FROM RELIABLE TEXT TO REAL VOICES: TRUST-AWARE PROGRESSIVE ADAPTATION FOR LOW-RESOURCE TTS

Jiayi Lu[1,2,*], *Yizhong Geng*[1,3,*], *Jinghan Yang*[3], *Tianhan Jiang*[4], *Boxun An*[5], *Yingming Gao*[3], *Ya Li*[3,†]

[1]Beijing Logic Intelligence Technology [2]University of Washington
[3]Beijing University of Posts and Telecommunications
[4]University of California, USA [5]Northwestern University, USA

ABSTRACT

Low-resource text-to-speech (TTS) adaptation is constrained by scarce paired data and costly manual transcription. Existing fixed-voice TTS systems can provide relatively accurate pronunciation, but their synthetic speech offers limited speaker diversity and may exhibit flat prosody. Real recordings provide natural prosody and diverse voices, yet their automatic speech recognition (ASR) pseudo-labels may contain transcription errors. We find that supervision order affects content accuracy and speaker similarity. We propose trust-aware progressive adaptation: synthetic-to-real adaptation first establishes text–speech correspondences, then restores reference-speaker control using real speech. Transcript-agreement weighting uses agreement between two fixed ASR systems as a proxy for pseudo-label reliability to limit noisy supervision. Experiments with FireRedTTS3 on Burmese and Lao and OmniVoice on Burmese show improved content accuracy with high naturalness and competitive speaker similarity. Jointly considering supervision order and pseudo-label reliability when combining synthetic and real speech offers a practical path to zero-shot voice cloning in low-resource languages with less manual transcription. Audio demos are available at https://insiderx-pro.github.io/S2R-Adaptation-TTS/.

***Index Terms*—** Low-resource TTS, progressive adaptation, reliability weighting, zero-shot voice cloning

1. INTRODUCTION

Multilingual TTS enables zero-shot voice cloning from short reference recordings [1, 2, 3, 4]. Recent systems broaden acoustic-domain and language coverage [5, 6, 7], but low-resource adaptation still lacks high-quality text–speech pairs [8, 9]. Manual transcription is costly, motivating the use of unpaired text and speech [10, 11].

These resources offer two routes to constructing training pairs. Existing TTS systems synthesize speech from text [12], providing relatively accurate pronunciation but potentially flat prosody [13]. Fixed voices also limit speaker diversity, as illustrated by the language-specific models in MMS [14]. Synthetic speech thus provides useful pronunciation supervision but limited acoustic diversity. Alternatively, ASR supplies pseudo-labels for real speech [15], preserving natural prosody and diverse speaker characteristics while introducing transcription noise.

Combining these sources is nontrivial: synthetic augmentation can degrade speaker similarity [16], while noisy pseudo-labels can erode content accuracy. This trade-off motivates studying supervision order and pseudo-label reliability jointly.

We propose trust-aware progressive adaptation. Synthetic-to-real adaptation learns text–speech correspondences from synthetic speech before using real speech to restore reference-speaker control. Transcript-agreement weighting uses agreement between two fixed ASR systems as a proxy for pseudo-label reliability. We evaluate FireRedTTS3 [17] on Burmese and Lao and OmniVoice on Burmese.

Our contributions are:

- **We show that supervision order changes the trade-off:** synthetic speech improves content accuracy but can weaken speaker similarity; real speech can restore it.
- **We develop trust-aware progressive adaptation:** synthetic-to-real adaptation uses transcript-agreement weighting on pseudo-labeled real speech to limit noisy supervision.
- **We test content gains and reference conditioning:** independent ASR corroborates gains across languages and backbones; FireRedTTS3 interventions probe acoustic history.

2. METHOD

In Fig. 1, the Base model, Synthetic-Stage-Model, and Final-Stage-Model are abbreviated Base, S, and S→R: synthetic supervision precedes reliability-weighted real supervision.

2.1. Synthetic supervision

Fixed Teacher TTS T maps text x_i and voice v_i to $y_i^S = T(x_i, v_i)$. Filtering (Section 3) and same-voice reference pairing yield $\mathcal{D}_S = \{(x_i, y_i^S, r_i^S)\}$, with reference speech r_i^S, example index i, and S/R denoting synthetic/real data. Training uses unit weights and original input text; filtering does not guarantee correct pronunciation.

2.2. Transcript agreement

For real recording y_i^R, fixed primary ASR_1, A_1, supplies $\tilde{x}_i = A_1(y_i^R)$; fixed secondary ASR_2, A_2, supplies only a comparison transcript. Training uses $\mathcal{D}_R = \{(\tilde{x}_i, y_i^R, r_i^R)\}$, without transcript fusion or correction. Reference speech r_i^R is a distinct 1–10 s utterance from the same local speaker within a recording.

We measure normalized character disagreement:

$$d_i = \frac{\text{ED}(\mathcal{N}(\tilde{x}_i), \mathcal{N}(A_2(y_i^R)))}{\max(1, |\mathcal{N}(\tilde{x}_i)|)}. \quad (1)$$

*Equal contribution. †Corresponding author: Ya Li.

