SAWT: Restoring Speech for Synthesis

Anchored Generation and Content Fidelity at 48 kHz

Mostafa Mahdi

Quran Lab University of Copenhagen rhk510@alumni.ku.dk

Preprint, September 2026. Not peer reviewed and not posted to arXiv. Weights, code and model card: https://huggingface.co/Quran-Lab/sawt-v4

Abstract. SAWT V4 restores speech at 48 kHz to provide cleaner acoustic targets for text-to-speech training. A deterministic content anchor conditions a 372M-parameter latent flow-matching generator, while inference-time checks test sampled windows for energy, content and speaker consistency. Training uses approximately 900 hours of dry speech in 35 languages; rented GPU time for development cost about US$1,200. Batched inference processes audio at 10 to 16 times real time on an RTX 4090. SAWT achieves the lowest recognised word error among the three evaluated restorers in all four reverberation conditions tested with baselines. At 3 s nominal RT60 and 20 dB SNR, word error is 0.520, compared with 0.815 for Sidon and 0.938 for the internal predecessor. On 200 LibriTTS-R utterances, SAWT reduces transcript drift by 38% relative to Miipher-1 (0.042 versus 0.068), increases speaker similarity (0.988 versus 0.962), and yields 65% fewer measured dropout frames. Reference word error is 0.077 versus 0.082, with no clear paired difference. On 120 archive recitations, SAWT reduces phoneme error by 19% relative to its predecessor (0.0876 to 0.0713) and raises DistillMOS by 0.123. On synthetic recitation development data, it reduces phoneme error by 48% relative to the damaged input (0.152 to 0.078). Optional recogniser-based selection further lowers archive phoneme error to 0.0684, with an uncertain paired improvement. A canonical-text audit and controlled reverberation sweep expose content errors that predicted quality alone can miss. Unprocessed inputs retain lower mean reference error on LibriTTS-R and the archive, severe reverberation leaves substantial errors, and Miipher-1 and Miipher-2 lead their respective public comparisons in predicted quality; Miipher-2 also leads the other reported aggregate measures on its 16 clips. SAWT combines practical restoration with an audit of content preservation; its effect on downstream TTS quality remains to be measured.

1 Introduction

Restored audio is becoming training data. Miipher [1] supplies restorations for LibriTTS-R [17], and Miipher-2 [2] targets restoration at million-hour scale. Sidon [3], DiTSE [4] and VoiceBridge [5] extend the available approaches. For a corpus with fixed transcripts, an actual content change can create a mismatch between a training target and its label. Acoustic improvement alone cannot establish that this risk is small. Figure 1 reports quality and reference word error across an entire evaluation set; Figure 4 illustrates how the two can disagree on an individual recording. We develop a restorer around this distinction and assess both the acoustic gains and the content errors that remain.

We began with Quran recitation. Muslims receive the Quran as the word of God, read on the page and heard in the human voice. Its invitation to contemplate its verses (Quran 38:29)[1] gives attentive listening a place in worship, learning and personal reflection. The practice of recitation joins that attention to a duty of accuracy: a learner listens, repeats and accepts correction before passing the words on. Someone unfamiliar with Islam can hear in these recordings the beauty of Arabic recitation and the seriousness with which its words are held. Preserving access to that encounter is one purpose of this work.

Tape hiss, echo and repeated compression now stand between many historical recitations and their listeners. Removing those obstacles can return performances to classrooms, homes and individual study. It also demands restraint. The restored voice must remain accountable to the recording, down to phonetic distinctions that a listener or a quality predictor might otherwise overlook. A fluent substitution is still a change to the recitation.

Recitation offers a useful evaluation setting: recordings can be aligned to a canonical text even when no clean acoustic reference survives. This permits phoneme-level comparison across recording conditions. The text is used for evaluation, not supplied to the restorer. SAWT estimates its conditioning from the audio, allowing use without a supplied transcript; performance outside the evaluated languages remains an empirical question.

The architecture assigns *reading* to a deterministic predictor, the anchor, and *rendering* to a conditional latent generator. The anchor estimates local speech structure and a clean latent from the damaged recording; the generator samples 48 kHz texture under those estimates. Every window is then decoded and checked for agreement with the anchor, for energy and for continuity with the preceding speaker representation; a failed check requests another sample, up to a fixed limit. This links conditioning and verification: the content representation guides the reconstruction and then tests the generated waveform. The optional phoneme selector extends that design to local linguistic distinctions.

SAWT V4 is the release version described here; its predecessor V3.1 was an internal system and appears here only as a paired reference on the archive. The paper makes six contributions.

1. **An anchored 48 kHz generative restorer with output checks.** Input-derived content estimates condition generation and provide a reference for checking the sampled

[1] See https://quran.com/38/29.

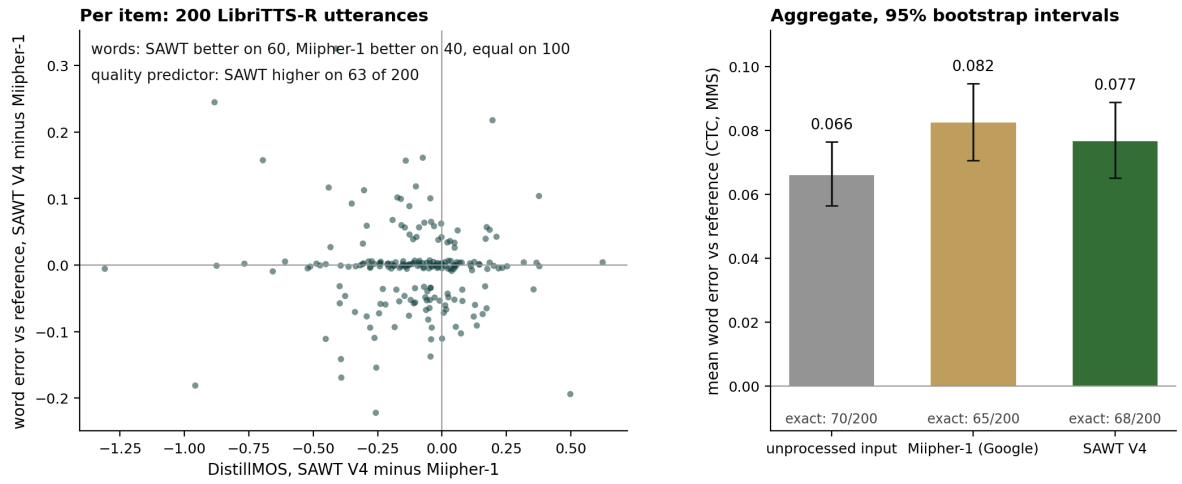

Figure 1: Quality and content across 200 utterances. The LibriTTS-R comparison in Table 4. Left: per-item differences between SAWT V4 and public Miipher-1 restorations in predicted quality and reference WER; 100 items tie on WER. Right: mean WER with 95% bootstrap intervals and counts of exact transcripts. The paired WER difference between restorers spans zero; this is not an equivalence test. The unprocessed input has the lowest mean WER, while DistillMOS favours Miipher-1 over SAWT.

waveform. Plain weights and client code are packaged for release; batched inference runs at 10 to 16 times real time on an RTX 4090 (Section 3).

