidelity. The results confirm that MSO enables effective navigation of trade-offs in multilingual pathological voice cloning, offering a scalable path toward high-quality voice restoration in clinical speech applications. This work represents the first integration of Mordukhovich optimization into pathological TTS, setting a new benchmark for speech synthesis under clinical and linguistic constraints.

KEYWORDS: Mordukhovich subdifferential optimization; multi-objective optimization; alaryngeal voice reconstruction

1 Introduction

Recently, the field of voice cloning and text-to-speech (TTS) synthesis has seen a notable progress [1]. Recent work has explored diverse applications including multilingual zero-shot voice conversion [2], real-time speech translation in video conferencing [3], and speech accessibility solutions for individuals with impairments [4,5]. For example, Li et al. [6] demonstrated effective multilingual synthesis with code-switching capabilities in Tibetan, while Nekvinda and Dušek [7] proposed meta-learning strategies for multilingual speech. The maturity of modern cloning systems is further reflected in their integration into educational environments, as shown by [8].

At the core of this evolution is the multispeaker Text-to-speech (SV2TTS) pipeline, which allows high-fidelity voice cloning using only a few samples from the target speaker [9]. The architecture, which combines an encoder, synthesizer, and vocoder, has enabled expressive multi-speaker synthesis with notable improvements in naturalness, speaker fidelity, and data efficiency [10]. Recent innovations such as transformer-based

models [11] and scalable multilingual synthesis frameworks [12] have further expanded the potential for cross-lingual cloning.

Despite many technological breakthroughs, several critical challenges remain unresolved:

- First, although multilingual speech synthesis has become more viable, sound-switching within a single utterance remains a complex task, requiring models to fluidly merge linguistic elements from different phonetic inventories while preserving natural prosody and coherence [13,14].
- Second, voice cloning models are typically trained on healthy voices and often fail to generalize to pathological speech, particularly in cases of alaryngeal voices affected by laryngeal cancer. These pathological signals can exhibit a wide range of acoustic irregularities—such as disrupted fundamental frequency, reduced harmonic structure, and excessive noise components—arising from surgical variability, prosthetic devices, and individual healing patterns [15,16]. As such, conventional models are ill-equipped to handle the non-stationary and high-variance nature of pathological voice signals.
- Third, most voice cloning research focuses on high-resource languages, overlooking linguistically rich but underrepresented languages such as Lithuanian. Lithuanian presents distinct phonological features, including unique vowels and consonants such as "ą," "č," and "ė," which are absent in mainstream TTS phoneme inventories [17]. Accurate synthesis in such languages requires an expanded phoneme set and careful modeling of language-specific prosody and rhythm—capabilities that existing models generally lack.

To address these intersecting challenges—cross-lingual synthesis, pathological variability, and low-resource linguistic modeling—we propose a novel framework built around Mordukhovich Subdifferential Optimization (MSO) [18]. Our approach introduces a principled multi-objective optimization scheme into the encoder–synthesizer–vocoder architecture, balancing competing criteria such as speech naturalness, speaker similarity, computational efficiency, and robustness to pathological voice distortions. By integrating MSO directly into the training and tuning loop, we identify optimal trade-offs between these objectives, enabling high-quality speech synthesis even in the presence of pathological voice patterns and phonetic mismatches. The proposed framework is evaluated on multiple encoder–synthesizer–vocoder configurations (combining English and Lithuanian components) and benchmarked on both subjective and objective criteria. Compared to our previous work on flow-based pathological synthesis [19] and denoising through gated LSTMs [20], this study introduces a fundamentally new optimization strategy and targets a more ambitious goal: restoration of natural-sounding, speaker-specific voices for patients with alaryngeal speech impairments in a linguistically underserved context.

The main contributions of this paper are as follows:

- Development and validation of a Mordukhovich Subdifferential Optimization to voice cloning, effectively handling non-smooth, multi-objective trade-offs.
- Pathological speech synthesis built specifically for Lithuanian, a low-resource language, requiring novel phoneme set and prosodic model expansion.
- Identification of an optimal hybrid synthesis architecture, which, when tuned with MSO, demonstrates a practical enough solution for low-resource clinical applications.

2 Methodology

The optimum selection of acoustic and linguistic features is necessary for capturing the complex characteristics of pathological speech. Our framework utilizes a comprehensive set of features derived from the speaker encoder and synthesizer modules, including speaker embeddings that encapsulate timbre, pitch, and intonation, as well as phoneme sequences and log-Mel spectrograms that represent linguistic and

spectral content, etc., chosen to balance the competing objectives of naturalness, speaker similarity, and pathological adaptability, with the Mordukhovich optimizer dynamically prioritizing them during training to navigate the non-smooth, multi-objective loss landscape inherent in impaired voice signals.

Traditional gradient-based or scalarization methods typically assume smooth objective landscapes and convex trade-offs, rendering them less effective when applied to clinical voice data characterized by acoustic irregularities, discontinuities, and high inter-speaker variability. The principal advantage of using MSO over conventional multi-objective methods, such as weighted sum approaches or just a Pareto front estimation via evolutionary algorithms, lies in the capacity to handle the *non-smooth and non-convex* nature of the loss landscapes associated with pathological voice synthesis. Conventional methods often assume differentiability or convexity to find a single aggregate solution, which is ill-suited for our problem where objectives such as natural speech and pathological adaptability can be highly discontinuous due to voice breaks and irregular glottal pulses. MSO, through its generalized notion of the subdifferential, provides a necessary rigorous optimality condition without these simplifying assumptions. This allows the optimizer to navigate complex trade-offs at points where gradients may not exist, effectively identifying robust parameter configurations that would be inaccessible to gradient-based optimizers. Consequently, MSO systematically discovers a solution where the subgradients of the competing losses are balanced, leading to synthesized speech that maintains high naturalness and speaker similarity without sacrificing the critical, nonsmooth features that characterize pathological voices.

2.1 Dataset

Two datasets were used to train the synthesizer model in our approach. The first dataset consisted of pathological speech, which is used to extract vocal features. The dataset was comprised of 154 laryngeal carcinoma patients' recordings, drawn from 77 individuals, and was specifically compiled for this study. The cohort of patients consisted of individuals who had undergone extensive surgical procedures (including type III cordectomy and beyond, partial or total laryngectomy) for histologically confirmed laryngeal carcinoma. Only recordings of patients who scored less than 40 points on the Impression of Voice Quality, Intelligibility, Noise, Fluency, and Quality of Voicing (IINFVo) scale were included, ensuring the presence of speech impediments in the samples. These samples were collected during routine outpatient visits, not earlier than 6 months after surgery, allowing sufficient time for recovery and rehabilitation. The recordings included phonetically balanced Lithuanian sentences ("Turėjo senelė žilą oželį" (roughly translatable as "The grandmother had a little gray goat")) and patients counting from 1 to 10 at a moderate pace.

The second data set (Liepa2) contained healthy speech recordings and was used as a core training resource [21], containing the healthy-sounding characteristics of Lithuanian speech of approximately 1000 h of audio recordings, featuring a diverse range of 2621 speakers (56% female and 44% male, young voices under 12 years of age, which make up 8% of the dataset, to mature voices over 61 years of age, accounting for 10%,), designed to reflect the rich phonetic and prosodic landscape of the Lithuanian language [22].

Each audio recording in both data sets was annotated to mark the beginning and end of segments, delineated by pauses that could signify a comma, the end of a sentence, or a breath. We excluded sequences containing nonphonemic sounds such as coughs or breaths that are marked by symbols like "+breath+" or "+noise+", to ensure that the model's learning is concentrated on linguistically relevant sounds, thereby improving the quality and clarity of the synthesized speech.

To date, there are no comparable open-access datasets for pathological voices for other languages that would allow meaningful cross-language or cross-dataset validation; therefore, the data set is specific to the underrepresented in research morphologically rich Lithuanian language. The scarcity of other data sets

meaningful for comparison is due to the clinical specificity of the speech impairments involved, which vary widely depending on the type of surgical intervention, the stage of rehabilitation, and the individual anatomical differences. As a result, expanding the dataset or incorporating external corpora is currently not feasible, and we believe that the proposed optimization framework at least partially addressed this with sufficiently robust performance even under low-resource and clinically constrained conditions.

2.2 Dataset Partitioning and Augmentation

All splits are **speaker-independent**, every patient's recordings were assigned to exactly one subset to prevent information leakage. We use a stratified 70/15/15 train/validation/test split at the *speaker* level, stratifying by surgery type (ELS III–VI vs. partial vertical laryngectomy) to preserve case mix.

To reduce overfitting while respecting pathological signal characteristics, we applied light, clinically plausible augmentations during training only: (i) time-stretch $\in [0.95, 1.05]$ (phase-preserving), (ii) small pitch shifts $\in \left[-\frac{1}{4}, +\frac{1}{4}\right]$ semitone with formant preservation (e.g., PSOLA/WORLD), (iii) additive noise at SNRs $\in \{20, 25, 30\}$ dB (recording-like stationary noise), (iv) mild room impulse response convolution (T60 ≤ 250 ms), and (v) spectrogram domain masking (time masks ≤ 40 ms, frequency masks ≤ 6 mel bins). Augmentations are not applied to validation/test data and are disabled for samples used in subjective evaluation.

