# VoxAetherion:
# Latent-Space Perturbation for Data-Scarce Speech Synthesis

**Anonymous ACL submission**


## Abstract

We present VoxAetherion, a novel methodology for enhancing text-to-speech (TTS) synthesis through targeted perturbation within the discrete latent space of neural audio codecs. While modern TTS systems achieve high fidelity, they often struggle to capture the unique acoustic signatures of underrepresented speakers or those with limited available data. Our approach exploits the hierarchical structure of Residual Vector Quantization (RVQ), where early layers encode phonetic content and later layers capture paralinguistic attributes like timbre and prosody. By selectively perturbing frames with high layer deviation in specific RVQ layers, we generate acoustically diverse variants that preserve linguistic intelligibility while expanding the model's exposure to vocal variability. Experimental results across multiple data scales (25%, 50%, and 75% of LibriTTS-R) demonstrate that the methodology significantly improves synthesis quality, particularly for speakers in a heavily impaired speech zone (Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) < 10 dB). Our findings suggest that latent-space perturbation is an effective strategy for addressing data scarcity and improving personalization in neural speech synthesis.


## 1 Introduction

The human voice is a vital pillar of identity. While recent TTS advancements enable high-fidelity vocal reconstruction, authentically personalized synthesis for low-resource or atypical speakers remains a challenge. This goal aligns with Calm Interfaces (Schmitz et al., 2019), which advocates for technology that provides seamless, non-intrusive support as a natural extension of the user. However, standard architectures often reach a "robotic" plateau when training data is sparse, failing to capture the prosodic nuances of unique identities.

Unlike traditional signal-level augmentations, we introduce a strategy operating directly within the RVQ latent space. By targeting "acoustically complex" frames where the codec exhibits high layer deviation, we introduce meaningful variations without compromising phonetic structure. This approach allows synthetic variants to refine TTS decision boundaries while preserving linguistic integrity.

Our contributions are threefold. First, we propose Selective Perturbation to target high-deviation latent regions. Second, through Layer Optimization, we identify the specific RVQ layers that maximize acoustic diversity. Finally, we provide Empirical validation of the method, which significantly boosts synthesis quality for low-resource speakers across varying data scales.

## 2 Literature Review

Neural audio codecs have emerged as effective tools for compressing speech into discrete token sequences. SoundStream (Zeghidour et al., 2021) introduced RVQ, where multiple quantization layers iteratively encode residual information, enabling scalable bitrate control. EnCodec (Défossez et al., 2022) incorporated transformer-based language modeling over RVQ codes, while the Descript Audio Codec (DAC) (Kumar et al., 2023b) addressed codebook collapse through factorized and L2-normalized codes. Analyses of RVQ structure reveal that early layers capture coarse acoustic features such as phonetic content, while later layers encode fine-grained paralinguistic attributes including timbre and prosody (Zhang et al., 2023). This hierarchical organization motivates our layer-selective perturbation strategy.

The discrete representations produced by neural codecs have enabled a new paradigm in speech synthesis. (Gökgöz et al., 2025a) utilized them to build voice activity detectors. AudioLM (Borsos et al., 2023a) first demonstrated that audio generation via language modeling over discrete tokens. Building

on this foundation, VALL-E (Wang et al., 2023) applied conditional language modeling over codec tokens to achieve zero-shot speaker adaptation, and VALL-E 2 (Chen et al., 2024) further improved robustness through repetition-aware sampling. NaturalSpeech 3 (Ju et al., 2024) combines factorized codec representations with diffusion-based generation for zero-shot synthesis. However, these approaches require reference audio to condition speaker identity. ParlerTTS (Lyth and King, 2024; Lacombe et al., 2024b) conditions synthesis on text descriptions of speaker attributes, though requiring substantial paired data with diverse speaker and style coverage.

Prior work addresses data scarcity through two complementary strategies. SPEARTTS (Kharitonov et al., 2023) decomposes TTS into two stages and generates synthetic parallel data via backtranslation, reducing parallel data requirements to as few as 15 minutes. SoundStorm (Borsos et al., 2023b) improves efficiency through parallel acoustic token generation. For Automatic Speech Recognition (ASR), synthetic speech from TTS or voice conversion expands training corpora (Casanova et al., 2023). A separate line focuses on signal-level transformations: speed and pitch perturbation (Ko et al., 2015), Vocal Tract Length Perturbation (VTLP) for speaker variability (Jaitly and Hinton, 2013), and SpecAugment (Park et al., 2019a) for time-frequency masking. These methods operate on waveforms or spectrograms rather than exploiting codec latent spaces.

Zheng et al. (2025) proposed simulating noise perturbations directly on RVQ codewords through probabilistic top-K sampling, applying perturbations progressively from the last quantizer to the first. While also operating in the discrete latent space, their method differs from ours in three respects: (1) it targets codec robustness under noisy conditions rather than TTS data augmentation; (2) it perturbs all frames uniformly rather than selectively targeting high-deviation regions; and (3) it does not incorporate quality filtering or downstream TTS evaluation. Our approach is designed specifically for augmentation: selective frame targeting combined with WER-based quality classification ensures that perturbated variants remain linguistically intelligible while maximizing acoustic diversity in the regions that benefit most from it.

The following sections present our approach to selective RVQ perturbation (Section 3) and its experimental validation (Section 4).

## 3 Methodology

We propose a method for generating acoustically diverse variants of speech samples by operating directly in the RVQ latent space. Unlike traditional augmentation methods that operate on waveforms or spectrograms, our approach exploits the hierarchical structure of neural audio codecs to introduce controlled acoustic variations while preserving linguistic content. This methodology was integrated into the SRLab (Cornaggia-Urrigshardt et al., 2022) experimental framework to validate its impact on personalized speech synthesis and expand the capabilities of task specific ASR (Gökgöz et al., 2025c) and keyword spotting (KWS) (Gökgöz et al., 2025b; Wilkinghoff et al., 2021) modules trained on synthetic data. Figure 1 illustrates the complete pipeline, comprising four stages: layer-wise acoustic perturbation (Stage 1, Section 3.1), data preprocessing with automatic annotation (Stage 2, Section 3.2), description-conditioned TTS fine-tuning (Stage 3, Section 3.3), and evaluation (Stage 4, Section 3.4).

### 3.1 Layer-wise Acoustic Perturbation

Our perturbation method 1 operates on discrete speech representations obtained through SoundStream, a neural audio codec comprising an encoder, RVQ module, and decoder. The encoder maps input waveforms to continuous latent representations, which the RVQ module discretizes through $L$ stacked quantization layers. Each layer maps vectors to their nearest codebook entries and passes the residual to subsequent layers. The final quantized representation accumulates contributions across all layers: $\mathbf{Q} = \sum_{\ell=1}^{L} \mathbf{Q}_\ell$ where $\mathbf{Q}_\ell$ denotes the quantized output of layer $\ell$. This hierarchical structure encodes coarse acoustic patterns in early layers and progressively finer details in later layers.

#### 3.1.1 Perturbation Process

The core insight of our method is that targeting frames with high layer deviation introduces meaningful acoustic variations while preserving linguistic content in well-reconstructed frames.

The perturbation operates on the final accumulated quantized representation $\mathbf{Q} \in \mathbb{R}^{T \times D}$, where $T$ is the number of temporal frames and $D = 256$ is the embedding dimension. For each RVQ layer $\ell \in \{\ell_{\text{start}}, \ldots, \ell_{\text{end}}\}$, we retrieve the layer-specific

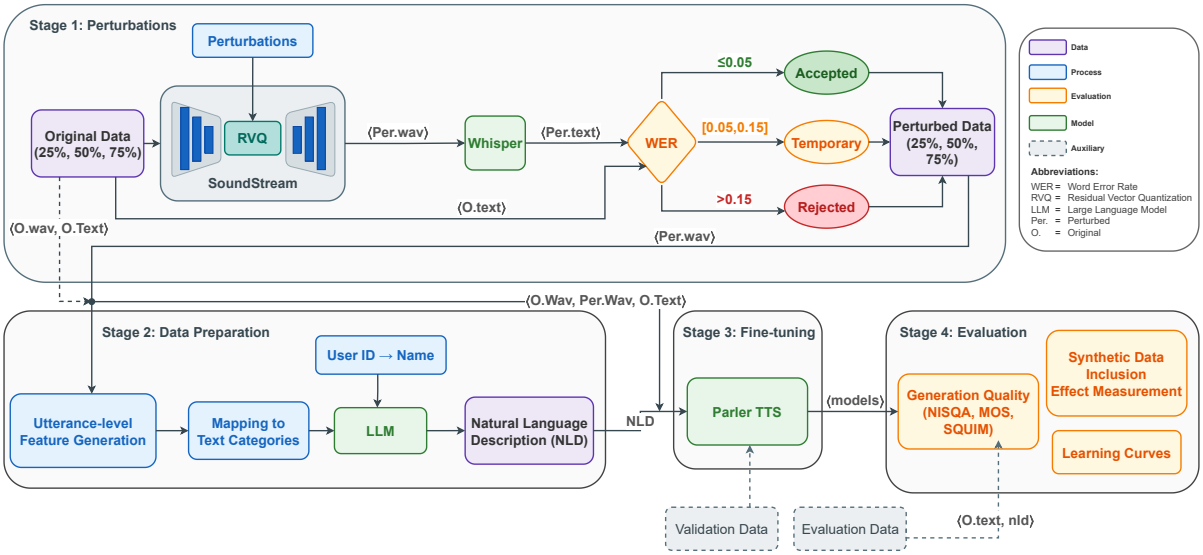


Figure 1: Pipeline overview. **Stage 1**: Audio is encoded via SoundStream, quantized through RVQ, perturbed via layer-wise noise accumulation, decoded, and classified by WER to assess how perturbations challenge ASR decision boundaries. **Stage 2**: Original and perturbed audio undergo feature extraction, binning, and LLM-based description generation. **Stage 3**: Paired audio-description data fine-tunes ParlerTTS. **Stage 4**: Models are evaluated using speech quality metrics.

quantized vectors $\mathbf{Q}_\ell \in \mathbb{R}^{T\times D}$ and compute the per-frame Euclidean distance $\mathbf{d}_\ell = \|\mathbf{Q}_\ell - \mathbf{Q}\|_2 \in \mathbb{R}^T$. This distance quantifies how much layer $\ell$'s contribution alone deviates from the full accumulated output. Large values indicate frames where the remaining layers must compensate substantially. In acoustic terms, these frames correspond to regions of high spectral complexity: transient boundaries, coarticulation transitions, and speaker-specific spectral detail that require fine-grained encoding across many layers. Perturbation at these frames produces shifts in spectral envelope, micro-temporal structure, and fine spectral texture that the neural codec decoder maps back onto the learned speech manifold, yielding naturally-sounding variants rather than arbitrary distortions.