2. **A canonical-text audit for historical recitation.** A fixed linguistic reference makes phoneme-level evaluation possible without a clean recording. We combine this audit with reference WER, transcript drift, speaker consistency, dropout counts and residual decay, measuring what restoration preserves as well as what it removes (Section 4).
3. **The lowest measured WER among the tested restorers in all four baseline reverberation conditions.** At 3 s nominal RT60 and 20 dB SNR, SAWT's WER is 36% lower than Sidon's and 45% lower than V3.1's. The wider sweep extends beyond the training decay range and measures content recovery alongside residual tails (Section 5.1).
4. **Quantified preservation gains on public speech and archives.** On LibriTTS-R 200, SAWT has 38% lower transcript drift, 65% fewer measured dropouts and higher speaker similarity than Miipher-1. On archive recitation, it reduces PER by 19% and raises DistillMOS by 0.123 relative to V3.1. These gains concern the named measures and comparisons; input-reference errors and quality deficits are reported alongside them (Sections 5.2 and 5.3).
5. **Phoneme-informed candidate selection with a measured development gain.** Choosing among three candidates by recognised phoneme drift reduces synthetic development PER by 14% relative to the default, from 0.0784 to 0.0677, with a paired interval excluding zero. It requires no supplied transcript. The archive follow-up gives a smaller, uncertain gain, and evaluation shares the selection recogniser (Section 5.4).
6. **An empirical demonstration of a quality metric's content blind spot.** DistillMOS remains between 3.8 and 4.3 across a reverberation sweep whose WER reaches 0.96. Error alignments further identify consonant deletions that pass feature checks. These results give concrete targets for evaluating and improving restoration for training data (Sections 5.6 and 7).

Dated evaluation records in the research repository support these comparisons. Appendix B describes the reproduction materials and their publication status.

2 Background

Feature prediction. Miipher [1] predicts clean features from speech and text representations and reconstructs audio through a vocoder. Miipher-2 [2] adapts features from a frozen Universal Speech Model and synthesises the waveform with WaveFit. Sidon [3] makes a related approach available with w2v-BERT 2.0 and a neural vocoder. SAWT draws on this family of methods for its anchor. Their deterministic operation gives repeatable estimates, but those estimates can still contain content errors.

Latent generation. DiTSE [4] reconstructs speech with a diffusion transformer operating in a learned latent space. VoiceBridge [5] couples latent bridge modelling to an energy-preserving variational autoencoder. Both seek high acoustic quality while retaining characteristics of the input. SAWT uses a separately trained predictor to condition latent generation and to examine the sampled waveform during inference.

Audio representations. Codec choice determines the signal the generator must model and the reconstruction errors it cannot avoid. We compared DAC-VAE [6], EAR-VAE and KVAE on reconstruction quality, periodicity, band-limited log-spectral distance, intelligibility and sensitivity to latent noise. DAC-VAE was selected: its 128-channel representation runs at 25 Hz for 48 kHz audio. Section 3.1 gives the reconstruction comparison and decoder adaptation procedure.

Preservation measures. DistillMOS [14] and DNSMOS [16] estimate aspects of perceived quality. Assessing what survives restoration requires additional measurements. We use recogniser error, speaker similarity, phoneme error against the recitation text, local energy dropouts and residual decay.

Together they reveal changes that a quality score alone can conceal; each remains subject to the errors of its underlying instrument.

3 The SAWT V4 system

SAWT processes the same recording along two paths (Figure 2). Resampling and level normalisation precede codec encoding, which supplies the damaged latent. A 16 kHz view enters the anchor to produce speech conditioning. For each 8 s window, the generator follows a learned flow for 64 Euler steps. The decoder maps the sampled latent to 48 kHz audio, which is checked before the window is committed; failing windows may be sampled again.

3.1 Representation

The codec must support both accurate reconstruction and generation in its latent space. Table 1 reports the initial English, multilingual and recitation probes used in selecting DAC-VAE. We freeze its encoder and adapt the decoder for 40,000 steps on dry audio, using sixteen 1 s clips per batch and latent-noise augmentation. This decoder, D2, uses the bare output projection without the additive watermark branch. Channel-wise standardisation is estimated from 2,000 clean clips.

The final decoder warrants a separate comparison from the initial codec probe. Table 2 reports D2 after 40,000 adaptation steps. Its reconstruction changes DistillMOS by +0.031, −0.048 and +0.046 on English, multilingual speech and recitation, respectively, with small changes in PQ and periodicity. On Miipher-2 outputs, codec reconstruction scores 4.376 DistillMOS and 7.36 PQ, against 4.372 and 7.38 for the reference waveforms. The probe also records STOI of 0.990 and a PER increase of 0.007. These measurements characterise codec reconstruction; restoration introduces the additional error of predicting the latent from damaged audio.

Spectral measurements qualify the quality scores. D2's masked 8 to 24 kHz LSD is 11.3, 14.8 and 12.1 dB, versus 7.5, 7.7 and 8.3 dB for the mel/BigVGAN reference path on the same three sets. Fricative LSD is closer: 7.5, 7.9 and 7.7 dB versus 7.2, 7.2 and 7.4 dB. The latent-noise standard deviations associated with a 0.10 DistillMOS loss are 0.10, 0.12 and 0.07. D2 therefore supports the chosen representation while retaining spectral and noise-sensitivity limits that a near-constant quality score would miss.

3.2 Anchor

A frozen 600M-parameter w2v-BERT 2.0 network [7] forms the anchor backbone. Rank-64 LoRA adapters [8] are inserted into the attention and feed-forward projections of encoder blocks 1 to 19, giving 42.3M trainable adapter parameters. The anchor returns cleaned layer-8 features at 50 Hz with 1,024 channels, the mean of layers 8 to 20 (the *stack*), a three-channel pitch estimate and a predicted clean codec latent. A six-block Conformer-lite produces the latent estimate from the stack, an 80-band mel representation of the 48 kHz input and a bandwidth bin. The file-level speaker vector is the duration-weighted mean stack. We train for 60,000 steps on paired degraded and clean views, minimising mean squared error against teacher features and L1 error on the latent estimate. The anchor's outputs are fixed for a given input.

3.3 Generator

Our generator has 372M parameters: a diffusion transformer [9] with 24 layers, width 1,024 and 16 attention heads. Blocks use adaLN-single modulation, rotary position embeddings [10] and query-key normalisation. The model receives the evolving sample x_t, the standardised damaged latent z and the anchor's clean-latent estimate. Local conditioning streams are pooled from 50 to 25 Hz, concatenated with z and the estimate, RMS-normalised, and projected through a gate into every block. Global conditioning contains the projected speaker vector, a bandwidth bin and a room token encoding RT60 and C50 bins. Training uses room measurements from the target; inference selects the *dry* token. Learned null vectors implement conditioning dropout and permit classifier-free guidance. The default evaluations use guidance scale 1; Section 5.7 separately reports the guidance sweep.

We learn the transformation from Gaussian noise to a clean latent by flow matching [11]. With target x_1, initial state $x_0 \sim \mathcal{N}(0, I)$ and interpolated state $x_t = (1-t)x_0 + tx_1$, the network predicts the clean endpoint $\hat{x}_1$. This prediction defines the velocity

$$v_\theta(x_t, t) = \frac{\hat{x}_1 - x_t}{1-t}. \tag{1}$$

The loss is mean squared error between v_θ and the target velocity $x_1 - x_0$. We sample training time from a logit-normal distribution and cap it at 0.98. Unit Gaussian initialisation was retained after comparison with a start centred on the anchor estimate, which performed less well during development.

3.4 Training and data

Clean targets. We screened the candidate recordings for background energy, speech-to-floor difference, decay, hum, clipping and bandwidth. The thresholds included a maximum floor of −50 dBFS, at least 30 dB between speech and floor, and no more than 15 dB of hum. Accepted recordings with limited bandwidth retain a corresponding tag. The merged screening manifest records 593,235 candidates, including rejections; the audit estimates approximately 900 retained hours from the corpora in Table 3. LibriTTS-R, FLEURS-R, CML and JVS do not supply 48 kHz generator targets, although lower-rate data are used for anchor training. The exclusion list covers 165 recitation identities and fingerprint-matched copies, as well as speakers used in Miipher-16, Phase C, LibriTTS-R 200 and Miipher-dev.