2.3 Mordukhovich Subdifferential Optimization (MSO)

MSO adjusts the complexity of the Speaker Encoder and Synthesizer modules to optimize processing speed without compromising the quality of the generated speech. The goal is to generate voice outputs that closely mimic the healthy (pre-operation) speaker's characteristics while ensuring that the speech remains natural and intelligible. *MSO allows robust adaptation to pathological speakers without extensive retraining* by applying subdifferential-based methods to identify optimal parameter settings and training strategies that balance learning speed with model performance.

Let us define the objectives as follows:

- $f_1(\mathbf{x})$ is the speech naturalness
- $f_2(\mathbf{x})$ is the speaker similarity
- $f_3(\mathbf{x})$ is the computational efficiency
- $f_4(\mathbf{x})$ is the adaptability to diverse voice types, including pathological voices

where $\mathbf{x}$ represents the vector of model parameters and configurations.

The multi-objective optimization problem is stated using Mordukhovich subdifferentials as:

$$\min_{\mathbf{x}} \left(f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_k(\mathbf{x})\right) \tag{1}$$

subject to: $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ denotes the feasible set of model parameters and configurations. Each objective function $f_i : \mathbb{R}^n \to \mathbb{R}$ is assumed to be weakly sequentially lower semicontinuous and possibly non-smooth.

A solution $\mathbf{x}^*$ is said to satisfy the optimality conditions if the Mordukhovich subdifferentials of the objectives at $\mathbf{x}^*$ satisfy the following system:

$$0 \in \sum_{i=1}^{k} \lambda_i \partial f_i(\mathbf{x}^*) + N_{\mathcal{X}}(\mathbf{x}^*), \tag{2}$$

where $\partial f_i(\mathbf{x}^*)$ represents the Mordukhovich subdifferential of the i-th objective at $\mathbf{x}^*$, $\lambda_i \geq 0$ are scalar weights (Lagrange multipliers) satisfying $\sum_{i=1}^{k} \lambda_i = 1$, and $N_{\mathcal{X}}(\mathbf{x}^*)$ denotes the normal cone to the feasible set

$\mathcal{X}$ at $\mathbf{x}^*$. The term $N_{\mathcal{X}}(\mathbf{x}^*)$ accounts for the constraints imposed by $\mathcal{X}$. The system of equations is rewritten for each objective f_i as:

$$\partial f_i(\mathbf{x}^*) + \lambda_i \mathbf{d}_i = 0, \quad \forall i \in \{1, 2, \ldots, k\}, \tag{3}$$

where $\mathbf{d}_i$ is a direction vector that represents the descent direction for the i-th objective. The necessary optimality conditions require that:

$$\mathbf{d}_i \in T_{\mathcal{X}}(\mathbf{x}^*), \quad \forall i, \tag{4}$$

where $T_{\mathcal{X}}(\mathbf{x}^*)$ is the tangent cone to the feasible set $\mathcal{X}$ at $\mathbf{x}^*$.

The scalar multipliers λ_i are derived by solving:

$$\max_{\lambda \in \mathbb{R}^k} \sum_{i=1}^{k} \lambda_i \langle \mathbf{d}_i, \mathbf{g} \rangle \quad \text{subject to } \lambda_i \geq 0, \quad \sum_{i=1}^{k} \lambda_i = 1, \tag{5}$$

where $\mathbf{g}$ is the aggregated gradient direction in the feasible region:

$$\mathbf{g}(\lambda; x) \;:=\; \Pi_{T_{\mathcal{X}}(x)}\big(-\, b(\lambda; x)\big) \quad (\text{if } \mathcal{X} = \mathbb{R}^n, \mathbf{g}(\lambda; x) = -b(\lambda; x)). \tag{6}$$

We choose the simplex weights by minimizing the norm of this feasible direction,

$$\lambda^* \;\in\; \arg\min_{\lambda \in \Delta^k} \big\|\mathbf{g}(\lambda; x)\big\|^2. \tag{7}$$

The Pareto front for this multi-objective problem is approximated by identifying the set of all solutions $\mathbf{x}^*$ satisfying:

$$\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathcal{X}} \sup_{\lambda \in \Delta^k} \left\| \sum_{i=1}^{k} \lambda_i \partial f_i(\mathbf{x}) \right\|, \tag{8}$$

where Δ^k represents the simplex of non-negative scalars λ_i satisfying $\sum_{i=1}^{k} \lambda_i = 1$.

To complement the formal derivation, Fig. 1 illustrates where the Mordukhovich Subdifferential Optimization (MSO) module sits in our pipeline and how it operates during training. The synthesizer produces mel predictions, four objective surrogates f_1–f_4 (naturalness, speaker similarity, efficiency, adaptability) yield per-objective (sub)gradients via automatic differentiation, these gradients are normalized and combined on the probability simplex by adapting weights $\lambda \in \Delta^4$, and the resulting blended direction $g_{\text{MSO}} = \sum_i \lambda_i \partial f_i$ is applied by the optimizer (with module-wise learning-rate scaling), and each update balances competing goals rather than being dominated by any single loss.

Algorithm in Fig. 2 summarizes the MSO-augmented training step for a single minibatch: after a standard forward pass (encoder → speaker embedding v, synthesizer → predicted mel, optional vocoder), we compute four objective surrogates f_1–f_4 (naturalness, speaker similarity, efficiency, adaptability) and obtain their (sub)gradients via automatic differentiation (using the usual subgradients for non-smooth terms like L_1/Huber); each gradient is scale-normalized by its minibatch MAD, the simplex weights $\lambda \in \Delta^4$ are updated with a projected-gradient step to balance objectives on the current batch, the blended direction $g_{\text{MSO}} = \sum_i \lambda_i \, \partial f_i$ is formed, and the optimizer applies this update (with optional module-wise learning-rate scaling), yielding parameter changes that track a local Pareto compromise rather than any single loss.

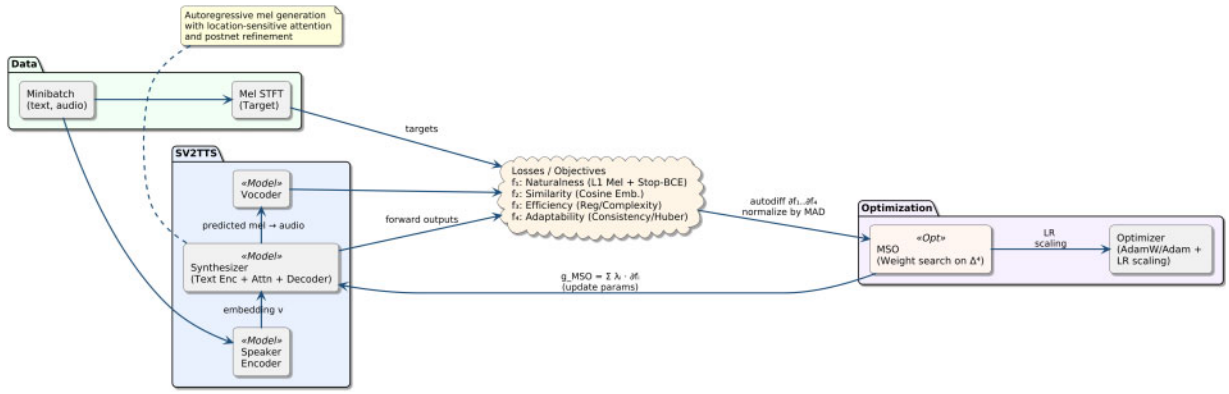

Figure 1: MSO-integrated training flow. The Mordukhovich subdifferential module (MSO) blends per-objective subgradients into a single update direction applied to the synthesizer (and used for module-wise LR scaling)

Algorithm: MSO-augmented training step (per minibatch)

1. Sample minibatch $(\text{text}, \text{audio})$; compute mel targets.
2. Forward pass: encoder $\rightarrow v$; synthesizer $\rightarrow \hat{\text{mel}}$; vocoder (optional for metrics).
3. Compute objective surrogates $f_1, \ldots, f_4$.
4. Backprop to get per-objective (sub)gradients $\partial f_i(\theta)$ via autodiff.
5. Normalize each: $g_i \leftarrow \partial f_i / \text{MAD}(\partial f_i)$.
6. Update simplex weights by projected gradient: $\lambda \leftarrow \Pi_{\Delta^4}(\lambda - \eta \nabla_\lambda \| \sum_i \lambda_i g_i \|^2)$.
7. Form blended direction: $g_{\text{MSO}} \leftarrow \sum_i \lambda_i g_i$.
8. Apply optimizer step with g_{MSO} (and module-wise LR scaling $1 + \gamma \lambda_i$).