We selectively target frames exhibiting extreme layer deviation rather than perturbing all frames uniformly. Given a quantile threshold $q \in [0, 1]$, we extract indices corresponding to frames exceeding the threshold: $\mathcal{I}_\ell = \{i : d_{\ell,i} \geq \text{Quantile}(\mathbf{d}_\ell, q)\}$. With $q = 0.95$, this identifies the top 5% of frames per layer. Consequently, a high quantile threshold (e.g., $q$=0.95) targets a small fraction of frames with the largest layer deviations, whereas a low threshold (e.g., $q$=0.05) selects the majority of frames, including those with minimal deviation. Frame selection is performed independently for each layer, so a given frame may appear in the selection set of multiple layers or of none.

For each layer $\ell$, we extract vectors at selected indices $\mathbf{Q}[\mathcal{I}_\ell] \in \mathbb{R}^{|\mathcal{I}_\ell|\times D}$ and fit a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$ using Ledoit-Wolf shrinkage (Ledoit and Wolf, 2021) for covariance regularization. We then draw $T$ independent noise vectors of dimension $D$ from this distribution, one for every frame in the utterance (not only the selected ones). The noise characteristics are thus shaped by the acoustically complex regions but applied to all frames. Noise contributions accumulate across layers: $\mathbf{Q}_{\text{perturbed}} = \mathbf{Q}_{\text{original}} + \alpha \cdot \sum_{\ell=\ell_{\text{start}}}^{\ell_{\text{end}}} \boldsymbol{\epsilon}_\ell$. Since sampling is stochastic, each call produces a different set of noise vectors, allowing generation of multiple diverse variants from a single source recording.

### 3.1.2 Quality-based Sample Classification

We employ ASR-based evaluation to assess how perturbations affect speech elements. This classification measures how modifications challenge the decision boundaries of an ASR system, which may differ from human perception. We transcribe each perturbed sample using Whisper (Radford et al., 2022) Base English and compute Word Error Rate (WER): samples are ACCEPTED if WER $\leq 0.05$, TEMPORARY if $0.05 <$ WER $\leq 0.15$, and REJECTED if WER $> 0.15$. This classification measures how modifications challenge ASR decision boundaries, which may diverge from human perception.

**Algorithm 1** Layer-wise Acoustic Perturbation (Stage 1)

**Require:** Audio $\mathbf{x}$, layer range $[\ell_{\text{start}}, \ell_{\text{end}}]$, quantile $q$, scale $\alpha$, transcript $t$
**Ensure:** Perturbed audio $\mathbf{x}'$, quality category $c$

$\mathbf{z} \leftarrow \text{Encode}(\mathbf{x})$; $\mathbf{Q} \leftarrow \text{RVQ}(\mathbf{z})$ $\{\mathbf{Q} \in \mathbb{R}^{T\times D}, D{=}256\}$
$\boldsymbol{\epsilon}_{\text{total}} \leftarrow \mathbf{0}_{T\times D}$ {Initialize noise accumulator}
**for** $\ell = \ell_{\text{start}}$ **to** $\ell_{\text{end}}$ **do**
  $\mathbf{d}_\ell \leftarrow \|\mathbf{Q}_\ell - \mathbf{Q}\|_2$ {Per-frame layer deviation, $\mathbf{d}_\ell \in \mathbb{R}^T$}
  $\mathcal{I}_\ell \leftarrow \{i : d_{\ell,i} \geq \text{Quantile}(\mathbf{d}_\ell, q)\}$ {Top $(1{-}q)$ fraction}
  $\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell \leftarrow \text{LedoitWolf}(\mathbf{Q}[\mathcal{I}_\ell])$ {Fit $\mathcal{N}$ to $|\mathcal{I}_\ell|$ frames}
  $\boldsymbol{\epsilon}_\ell \sim \mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)^T$ {$T$ independent draws $\in \mathbb{R}^{T\times D}$}
  $\boldsymbol{\epsilon}_{\text{total}} \leftarrow \boldsymbol{\epsilon}_{\text{total}} + \boldsymbol{\epsilon}_\ell$ {Accumulate across layers}
**end for**
$\mathbf{Q}' \leftarrow \mathbf{Q} + \alpha \cdot \boldsymbol{\epsilon}_{\text{total}}$
$\mathbf{x}' \leftarrow \text{Decode}(\mathbf{Q}')$
$c \leftarrow \text{ClassifyByWER}(\text{Whisper}(\mathbf{x}'), t)$
**return** $\mathbf{x}', c$

The ACCEPTED category indicates perturbations that remain within ASR decision boundaries; TEMPORARY samples push against these boundaries; REJECTED samples indicate perturbations substantial enough to alter transcription output. Notably, these labels are used for meta-classification purposes rather than as criteria for discarding data. As we demonstrate in Section 4, speech often remains humanly intelligible even in REJECTED samples.

#### 3.1.3 Hyperparameters

Three hyperparameters control the perturbation process:

- **Layer range** $[\ell_{\text{start}}, \ell_{\text{end}}]$: Earlier layers (0-7) ; later layers (8-15).
- **Quantile threshold** $q$: Higher values select fewer frames with largest layer deviations.
- **Scaling factor** $\alpha$: Controls perturbation magnitude.

Based on systematic optimization (Section 4), we identify optimal parameters: $\alpha = 0.15$, $q = 0.95$, and layer range 8-15. Algorithm 1 summarizes the procedure.

### 3.2 Data Preparation

Stage 2 of Figure 1 shows data preparation. The experiments use LibriTTS-R (Koizumi et al., 2023) train-clean-360, containing approximately 182 hours of English speech. The corpus is partitioned into subsets of 25%, 50%, and 75% of the training data to simulate progressively resourced conditions. For each subset we run Algorithm 1 three times with identical hyperparameters, producing three independent perturbation variants denoted P1, P2, and P3. Because the Gaussian sampling is stochastic, each run yields different perturbated samples. This results in nine perturbed datasets (three per subset) and allows us to measure both average performance and variance across runs. Both original and perturbed audio proceed through preprocessing.

Features are extracted using the same methodology described by Lyth and King (2024). We used the open-source library DataSpeech (Lacombe et al., 2024a) to automatically extract speaking rate, Signal-to-Noise Ratio (SNR), reverberation (C50) (Lavechin et al., 2023), pitch variation, and SQUIM metrics (Kumar et al., 2023a) including STOI (Taal et al., 2011, 2010), SI-SDR (Le Roux et al., 2018), and PESQ (Wang et al., 2022). Continuous values are discretized into bins using quantile-based boundaries, with descriptive phrases assigned to each bin (*e.g.,* speaking rate into 7 bins ranging from *very slowly* to *very fast*). Mistral-7B-v0.1 (Jiang et al., 2023) then generates natural language descriptions from these binned keywords.

### 3.3 ParlerTTS Fine-tuning

Stage 3 shows fine-tuning of ParlerTTS (Lyth and King, 2024), a description-conditioned TTS system comprising DAC for tokenization, frozen Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020) encoder for description embeddings, autoregressive transformer decoder, and DAC decoder for waveform reconstruction.

Validation and evaluation sets (4,000 samples each) are constructed from held-out data. For each data scale, four fine-tuning runs use ParlerTTS Mini v1 (0.9B parameters, pre-trained on over 40,000 hours of speech): baseline (original only), +Pert.1 (original plus first variant), +Pert.2 (plus second variant), and +Pert.3 (plus all three variants). The perturbation operates at the codec level and is not tied to any particular TTS architecture: Algorithm 1 requires only a multi-layer RVQ code-

book with per-layer access to quantized vectors, a property shared by SoundStream, EnCodec, and DAC. The layer range and quantile threshold may need re-tuning for codecs with different hierarchical organizations. For evaluation, we selected the final checkpoint from each model to capture the full effect of perturbation exposure.

### 3.4 Evaluation Protocol

Stage 4 shows the evaluation. For each model, the speech is synthesized using evaluation set transcripts and descriptions from the last checkpoint. Three metrics are calculated: SQUIM (Kumar et al., 2023a), NISQA (Mittag et al., 2021) for multidimensional quality assessment, and DNSMOS (Reddy et al., 2021) for noise-sensitive perceptual assessment.

For perturbed audio, VISQOL (Hines et al., 2012, 2013) is calculated to measure the perceptual distance between the perturbed and original speech, quantifying the acoustic diversity introduced by the perturbation process.

## 4 Experiments

We conduct systematic experiments to validate the perturbation methodology, represented as Stage 1 in Figure 1, through three phases: (1) initial perturbation analysis examining configuration effects, (2) phoneme-level validation using forced alignment, and (3) hyperparameter optimization combining human evaluation with automated experiments.