Training damage. We generate degraded inputs on the GPU as training proceeds. Room impulse responses span RT60 values of 0.15 to 1.6 s and include 595 measured responses. A separate hall bank combines octave-band decay profiles, a band-limited public-address path and delayed copies. Additive noise is drawn from 239,032 non-speech AudioSet clips [20], DNS5 and FSD50K, with SNR sampled uniformly between −5 and 20 dB. Mains hum, MP3 compression at 65 to 245 kbps, packet loss and archive-style bandwidth restrictions supply further damage. Clean latents are cached; the frozen encoder and anchor run on each newly degraded input,

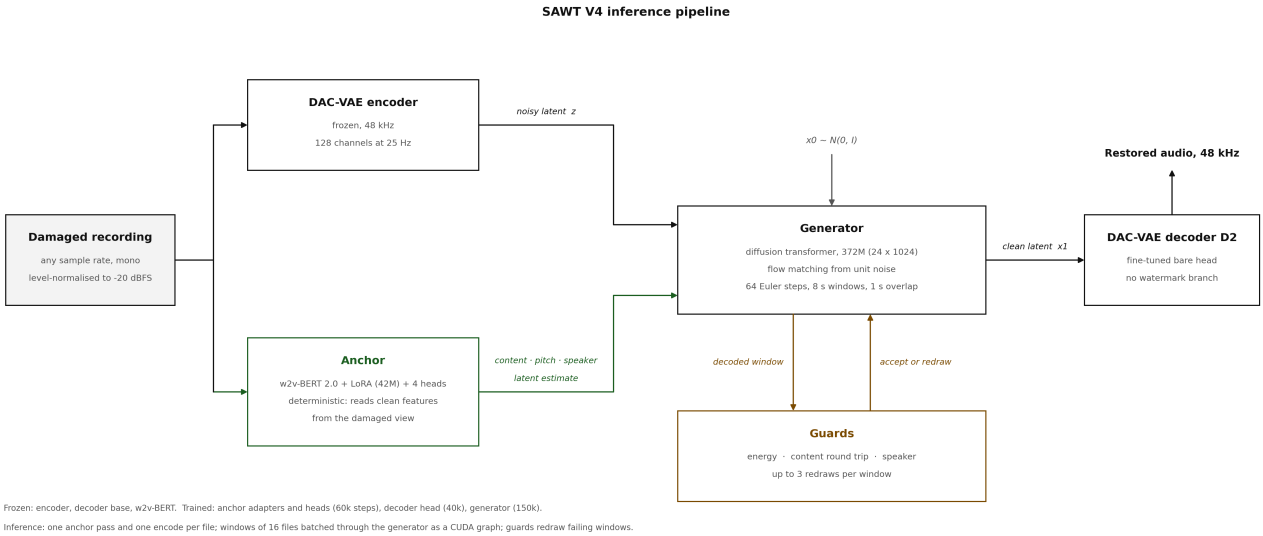

Figure 2: From damaged audio to a checked reconstruction. The input supplies a codec latent z and anchor features. Conditioned on both, the generator integrates from Gaussian noise x_0 towards a clean latent x_1. Decoder D2 reconstructs the 48 kHz waveform. Checks compare the result with input-derived features and the preceding window, requesting redraws where needed. Windows that remain below threshold are flagged.

latent	English 40	multilingual 40	recitation 40	Miipher-2 renders	noise σ at −0.10
DAC-VAE (Meta)	4.303 → **4.395**	4.139 → **4.083**	3.793 → **3.712**	4.372 → **4.366**	0.06 to 0.13
KVAE	4.303 → 4.108	4.139 → 4.055	3.793 → 3.704	4.372 → 4.294	**0.11 to 0.15**
mel + BigVGAN (V3 path)	within 0.01				0.01 to 0.04

Table 1: Codec reconstruction before decoder adaptation. Entries pair the DistillMOS of a clean reference with that of its reconstruction. The final column gives the latent-noise standard deviation producing a 0.10 score loss. Periodicity changes stay within 0.009 for both VAEs, compared with 0.012 to 0.058 for the mel path. Noise tolerances are specific to the representation and its scaling.

set	ΔDistillMOS	Δperiodicity	PQ: clean → D2	LSD 8 to 24 kHz	fricative LSD
English	+0.031	−0.004	7.50 → 7.52	11.3	7.5
Multilingual	−0.048	+0.001	6.99 → 6.99	14.8	7.9
Recitation	+0.046	−0.005	7.25 → 7.28	12.1	7.7

Table 2: Reconstruction with the final decoder D2. Differences subtract the clean-reference value; LSD denotes signal-masked log-spectral distance in dB, with lower values preferable. Small changes in predicted quality and periodicity coexist with measurable spectral error. These are codec probes, not complete restoration results.

corpus	languages
Bible-TTS [19]	Hausa, Yoruba, Ewe, Asante Twi, Akuapem Twi, Kikuyu, Lingala, Luganda, Luo, Chichewa
OpenSLR high-quality sets	Spanish (five Latin American varieties), Catalan, Basque, Galician, Nigerian English, English dialects, Tamil, Telugu, Kannada, Malayalam, Marathi, Gujarati, Nepali, Javanese, Sundanese, Khmer, Burmese
EARS [18]	English, anechoic and expressive
AISHELL-3	Mandarin
VCTK	English, 109 speakers
GTSinger	singing in nine languages
Arabic recitation, Arabic Speech Corpus	Arabic
DAPS clean, SIWIS, Expresso, JSUT, Opencpop, VocalSet, M4Singer	English, French, Japanese, Mandarin singing

Table 3: Sources of screened training targets. The merged audit estimates 900 hours in 35 languages. Accepted recordings with restricted bandwidth carry a tag describing their cutoff.

so degraded inputs vary across training steps.

Optimisation and cost. Eight H100 GPUs train the main model for 150,000 steps. Each GPU receives 32 five-second clips, for a global batch of 256 and a reported step time of 0.48 s. The learning rate reaches 10^{-4} after 5,000 warm-up steps and decays over the last 20% of training. An exponential

moving average uses coefficient 0.9999. Development comparisons of reconstruction, temporal variation and content preservation guide model selection. Rented GPU time for the full programme, covering the codec probes, decoder adaptation, anchor, generator and ablations, cost approximately US$1,200.

Acoustic simulation. Inspection of the impulse responses exposed a defect in the original hall-bank generator, which we corrected. A classifier could still distinguish real and simulated 2 s excerpts. Other recording differences confounded that comparison, leaving the realism of the reverberation model unresolved. We therefore use the simulations to control damage conditions and evaluate real recordings separately.

3.5 Inference

Windowed sampling. The evaluated sampler takes 64 Euler steps from seed 0 with guidance scale 1. Windows are 8 s long and overlap by 1 s. Indexing a shared noise field by absolute latent frame gives adjacent windows consistent noise. At each integration step, the overlap is replaced by a noised version of the latent already committed for the preceding window.

Output checks. Full-recording development tests exposed low-energy collapse that short evaluation crops had missed. We apply three checks to each sampled window. The standard deviation of its standardised latent must reach 0.70; anchor features recomputed from the decoded audio must have cosine agreement of at least 0.85 with the input-derived features; and WavLM-SV [12] cosine similarity to the preceding window must reach 0.90. Failure triggers a redraw, with at most three retries. Development probes, including the reverberation sweep, informed the content threshold. The final waveform follows the input's short-term dynamics, and a per-file sidecar identifies windows still below threshold, which gives archives a refuse-or-review path. Section 5.4 adds an optional fourth check that reads phones.