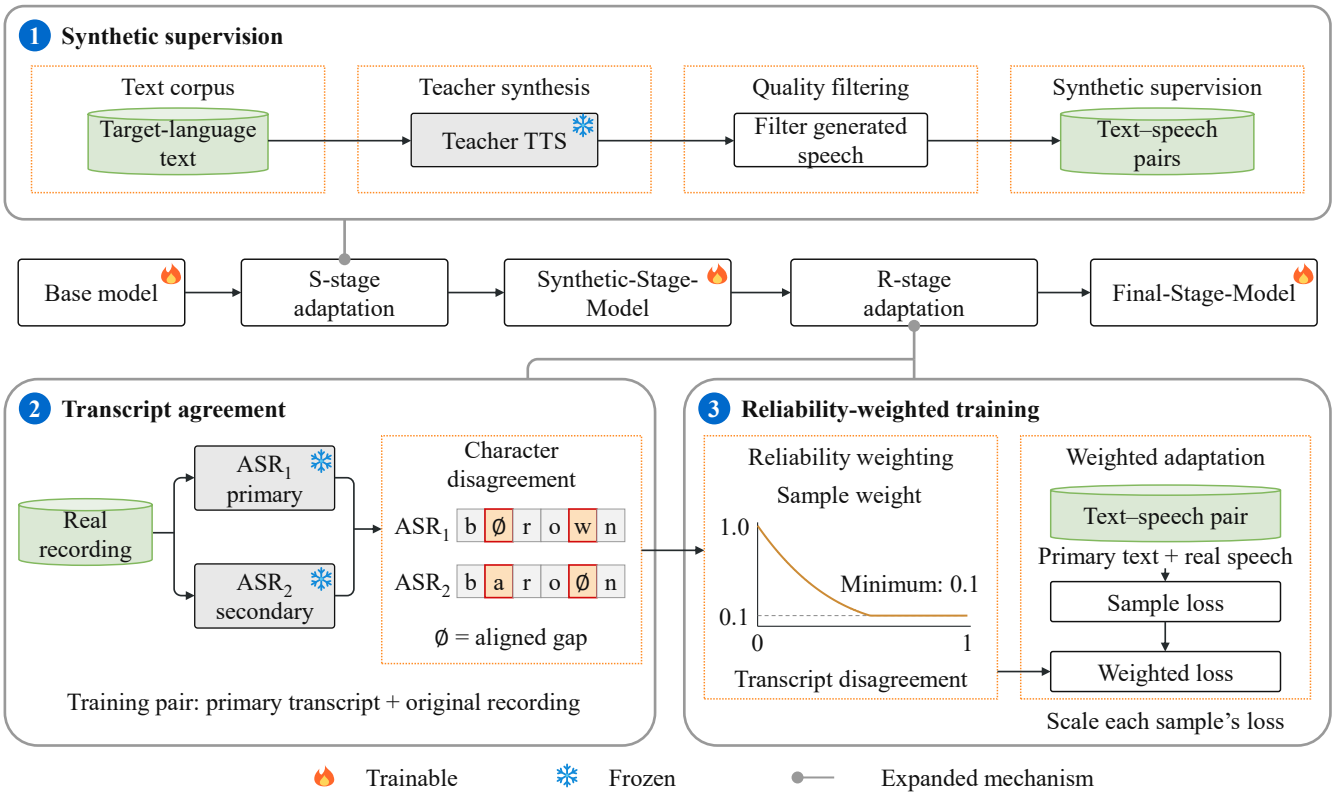

Fig. 1. Trust-aware progressive adaptation. Synthetic supervision establishes text–speech correspondences in S-stage adaptation; transcript agreement guides reliability-weighted training on real recordings in R-stage adaptation. The primary transcript remains the training text.

ED and $|\cdot|$ denote code-point Levenshtein distance and text length. Scoring-only normalization $\mathcal{N}$ applies Unicode NFC, removing whitespace and punctuation/symbol/control categories but retaining letters, digits, and combining marks, without case, numeral, or Zawgyi conversion. Empty primary labels are excluded; valid empty secondary transcripts give $d_i = 1$, distinct from failed ASR requests.

Agreement does not verify correctness: recognizers may share errors. In Fig. 1, character/transcript disagreement denotes d_i; $\emptyset$ marks an alignment gap, not a training character.

2.3. Reliability-weighted training

Reliability weighting. To limit noisy supervision's contribution [18], we assign

$$w_i^{(\gamma)} = \max(w_{\min}, [1 - \min(d_i, 1)]^{\gamma}), \tag{2}$$

where $\gamma > 0$ controls decay and $w_{\min}$ floors weights. Cubic uses $\gamma = 3$, $w_{\min} = 0.10$ (Section 3): perfect agreement gets unit weight; disagreement reduces influence.