### 4.1 Experimental Setup

For perturbation analysis and hyperparameter optimization, we use MLS-English (Pratap et al., 2020) 10k, selecting 2 hours of audio (483 samples) from a single female speaker. This controlled single-speaker setup isolates perturbation effects from speaker variability. All experiments employ a pretrained SoundStream model with 27M parameters, 16 RVQ layers, a codebook size of 1,024, and 256-dimensional vectors. Perturbed samples are transcribed using Whisper Base English for WER-based quality classification, with an English-only mode enforced to prevent spurious non-English output. This single-speaker setup serves exclusively for hyperparameter optimization; the full multi-speaker evaluation across all speakers in each LibriTTS-R split is reported in Section 5. Audio is processed at the utterance level as segmented by the corpus; no further segmentation is applied. At 16 kHz with SoundStream's 320:1 downsampling, a 5-second utterance yields $T \approx 250$ frames, each of dimension $D = 256$.

### 4.2 Initial Perturbation Analysis

The WER-based classification assesses whether perturbations meaningfully affect speech elements and how configurations challenge ASR decision boundaries. Full results across all 21 configurations are reported in Table 2 (Appendix A).

The experimental results reveal that both the scope and nature of perturbation drive model performance. Acceptance rates decrease monotonically as more layers are included: from 55.7% with layers 8-15 to 35.0% with all 16 layers. However, sensitivity is most strongly determined by quantile strategy rather than layer depth alone. The model exhibits high robustness to perturbations in low-quantile regions, maintaining stability even as the perturbation magnitude increases: at $\alpha$=0.3, low-quantile configurations retain 67.9% acceptance. In contrast, high-quantile perturbations ($q$=0.95) are significantly more destructive, with acceptance dropping to 3.1% at the same $\alpha$ value. Notably, combining both quantile extremes produces results identical to low-quantile alone.

### 4.3 Phoneme-level Validation

Random listening tests across quality categories reveal that while perturbations can introduce audible artifacts, speech remains highly intelligible even in REJECTED samples. Phonetic integrity is preserved; rejection based on WER primarily captures acoustic modifications that stress ASR performance rather than undermining fundamental human intelligibility. To verify that these perturbations genuinely modify speech elements rather than merely adding obscuring noise, we performed a temporal analysis. While the forced aligner provided by Gökgöz et al. (2024) within the **SRLab** context operates on word tokens, we deployed Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) to achieve the granularity required for phoneme-level timing analysis (see Appendix A). Our analysis shows that most phonemes exhibit median duration differences near zero, indicating no systematic bias toward temporal lengthening or shortening. While interquartile ranges reveal substantial instance-level variability, duration distributions remain consistent across all quality categories. This suggests that WER-based classification is driven by spectral properties and acoustic

distortions rather than systematic shifts in temporal alignment.

### 4.4 Hyperparameter Optimization

We systematically optimize three hyperparameters: **scaling factor $\alpha$, quantile threshold $q$, and layer range** through automated progressive experiments, corroborated by a preliminary listening study (Appendix B.2).

Progressive experiments tested six $\alpha$ values ($\{0.07, 0.10, 0.15, 0.20, 0.25, 0.30\}$) across layers 0-15 with $q$=0.95. Acceptance rates decrease with both increasing layer count and $\alpha$ values, reaching zero at $\alpha$=0.30 for the full 16-layer configuration. The $\alpha$=0.15 configuration shows uniform decline across layer ranges, representing an optimal balance between conservative ($\alpha \leq 0.10$) and aggressive ($\alpha \geq 0.20$) regimes (Appendix B). A preliminary listening study corroborates this finding, with $\alpha$=0.15 achieving the highest human agreement rate and 95.1% of REJECTED samples judged intelligible, confirming WER-based classification is more restricted than human perception.

For quantile selection, five high-quantile values ($q_{\text{high}} \in \{0.55, 0.65, 0.75, 0.85, 0.95\}$) reveal two distinct regimes: a *stable* regime ($q \leq 0.75$) maintaining consistent acceptance rates, and a *sensitive* regime ($q \geq 0.85$) showing sharp degradation. Low-quantile perturbations show negligible acoustic effect regardless of threshold or layer configuration, and combined high/low selection mirrors high-quantile-only patterns (Appendix C).

Layers 0-7 maintain high acceptance rates regardless of settings, whereas layers 8-15 exhibit greater parameter sensitivity while preserving intelligibility; notably, layer 15 introduces acoustic degradation without a proportional gain in diversity or quality.

### 4.5 Final Configuration

Based on systematic optimization, the final parameters are set to $\alpha = 0.15$, $q = 0.95$, and layers 8–15. This configuration achieves meaningful acoustic diversity, validated through MFA phoneme analysis, while maintaining high human intelligibility as confirmed by our listening study. Crucially, the WER-based classification effectively captures ASR sensitivity to specific acoustic modifications rather than a loss of fundamental intelligibility.

## 5 Evaluations and Discussion

This section evaluates perturbation-based training across three data scales (25%, 50%, and 75% of LibriTTS-R train-clean-360). Results are organized in two stages: Section 5.1 assesses perturbated audio directly against original recordings, quantifying the trade-off between diversity and intelligibility; the remaining subsections evaluate the downstream impact on TTS quality by comparing models trained with and without perturbated data. We focus on the 25% split as the primary analysis, as it represents the most resource-constrained scenario. Results for 50% and 75% splits appear in Appendix D.

### 5.1 Perturbation Variant Diversity vs. Intelligibility

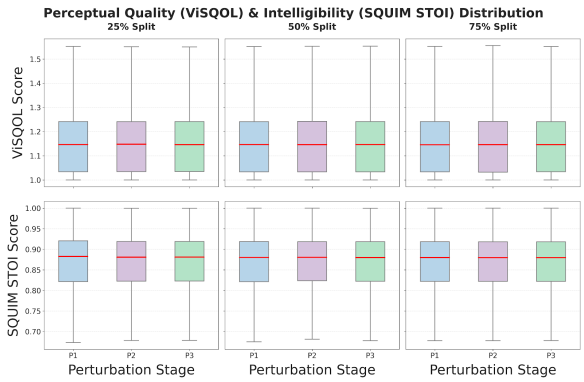


Figure 2: Virtual Speech Quality Objective Listener (ViSQOL) (top) and SQUIM Short-Time Objective Intelligibility (STOI) (bottom) distributions across data splits and perturbation variants. Low ViSQOL scores (approximately 1.15) indicate substantial acoustic diversity, while high SQUIM STOI scores (above 0.9) confirm preserved intelligibility.

Quantitative analysis using ViSQOL and SQUIM STOI reveals a distinct decoupling between acoustic texture and linguistic content (Figure 2). ViSQOL scores around 1.15 confirm substantial acoustic diversity, while SQUIM STOI scores above 0.9 demonstrate that intelligibility remains intact. This combination of low perceptual similarity with high intelligibility provides the mechanism underlying the Signal Rescue effect: by populating the training distribution with acoustically diverse but linguistically accurate variants, the TTS model learns to disentangle speaker identity from channel characteristics, effectively expanding the decision boundary for underrepresented speakers.

Speaker embedding analysis using SpeechBrain's ECAPA-TDNN model (Ravanelli et al., 2021, 2024), confirms that perturbated audio shifts substantially in the speaker embedding space. Since the ParlerTTS fine-tuning maps both original and perturbated audio to the same speaker IDs, the model learns to associate acoustically diverse variants with their respective speakers. The downstream synthesis improvements reported in Sections 5 through 6.6, where quality gains hold across all data scales and perturbation runs, confirm that the model benefits from this diversity without losing speaker specificity. Full speaker embedding results are reported in Appendix F.

### 5.2 Speaker-Level Performance Analysis

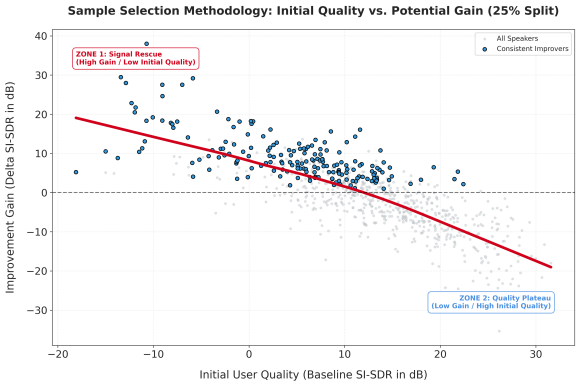


Figure 3: SI-SDR gain versus baseline quality for the 25% split. Blue points indicate consistent improvers (180 speakers). The trend line reveals an inverse relationship between initial quality and improvement potential. Zone 1 (Signal Rescue) shows high gains for low-quality speakers; Zone 2 (Quality Plateau) shows diminishing returns for high-quality speakers.

To assess the impact of perturbation-based training, we computed SI-SDR gain relative to the baseline for each speaker. Figure 3 presents a scatter plot of improvement gain against baseline SI-SDR for the 25% split. The analysis identifies two distinct performance zones based on initial speaker quality. Our analysis identified 180 speakers who demonstrated consistent SI-SDR improvement across all three independently generated perturbation variants. The consistency of improvement across independent perturbation datasets suggests that the observed gains are attributable to the perturbation methodology rather than random variation.

### 5.3 SI-SDR Threshold Analysis

Analysis of baseline SI-SDR values reveals clear performance thresholds that predict the effectiveness of perturbation-based training. Speakers below 10 dB (**Signal Rescue**) exhibited substantial, near-guaranteed improvement: characterized by limited data representation in the original corpus, these speakers consistently benefited from the acoustic diversity introduced by perturbation variants. Speakers in $[10, 15]$ dB (**Transition Zone**) showed marginal improvements; monitoring SI-SDR during training can inform whether to continue incorporating perturbation data for individual speakers. Speakers above 15 dB (**Quality Plateau**) showed inconsistent results, as they already achieve high synthesis quality with original data alone, leaving limited room for improvement. This stratification provides a practical criterion for applying perturbation-based training: prioritize speakers in the Signal Rescue zone, apply with monitoring in the Transition Zone, and consider alternative strategies for selecting samples to perturb for speakers in the Quality Plateau zone.