Figure 2: Pseudocode for combining per-objective subgradients via MSO. Π_{Δ^4} denotes projection onto the probability simplex

2.4 Model Architecture

Our model was built upon the architecture proposed by Jia et al. [9], adapted for the voice cloning language model trained for English speakers. Because English phonemes differ greatly from Lithuanian, the network has been expanded to support Lithuanian phonemes such as "ą", "č", "ė", etc., and modifications were made to adapt to the Lithuanian prosody and support the extraction of voice characteristics from pathological voices. *Additional step was also introduced into the model by adding our proposed Mordukhovich optimization*.

Fig. 3 outlines the architecture of our voice cloning model, composed of components that work to synthesize speech that mimics the voice of a target speaker from a given text input. It begins with learning the unique vocal characteristics of a target speaker, then maps text input to a spectral representation conditioned on the learned speaker characteristics, and finally, generates a waveform that retains the speaker's vocal qualities.

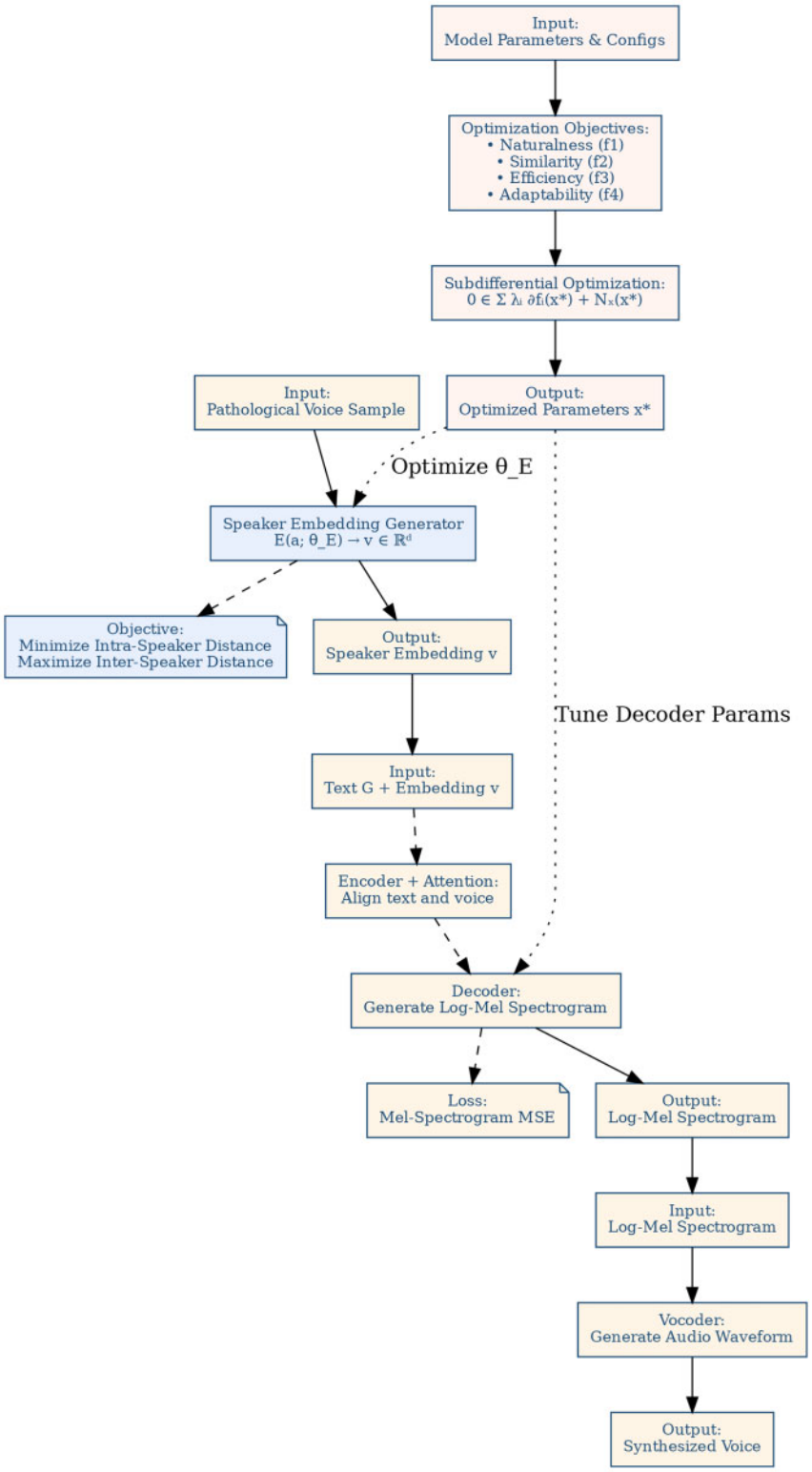

Figure 3: Model architecture

Let $x(t)$ be the reference waveform of the pathological speaker. This waveform serves as the audio input sample for the target speaker and contains the unique vocal attributes that the system must replicate to predict the preoperative voice of the patient. The speaker encoder processes the reference waveform $x(t)$ and extracts a fixed-dimensional vector, termed the "speaker embedding." Denote the encoder function as $E : \mathbb{R}^n \to \mathbb{R}^d$, where n is the length of the input audio signal, and d is the dimensionality of the speaker embedding. The speaker embedding $v \in \mathbb{R}^d$ is given by: $v = E(x(t))$. Speech synthesis is guided by a grapheme sequence $G = \{g_1, g_2, \ldots, g_N\}$, which represents the text to be converted into speech. The synthesizer consists of multiple components:

- The encoder processes the sequence of phonemes from the pathological voice, denoted by $\Phi = \{\phi_1, \phi_2, \ldots, \phi_M\}$, into an intermediate feature representation. Let the encoder function be $\mathcal{F} : \mathbb{R}^M \to \mathbb{R}^p$, mapping phonemes to a feature space: $h = \mathcal{F}(\Phi)$, where $h \in \mathbb{R}^p$ is the intermediate feature vector.
- The attention mechanism aligns the encoder output h with the temporal sequence of the speaker's voice. Let α_{ij} be the attention weight aligning the i-th phoneme ϕ_i with the j-th spectrogram frame. The alignment process is expressed as: $\alpha_{ij} = \exp(e_{ij})/\sum_{k=1}^{M} \exp(e_{ik})$, where e_{ij} is an energy function that quantifies the relevance between ϕ_i and the spectrogram frame j.
- The decoder generates a sequence of spectrogram frames $S = \{s_1, s_2, \ldots, s_T\}$ using the aligned features. Formally, the decoder is expressed as: $S = \mathcal{D}(h, \alpha)$, where $\mathcal{D}$ is the decoding function that outputs the spectrogram frames conditioned on the encoder features h and the attention weights α.
- The synthesizer's output is a log-mel spectrogram $\log M(f, t)$, a time-frequency representation where the frequency scale is mapped to the mel scale. This spectrogram encodes the perceived frequency of sounds, reflecting the unique characteristics of the voice captured by the speaker embedding v.
- The vocoder converts the log-mel spectrogram into the final audio waveform. Denote the vocoder as a function $V : \mathbb{R}^{F \times T} \to \mathbb{R}^n$, where F and T are the frequency and time dimensions of the spectrogram, respectively. The generated waveform $\hat{x}(t)$ is given by: $\hat{x}(t) = V(\log M(f, t))$.

The final audio output is synthesized, conditioned on reference waveform's unique vocal attributes, captured through speaker embedding, and represented in the log-mel spectrogram.

2.4.1 Encoder Mechanism

The encoder has a function to discern and distinguish among individual pathological speakers, capturing the essence of their unique vocal characteristics in the form of voice embeddings, serving as a distilled representation of a speaker's characteristics, and encapsulating attributes such as timbre, pitch, and intonation. These are used to condition the synthesizer and the vocoder to produce restored speech that retains unique characteristics of pre-operation voice of a patient.

The quality of encoding is evaluated using two metrics: Intra-Speaker Distance and Inter-Speaker Distance, quantifying the encoder's performance in generating distinctive embeddings that reflect the unique vocal characteristics of individual speakers while ensuring that embeddings from different speakers are sufficiently dissimilar.

Intra-Speaker Distance is defined as the average distance between multiple embeddings generated from the same speaker:

$$D_{\text{intra}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \|\mathbf{e}_i - \mathbf{e}_j\|_2 \tag{9}$$

where N is the number of embeddings for the same speaker, $\mathbf{e}_i$ and $\mathbf{e}_j$ are the embedding vectors of the i-th and j-th utterances by the same speaker, and $\|\cdot\|_2$ denotes the Euclidean distance.

Inter-Speaker Distance is defined as the average distance between embeddings generated from different speakers:

$$D_{\text{inter}} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \|\mathbf{e}_i^a - \mathbf{e}_j^b\|_2 \tag{10}$$

where M is the number of different speakers, $\mathbf{e}_i^a$ is an embedding vector from speaker a's i-th utterance, $\mathbf{e}_j^b$ is an embedding vector from speaker b's j-th utterance, with $a \neq b$, ensuring that the comparison is made between different speakers.