Throughput. On the evaluation workstation, an unbatched step takes approximately 65 ms despite requiring only 2 ms of arithmetic. To reduce launch overhead, the fast implementation processes windows from 16 files together and executes the generator step as a CUDA graph. Each file is encoded once. An RTX 4090 then processes 10 to 16 seconds of audio per second using 14.5 GB of GPU memory. Measured differences from the unbatched reference are consistent with bf16 rounding. At that measured range, processing 1,000 hours would take approximately 63 to 100 GPU hours, excluding data transfer and scoring. Recording length and retry frequency affect throughput.

4 Experimental setup

Evaluation material. The recitation archive contains 120 clips and 9 long pilot tapes. We assess generalisation using Google's 16 demonstration inputs and Miipher-2 outputs; 200 LibriTTS-R test-other utterances with public Miipher-1 restorations [17]; 120 real recordings spanning varied channel damage (general real-120); and the Phase C set of 200 read-speech items. The two 120-item collections are separate. Development uses Miipher-dev 48, English 40, multilingual 40 and synthetic recitation 120, supplemented by the reverberation sweep and Danish listening probe. V3.1 denotes the lab's internal predecessor, which was never released; it appears as a paired reference where it was measured.

Paired analysis. We align the outputs and match their speech loudness to the input before scoring. The sampling seed and final configuration were fixed ahead of the final archive and general speech evaluations. Paired uncertainty estimates use percentile bootstrapping with 10,000 resamples and seed 0. Unless stated otherwise, intervals use the 2.5th and 97.5th percentiles (95% coverage). The DistillMOS comparisons with Miipher-1 and Miipher-2 retain the original 5th and 95th percentile bounds (90% intervals, or separate one-sided 95% limits). Recognition rates are means of per-file rates rather than pooled token counts. Results from configuration development are marked as such.

Words and voice. The lab's zipformer phoneme recogniser (CTC, greedy decoding, run through sherpa-onnx) transcribes recitation as phonemes, which are compared with the canonical text to obtain phoneme error rate (PER). For LibriTTS-R and Phase C, word error rate (WER) compares recognised speech with the corpus text. For the synthetic development sets and reverberation sweep, it compares recognised restored speech with the transcript recognised from the clean recording; these references themselves can contain recognition errors. The CTC path uses MMS-1B, language-specific adapters and greedy decoding. Whisper supplies transcript drift on the demonstration set and WER for Phase C. Drift compares recognised output with recognised input, so it measures change rather than correctness against human annotation. Speaker consistency is the cosine similarity of WavLM embeddings from the input and output; clean-reference similarity is also used when available. Interpretation depends on both the recogniser and the chosen reference.

Local signal loss. We measure dropouts in 50 ms frames. Input frames are retained when their energy exceeds the file median by more than 6 dB. After correcting the median output/input level difference on these frames, we count those attenuated by more than 15 dB in the output. The reported dropout percentage is their share of the retained frames. It detects local energy loss, including possible noise removal, and must not be read as a percentage of deleted words.

Acoustic quality. Predicted quality is measured with DistillMOS [14] and the production-quality output (PQ) of Audiobox Aesthetics [13]. The fifth percentile of 20 ms frame energy estimates the floor; subtracting it from median active-frame energy gives speech-to-floor. To measure residual decay, we select speech offsets followed by at least 500 ms of detected inactivity and measure energy 100 to 300 ms after the offset. We report that tail both in absolute level and relative to the floor.

5 Results

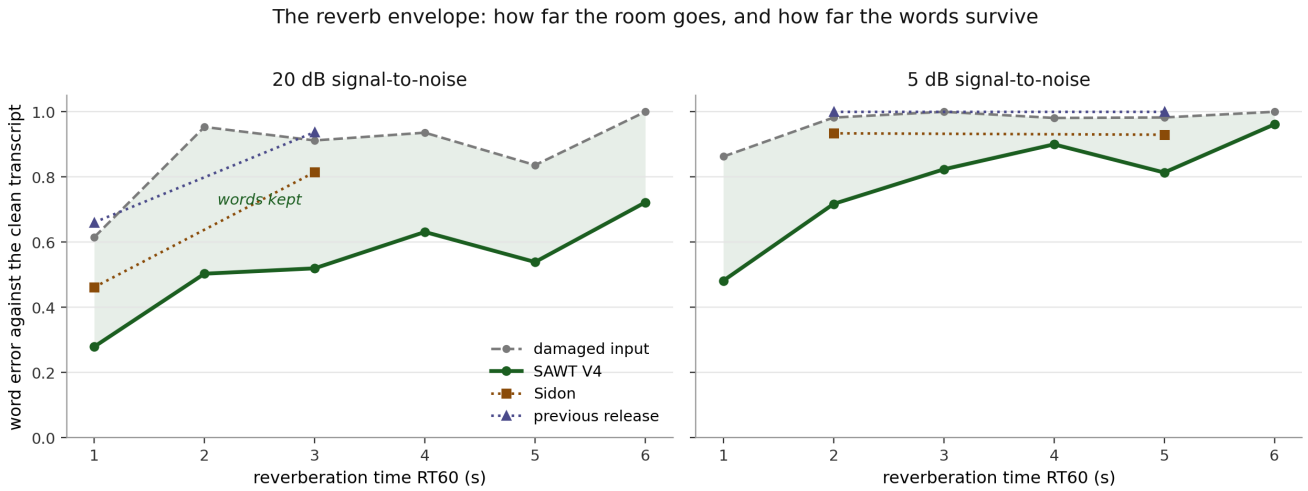

Figure 3: The reverb envelope: how far the room goes, and how far the words survive. Each of 24 utterances is degraded at six nominal RT60 values from 1 to 6 s and at SNR 20 or 5 dB, yielding 288 inputs. SAWT is solid green, the damaged input dashed grey, and the shading marks the intelligibility SAWT gives back. Markers locate the four conditions evaluated with Sidon and V3.1. The utterances were excluded from training. The sweep used the earlier inference recipe (content threshold 0.80, without the speaker check or dynamics matching) and subsequently informed guard calibration.

5.1 Reverberation beyond the training range

SAWT ranks first in recognised word preservation among the three tested restorers in all four baseline comparison conditions. At 3 s nominal RT60 and 20 dB SNR, its WER is 36% below Sidon and 45% below V3.1. Figure 3 covers six nominal RT60 values from 1 to 6 s, reaching 7.3 s measured decay, beyond the training hall bank's 4.5 s maximum. Absolute output tails in the four baseline conditions are −64 to −69 dBFS. Low tail energy, however, does not establish content recovery. At 20 dB SNR, SAWT reduces WER from 0.62 to 0.28 at 1 s, from 0.91 to 0.52 at 3 s, and to 0.72 at 6 s. At 3 s and 20 dB, WER is 0.520 for SAWT, 0.815 for Sidon and 0.938 for V3.1. At 3 s and 5 dB, input WER is 1.016 and SAWT WER is 0.823; the two baselines were not evaluated at that cell. This is substantial recovery relative to the damaged input, with substantial residual error. Such outputs need content review before use as labelled training targets. DistillMOS remains between 3.8 and 4.3 across the sweep, despite the wide range of WER. These results describe the simulated conditions; RT60 and SNR alone do not establish performance in a particular mosque or hall.

5.2 Content and speaker preservation on general speech

SAWT preserves more of the input representation than Miipher-1 on three measured axes: transcript stability, speaker similarity and local energy continuity. On 200 LibriTTS-R utterances, transcript drift is 38% lower and measured dropout frames are 65% fewer (Table 4, Figure 1). Against the reference text, CTC WER is 0.077 for SAWT, 0.082 for Miipher-1 and 0.066 for the unprocessed input. The paired SAWT-minus-Miipher-1 difference is −0.0058, with interval [−0.0149, +0.0039]. This comparison does not establish either superiority or equivalence. SAWT has lower WER on 60 items, Miipher-1 on 40, with 100 ties; exact transcripts number 68, 65 and 70 for SAWT, Miipher-1 and input, respectively. SAWT-minus-input WER is +0.0106, with interval [+0.0038, +0.0182]. Lower drift should therefore not be mistaken for lower reference error: SAWT's transcript drift is 0.042 against Miipher-1's 0.068, speaker similarity is 0.988 against 0.962, and measured dropouts are 0.67% against 1.90%. DistillMOS favours Miipher-1 by 0.106 (SAWT-minus-Miipher-1 90% interval [−0.135, −0.078]). On the Miipher-dev 48 development set, the quality difference from Sidon is −0.147.