### 5.4 Multi-Metric Evaluation

For speakers in the Signal Rescue and Transition zones who exhibited consistent SI-SDR improvement, we conducted a comprehensive evaluation across five objective metrics: SI-SDR, STOI, PESQ, MOS (DNSMOS Signal), and NISQA Noisiness. Figure 4 presents box plots comparing metric distributions between the baseline (B) and three perturbation stages (P1, P2, P3).

Speakers in these zones showed improvement across all five evaluation metrics compared to the baseline trained exclusively on original data. The upward shift in median values and compact distribution of box plots indicate that perturbation-based training provides systematic quality improvements for underrepresented speakers. Notably, the improvement occurs primarily from B to P1, with P1, P2, and P3 exhibiting similar distributions. This pattern suggests that the benefit derives from introducing perturbation-based acoustic diversity into training, rather than from the cumulative addition of multiple variants. A single perturbation variant may be sufficient to achieve the observed improvements.

A natural question is whether the observed gains stem from the perturbation methodology or simply from having more training data. Three observations

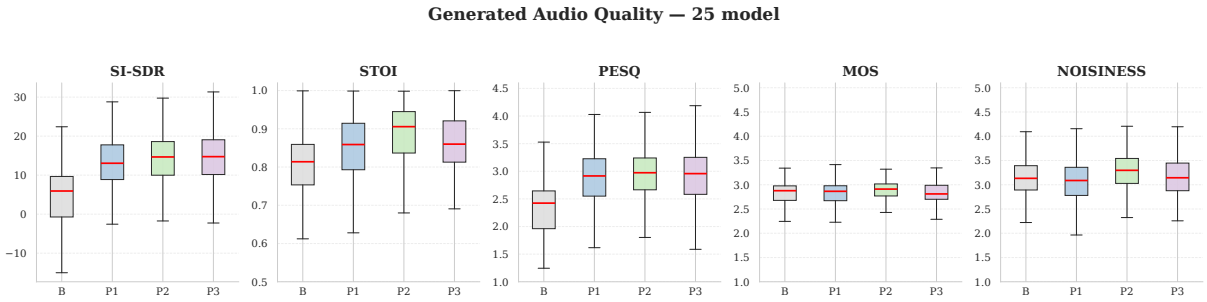


Figure 4: Quality metrics for consistent improvers in Signal Rescue and Transition zones (25% split). B: baseline; P1, P2, P3: perturbation stages.

argue against a pure volume explanation: performance plateaus after P1 despite progressive data addition; the effect is speaker-specific rather than uniform, benefiting only Signal Rescue speakers; and perturbated samples populate new acoustic regions (ViSQOL $\approx$ 1.15) rather than duplicating existing coverage. The *character* of the added data, its acoustic diversity, rather than its volume, drives the improvement.

### 5.5 Cross-Split Comparison

| Metric | 25% | 50% | 75% |
|---|---|---|---|
| Consistent improvers | 180 | 196 | 212 |
| Consistency rate (%) | 35.2 | 38.4 | 41.5 |
| Hours/split | 46.83 | 93.59 | 140.40 |

Table 1: Larger training splits yield higher consistency rates, with SI-SDR threshold patterns consistent across all splits.

To validate the generalizability of our findings, we conducted identical analyses on 50% and 75% data splits (Appendix D). Table 1 summarizes key findings. The increasing number of consistent improvers (180 $\rightarrow$ 196 $\rightarrow$ 212) indicates that perturbation-based training scales effectively, with SI-SDR threshold patterns remaining consistent across splits: speakers below 10 dB show robust improvements regardless of data scale, while high-quality speakers exhibit similar variability. To ensure fair comparison, we verified via ANOVA (Girden, 1992) that consistent improver groups do not differ in available training data ($p = 0.2511$), confirming that performance differences reflect model behavior rather than selection bias (Appendix E).

### 5.6 Discussion

The results demonstrate that RVQ latent space perturbation effectively improves TTS quality for speakers with limited data representation. The inverse relationship between baseline quality and improvement gain suggests that perturbation-based training addresses a fundamental limitation of neural TTS systems: their tendency to underperform on underrepresented speakers. The mechanism underlying this improvement is revealed by the decoupling of perceptual similarity and intelligibility: ViSQOL scores near 1.15 indicate substantial acoustic modification, while Speech Quality Understanding and Intelligibility Metric (SQUIM) STOI scores above 0.9 confirm that linguistic content remains intact, enabling the model to learn speaker characteristics from acoustically diverse variants without compromising phonetic accuracy.

The consistency of the SI-SDR threshold pattern across data scales indicates that this relationship reflects an intrinsic property of the perturbation methodology rather than an artifact of specific training conditions, establishing latent-space perturbation as a scale-invariant strategy for addressing data scarcity in personalized TTS.

## 6 Conclusion

We introduce a method for generating acoustic diversity through layer-wise RVQ latent-space perturbation to improve synthesis quality for underrepresented speakers. By producing substantial acoustic variation while preserving linguistic content, this mechanism facilitates significant recovery of synthesis performance for speakers with limited training data. The effect is most pronounced for speakers with low initial baselines (SI-SDR $<$ 10 dB), where data scarcity prevents learning robust speaker representations. These gains remain consistent across all data scales, establishing latent-space perturbation as a scale-invariant solution for data scarcity in personalized TTS.

## Limitations

Performance plateaued for speakers with high baseline fidelity (SI-SDR $> 15$ dB). For these speakers, perturbation adds variants that fall within the model's already-learned decision boundary, offering little additional training signal. In some cases the added variants may conflict with an already adequate speaker representation, causing marginal degradation. Future work should explore adaptive perturbation conditioned on per-speaker quality: applying it only below a quality threshold, while using alternative strategies (e.g., style-conditioned sampling or targeted prosodic augmentation) for well-represented speakers.

The optimal layer range (8-15) and hyperparameters ($\alpha$, $q$) have been validated only for SoundStream. Recalibration is likely needed for other codec architectures with different latent hierarchies, although the algorithm's requirements (multi-layer RVQ with per-layer codebook access) are satisfied by all major neural audio codecs.

We do not compare against waveform-level augmentation methods (pitch shifting, time stretching, SpecAugment (Park et al., 2019b)) or the RVQ-level perturbation of Zheng et al. (2025). A fair comparison would require training separate TTS models under identical conditions for each method. Our latent-space approach and signal-level methods are complementary rather than competing: the neural codec decoder constrains perturbations to the learned speech manifold, a property that waveform-level transformations do not share.

Our subjective evaluation is limited to a preliminary listening study with five participants. A MUSHRA-based (ITU-R, 2001) study with a larger listener pool would give stronger evidence regarding the perceptual quality of perturbation-based training.

## References


Paul Boersma and David Weenink. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.38, retrieved 2 January 2021 , http://www.praat.org/.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023a. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023b. SoundStorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*.

Edresson Casanova, Christopher Shulby, Alexander Korolev, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Aluísio, and Moacir Antonelli Ponti. 2023. ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion. In *Interspeech*, pages 1244–1248.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024. VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Alessia Cornaggia-Urrigshardt, Fahrettin Gökgöz, Frank Kurth, Hans-Christian Schmitz, and Kevin Wilkinghoff. 2022. Speech recognition lab. *Procedia Comput. Sci.*, 205(C):218–228.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Ellen R Girden. 1992. *ANOVA: Repeated measures*. 84. Sage.

Fahrettin Gökgöz, Hisham Ali, and Priya Pal. 2025a. Discrete audio representations from soundstream: A dual approach to efficient transmission and speech detection.

Fahrettin Gökgöz, Alessia Cornaggia-Urrigshardt, and Kevin Wilkinghoff. 2024. Strong label generation for preparing speech data in military applications using CTC loss.

Fahrettin Gökgöz, Alessia Cornaggia-Urrigshardt, and Kevin Wilkinghoff. 2025b. Personalized speech synthesis for zero-shot keyword spotting. In *Speech Communication; 16th ITG Conference*, pages 61–65.

Fahrettin Gökgöz, Hussein Hasso, and Hans-Christian Schmitz. 2025c. Task-specific, personalized automatic speech recognition.

Andrew Hines, Jan Skoglund, Anil Kokaram, and Naomi Harte. 2012. Visqol: The virtual speech quality objective listener. In *IWAENC 2012; international workshop on acoustic signal enhancement*, pages 1–4. VDE.

Andrew Hines, Jan Skoglund, Anil Kokaram, and Naomi Harte. 2013. Robustness of speech quality metrics to background noise and network degradations: Comparing visqol, pesq and polqa. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3697–3701.

ITU-R. 2001. *Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems*. International Telecommunications Union.

Paul Jaccard. 1912. The Distribution of the Flora in the Alpine ZONE.[1]. *New Phytologist*, 11(2):37–50.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

Navdeep Jaitly and Geoffrey E Hinton. 2013. Vocal tract length perturbation (VTLP) improves speech recognition. In *ICML Workshop on Deep Learning for Audio, Speech and Language*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, and 1 others. 2024. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *arXiv preprint arXiv:2302.03540*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proc. Interspeech*, pages 3586–3589.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023. LibriTTS-R: A restored multi-speaker text-to-speech corpus. *arXiv preprint arXiv:2305.18802*.

Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. 2023a. TorchAudio-Squim: Reference-less speech quality and intelligibility measures in TorchAudio. *arXiv preprint arXiv:2304.01448*.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023b. High-fidelity audio compression with improved RVQGAN. *arXiv preprint arXiv:2306.06546*.

Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. 2024a. Data-speech. https://github.com/ylacombe/dataspeech.

Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. 2024b. Parler-tts. https://github.com/huggingface/parler-tts.