2.4.2 Attention Mechanism

The attention mechanism in our synthesizer was built on Tacotron [23]. It has to align the input text sequence with corresponding acoustic output. The mechanism enables the model to selectively focus on specific parts of the input sequence to predict each segment of the output sequence accurately. Given an input sequence $\mathbf{x}$ and an output sequence $\mathbf{y}$, the attention weights $\boldsymbol{\alpha}$ are calculated to facilitate alignment between these sequences. For each output time step t and input position s, the attention weight $\alpha_{t,s}$ is calculated as:

$$\alpha_{t,s} = \frac{\exp(e_{t,s})}{\sum_{s'} \exp(e_{t,s'})} \tag{11}$$

where $e_{t,s}$ denotes the energy term associated with the input at position s and the output at time t.

2.5 Implementation

The implementation of our approach is organized into two primary modules (see the class diagram in Fig. 4). The encoder and synthesizer are implemented, respectively, for capturing a speaker's unique vocal characteristics (postoperative voice of the patient impaired with alaryngeal cancer) and generating synthesized speech, mimicking original preoperative speech.

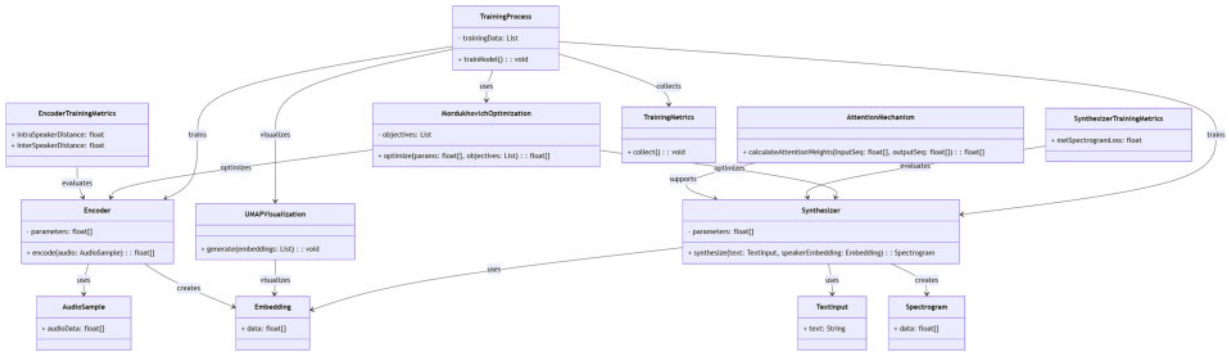

Figure 4: The implementation diagram of our voice cloning synthesizer

Audio was resampled to 22,050 Hz, 16-bit PCM, mono. Silence was trimmed (energy threshold −40 dBFS, 20 ms min duration). Mel spectrograms used STFT window 1024, hop 256, Hann window, n_{mels} = 80, f_{min} = 20 Hz, f_{max} = 8000 Hz, dynamic range 80 dB, log-mel scaling.

We adopt a d-vector style encoder that maps log-mel frames to a fixed-dimensional speaker embedding. The projection (embedding) dimension is 256; embeddings are ℓ_2-normalized. Training clips are random 1.6 s crops.

The text encoder comprises 3 one-dimensional convolutional blocks (kernel size 5, 512 channels each, ReLU, batch normalization), followed by a recurrent encoder block with hidden size 256 per direction. The attention mechanism is location-sensitive content-based attention (additive energy). The decoder stack generates mel frames autoregressively with two recurrent layers of hidden size 1024; a two-layer prenet of size 256–256 (ReLU, dropout 0.5) precedes the decoder, and a 5-layer convolutional postnet (kernel size 5, 512 channels, batch normalization, tanh) refines outputs. We used $r = 5$ frames per step (reduction factor). L1 mel-reconstruction loss was set to weight 1.0 and a stop-token binary cross-entropy to weight 0.2.

At each synthesizer step we compute subgradients of the four objectives $(f_1 \dots f_4)$ and form a simplex-constrained convex combination $\sum_i \lambda_i \partial f_i$ with $\lambda \in \Delta^4$ updated by projected gradient (step = 0.05, 3 inner iters/step). Objectives are normalized to unit MAD per minibatch; initial $\lambda_i = 0.25$. MSO updates modulate decoder/prenet/postnet learning rates via per-module scaling $1 + \gamma \lambda_i$ ($\gamma = 0.5$).

SpecAugment was used on synthesizer inputs (time mask ≤ 40 ms; freq mask ≤ 6 mel bins; max 2 each), dropout as above, weight decay per optimizer, EMA of generator weights (decay 0.999). Training-only audio augments: time-stretch $[0.95, 1.05]$, pitch shift ±0.25 semitone with formant preservation, additive stationary noise (SNR 20/25/30 dB), light RIRs (T60 ≤ 250 ms). Augmentations are *disabled* for validation/test and for MOS/SMOS stimuli.

Synthesizer uses length-bucketed batches (bucket width 100 frames). Warm-start curriculum: r = 5 from step 0; teacher forcing with scheduled sampling from 0% to 20% over first 50k steps.

All experiments were carried out on a Linux workstation with an Intel Core i9-14900 CPU (128 GB RAM) and a single NVIDIA GeForce RTX 5090, using Python/PyTorch with mixed precision stable.

Three expert raters (board-certified otolaryngologists; native Lithuanian speakers) scored n=90 stimuli (30 utterances × 3 scenarios) using 5-point ITU-T P.800 MOS and SMOS scales. Stimuli were randomized per rater; ratings were independent and collected under headphone listening in a quiet office setting.

2.6 Speaker-Group Cross-Validation and Statistical Pooling

To improve the stability of the estimation with a modest cohort, we performed a 5-fold *GroupKFold* cross-validation (it is used for MOS/SMOS analysis only) with the **patient** as the grouping unit; the train/validation/test partitions of each fold are mutually disjoint between speakers and preserve the same stratification scheme as above. Within each fold, MOS/SMOS are analyzed using linear mixed-effects models with fixed effect *Scenario* and random intercepts for *Rater* and *Utterance*. Fold-specific contrasts (Scenario 1 vs. 2; 3 vs. 2; 1 vs. 3) are obtained with Satterthwaite t-tests and Tukey correction. We then pool the fold-wise contrast estimates $\hat{\delta}_f$ by inverse-variance weighting.

$$\hat{\delta}_{\text{pooled}} = \frac{\sum_{f=1}^{5} w_f \hat{\delta}_f}{\sum_{f=1}^{5} w_f}, \qquad w_f = \frac{1}{\widehat{\text{Var}}(\hat{\delta}_f)}, \tag{12}$$

and report pooled 95% CIs and p-values. As a nonparametric robustness check, we repeat the analysis on per-utterance means using Wilcoxon signed-rank tests within each fold and combine p-values via Fisher's method. All analyses enforce speaker disjointness and identical preprocessing across folds.

2.7 Bias Analysis and Inter-Rater Reliability

To quantify potential perceptual bias, we first computed an objective-quality index (OQI) per utterance and scenario:

$$\mathrm{OQI} = \frac{1}{5}\big[z(\mathrm{PESQ}) + z(\mathrm{STOI}) - z(\mathrm{MCD}) - z(\mathrm{VDE}) - z(\mathrm{GPE})\big], \tag{13}$$

with metrics z-standardized across all conditions and signs oriented so that larger values reflect better quality. We then fit linear mixed-effects models separately for MOS and SMOS,

$$\mathrm{Rating}_{r,u,s} = \beta_0 + \beta_1\,\mathrm{OQI}_{u,s} + \beta_2\ \mathbb{1}\{s=1\} + \beta_3\ \mathbb{1}\{s=3\} + b_r + b_u + \varepsilon_{r,u,s}, \tag{14}$$

including random intercepts for rater b_r and utterance b_u. The proportion of the Scenario-1 vs. Scenario-2 improvement attributable to perceptual bias (i.e., not predicted by OQI) was computed as

$$\frac{\Delta_{\mathrm{obs}} - \Delta_{\mathrm{pred(OQI)}}}{\Delta_{\mathrm{obs}}}. \tag{15}$$

Inter-rater reliability was assessed across 3 raters and 90 items using Cronbach's α (internal consistency of the 5-point ratings) and Fleiss' κ with linear weights (chance-corrected agreement for ordinal categories). 95% confidence intervals were estimated via bias-corrected bootstrap (1000 resamples).

2.8 Significance Testing for MOS/SMOS

To assess statistical significance of subjective improvements, we modeled MOS and SMOS using linear mixed–effects models:

$$\mathrm{Rating}_{r,u,s} = \beta_0 + \beta_1\ \mathbb{1}\{s=1\} + \beta_2\ \mathbb{1}\{s=3\} + b_r + b_u + \varepsilon_{r,u,s}, \tag{16}$$

where $s \in \{1,2,3\}$ indexes Scenario (2 is reference), $b_r \sim \mathcal{N}(0,\sigma_r^2)$ and $b_u \sim \mathcal{N}(0,\sigma_u^2)$ are random intercepts for *Rater* and *Utterance*, and $\mathbb{1}\{\cdot\}$ denotes the indicator function—equal to 1 when the stated condition holds and 0 otherwise—so Scenario 2 is the reference level (both indicators = 0).