Miipher-2 leads on the 16 public demonstration clips. SAWT scores lower in DistillMOS on every item: the mean difference is −0.534, with a 90% interval of [−0.857, −0.285]. Aggregate transcript drift, speaker similarity and dropouts also favour Miipher-2. Inspection identifies a persistent collapse on a 5.6 s recording, little change to two clips of 2.3 to 2.5 s, and a poor Maori restoration that passes the checks. Excluding those four post hoc still leaves a median quality deficit of about 0.30 on the remaining twelve; the headline comparison includes all 16. Miipher-2's restoration components were trained on 3,000 hours of speech [2]; its million-hour description concerns intended processing scale, not restoration-training duration. The public examples establish a deficit on this set, without estimating performance across all archival conditions. Section 7 reports a short-input diagnostic and proposed follow-up work. All 16 inputs, the Miipher-2 restorations and the SAWT restorations are published at `https://comparison.quranlab.ai`, loudness levelled, with the per-clip measurements and a blind listening mode.

On general real-120, the hardest real recordings (halls, telephone, radio), SAWT attenuates 7.94% of selected frames beyond the dropout threshold against 4.90% for the internal predecessor, upper confidence limit +5.4 percentage points on the 100 files with paired non-missing dropout values. The displayed arm means summarise the 120-item set. On Phase C 200 the two read 0.0721 and 0.0738 word error (input 0.0710), a difference without clear evidence either way.

set	against	DistillMOS	faithfulness
LibriTTS-R 200	Miipher-1 outputs	−0.106	word error vs reference 0.077 vs 0.082 (input 0.066); transcript drift **0.042** vs 0.068; speaker **0.988** vs 0.962; dropouts **0.67** vs 1.90%
Google's 16 clips	Miipher-2 outputs	3.83 vs 4.36	drift 0.385 vs 0.375; speaker 0.957 vs 0.974; dropouts 1.18 vs 0.87%
Miipher-dev 48	Sidon	−0.147	lower CTC drift in development comparisons
Phase C 200	V3.1 (internal)		WER 0.0721 vs 0.0738 (input 0.0710)
General real-120	V3.1 (internal)		dropouts 7.94 vs 4.90%; difference upper limit +5.4 percentage points

Table 4: Paired general speech comparisons, with SAWT listed first. DistillMOS differences subtract the reference-system score. Lower word error, drift and dropout values are preferable; higher speaker similarity is preferable. Word error is measured against the reference text; drift is the change in recognised words relative to the damaged input. Miipher-dev informed development.

5.3 The recitation archive

On archive recitation, predicted quality, speaker consistency and phoneme error all improve over V3.1. The 19% relative PER reduction accompanies an increase in predicted quality (Table 5). The recordings include low-bitrate uploads, tape and halls. The DistillMOS difference is +0.123, with interval $[0.078, 0.171]$; speaker consistency is 0.941, an increase of 0.018 with lower confidence limit 0.013. The measured floor is −69.6 dBFS, with 47.2 dB speech-to-floor. Absolute tails on the subset with digitally silent pauses lie between −66 and −101 dBFS. These acoustic measures show suppression of background and decay, without certifying the words. On heavily damaged synthetic recitation with clean-reference transcripts, PER falls from 0.152 at the input to 0.078 after restoration: a 48% relative reduction. This shows substantial content recovery in the synthetic development setting, where clean-reference audio is available.

Against the canonical text, archive PER is 0.0713 for SAWT, 0.0876 for the predecessor and 0.0618 for the source recordings. The paired difference from the predecessor is −0.016, with interval $[-0.028, -0.005]$; the difference from the input is +0.0095, with interval $[-0.0001, +0.0183]$. The input thus retains the lowest mean PER, and the interval does not establish an absence of degradation. Clean synthetic recitation scores 0.0589 against its canonical text, but this is a different collection and does not establish an irreducible recogniser floor for the archive. By channel, input-to-SAWT PER is 0.060 to 0.080 for low bitrate, 0.072 to 0.070 for mid bitrate and 0.050 to 0.060 for high bitrate. Two restored items exceed their input error by more than 0.10. PER measures recogniser output: channel-dependent recognition errors and actual speech changes can both affect it. Independent transcription or listening is needed to separate them; a quality gain does not cancel the observed PER increase.

The relative tail statistic needs one caveat. Its median is +1.33 dB, its 90th percentile +23.3 dB, and the spread comes from the denominator: 18 of the 77 recordings with a measurable tail contain digitally silent pauses at the numerical floor of −120 dBFS, so their tails read far above a floor no listener can hear. Their absolute tails lie between −66 and −101 dBFS. Removing those recordings lowers the 90th percentile to +11.6 dB. Both statistics are reported because they describe different things.

Where the added phone error comes from. A separate alignment audit of unlevelled outputs records 48 additional errors over the input across 3,772 reference units. Consonant deletions increase from 25 to 62, consonant substitutions from 81 to 85, marker deletions from 2 to 8, and insertions from 55 to 59; length errors remain at 22 and short-vowel errors fall from 17 to 13. These counts use a different preprocessing condition from the level-matched headline PER and should not be used to reconstruct it. Of 120 items, 45 worsen, 16 improve and 59 tie; the ten worst account for 77% of the added error. None is flagged by the guards, whose semantic cosines lie between 0.85 and 0.95. Informal listening suggests weak or missing consonants in affected passages. A fine-dynamics correction using 20 ms envelopes and boosts of at most 6 dB changes development PER by +0.0008, with interval $[-0.0033, +0.0057]$, on 40 clips. This null result does not isolate a cause, but suggests that simple level correction is insufficient. The alignment audit identifies local consonant preservation as a target for further training and independent listening.

5.4 Candidate selection by a recogniser

Feature agreement can miss local phoneme errors. We therefore draw three candidates per window and select the guard-clean candidate with the smallest phoneme drift from the input, breaking ties by semantic cosine. The same phoneme recogniser supplies the input and candidate transcripts. No externally supplied transcript or canonical text is required. A control selects among the same three draws using cosine alone. Because the selection recogniser also supplies the evaluation transcripts, the resulting gains may partly reflect optimisation for that recogniser; an independent recogniser and human assessment are needed to establish broader content preservation.

measure	SAWT V4	V3.1 (internal)	input
Phoneme error vs canonical text	**0.0713** (0.0684 with recogniser selection)	0.0876	0.0618
Tail relative to floor: median / 90th percentile	+1.33 / +23.3 dB		
Supplementary absolute tail, 18 silent-pause files	−66 to −101 dBFS		
Floor / speech-to-floor	**−69.6** dBFS / **47.2** dB		
Speaker consistency	**0.941**	0.923	
DistillMOS / PQ difference from V3.1	**+0.123** / −0.016		

Table 5: Measurements on the recitation archive of 120 clips and 9 long tapes. Metric availability varies: 77 recordings have measurable relative tails, while 129 contribute background levels. For recordings with silent pauses, absolute tail levels clarify the large ratios to the numerical floor.

	one draw	best of three, chosen by	
	(recipe D)	cosine	recogniser
synthetic 120 (clean transcript), development			
PER vs transcript	0.078	0.073	**0.068**
DistillMOS	3.68	3.68	3.69
real 40 (canonical text), development			
PER vs canonical text	0.080	0.079	**0.076**
drift from input	0.064	0.065	**0.055**
DistillMOS	3.82	3.79	3.80
archive 120 (canonical text), fixed-setting follow-up			
PER vs canonical text	0.071	not run	**0.068**
drift from input	0.053	not run	**0.048**
DistillMOS	3.83	not run	3.81

Table 6: Candidate selection. Three draws per window, the kept one chosen by the semantic cosine or by phoneme drift from the input window under the lab recogniser; the frozen recipe D draws once. Lower PER and drift are preferable; the archive source reads 0.062. The archive run was made once, after the development sets had chosen the setting.