Marvin Lavechin, Marianne Métais, Hadrien Titeux, Alodie Boissonnet, Jade Copet, Morgane Rivière, Elika Bergelson, Alejandrina Cristia, Emmanuel Dupoux, and Hervé Bredin. 2023. Brouhaha: Multi-task training for voice activity detection, speech-to-noise ratio, and C50 room acoustics estimation. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. 2018. SDR - half-baked or well done? *arXiv preprint arXiv:1811.02508*.

Olivier Ledoit and Michael Wolf. 2021. Shrinkage estimation of large covariance matrices: Keep it simple, statistician? *Journal of Multivariate Analysis*, 186:104796.

Dan Lyth and Simon King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner [computer program]. http://montrealcorpustools.github.io/Montreal-Forced-Aligner/. Version 0.9.0, retrieved 17 January 2017.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Proc. Interspeech*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019a. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*, pages 2613–2617.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019b. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, interspeech_2019, page 2613–2617. ISCA.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Proc. Interspeech*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra,

Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, and 14 others. 2024. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333).

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624. ArXiv:2106.04624.

Chandan K A Reddy, Vishak Gopal, and Ross Cutler. 2021. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. *arXiv preprint arXiv:2010.15258*.

Hans-Christian Schmitz, Alessia Cornaggia-Urrigshardt, Fahrettin Gökgöz, Samantha Kent, and Kevin Wilkinghoff. 2019. Calm interfaces for integrated C2 systems.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Miao Wang, Christoph Boeddeker, Rafael G. Dantas, and Ananda Seelan. 2022. PESQ (perceptual evaluation of speech quality) wrapper for Python users.

Kevin Wilkinghoff, Alessia Cornaggia-Urrigshardt, and Fahrettin Gökgöz. 2021. Two-dimensional embeddings for low-resource keyword spotting based on dynamic time warping. In *14th ITG Conference on Speech Communication (ITG Speech)*, pages 9–13. VDE-Verlag.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. SoundStream: An end-to-end neural audio codec. *arXiv preprint arXiv:2107.03312*.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechTokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*.

Rui-Chen Zheng and 1 others. 2025. Enhancing noise robustness for neural speech codecs through resource-efficient progressive quantization perturbation simulation. *arXiv preprint arXiv:2509.19025*.

# A Phoneme-level Analysis

The 21 experiments from Section 4 were analyzed using MFA to assess perturbation effects on speech elements at the phoneme level. MFA provides automatic time-alignment of speech with corresponding text transcriptions, enabling precise comparison of duration differences between perturbed and original phonemes. Table 2 summarizes the WER-based classification counts for all 21 configurations; the following subsections present the corresponding phoneme-level duration analysis.

| Layers | Quantile / $\alpha$ | Acc. | Rej. | Temp. |
|---|---|---|---|---|
| *Layer range variation (high $q$=0.95, $\alpha$=0.07)* | | | | |
| 8-15 | - | 269 | 41 | 173 |
| 7-15 | - | 246 | 45 | 192 |
| 6-15 | - | 228 | 47 | 208 |
| 5-15 | - | 237 | 62 | 184 |
| 3-15 | - | 204 | 56 | 223 |
| 1-15 | - | 179 | 76 | 228 |
| 0-15 | - | 169 | 79 | 235 |
| *Low quantile variations (layers 8-15)* | | | | |
| 8-15 | low 0.05, $\alpha$=0.07 | 347 | 24 | 112 |
| 8-15 | low 0.15, $\alpha$=0.07 | 363 | 19 | 101 |
| 8-15 | low 0.05, $\alpha$=0.1 | 340 | 22 | 121 |
| 8-15 | low 0.05, $\alpha$=0.3 | 328 | 28 | 127 |
| 8-15 | low 0.15, $\alpha$=0.3 | 339 | 25 | 119 |
| 8-15 | low 0.35, $\alpha$=0.3 | 334 | 22 | 127 |
| *High quantile with varying $\alpha$ (layers 8-15)* | | | | |
| 8-15 | high 0.95, $\alpha$=0.1 | 203 | 65 | 215 |
| 8-15 | high 0.95, $\alpha$=0.3 | 15 | 416 | 52 |
| *Combined quantile (both low and high)* | | | | |
| 8-15 | both 0.15/0.85, $\alpha$=0.07 | 347 | 23 | 113 |
| 0-15 | both 0.15/0.85, $\alpha$=0.07 | 331 | 31 | 121 |
| *Layers 0-7 variations* | | | | |
| 0-7 | low 0.05, $\alpha$=0.1 | 350 | 19 | 114 |
| 0-7 | high 0.95, $\alpha$=0.1 | 221 | 64 | 198 |
| 0-7 | low 0.05, $\alpha$=0.3 | 310 | 26 | 147 |
| 0-7 | high 0.95, $\alpha$=0.3 | 17 | 389 | 77 |

Table 2: Results from 21 perturbation experiments (483 samples). Acc.=Accepted, Rej.=Rejected, Temp.=Temporary. Default settings: high quantile ($q$=0.95), $\alpha$=0.07.

## A.1 Phoneme Duration Difference Heatmap

Figure 5 presents a heatmap visualization of average phoneme duration differences across all 21 experiments.

The heatmap reveals distinct patterns in phoneme duration variations. Experiments using

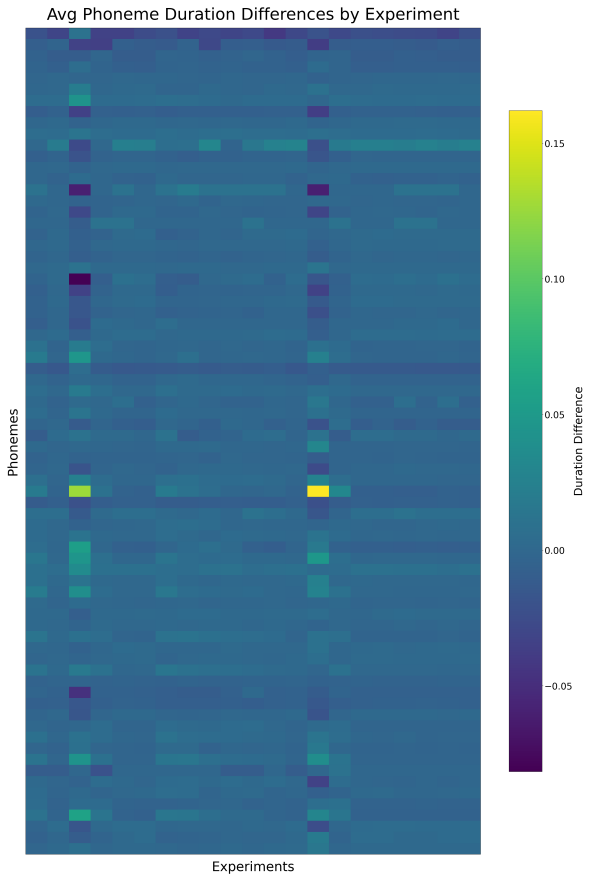


Figure 5: Heatmap of average phoneme duration differences across all 21 experiments. Experiments with high-quantile selection and $\alpha = 0.3$ (layers 0-7 and layers 8-15) exhibit pronounced color variation, indicating substantial duration modifications. Low-quantile configurations show minimal variation regardless of $\alpha$ values.

high-quantile selection with $\alpha = 0.3$ exhibit pronounced color variation across all phonemes, indicating substantially higher average duration differences. In contrast, low-quantile configurations show minimal color variation despite identical $\alpha$ values, confirming that high-quantile frame selection targets acoustically complex regions where perturbations produce substantial effects.

### A.2 Phoneme Duration Difference Box Plot

Figure 6 presents a box plot analysis of phoneme duration differences aggregated across all experiments.

Most phonemes exhibit median duration differences centered near zero, indicating no systematic directional bias. However, interquartile ranges and outliers reveal substantial instance-level variability. Silence markers showed the greatest susceptibility to duration changes, aligning with phonetic

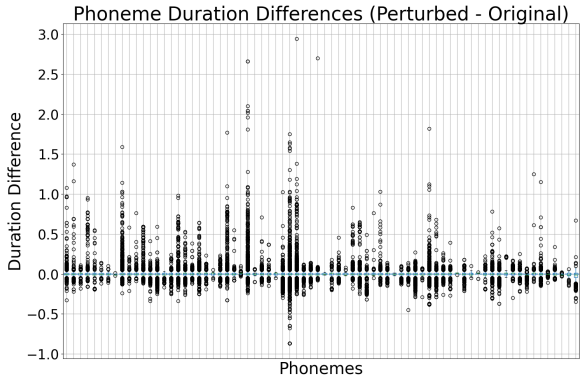


Figure 6: Box plot of phoneme duration differences (perturbed minus original). Most phonemes exhibit median differences near zero with substantial instance-level variability.

expectations. Asymmetric distributions for certain phonemes suggest specific sound classes are more prone to lengthening or shortening under perturbation.

### A.3 Status-Based Duration Difference Analysis

Figure 7 examines whether phoneme-level duration perturbations influence status classification.

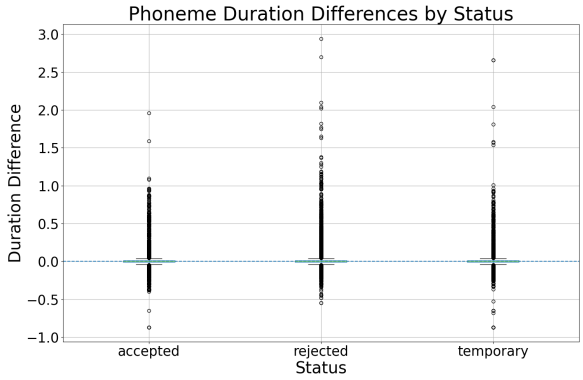


Figure 7: Duration difference distributions stratified by quality category. All three categories exhibit similar medians and interquartile ranges, indicating WER-based classification captures spectral modifications rather than temporal distortions.