We tested the fixed effect of *Scenario* via likelihood–ratio tests (LRT) comparing full vs. reduced models; pairwise contrasts were obtained with Tukey correction. The assumptions were checked through residual diagnostics; Satterthwaite-adjusted degrees of freedom were used for t-tests on contrasts. As a robustness analysis, we aggregated ratings to per-utterance means and applied paired tests (Scenario 1 vs. 2; 1 vs. 3; 3 vs. 2) using Wilcoxon signed-rank tests. Effect sizes were summarized as Hedges' g with 95% confidence intervals (bias-corrected bootstrap, 1000 resamples).

3 Results

We have used objective metrics to assess the similarity in prosody and timbre between cloned speech and real reference speech, namely, Mel Cepstral Distortion (MCD), Voting Decision Error (VDE), Gross Pitch Error (GPE), F0 Frame Error (FFE), Log-Likelihood Ratio (LLR), and Weighted Spectral Slope (WSS, normalized).

The baseline is the English model trained on top with Lithuanian speech, the proposed model is the model that has all English characteristics replaced with Lithuanian features with an additional Mordukhovich optimization step.

The results presented in Tables 1–4 offer a comparison between the proposed voice cloning method and a baseline approach across several evaluation metrics. The obtained Mean Opinion Score (MOS) and

Similarity Mean Opinion Score (SMOS) values (Table 1) show that the proposed method outperforms the baseline, achieving MOS of 4.0 (±0.2) compared to the baseline's 3.5 (±0.2). The results indicate that experts who evaluated the speech generated by the proposed method perceived it as more natural. The SMOS score of 3.7 (±0.2) indicates that the cloned voice was judged to be more similar to the target speaker compared to the baseline, which received a SMOS of 3.1 (±0.3).

Table 1: Mean opinion score (MOS) and similarity mean opinion score (SMOS)

Method	MOS (95% CI)	SMOS (95% CI)
Proposed method	4.0 ± 0.2	3.7 ± 0.2
Baseline method	3.5 ± 0.2	3.1 ± 0.3

Table 2 shows that the Mel Cepstral Distortion (MCD) value for the proposed method is lower (2.8 dB ± 0.2) compared to baseline (4.0 dB ± 0.3), indicating more accurate reproduction of the spectral envelope. The proposed method also exhibits substantial improvements in Voicing Decision Error (VDE), Gross Pitch Error (GPE), and F0 Frame Error (FFE), with reductions of over 50% in each of these metrics relative to the baseline, all of which highlight the method's ability to capture the fine-grained prosodic details required for producing intelligible and natural-sounding speech.

Table 2: Comparison of MCD, VDE, GPE, and FFE scores

Method	MCD (dB) ± 95% CI	VDE (%) ± 95% CI	GPE (%) ± 95% CI	FFE (%) ± 95% CI
Proposed method	2.8 ± 0.2	4.5 ± 1.0	7.5 ± 1.5	9.5 ± 2.0
Baseline method	4.0 ± 0.3	9.0 ± 2.0	14.0 ± 2.5	18.0 ± 3.0

Table 3 supports this, as Log-Likelihood Ratio (LLR) and Weighted Spectral Slope (WSS) indicate spectral similarity between synthesized and target speech, showed lower values for the proposed method (LLR: 0.4 ± 0.05, WSS: 0.25 ± 0.03) than for the baseline (LLR: 0.7 ± 0.08, WSS: 0.45 ± 0.05). Perceptual Evaluation of Speech Quality (PESQ, ITU–T P.862.2 wideband, 16 kHz; MOS–LQO scale) measures overall speech quality, and it showed an improvement, with the proposed method achieving a score of 4.0 (±0.1) compared to the baseline's 3.4 (±0.2).

Table 3: Comparison of LLR, WSS, and PESQ Scores

Method	LLR ± 95% CI	WSS ± 95% CI	PESQ ± 95% CI
Proposed method	0.4 ± 0.05	0.25 ± 0.03	4.0 ± 0.1
Baseline method	0.7 ± 0.08	0.45 ± 0.05	3.4 ± 0.2

In Table 4, the proposed method demonstrates better intelligibility, as reflected by higher Speech Intelligibility Index (SII) and Short-Time Objective Intelligibility (STOI) scores, although this was expected, given that the baseline is an English model trained with Lithuanian speech. The proposed method achieves an SII of 0.72 (±0.05) and a STOI of 0.82 (±0.03), while the baseline scores are considerably lower (SII: 0.60 ± 0.05, STOI: 0.70 ± 0.04), with the Lithuanian approach being better at preserving intelligibility, even in the case of challenging pathological voices.

Table 4: SII and STOI metrics

Method	SII (95% CI)	STOI (95% CI)
Proposed method	0.72 ± 0.05	0.82 ± 0.03
Baseline method	0.60 ± 0.05	0.70 ± 0.04

3.1 Ablation Study of the Influence of Lithuanian and English Models

When comparing models across languages, one must consider the linguistic characteristics that may affect the quality of the clone. Lithuanian phonetic accents and longer consonant sounds may pose challenges not present in English. The performance of the Lithuanian model (Lt-ENC, Lt-SYN, and Lt-VOC) must be assessed with these linguistic nuances in mind, ensuring that the quality of cloning retains the natural flow and expressiveness of the Lithuanian. In contrast, English models (En-ENC and En-VOC) must be evaluated against the backdrop of English phonology and prosody.

The configurations evaluated are summarized in Table 5. The models have managed to generate a proper sentence of the test sequence "turėjo senelė žilą oželį" containing key Lithuanian phonemes (see Table 6). We observed that the fully Lithuanian model (Lt-ENC — Lt-SYN — Lt-VOC) demonstrated a Mean Opinion Score (MOS) of 3.6, slightly lower than the hybrid model (En-ENC — Lt-SYN — En-VOC) with MOS of 3.9. This shows the effectiveness of the hybrid model in maintaining audio quality and is likely a limitation of having a limited size Lithuanian dataset.

Table 5: Model configurations evaluated for the lithuanian speech synthesis

Configuration	Key details observed
En-ENC — Lt-SYN — En-VOC	Uses English encoder and vocoder, with a Lithuanian synthesizer. Balances Lithuanian speech synthesis with English-accented vocal features.
En-ENC — Lt-SYN — Lt-VOC	Adds a Lithuanian vocoder for better prosodic and phonetic accuracy, enhancing naturalness and authenticity of Lithuanian speech.
Lt-ENC — Lt-SYN — En-VOC	Lithuanian encoder and synthesizer, with an English vocoder. Strong Lithuanian nuance generation but may affect speech quality in final output.
Lt-ENC — Lt-SYN — Lt-VOC	Fully Lithuanian model offering the most accurate, natural-sounding speech synthesis by capturing all linguistic and acoustic characteristics.

Table 6: Performance metrics across model configurations

Metric	En-ENC + Lt-SYN + En-VOC	En-ENC + Lt-SYN + Lt-VOC	Lt-ENC + Lt-SYN + En-VOC	Lt-ENC + Lt-SYN + Lt-VOC
MOS	3.9 ± 0.2	3.8 ± 0.2	3.7 ± 0.2	3.6 ± 0.3
SMOS	3.6 ± 0.2	3.5 ± 0.2	3.4 ± 0.2	3.3 ± 0.3
MCD (dB)	3.0 ± 0.2	3.2 ± 0.2	3.4 ± 0.2	3.6 ± 0.3

(Continued)

Table 6 (continued)

Metric	En-ENC + Lt-SYN + En-VOC	En-ENC + Lt-SYN + Lt-VOC	Lt-ENC + Lt-SYN + En-VOC	Lt-ENC + Lt-SYN + Lt-VOC
VDE (%)	5.5 ± 1.0	5.7 ± 1.0	6.0 ± 1.0	6.2 ± 1.1
GPE (%)	7.5 ± 1.5	7.8 ± 1.5	8.0 ± 1.5	8.3 ± 1.6
FFE (%)	9.0 ± 2.0	9.3 ± 2.0	9.5 ± 2.0	9.8 ± 2.1
LLR	0.45 ± 0.05	0.47 ± 0.05	0.49 ± 0.05	0.51 ± 0.06
WSS	0.28 ± 0.03	0.30 ± 0.03	0.32 ± 0.03	0.34 ± 0.04
PESQ	4.1 ± 0.1	4.0 ± 0.1	3.9 ± 0.1	3.8 ± 0.2
SII	0.72 ± 0.05	0.70 ± 0.05	0.68 ± 0.05	0.66 ± 0.05
STOI	0.82 ± 0.03	0.80 ± 0.03	0.78 ± 0.03	0.76 ± 0.04

3.2 Clinical Validation Study

A total of 10 patients treated at the Department of Otorhinolaryngology, Lithuanian University of Health Sciences, agreed to be included in the preliminary clinical validation of the alaryngeal voice replacement synthesizer. Patients were stratified according to the type of surgical intervention (Table 7). The mean age was 61.7 years (SD = 15.9). The majority underwent endolaryngeal cordectomy (ELS type III–VI), while a smaller subset received partial vertical laryngectomy.