Phoneme-based selection adds a 14% relative PER reduction on synthetic development data without retraining the generator. Table 6 reports the comparison. On synthetic recitation, PER against the clean transcript falls from 0.078 to 0.068 (paired difference −0.011, interval [−0.018, −0.004]; 34 items improve and 15 worsen). DistillMOS changes by +0.010, with interval [−0.011, +0.031]. Cosine selection recovers roughly half the point-estimate gain; a direct paired comparison between selectors would be needed to isolate the recogniser's contribution beyond additional draws. On 40 real development clips, the recogniser-selected PER difference is −0.004, with interval [−0.010, +0.002].

The archive follow-up uses the setting fixed after development. PER falls from 0.0713 to 0.0684, a paired difference of −0.0029 with interval [−0.0074, +0.0014]; 21 items improve, 14 worsen and 85 tie. Drift falls from 0.053 to 0.048. The point-estimate gap to the input shrinks by about 31%, from +0.0095 to +0.0066, but both input-comparison intervals include zero: [−0.0001, +0.0183] before selection and [−0.0029, +0.0155] after it. Neither establishes equivalence to the input. DistillMOS changes by −0.025, with interval [−0.054, +0.002]; speaker similarity changes from 0.941 to 0.942 and dropouts remain at 0.65%. This archive had already been inspected for default-model errors, so the follow-up is not an untouched test of the whole development process.

Selection requires three candidate generations per window and four CPU transcriptions. Throughput falls to approximately three seconds of audio per second, compared with 10 to 16 for the default, making this an optional recitation mode. In 45 of 240 archive windows, all three candidates still exceed 0.10 phoneme drift from the input. Selection can rank available renders; it cannot guarantee a correct candidate.

5.5 Listening

In a blind comparison of 24 items and three systems, one listener selected SAWT on 12.0 items, Sidon on 8.5 and V3.1 on 3.5, splitting ties. System letters were shuffled per item and the key opened after responses. SAWT led on English (5 of 6) and archive recitation (5 of 8); Sidon led on multilingual speech unfamiliar to the listener (4 of 6). On four mildly damaged Miipher-style inputs, V3.1 received two selections and the other systems one each. This is an exploratory listening result from one person, without a population-level preference estimate or an independent content assessment in unfamiliar languages.

A native speaker also assessed a Danish broadcast probe, a language absent from the target pool. Moderately damaged examples remained correct after restoration; heavily damaged examples sounded fluent but contained unacceptable content changes. That feedback set the content threshold, so it does not independently validate the final recipe; it is the reason the content guard exists.

5.6 Why DistillMOS cannot establish fidelity

DistillMOS is inadequate as the sole endpoint for speech restoration. Across our reverberation sweep it assigns scores of 3.8 to 4.3 while recogniser-based WER reaches 0.96. Other development outputs score above 3.0 despite severe content errors. Figure 4 gives a concrete example: a score of 3.42 accompanies substantial changes in the recognised words. A high predicted quality score therefore cannot certify that a recording has survived restoration intact. The metric can rate the sound favourably while missing the error that matters most to an archive, and to any model trained on the output.

Its input also limits what it can assess. DistillMOS accepts 16 kHz audio, and our scoring code resamples every output to that rate, following the released implementation [15]. The resulting score cannot directly assess reconstruction above the 8 kHz Nyquist limit. It therefore leaves the 8 to 24 kHz band of a 48 kHz restoration outside its observation. In addition, the model card lists English as the supported language and cautions that unseen languages and tasks may yield misleading estimates. Calibration on multilingual

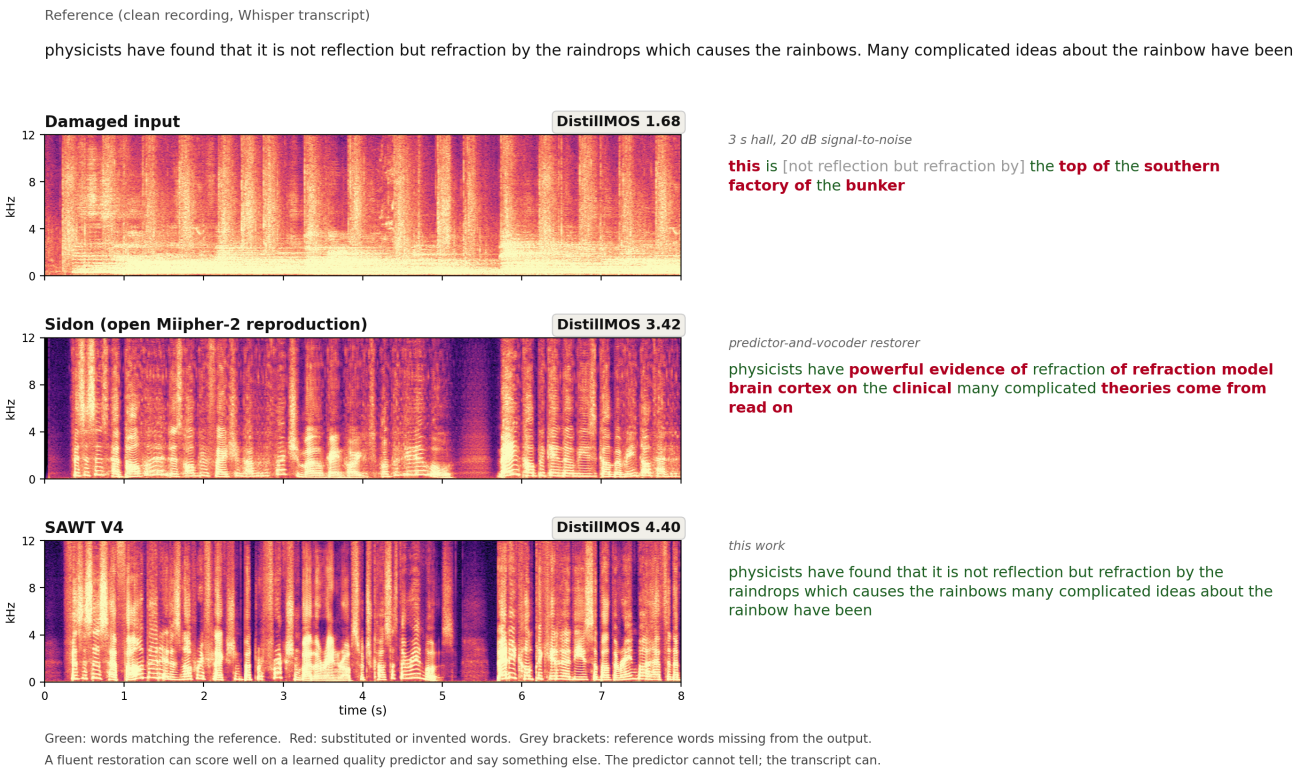

Figure 4: Predicted quality can accompany incorrect recognised words. A selected utterance at 3 s nominal RT60 and 20 dB SNR, followed by Sidon and SAWT V4 restorations. Spectrograms span 0 to 12 kHz; badges report DistillMOS. Whisper transcripts are aligned to the clean reference: green marks agreement, red substitutions or insertions, and grey brackets deletions. Sidon scores 3.42 despite transcript changes. SAWT scores 4.40 with an exact Whisper transcript, but its CTC WER is 0.20. This example illustrates metric disagreement and recogniser dependence; it is not an aggregate performance estimate. Figures 1 and 3 report aggregate comparisons.

speech and Quran recitation cannot be assumed from its performance on other material.