All three status categories exhibit similar distributional characteristics. Median duration differences center near zero across categories, indicating phoneme-level perturbations do not introduce systematic temporal biases. REJECTED samples exhibit slightly higher concentrations of extreme outliers, suggesting extreme duration changes may contribute to rejection when combined with other factors. The similarity between ACCEPTED and

TEMPORARY categories indicates duration stability alone is insufficient for classification, suggesting spectral properties play crucial roles.

# B Scaling Factor Analysis

## B.1 Progressive Analysis

We tested six $\alpha$ values progressively across layers 0-15 using high 0.95 quantile. Figure 8 presents acceptance rates as functions of layer range and scaling factor.

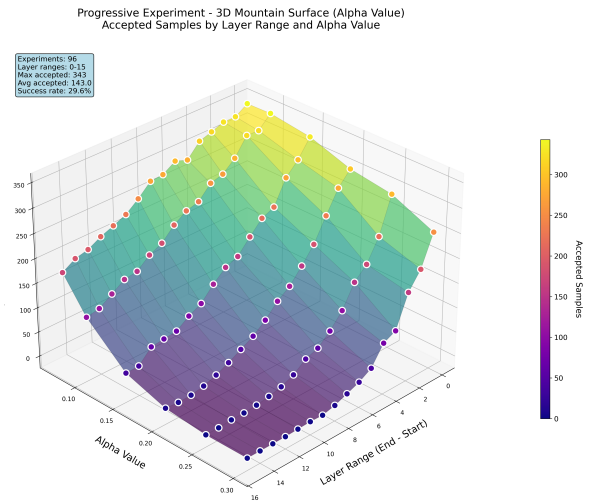


Figure 8: 3D surface plot showing acceptance rates vs. layer range and $\alpha$. Acceptance decreases with increasing layers and $\alpha$, reaching zero at $\alpha = 0.30$ for all layers. The $\alpha = 0.15$ configuration exhibits uniform decline, balancing perturbation effectiveness with quality preservation.

The progressive experiment reveals distinct behavioral regimes: **Conservative** ($\alpha \leq 0.10$) maintains high acceptance but may produce insufficient perturbation; **Optimal** ($\alpha = 0.15$) balances effectiveness with quality; **Aggressive** ($\alpha \geq 0.20$) causes dramatic degradation.

## B.2 Listening Study

Six scaling factors were evaluated using 123 REJECTED samples per configuration. Eight participants judged transcript-audio agreement; after filtering, five remained.

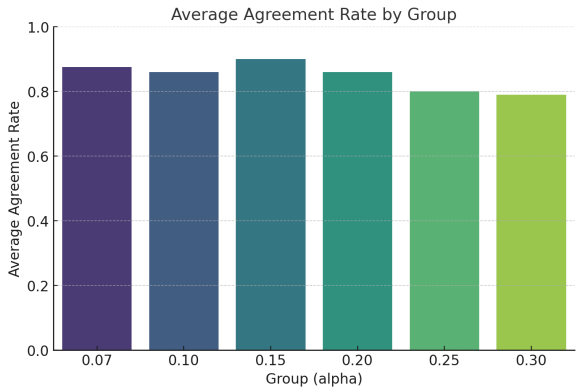


Figure 9: Agreement rates by $\alpha$ (n=5). Results suggest $\alpha = 0.15$ achieves the highest agreement.

Notably, the sample size of five participants is insufficient for robust statistical inference. We report results for completeness but rely primarily on automated metrics.

Majority voting revealed 116 of 122 REJECTED samples (95.1%) were judged intelligible by humans, suggesting WER-based classification is more restrictive than human perception. Surprisingly, $\alpha = 0.15$ (Figure 9) achieved the highest agreement, corroborating automated results.

# C Quantile Threshold Analysis

Three conditions were evaluated: high-quantile only, low-quantile only, and combined. All used $\alpha = 0.07$ with progressive layer accumulation.

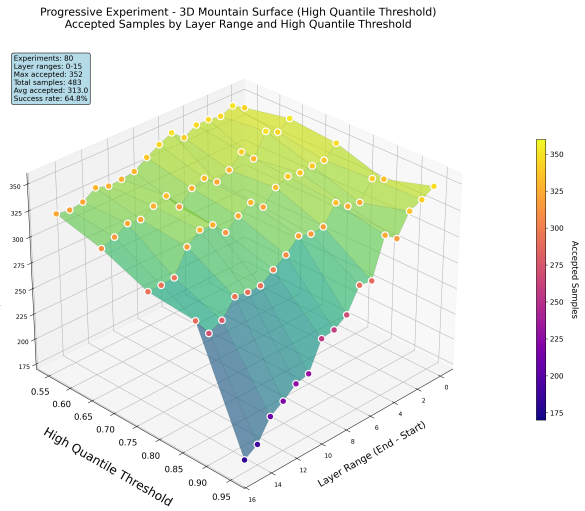


Figure 10: Acceptance rates across layer ranges and high-quantile thresholds. Values 0.55-0.75 maintain stable acceptance; sharp degradation at 0.85-0.95 indicates effective perturbation boundaries.

High-quantile results reveal two regimes (Figure 10): **Stable** ($q \leq 0.75$) with consistent accep-

tance, and **Sensitive** ($q \geq 0.85$) with sharp degradation. Low-quantile (Figure 11) shows uniform high acceptance (71.0%) regardless of threshold, confirming minimal acoustic effect.

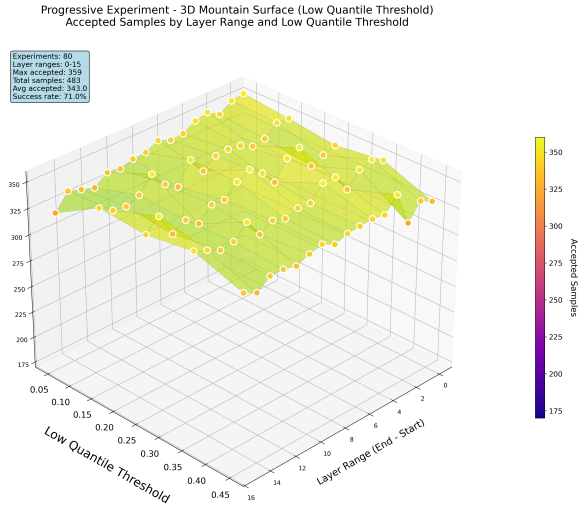


Figure 11: Acceptance rates across layer ranges and low-quantile thresholds. Values show uniform high acceptance regardless of threshold.

After verifying that each quantile is uniformly distributed and has the same number of vectors, we conducted a combined analysis of 25 pairs. Results reveal that high-quantile selection dominates; low-quantile contributions are diluted. For optimal effectiveness, high-quantile selection alone ($q = 0.95$) is sufficient and preferable.

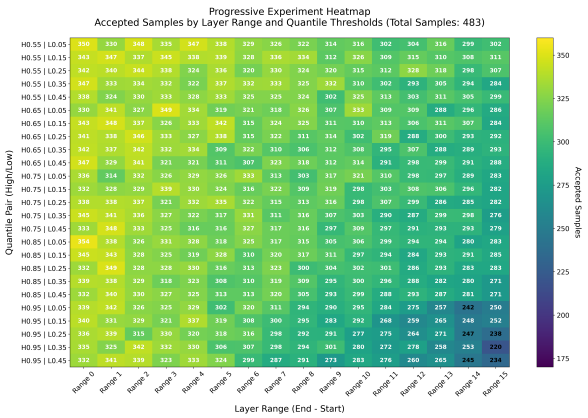


Figure 12: Acceptance rates for combined high/low quantile pairs. Results mirror high-quantile-only patterns, confirming high-quantile dominance.

## D Additional Split Results

Figure 13 presents SI-SDR gain analysis identifying 196 consistent improvers for the 50% split. Figure 14 presents an analysis identifying 212 consistent improvers for the 75% split.

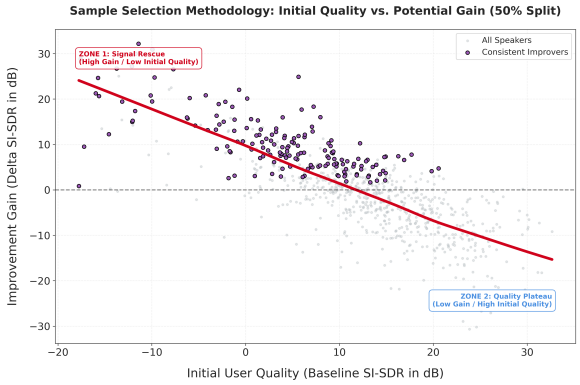


Figure 13: SI-SDR gain vs. baseline quality (50% split). Purple points indicate 196 consistent improvers. SI-SDR threshold patterns consistent with 25% split.

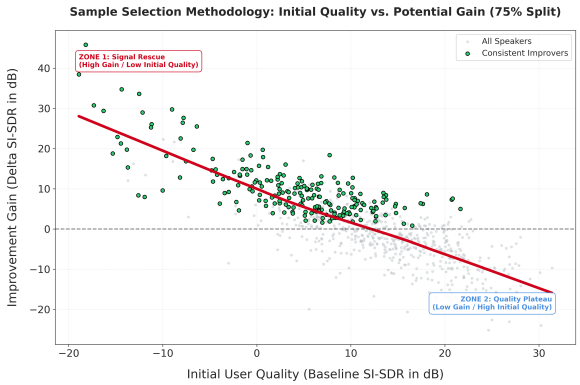


Figure 14: SI-SDR gain vs. baseline quality (75% split). Green points indicate 212 consistent improvers.

Figure 15 presents quality metrics for consistent improvers across both splits, confirming the scalability of perturbation-based training.