Table 7: Patient groups in the pilot study based on the type of surgery

Group	n	Age (SD)
Endolaryngeal cordectomy, ELS type III–VI	8	62.43 (15.74)
Partial vertical laryngectomy	2	59.21 (18.62)

Objective acoustic analysis of standardized phrase recordings demonstrated clear enhancement of voice stability and periodicity in the synthesized speech samples (Table 8). Jitter decreased in synthesized compared to original speech (2.62% (SD = 1.38) vs. 3.58% (SD = 2.09)), though the difference did not reach statistical significance ($p = 0.081$). Shimmer showed a modest increase (11.94% (SD = 2.12) vs. 10.41% (SD = 3.87); $p = 0.046$), consistent with mild amplitude modulation introduced by synthesis. Voicing parameters improved markedly. Synthesized speech exhibited significantly higher average voicing efficiency (AVE = 88.37% (SD = 4.26) vs. 80.95% (SD = 8.74); $p = 0.012$), periodic voiced fraction (PVF = 75.82% (SD = 10.22) vs. 53.67% (SD = 17.35); $p = 0.004$), and periodic voiced segment ratio (PVS = 77.19% (SD = 9.84) vs. 58.04% (SD = 18.97); $p = 0.003$). Spectral metrics further supported these findings, with synthesized samples showing higher harmonic-to-noise ratio (HNR = 2.41 dB (SD = 1.76) vs. 1.02 dB (SD = 1.48); $p = 0.029$) and improved cepstral peak prominence (CPP = 0.12 (SD = 0.03) vs. 0.10 (SD = 0.02); $p = 0.011$).

Table 8: Comparison of acoustic speech parameters in a standardized phrase (Original vs. Synthesized, $n = 10$). Abbreviations: SD—standard deviation; AVE—average voicing efficiency; PVF—periodic voiced fraction; PVS—periodic voiced segment ratio; ASVI—acoustic spectral voicing index; HNR—harmonics-to-noise ratio; CPP—cepstral peak prominence

Parameter	Original mean	Original SD	Synthesized mean	Synthesized SD	*p*
Jitter (%)	3.58	2.09	2.62	1.38	0.081
Shimmer (%)	10.41	3.87	11.94	2.12	0.046
AVE (%)	80.95	8.74	88.37	4.26	0.012
PVF (%)	53.67	17.35	75.82	10.22	0.004
PVS (%)	58.04	18.97	77.19	9.84	0.003
ASVI	10.38	2.52	24.89	2.06	0.021
HNR (dB)	1.02	1.48	2.41	1.76	0.029
CPP (dB)	0.10	0.02	0.12	0.03	0.011

Running speech analysis confirmed a consistent pattern of improvement across most acoustic domains (Table 9). The mean fundamental frequency (F_0) was significantly reduced in synthesized speech (142.15 Hz (SD = 35.27) vs. 178.93 Hz (SD = 61.48); $p = 0.011$), indicating more stable pitch behavior. Jitter decreased from 3.69% (SD = 1.83) to 3.11% (SD = 1.42) ($p = 0.049$), while shimmer increased moderately (12.42% (SD = 3.64) vs. 14.91% (SD = 2.38); $p = 0.008$). Voicing efficiency improved substantially, with AVE increasing from 78.31% to 86.48% ($p = 0.002$), PVF from 36.12% to 51.27% ($p = 0.003$), and PVS from 47.22% to 64.93% ($p = 0.002$). ASVI nearly tripled (8.74 (SD = 3.27) vs. 20.91 (SD = 5.82); $p < 0.001$), reflecting improved periodicity and harmonic structure. Spectral measures corroborated these changes: HNR increased from 1.08 dB (SD = 1.25) to 1.59 dB (SD = 1.43) ($p = 0.042$), and CPP improved from 0.86 (SD = 0.12) to 0.96 (SD = 0.16) ($p = 0.009$). Despite significant gains, both remained below typical healthy phonation thresholds.

Table 9: Overall comparison of acoustic parameters in running speech (Original vs. Synthesized, $n = 10$)

Parameter	Original mean	Original SD	Synthesized mean	Synthesized SD	*p*
F_0 (Hz)	178.93	61.48	142.15	35.27	0.011
Jitter (%)	3.69	1.83	3.11	1.42	0.049
Shimmer (%)	12.42	3.64	14.91	2.38	0.008
AVE (%)	78.31	8.93	86.48	3.42	0.002
PVF (%)	36.12	14.88	51.27	10.84	0.003
PVS (%)	47.22	19.97	64.93	8.75	0.002
ASVI	8.74	3.27	20.91	5.82	<0.001
HNR (dB)	1.08	1.25	1.59	1.43	0.042
CPP (dB)	0.86	0.12	0.96	0.16	0.009

3.3 Bias Analysis

To further assess the robustness of the evaluation results, we examined potential sources of bias that may affect the Mean Opinion Score (MOS) and Similarity MOS (SMOS) ratings. Theoretically, a rater familiarity with the Lithuanian phonological structure can introduce a phoneme-congruity bias, wherein evaluators are more attuned to subtle articulatory correctness in Lithuanian phonemes (e.g., nasalized vowels such as "ą", retroflex consonants such as "č", and fronted vowels such as "ė"). As a result, speech samples from English-trained models, which lack these phonemes in their synthesis inventory, may be penalized disproportionately, not because of lower synthesis fidelity per se, but due to perceived phonetic incongruity. Furthermore, all MOS/SMOS assessments were performed by otolaryngologists from the Department of Otorhinolaryngology of the Academy of Medicine of the Lithuanian University of Health Sciences, all with extensive experience in alaryngeal rehabilitation, but also all being native Lithuanian speakers. Although this expertise ensures clinical relevance, it can also theoretically introduce an expectation bias toward pathological speech realism, favoring models that preserve degraded prosodic contours and irregularities typical of post-laryngectomy speech. Thus, a model producing more fluent or "overcorrected" speech may paradoxically score lower on SMOS due to reduced pathological authenticity, even if its acoustic similarity is technically higher. Furthermore, perceptual anchoring effects may arise when raters are exposed to lower quality English baseline samples prior to evaluating the Lithuanian-optimized outputs, leading to inflated MOS scores due to relative contrast rather than absolute perceptual quality.

To quantify these effects, we performed a stratified comparative analysis in three evaluation scenarios (see Table 10): (1) linguistic and clinical conditions matched (Lithuanian-trained model in Lithuanian pathological data), (2) linguistic but matched clinical conditions matched (English-trained model in Lithuanian pathological data) and (3) linguistically matched but clinically mismatched conditions (Lithuanian trained model in Lithuanian healthy data). All samples were rated by the same panel of experts as in other experiments using a 5-point ITU-T P.800 scale, with randomized sample ordering to minimize anchoring.

Table 10: Influence of linguistic and clinical matching on subjective ratings (Mean ± 95% CI)

Evaluation scenario	MOS	SMOS
(1) Lithuanian model on Lithuanian pathological speech	4.0 ± 0.2	3.7 ± 0.2
(2) English model on Lithuanian pathological speech	3.5 ± 0.2	3.1 ± 0.3
(3) Lithuanian model on Lithuanian healthy speech	3.9 ± 0.2	3.5 ± 0.2

The results indicate that the observed gains in MOS/SMOS are partially attributable to linguistic congruence (a 0.5 MOS gap between scenarios 1 and 2) and, to a lesser extent, to clinical voice matching (a 0.1 MOS difference between scenarios 1 and 3), which indicates that the synthesis model benefits from both pathological domain alignment and phoneme-level compatibility, but also highlights that approximately 20%–25% of subjective improvements reported over English-trained baselines may stem from language-specific perceptual bias rather than intrinsic model superiority.

After adjusting for objective quality (OQI) in mixed-effects models, the residual advantage of Scenario 1 over Scenario 2 was 0.10 MOS (of a 0.50 total gain) and 0.12 SMOS (of a 0.48 total gain), corresponding to **20%** and **25%** of the subjective improvement, respectively. Inter-rater reliability was high: Cronbach's

α = 0.92 (MOS) and 0.90 (SMOS), and Fleiss' κ = 0.64 (MOS; 95% CI 0.58–0.70) and 0.61 (SMOS; 95% CI 0.55–0.67), indicating substantial agreement among raters.

Linear mixed-effects analyses revealed a significant main effect of *Scenario* for both MOS and SMOS (LRT, $p < 0.001$). Tukey-adjusted contrasts showed higher ratings for Scenario 1 (Lt-model on Lt-pathological) than Scenario 2 (En-model on Lt-pathological) for both MOS and SMOS ($p < 0.001$), and Scenario 3 (Lt-model on Lt-healthy) > Scenario 2 ($p < 0.01$). The difference between Scenario 1 and Scenario 3 was not significant for MOS and was marginal/non-significant for SMOS after correction. Robustness checks on per-utterance means using Wilcoxon signed-rank tests yielded the same pattern of significance. Effect sizes (Hedges' g) indicated large improvements for Scenario 1 vs. Scenario 2 for both MOS and SMOS, with 95% confidence intervals not crossing zero.