The excluded band is not empty, and the two systems differ in it. Measuring the 8 to 20 kHz energy under the frames the input identifies as speech, relative to the 200 to 4,000 Hz voice band, on four of the 16 public clips: the inputs sit at −47 to −68 dB, the Miipher-2 restorations at −32 to −38 dB, and SAWT's at −32 to −34 dB. Both systems reintroduce high-frequency content that the input does not contain, and on two of the four clips SAWT places 3 to 4 dB more there than Miipher-2 does. A listener identified that content as audible background noise on those two clips, although the levelled floor between phrases was lower for SAWT than for Miipher-2 on both (for example −61.0 against −54.1 dBFS). The band in which two 48 kHz restorers differ most is therefore outside the observation of the predictor used to rank them. Four clips and one listener identify a measurement gap; they do not quantify a perceptual effect.

These limitations concern three different claims: whether speech sounds good, whether it preserves the utterance, and whether the full output bandwidth is reconstructed accurately. One predicted score does not establish all three. We retain DistillMOS as an acoustic proxy, including the comparisons in which SAWT scores lower, and interpret it alongside PQ, recognition error, speaker measures, dropout measurements and the listening evidence. A restoration benchmark should expose disagreement between these measures rather than hide it in a single ranking.

5.7 Ablations and alternatives

Table 7 compares four successive inference configurations on the same development recordings. Generator and decoder weights, 64-step Euler sampling, seed 0, the energy threshold of 0.70 and the three-redraw limit are fixed. Configuration A uses a content threshold of 0.80; B raises it to 0.85; C adds the 0.90 speaker-continuity check; and D adds dynamics matching. D is the released configuration.

Raising the content threshold reduces multilingual WER by 0.012, with a DistillMOS change of −0.007. Adding the speaker check leaves both averages unchanged on these short clips; its case is the 12-minute file, where it removes seam artefacts. Dynamics matching then reduces multilingual WER by a further 0.023 and English WER by 0.004. From A to D, multilingual WER falls by 0.035 while DistillMOS falls by 0.011. Because these recordings informed configuration selection, the table supports the choice of recipe rather than an independent estimate of its benefit.

Alternatives were measured and closed. Adapting the decoder to generated latents reduced residual tail by 2.4 dB in two trials but cost 0.10 to 0.27 DistillMOS on identical latents; D2 stays. Heun integration with 16 and 24 steps left 16 and 11 multilingual windows below the semantic threshold against 3 for 64-step Euler. Strong semantic guidance (2.0), evidence guidance and partial starts from the anchor estimate or the input latent were swept at nine settings and none beat the pure-noise start on both quality and faithfulness. Five

	English 40		Multilingual 40	
configuration	CTC WER	DistillMOS	CTC WER	DistillMOS
A: content threshold 0.80	0.220	4.099	0.266	4.382
B: content threshold 0.85	0.220	4.099	0.254	4.375
C: B + speaker check	0.220	4.099	0.254	4.375
D: C + dynamics matching	0.216	4.095	0.231	4.371

Table 7: Sequential inference comparison on development data. Lower CTC WER is preferable; higher DistillMOS is preferable. CTC WER covers 37 English and 35 multilingual clips after scorer exclusions; the set names denote 40-item collections. Each row adds to the preceding configuration, so the comparisons measure incremental effects in that order rather than independent effects of all components.

shorter decoder adaptations reconstructed worse than the chosen 40k-step run. Each is a dated row with its numbers.

Gentle guidance. This sweep uses a separately rendered default baseline: English WER is 0.2208, compared with 0.2161 in the earlier guard matrix. Effects below are paired within the guidance sweep; the discrepancy is not resolved by the available records. Semantic guidance scales of 1.15 and 1.30 leave DistillMOS within approximately 0.01 of the default across four development sets. At 1.30, English WER falls from 0.221 to 0.195 (paired difference interval $[-0.046, -0.006]$). Mild-set CTC drift falls by about 18%, although its difference interval, $[-0.0158, +0.0002]$, spans zero. At 1.15, synthetic recitation PER falls from 0.078 to 0.074 (interval $[-0.009, -0.001]$). Real-recitation PER is 0.0799 at the default and 0.0829 and 0.0834 at 1.15 and 1.30, respectively, with no detected improvement. Guidance requires an extra generator pass per step and yields approximately 6 to 7.5 seconds of audio per second. It is an optional setting; default guidance remains 1.0.

6 Discussion

Anchoring constrains generation. The anchor supplies deterministic estimates of speech structure, while the generator models the waveform through its latent representation. Recomputing features from the output provides a practical check on their agreement. The comparisons show that this system can recover content under severe degradation, but they do not isolate the anchor's causal contribution from training data, representation or sampling choices. Sequential guard ablations quantify narrower inference changes. Local phoneme errors that pass the feature checks expose a remaining weakness; recogniser-based selection is a partial response with additional cost and evaluation dependence.

Restoration for downstream TTS. SAWT is intended to improve the acoustic quality of training audio for downstream text-to-speech: less noise and reverberation, with clearer speech detail. The aim is a synthesiser that learns cleaner acoustic targets and produces crisper speech. LibriTTS-R provides prior evidence that restoration can support natural-sounding TTS [17]; we have not measured that benefit for SAWT. With fixed transcript labels, actual phonetic changes can create training mismatches, while removing noise and reverberation may improve the acoustic targets. Our recognition measures expose part of this trade-off without predicting its net effect on a synthesiser. A future controlled comparison could train the same TTS system on matched raw, restored and restored-with-review datasets, hold the training budget fixed, and report retained hours as well as independent intelligibility and listening assessments on held-out text and speakers. Until then, quality scores and guard acceptance support review of candidate data; they do not establish improved TTS. Original recordings and their correspondence to restored segments should remain available.

Why canonical text matters. Historical audio often has no clean acoustic counterpart, which makes waveform reconstruction error impossible to measure directly. A known recitation text supplies a different reference: it lets us assess the recognised phonemes while separately measuring changes in noise, decay and voice. Since the text is withheld from the restoration model, this evaluation reveals linguistic changes without instructing the generator to produce the expected words. It still depends on the accuracy of the phoneme recogniser, and it cannot by itself verify timing or delivery.

From restoration to access. Clearer recitations can return individual performances to listening and study. Canonical text supports a linguistic audit, while the original recording supplies the evidence for voice, timing and delivery. Other oral traditions and language archives face related needs, often with less training data or no reliable transcript; the Hausa and Danish examples in the model card were chosen for that reason. At 15 times real time, 1,000 hours take roughly 67 GPU hours to process.

7 Limitations and future work

The following limitations distinguish measured deficits from proposed changes that remain untested.

Quality on the public comparisons. SAWT trails Miipher-2 by 0.53 DistillMOS on its 16 clips and trails Miipher-1 and Sidon by 0.1 to 0.15 on the other reported sets. SAWT uses roughly 900 hours of dry targets, compared with Sidon's reported 2,219 hours of predictor data and Miipher-2's 3,000 hours of restoration training [3, 2]. These are different datasets and architectures with different pretrained backbones; their sizes do not explain the score gap by themselves. Expanding the dry-target pool and adapting the decoder jointly to generated latents are plausible follow-ups. The screening audit identifies about 900 additional hours, but the decoder adaptations tried so far reduced predicted quality.

Local phoneme errors. The archive audit identifies consonant deletions in windows that pass feature checks. A consonant-sensitive training objective is a candidate response,

while recogniser selection offers a measured inference alternative. Neither the current guards nor selection certifies phonetic accuracy; independent listening and a separate recognition model are needed.