## E Token Distribution Uniformity

We verified that consistent improver groups do not differ in available evaluation data via Analysis of Variance (ANOVA).

**ANOVA Result:** $p = 0.2511$ (no significant difference in token counts).

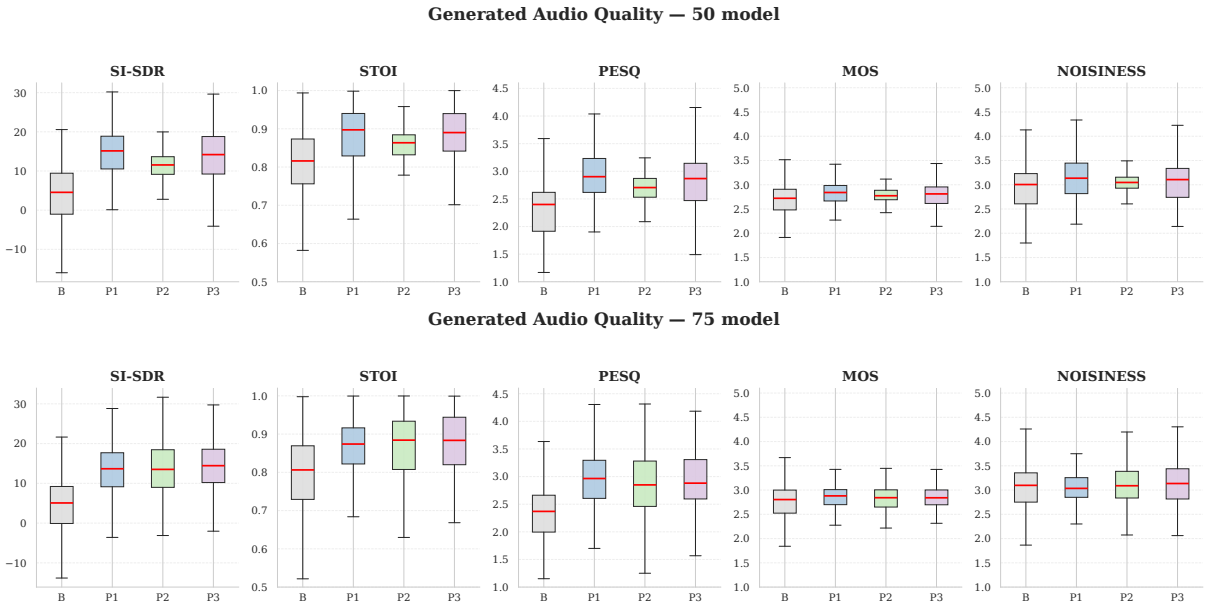


Figure 15: Quality metrics for consistent improvers across 50% split (196 speakers) and 75% split (212 speakers). Both scales show improvement from B to P1 with similar distributions across P1, P2, and P3, demonstrating the robustness and scalability of perturbation-based training.

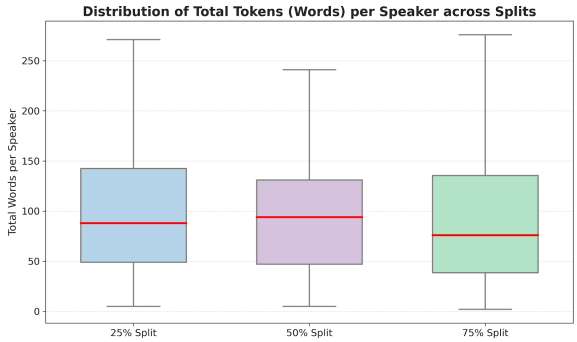


Figure 16: Token distribution across splits. Similar medians (77-98 words) and interquartile ranges indicate no selection bias.

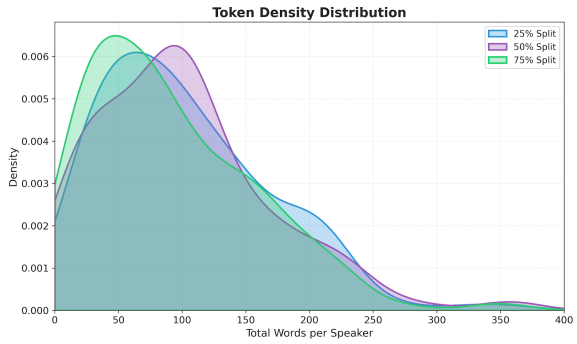


Figure 17: Token density distributions. Overlapping curves confirm equivalent data availability across groups.

This uniformity validates cross-split comparison (Figure 16, 17): the 75% split's superior performance reflects evaluation set size and model generalization, not selection bias toward speakers with more data.

## F Speaker Embedding Similarity

To quantify how perturbation affects speaker identity in the embedding space, we performed closed-set speaker identification using SpeechBrain's ECAPA-TDNN model (Ravanelli et al., 2021, 2024). For each speaker, a single reference embedding was computed from one randomly selected original utterance. For every perturbated utterance, we computed its ECAPA-TDNN embedding and compared it against all reference embeddings using the cosine similarity function provided by SpeechBrain's `SpeakerRecognition` module. The predicted speaker is the one whose reference embedding yields the highest cosine similarity.

**Accuracy** is the fraction of perturbated samples whose predicted speaker matches their true speaker:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{s}_i = s_i] \qquad (1)$$

where $s_i$ is the true speaker of sample $i$, $\hat{s}_i$ is the predicted speaker, and $N$ is the total number of perturbated samples.

**Jaccard index** (Jaccard, 1912) is computed per speaker and then averaged. For speaker $s$, let $\text{TP}_s$

be the number of samples correctly identified as $s$, $\text{FP}_s$ the number of samples from other speakers incorrectly assigned to $s$, and $\text{FN}_s$ the number of samples from $s$ incorrectly assigned to other speakers. The per-speaker Jaccard index is:

$$J_s = \frac{\text{TP}_s}{\text{TP}_s + \text{FP}_s + \text{FN}_s} \tag{2}$$

The reported Jaccard index is the mean across all speakers: $J = \frac{1}{|S|} \sum_{s \in S} J_s$. Unlike accuracy, Jaccard penalizes both misses and false alarms per speaker, providing a stricter measure of identification quality.

Table 3 reports results across all splits and perturbation runs.

| Split | Run | Accuracy | Jaccard |
|---|---|---|---|
| 25% | P1 | 0.1652 | 0.1002 |
| | P2 | 0.1645 | 0.0972 |
| | P3 | 0.1670 | 0.0990 |
| 50% | P1 | 0.1588 | 0.0926 |
| | P2 | 0.1604 | 0.0939 |
| | P3 | 0.1553 | 0.0909 |
| 75% | P1 | 0.1509 | 0.0882 |
| | P2 | 0.1517 | 0.0896 |
| | P3 | 0.1498 | 0.0877 |

Table 3: Closed-set speaker identification results for perturbated audio. Low accuracy (15-17%) and Jaccard index (9-10%) confirm that perturbation modifies speaker embeddings substantially. Values are consistent across P1, P2, and P3, indicating that the three independent runs produce comparable levels of speaker embedding shift.

Accuracy ranges from 15% to 17%, far below what would be expected if perturbation preserved speaker identity in the embedding space. The Jaccard index is lower still (9-10%) because it additionally penalizes cases where other speakers' samples are misattributed to a given speaker. Together, these metrics confirm that the perturbation introduces substantial acoustic variation at the speaker-characteristic level.

Two observations are relevant for interpreting these results. First, the ParlerTTS fine-tuning operates as a closed-set system: both original and perturbated audio are mapped to the same speaker IDs during training. The TTS model therefore learns to associate acoustically diverse input variants with a single speaker identity. The downstream quality improvements reported in Section 5 confirm that this closed-set mapping is effective: the model produces speaker-consistent output despite the embedding-level shift in its training data.

Second, accuracy is consistent across perturbation runs (P1, P2, P3) and across splits (25%, 50%, 75%), with a slight decrease for larger splits. This consistency mirrors the downstream TTS results, where P1, P2, and P3 yield similar quality distributions, reinforcing the conclusion that the three independent runs produce comparable perturbation characteristics.

## G Spectrogram Analysis

Figures 18 and 19 present mel spectrogram comparisons between original and perturbated audio with F0, F1, and F2 overlays extracted using Praat (Boersma and Weenink, 2021; Jadoul et al., 2018). Each figure contains four rows showing the same utterance from the same speaker: the original recording followed by three independent perturbation variants (P1, P2, P3). F0 contours (cyan) indicate fundamental pitch, F1 (green) and F2 (magenta) track the first two formant frequencies. Speakers were selected at random from each split, and no speaker is repeated across figures: every figure shows a unique speaker, ensuring that the observed patterns reflect general properties of the perturbation rather than speaker-specific artifacts.

### G.1 General Acoustic Comparison

Figure 18 shows spectrograms for selected male and female speakers across data splits.

Several patterns are consistent across all speakers and splits. F0 contours remain stable between original and perturbated versions, with voiced segments appearing at nearly identical frequency values and with the same temporal patterning. This confirms that the perturbation does not systematically shift fundamental frequency.

F1 and F2 trajectories follow broadly similar paths in original and perturbated audio, preserving the vowel-level articulatory structure. However, the perturbated variants exhibit increased frame-to-frame variability in formant tracks, particularly in F1, which shows more frequent local excursions. This is consistent with the perturbation acting on fine spectral detail within formant regions rather than shifting formant center frequencies.

The most visible difference between original and perturbated spectrograms is in harmonic structure. Original recordings display well-defined harmonic bands with clear inter-harmonic gaps. Perturbated variants show a more diffuse energy distribution: harmonic bands appear smeared, with energy re-

distributed into adjacent frequency regions. This effect is visible across all speakers and intensifies from P1 through P3, though all three variants produce qualitatively similar spectral patterns. The mid-to-high frequency range (1024-4096 Hz) shows the strongest perturbation effects.