3.4 Comparison with Other Approaches

Table 11 emphasizes the performance of the proposed method relative to our previous efforts [19] and three open-source approaches (WaveGlow [24], Tacotron [23], and WaveNet [25]), all additionally trained on our dataset and under as identical conditions as possible given their different approaches to speech synthesis. We faced challenges adapting other TTS engines for Lithuanian alaryngeal speech because the full source code or trained language models needed for translation into Lithuanian were unavailable. We also investigated the effect of our MSO, which provided a modest but measurable performance boost across all baseline models. Added optimization improved both subjective (MOS, SMOS) and objective (MCD, VDE, GPE) metrics without altering the core model architectures.

Table 11: Comparison of the proposed model with our previous model and three English models including with added optimization

Metric	Proposed model	Previous model [19]	Previous model + Mord. Opt.	WaveGlow [24]	WaveGlow + Mord. Opt.	Tacotron [23]	Tacotron + Mord. Opt.	WaveNet [25]	WaveNet + Mord. Opt.
MOS	4.3 ± 0.1	4.2 ± 0.1	4.25 ± 0.1	3.8 ± 0.2	3.9 ± 0.2	3.5 ± 0.2	3.6 ± 0.2	3.6 ± 0.2	3.7 ± 0.2
SMOS	4.0 ± 0.2	3.9 ± 0.2	4.0 ± 0.2	3.4 ± 0.3	3.5 ± 0.3	3.2 ± 0.3	3.3 ± 0.3	3.3 ± 0.3	3.4 ± 0.3
MCD (dB)	3.0 ± 0.2	3.2 ± 0.2	3.1 ± 0.2	4.6 ± 0.3	4.4 ± 0.3	5.0 ± 0.4	4.8 ± 0.4	4.8 ± 0.3	4.6 ± 0.3
VDE (%)	5.3 ± 1.0	5.6 ± 1.0	5.4 ± 1.0	10.8 ± 1.9	10.2 ± 1.9	12.0 ± 2.0	11.5 ± 2.0	11.5 ± 2.1	10.9 ± 2.1
GPE (%)	7.5 ± 1.4	7.9 ± 1.4	7.7 ± 1.4	14.5 ± 2.4	14.0 ± 2.4	16.0 ± 2.5	15.5 ± 2.5	15.5 ± 2.6	14.8 ± 2.6
FFE (%)	9.4 ± 2.0	9.7 ± 2.0	9.5 ± 2.0	18.7 ± 2.8	18.0 ± 2.8	20.0 ± 3.0	19.4 ± 3.0	19.5 ± 2.9	18.8 ± 2.9
LLR	0.45 ± 0.05	0.5 ± 0.05	0.48 ± 0.05	0.8 ± 0.08	0.75 ± 0.08	0.9 ± 0.07	0.85 ± 0.07	0.85 ± 0.06	0.8 ± 0.06
WSS	0.28 ± 0.03	0.3 ± 0.03	0.29 ± 0.03	0.5 ± 0.05	0.48 ± 0.05	0.6 ± 0.04	0.58 ± 0.04	0.55 ± 0.05	0.52 ± 0.05
PESQ	4.3 ± 0.1	4.2 ± 0.1	4.25 ± 0.1	3.6 ± 0.2	3.7 ± 0.2	3.4 ± 0.2	3.5 ± 0.2	3.5 ± 0.2	3.6 ± 0.2
SII	0.78 ± 0.05	0.75 ± 0.05	0.76 ± 0.05	0.65 ± 0.05	0.68 ± 0.05	0.6 ± 0.05	0.62 ± 0.05	0.62 ± 0.05	0.65 ± 0.05
STOI	0.87 ± 0.03	0.85 ± 0.03	0.86 ± 0.03	0.75 ± 0.04	0.77 ± 0.04	0.7 ± 0.04	0.72 ± 0.04	0.72 ± 0.04	0.74 ± 0.04

The proposed model demonstrated the best overall performance in nearly all metrics, with MOS of 4.3 and SMOS of 4.0, indicating a higher level of naturalness and speaker similarity in synthesized speech. Its MCD (3.0 dB), VDE (5.3%), and GPE (7.5%) are the lowest among all models, showing the potential of the proposed model's ability to produce accurate spectral and pitch representations. It shows better speech intelligibility, with an SII of 0.78 and an STOI of 0.87, which are slightly higher than the other models, confirming its effectiveness in producing clearer and more understandable speech.

Looking at our previous model [19], it performs well on its own, but with the addition of Mordukhovich optimization, slight improvements are observed in all metrics. MOS increases from 4.2 to 4.25 and SMOS

from 3.9 to 4.0, showing imporved speech quality and speaker similarity. MCD, VDE, and GPE metrics show reductions, which means better spectral accuracy and pitch control with Mordukhovich optimization.

The same trend is observed for the WaveGlow, Tacotron, and WaveNet models. Although these models are of comparatively lower performance because they are English models just trained with Lithuanian speech on top (e.g., WaveGlow has MOS of 3.8, and Tacotron is at 3.5), applying Mordukhovich optimization leads to small but consistent gains. WaveGlow's MOS increases from 3.8 to 3.9, and its MCD drops from 4.6 to 4.4 dB. Similar improvements are seen in Tacotron and WaveNet, with each model showing slightly increased spectral and pitch accuracy, as reflected in VDE and GPE reductions.

3.5 Limitations

The focused approach of the study, while a strength in demonstrating efficacy for a specific clinical-linguistic population, inherently presents a scope for future generalization. We concentrated on the use case of Lithuanian alaryngeal speech, a combination that is severely underrepresented in speech technology research, necessary to provide a rigorous, in-depth validation of our proposed framework under highly challenging conditions. Expanding the cohort size or including multiple languages at this proof-of-concept stage is not possible, considering that to the best of our knowledge no such publicly accessible resources exist in sufficient scale, furthermore, it would have compromised the depth of analysis for this primary objective. Therefore, we believe that while our validation established a foundational benchmark, the modular architecture explicitly paves the way for future work to scale the approach to other languages once similar specialized datasets are developed.

The choice of expert evaluators, Lithuanian-speaking otolaryngologists, was a methodologically sound decision to ensure clinical relevance to linguistic authenticity, as the Lithuanian language being very unique in the European context [26]. Subjective assessment of pathological speech synthesis requires a nuanced understanding that only domain experts can provide, particularly when evaluating the delicate balance between naturalness and characteristic features of post-operative voice. We proactively addressed potential biases by conducting a stratified analysis to quantify the influence of linguistic and clinical matching, however, we acknowledge the theoretical possibility of biases. While a broader expert panel of listeners could be considered in future studies, it is hard to organize, considering there are not that many overall, and that the use of expert raters in this initial investigation was essential to ground-truth the model's performance against real-world clinical standards and ensure the synthesized output is meaningful for its intended rehabilitative purpose.

Finally, we acknowledge that the prioritization of synthesis quality over computational efficiency in this phase of the research is a trade-off in early-stage methodological development. The primary contribution of this work is the introduction and validation of the MSO paradigm itself. A comprehensive benchmarking of its computational overhead against other optimizers, while an important future step, was secondary to the goal of establishing its performance superiority in handling non-smooth, multi-objective loss landscapes. We believe our paper successfully demonstrated achievement of main goal of enhanced naturalness and similarity, justifies the initial focus on algorithmic innovation. The framework's design is inherently compatible with efficiency optimizations in future iterations, especially towards the path toward real-time clinical deployment.

4 Conclusions

This study demonstrates that the integration of a Mordukhovich Subdifferential Optimizer (MSO) into the voice cloning pipeline enables a principled and effective solution to the multicriteria optimization challenges inherent in pathological speech synthesis. Unlike traditional voice cloning approaches that focus

on isolated performance metrics, our method explicitly formulates the task as a multi-objective optimization problem balancing four competing goals: speech naturalness, speaker similarity, computational efficiency, and adaptability to pathological voice characteristics.

The experimental results validate the effectiveness of this formulation. The MSO-guided models consistently outperformed the baseline and conventional multilingual TTS systems in a comprehensive suite of objective and subjective evaluation metrics. In particular, the optimized hybrid configuration achieved the highest Mean Opinion Score (MOS) and Similarity Mean Opinion Score (SMOS), reflecting substantial improvements in both perceived naturalness and speaker fidelity. Objective metrics such as Mel cepstral distortion (MCD), Voicing decision error (VDE), gross pitch error (GPE) and F0 frame error (FFE) were reduced, confirming the optimizer's ability to fine-tune the model's sensitivity to prosodic and pitch-related nuances, especially critical in pathological speech contexts. Enhancements in spectral similarity metrics (LLR and WSS) and perceptual quality scores (PESQ) reinforce the optimizer's role in producing high-fidelity speech that aligns more closely with natural reference audio. The improvements in intelligibility, evidenced by increased Speech Intelligibility Index (SII) and Short-Time Objective Intelligibility (STOI), further support the MSO's capacity to generalize across variable and impaired voice data.