Long reverberation and noise. At 3 s and 20 dB SNR, WER is already 0.52; at 5 dB it is 0.82. Low residual decay therefore coexists with substantial recognised content error. On general real-120, the predecessor also has fewer measured dropouts (4.90% versus 7.94%). The anchor is itself estimated from damaged audio. Re-reading a first restoration with a second anchor pass and training with measured long room responses are hypotheses to test, with particular attention to whether a second pass reinforces earlier mistakes. The realism of the simulated hall bank remains unresolved.

Short inputs and collapse. Two inputs under 3 s changed little and one 5.6 s input collapsed repeatedly. A post hoc diagnostic using 3 s windows raises the collapsed clip from 1.59 to 4.37 DistillMOS (Miipher-2: 4.60) and a 2.3 s clip from 2.85 to 3.12. A second short clip falls from 3.30 to 3.21, and a normal-length control loses 0.22. Reflected padding worsens the short clips and leaves the collapse unresolved. These four examples implicate window handling, but do not establish a universally beneficial replacement for the 8 s setting. Adaptive window lengths and fallback to the anchor estimate require broader evaluation. The Maori example presents a different failure: content changes pass the feature checks, demonstrating that guard acceptance is not proof of accuracy.

Evaluation coverage. Listening evidence comes from one listener, the main sampling recipe uses one seed, and the reverberation sweep contains 24 source utterances. Recognition accuracy varies by language, and the optional selector shares its recogniser with evaluation. Multiple seeds, independent native listeners and separate scoring models would strengthen the estimates. The effect on the quality of a downstream TTS model remains future work.

Cost accounting. The US$1,200 figure is rented GPU time for development. Labour, data preparation, the local workstation and listening costs sit outside it.

8 Broader impact

Synthesis can remove the audible signs that a recording has been processed. Archives should retain the original, label restored versions and review flagged passages, since clear output can still contain altered words or speaker characteristics; the per-file sidecar exists for that review. Decisions about consent and access belong to the collections responsible for the recordings.

The release package contains generator, anchor and decoder checkpoints and inference code under the SAWT License 2.0: the work is held as a permanent public trust for non-commercial use. It is not licensed to for-profit organisations, and it and any feature it powers may never be sold. Audio restored with it, any corpus built from that audio and any model trained on that corpus are derivatives carrying the same terms, so the model may not be used to prepare training data for a commercial system. Hosting may recover its direct costs only, attribution is required, and it may not be used to make a recording say what was not said or to deceive or impersonate.

9 Conclusion

SAWT V4 brings together deterministic content anchoring, 48 kHz latent generation and checks on the sampled waveform. It achieves the lowest WER among the three restorers in every reverberation condition with baseline measurements, reduces drift and dropouts relative to Miipher-1 on LibriTTS-R, and improves both predicted quality and phoneme error over V3.1 on archive recitation. These gains come from a system trained on approximately 900 hours and capable of processing audio at 10 to 16 times real time on one RTX 4090. The canonical-text audit and local error analysis make its remaining weaknesses measurable, while recogniser-based selection offers an additional improvement on synthetic development data. Higher mean reference error than unprocessed archive and LibriTTS-R inputs, and the deficit on Miipher-2's public examples, bound the claims. SAWT provides a practical system for preparing cleaner candidate TTS audio; a downstream synthesis comparison remains future work. For the recitations that motivated the project, the aim is renewed access to historical voices with their words, timing and delivery open to scrutiny.

Acknowledgements

This work uses Meta's w2v-BERT 2.0 and DAC-VAE, the datasets listed in Table 3, Sidon's released baseline, and Google's Miipher demonstration audio and LibriTTS-R restorations. We thank their creators for making these resources available. Model training ran on 8x NVIDIA H100 GPUs, with local evaluation on one workstation.

References

[1] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, Y. Zhang, W. Han, A. Bapna, M. Bacchiani. Miipher: A robust speech restoration model integrating self-supervised speech and text representations. *WASPAA*, 2023.

[2] S. Karita, Y. Koizumi, H. Zen, H. Ishikawa, R. Scheibler, M. Bacchiani. Miipher-2: A universal speech restoration model for million-hour scale data restoration. arXiv:2505.04457, 2025. `https://arxiv.org/abs/2505.04457`

[3] W. Nakata et al. Sidon: Fast and robust open-source multilingual speech restoration for large-scale dataset cleansing. arXiv:2509.17052, 2025. `https://arxiv.org/abs/2509.17052`

[4] H. R. Guimarães, J. Su, R. Kumar, T. H. Falk, Z. Jin. DiTSE: High-fidelity generative speech enhancement via latent diffusion transformers. arXiv:2504.09381, 2025. `https://arxiv.org/abs/2504.09381`

[5] C. Zhang, K. Zheng, Z. Chen, J. Zhu. VoiceBridge: General speech restoration with one-step latent bridge models. arXiv:2509.25275, revised 2026. `https://arxiv.org/abs/2509.25275`

[6] Meta AI. DAC-VAE: a 48 kHz variational audio codec with watermarking. Model release, 2025.

[7] Seamless Communication, Meta AI. Seamless: Multilingual expressive and streaming speech translation (w2v-BERT 2.0). arXiv, 2023.

[8] E. J. Hu et al. LoRA: Low-rank adaptation of large language models. *ICLR*, 2022.

[9] W. Peebles, S. Xie. Scalable diffusion models with transformers. *ICCV*, 2023.

[10] J. Su et al. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

[11] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, M. Le. Flow matching for generative modeling. *ICLR*, 2023.

[12] S. Chen et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE JSTSP*, 2022.

[13] A. Tjandra et al. Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound. arXiv:2502.05139, 2025. `https://arxiv.org/abs/2502.05139`

[14] B. Stahl, H. Gamper. Distillation and pruning for scalable self-supervised representation-based speech quality assessment. *ICASSP*, 2025. `https://arxiv.org/abs/2502.05356`

[15] Microsoft. Distill-MOS: implementation and model card. Accessed September 2026. `https://github.com/microsoft/Distill-MOS`

[16] C. K. A. Reddy, V. Gopal, R. Cutler. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. *ICASSP*, 2021.

[17] Y. Koizumi et al. LibriTTS-R: A restored multi-speaker text-to-speech corpus. *Interspeech*, 2023.

[18] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, T. Gerkmann. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. *Interspeech*, 2024.

[19] J. Meyer et al. BibleTTS: A large, high-fidelity, multilingual, and uniquely African speech corpus. *Interspeech*, 2022.

[20] J. F. Gemmeke et al. AudioSet: An ontology and human-labeled dataset for audio events. *ICASSP*, 2017.

A Inference configuration

The evaluated weights are the generator's exponential moving average at step 150,000, the final anchor and decoder D2. The sampler starts from unit Gaussian noise and applies 64 Euler steps with guidance scale 1, the dry room token and seed 0. It uses 8 s windows, 1 s overlap and 1 s padding. Thresholds are 0.70 for energy, 0.85 for content and 0.90 for speaker continuity; at most three redraws are allowed. Dynamics matching is enabled. Batched execution uses `restore_v4r_fast.py` with 16 lanes and a CUDA graph. The optional recitation mode of Section 5.4 adds `-select-k 3 -select-by asr`; the optional gentle guidance for general speech (Section 5.7) is `-cfg 1.15`.

B Reproduction

Weights, inference code and the model card are public at `https://huggingface.co/Quran-Lab/sawt-v4`. The 16 public demonstration clips, with all three versions of each and the per-clip measurements, are at `https://comparison.quranlab.ai`. The research repository holds the dated measurements, evaluation programs and configuration records, with publication planned alongside the camera-ready paper. Reproduction also depends on the evaluation manifests, scoring settings and exact reference restorations; the inference settings alone do not specify the comparison.