### G.2 Quality-Based Classification Comparison

Figure 19 shows spectrograms grouped by WER-based quality category (accepted, temporary, rejected) to illustrate the acoustic differences that drive classification outcomes.

The three status categories show a clear gradient of spectral modification. In accepted samples, perturbated variants are nearly indistinguishable from the original in the mel domain. Harmonic structure is well preserved, F1 and F2 tracks closely follow the original trajectories, and the overall energy distribution remains intact. The perturbation produces subtle changes in fine spectral texture that are sufficient to generate acoustic diversity (ViSQOL $\approx$ 1.15) without challenging ASR boundaries.

Temporary samples show moderate spectral changes. The harmonic structure is partly preserved but with visible energy redistribution, particularly in the 1024-2048 Hz range. F1 and F2 tracks exhibit more frame-level jitter compared to the original, with occasional deviations of 100-200 Hz before returning to the original trajectory. F0 remains stable. These modifications push against ASR decision boundaries ($0.05 < \text{WER} \leq 0.15$) while preserving the overall phonetic structure.

Rejected samples display the most substantial spectral changes. Energy is visibly redistributed across frequency bands, harmonic definition is reduced, and F1/F2 trajectories become noticeably noisier with more frequent and larger excursions from the original paths. Despite these changes, two properties are preserved: F0 contours remain largely intact, and the broad formant regions (the frequency bands where F1 and F2 concentrate) maintain their general location. This visual evidence corroborates the quantitative finding from Section 4 that WER-based rejection reflects spectral and timbral modifications rather than a collapse of phonetic structure. The 95.1% human intelligibility rate for rejected samples (Section B.2) is consistent with this observation: Listeners decode linguistic content as temporal and formant structures persist despite altered spectral detail.

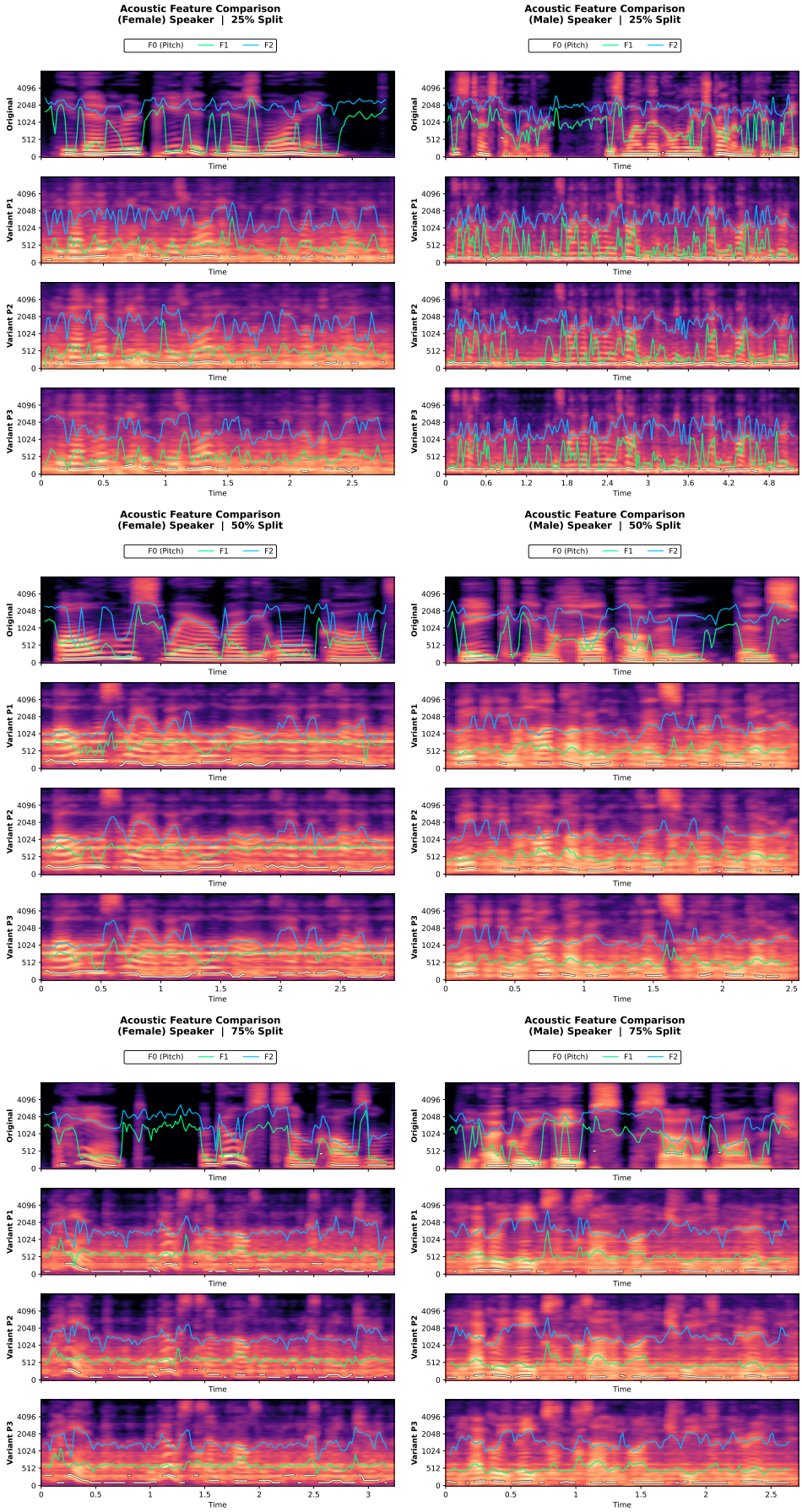


Figure 18: Mel spectrograms with F0 (cyan), F1 (green), and F2 (magenta) overlays comparing original and perturbated audio (P1, P2, P3) for male and female speakers across 25%, 50%, and 75% splits.

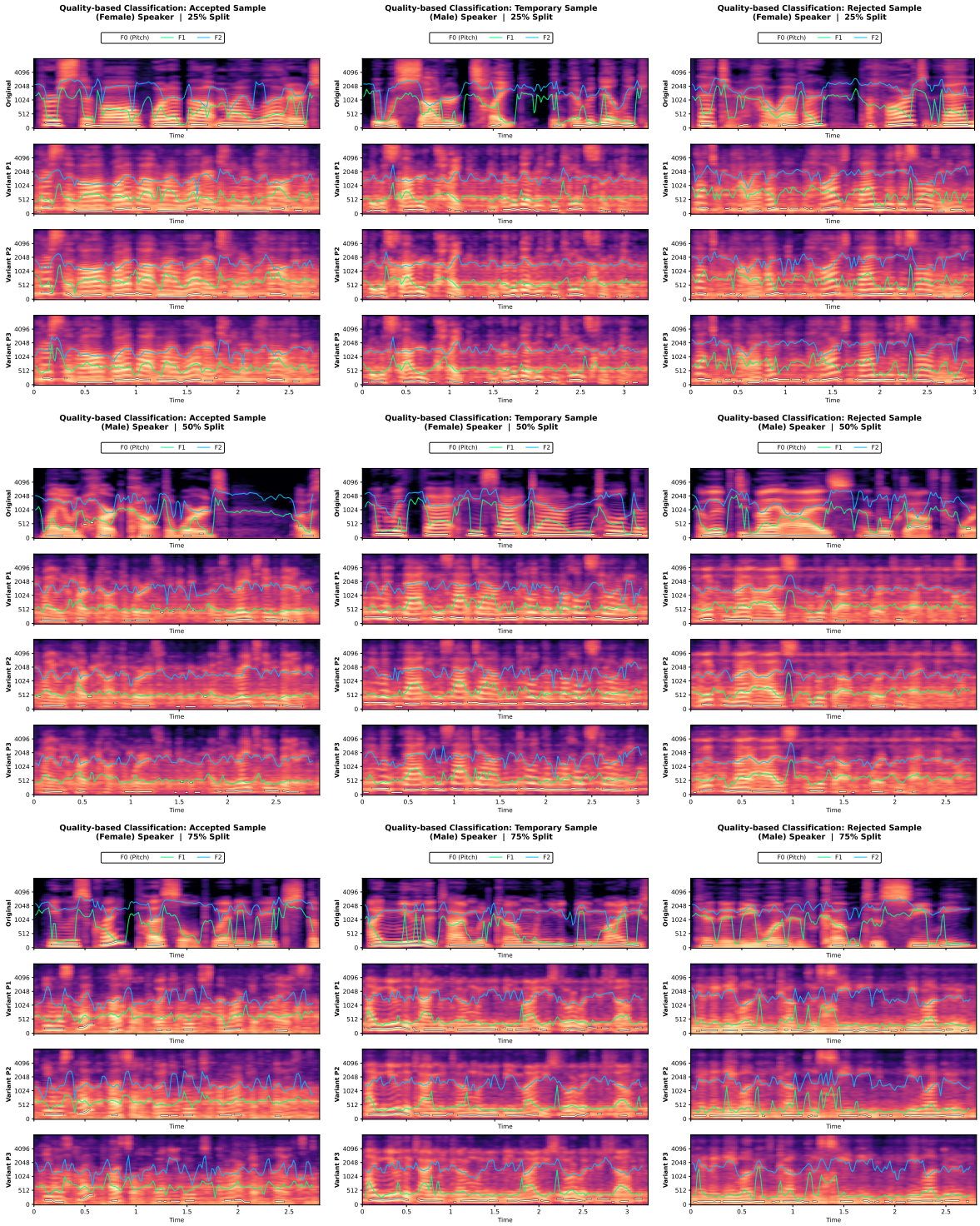


Figure 19: Mel spectrograms with acoustic feature overlays grouped by quality classification (accepted, temporary, rejected) across all three data splits. Each column corresponds to one status category; each row to one data split.