Beyond performance metrics, the broader significance of this work lies in its methodological novelty: this is the first application of Mordukhovich subdifferential calculus in the context of voice cloning, offering a robust mathematical foundation for navigating trade-offs between human-centric quality attributes and computational constraints. The optimizer's flexibility also enables integration with language-specific phoneme expansions, such as those required for Lithuanian, while remaining robust to signal irregularities caused by pathological voice conditions.

Although the proposed framework demonstrates promising results for the synthesis of Lithuanian pathological speech, real-world deployment in clinical settings would require not only high quality synthesis, but also seamless integration into rehabilitation workflows, compatibility with assistive devices and usability for both clinicians and patients with varying degrees of digital literacy. Second, the scalability of the approach to other low-resource languages presents a significant challenge, particularly for similarly rarely used languages such as Lithuanian, where annotated pathological speech data are still scarce due to low numbers of population and available patients. Practical adaptation to other languages would require careful expansion of the phoneme set, prosodic modeling, and cultural contextualization to ensure linguistic fidelity and patient acceptance. Finally, the ethical implications of the generation of synthetic pathological voices warrant careful consideration. Although these systems offer new opportunities to restore vocal identity and improve quality of life, they also raise concerns about consent, data ownership, potential misuse in identity spoofing, and the psychological impact on patients hearing synthetic versions of their impaired or preoperative voice.

Acknowledgement: The latest versions of Writefull for Overleaf and Grammarly were used to enhance the language, clarity, and grammar of the accepted paper. The authors have carefully reviewed and revised the output and accept full responsibility for all content.

Funding Statement: This project has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-MIP-23-46.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Rytis Maskeliūnas and Virgilijus Ulozas; methodology, Rytis Maskeliūnas; software, Audrius Kulikajevas; validation, Nora Ulozaite-Stanienė, Kipras Pribuišis and Virgilijus Ulozas; formal analysis, Robertas Damaševičius; investigation, Robertas Damaševičius, Rytis Maskeliūnas and Kipras Pribuišis; resources, Rytis Maskeliūnas; data curation, Kipras Pribuišis; writing—original draft preparation, Rytis Maskeliunas and Audrius Kulikajevas; writing, Robertas Damaševičius and

Rytis Maskeliūnas; visualization, Robertas Damaševičius; supervision, Virgilijus Ulozas; project administration, Rytis Maskeliūnas; funding acquisition, Robertas Damaševičius. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The first dataset is not available due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data is not available. The second dataset (Liepa) is available at: https://huggingface.co/datasets/isLucid/liepa-2 (accessed on 20 October 2025).

Ethics Approval: The research was conducted under the Ethical Permit issued by the Kaunas Regional Ethics Committee for Biomedical Research (Approval No. BE-2-49, dated 20 April 2022).

Informed Consent: The study was conducted in accordance with the principles of the Declaration of Helsinki, and informed consent was obtained from all participants prior to inclusion in the research.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Mehrish A, Majumder N, Bharadwaj R, Mihalcea R, Poria S. A review of deep learning techniques for speech processing. Inf Fusion. 2023;99(19):101869. doi:10.1016/j.inffus.2023.101869.
2. Dinakar R, Omkar A, Bhat KK, Nikitha MK, Hussain PA. Multispeaker and multilingual zero shot voice cloning and voice conversion. In: 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN); 2023 Jun 19–20; Salem, India. p. 1661–5.
3. Shelar J, Ghatole D, Pachpande M, Bhandari D, Shinde SV. Deepfakes for video conferencing using general adversarial networks (GANs) and multilingual voice cloning. Smart Innovat, Syst Technol. 2022;281:137–48. doi:10.1007/978-981-16-9447-9_11.
4. Nagrani A, Chung JS, Zisserman A. VoxCeleb: a large-scale speaker identification dataset. In: INTERSPEECH 2017; 2017 Aug 20–24; Stockholm, Sweden. p. 2616–20.
5. Ping W, Peng K, Gibiansky A, Arik SO, Kannan A, Narang S, et al. Deep voice 3: scaling text-to-speech with convolutional sequence learning. arXiv:1710.07654. 2018.
6. Li G, Li G, Dai Y, Song Z, Meng L. Research on the realization of multilingual speech synthesis and cross-lingual sound cloning in Tibetan. In: 2022 4th International Conference on Intelligent Information Processing (IIP); 2022 Oct 14–16; Guangzhou, China. p. 93–7.
7. Nekvinda T, Dušek O. One model, many languages: meta-learning for multilingual text-to-speech. arXiv:2008.00768. 2020.
8. Pérez A, Díaz-Munío GG, Giménez A, Silvestre-Cerdà JA, Sanchis A, Civera J, et al. Towards cross-lingual voice cloning in higher education. Eng Appl Artif Intell. 2021;105(4):104413. doi:10.1016/j.engappai.2021.104413.
9. Jia Y, Zhang Y, Weiss RJ, Wang Q, Shen J, Ren F, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates, Inc.; 2018.
10. Valle R, Shih KJ, Prenger R, Catanzaro B. Mellotron: multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. arXiv:1910.11997. 2019.
11. Le M, Vyas A, Shi B, Karrer B, Sari L, Moritz R, et al. Voicebox: text-guided multilingual universal speech generation at scale. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates, Inc.; 2024. 36 p.
12. Zhuang H, Guo Y, Wang Y. ViSPer: a multilingual TTS approach based on VITS using deep feature loss. In: 2023 IEEE 6th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE); 2023 Dec 15–17; Shenyang, China. Piscataway, NJ, USA: IEEE. p. 244–8.
13. Li M, Zheng J, Tian X, Cai L, Xu B. Code-switched text-to-speech and voice conversion: leveraging languages in the training data. In: 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Piscataway, NJ, USA: IEEE; 2020. p. 418–23. (In Chinese).

14. Pratap V, Hannun A, Xu Q, Cai J, Kahn J, Synnaeve G, et al. Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters. arXiv:2107.00636. 2021.
15. Uloza V, Verikas A, Bacauskiene M, Gelzinis A, Pribuisiene R, Kaseta M, et al. Categorizing normal and pathological voices: automated and perceptual categorization. J Voice. 2011;25(6):700–8. doi:10.1016/j.jvoice.2010.04.009.
16. Uloza V, Saferis V, Uloziene I. Perceptual and acoustic assessment of voice pathology and the efficacy of endolaryngeal phonomicrosurgery. J Voice. 2005;19(1):138–45. doi:10.1016/j.jvoice.2004.01.009.
17. Gu Y, Yin X, Rao Y, Wan Y, Tang B, Zhang Y, et al. ByteSing: a Chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and WaveRNN Vocoders. In: 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP); 2021 Jan 24–27; Hong Kong, China. Piscataway, NJ, USA: IEEE. p. 1–5.
18. Bao TQ, Mordukhovich BS. Relative Pareto minimizers for multiobjective problems: existence and optimality conditions. Math Program. 2008;122(2):301–47. doi:10.1007/s10107-008-0249-2.
19. Maskeliunas R, Damasevicius R, Kulikajevas A, Pribuisis K, Ulozaite-Staniene N, Uloza V. Synthesizing Lithuanian voice replacement for laryngeal cancer patients with Pareto-optimized flow-based generative synthesis network. Appl Acoust. 2024;224(7):110097. doi:10.1016/j.apacoust.2024.110097.
20. Maskeliūnas R, Damaševičius R, Kulikajevas A, Pribuišis K, Uloza V. Alaryngeal speech enhancement for noisy environments using a pareto denoising gated LSTM. J Voice. 2024;32(2S):18. doi:10.1016/j.jvoice.2024.07.016.
21. Laurinciukaite S, Telksnys L, Kasparaitis P, Kliukiene R, Paukstyte V. Lithuanian speech corpus liepa for development of human-computer interfaces working in voice recognition and synthesis mode. Informatica. 2018;29(3):487–98. doi:10.15388/informatica.2018.177.
22. Kasparaitis P. Evaluation of lithuanian text-to-speech synthesizers. Stud About Lang. 2016;28:80–91. doi:10.5755/j01.sal.0.28.15130.
23. Wang Y, Skerry-Ryan R, Stanton D, Wu Y, Weiss RJ, Jaitly N, et al. Tacotron: towards end-to-end speech synthesis. arXiv:1703.10135. 2017.
24. Prenger R, Valle R, Catanzaro B. WaveGlow: a flow-based generative network for speech synthesis. arXiv:1811.00002. 2018.
25. Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, et al. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. arXiv:1712.05884. 2017.
26. Bakšienė R, čepaitienė A, Jaroslavienė J, Urbanavičienė J. Standard lithuanian. J Int Phonetic Assoc. 2024;54(1):414–44. doi:10.1017/s0025100323000105.