Weighted adaptation. For minibatch $\mathcal{B}$ and sample loss ℓ_θ at trainable parameters θ,

$$\mathcal{L}_R(\theta) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_i^{(\gamma)} \ell_\theta(y_i^R, \tilde{x}_i, r_i^R), \tag{3}$$

averages by batch size $|\mathcal{B}|$, not weight sum. Weighting changes loss scale but need not proportionally scale AdamW updates. Offline scoring adds no inference-time ASR. Let $\mathcal{A}_B(\theta; \mathcal{D}, w)$ perform B updates from θ on data $\mathcal{D}$ weighted by w. From Base θ_0, $\theta_S = \mathcal{A}_{B_S}(\theta_0; \mathcal{D}_S, \mathbf{1})$ and $\theta_{SR}^{(\gamma)} = \mathcal{A}_{B_R}(\theta_S; \mathcal{D}_R, w^{(\gamma)})$, where B_S/B_R are update budgets and $\mathbf{1}/w^{(\gamma)}$ collect unit/real-example weights. Backbones/objectives stay unchanged; stages restart optimizers/schedules. R starts from Base; R→S reverses stages.

2.4. Backbone objectives and reference conditioning

FireRedTTS3 [17] trains its core with RedAE/CAM++ [19] frozen. We weight its full flow-matching [20]/stop loss, $\ell_i = \ell_{\text{flow},i} + 0.1\ell_{\text{stop},i}$, or OmniVoice's masked acoustic-token loss [6].

RedAE/CAM++ (omitted from Fig. 1) encode reference r as acoustic prompt $z^r = E_{\text{AE}}(r)$ and speaker embedding $e^r = E_{\text{spk}}(r)$. Controls fix pairings/reference text, exclude ineligible targets, and pair mixture references within domains. For text x and K patches [21, 22],

$$p_\theta(z_{1:K} \mid x, z^r, e^r) = \prod_{k=1}^{K} p_\theta(z_k \mid x, z^r, e^r, z_{<k}), \tag{4}$$

History $z_{<k}$ is ground truth in training and generated at inference; frozen encoders permit core-conditioning changes. For 16 texts, independent prompt/embedding swaps use two same-text references. For generated g, define $q = \text{SIM}(g, r_A) - \text{SIM}(g, r_B)$ (SIM-O cosines); Δ_{emb} averages the r_B-to-r_A embedding-induced change in q over prompts/texts. History tests use 8/23 prefix pairs (normalized ASR CER $\leq 5\%$), fixing reference, boundary, and subsequent random state. Suffix scoring excludes prefixes and the first 250 ms; unchanged-history replay reproduces outputs (Section 4). These tests probe reference response and history dependence.

Table 1. Main results. (a) CER (%), SIM-O, and H (0–100); adapted CER/SIM-O: mean ± sample SD over three seeds. (b) MOS (1–5): mean [approximate 95% crossed-bootstrap CI]. (c) Independent-ASR CER (%). Bold: numerical bests, not significance.

	FireRedTTS3 / Burmese			FireRedTTS3 / Lao			OmniVoice / Burmese		
(a) Strategy	CER ↓	SIM-O ↑	H ↑	CER ↓	SIM-O ↑	H ↑	CER ↓	SIM-O ↑	H ↑
Base	142.52	0.7244	0.00	100.45	0.7343	0.00	10.35	0.7278	80.34
S	17.43 ± 0.45	0.4513 ± 0.0041	58.36	13.49 ± 0.26	0.6108 ± 0.0006	71.61	8.21 ± 0.20	0.6864 ± 0.0010	78.54
R	54.09 ± 0.68	$\mathbf{0.7314} \pm 0.0005$	56.41	37.06 ± 0.36	$\mathbf{0.7537} \pm 0.0019$	68.60	10.74 ± 0.41	$\mathbf{0.7328} \pm 0.0013$	80.48
R→S	$\mathbf{14.71} \pm 0.62$	0.3636 ± 0.0038	50.98	$\mathbf{10.64} \pm 0.09$	0.5959 ± 0.0014	71.50	7.36 ± 0.12	0.6948 ± 0.0035	79.41
S→R: 1.0	23.27 ± 0.15	0.6942 ± 0.0014	72.89	14.99 ± 0.25	0.6993 ± 0.0010	76.74	10.21 ± 0.22	0.7142 ± 0.0017	79.56
S→R: 0.5	20.91 ± 0.37	0.6996 ± 0.0005	74.24	14.60 ± 0.14	0.6968 ± 0.0026	76.74	8.03 ± 0.33	0.7094 ± 0.0009	80.10
S→R: cubic	16.90 ± 0.26	0.6997 ± 0.0016	**75.97**	13.97 ± 0.36	0.6968 ± 0.0045	**77.00**	$\mathbf{6.85} \pm 0.08$	0.7098 ± 0.0026	**80.57**

	FireRedTTS3 / Burmese		FireRedTTS3 / Lao		OmniVoice / Burmese	
Strategy	(b) MOS ↑	(c) CER ↓	(b) MOS ↑	(c) CER ↓	(b) MOS ↑	(c) CER ↓
Base	2.17 [1.97, 2.38]	103.56	2.55 [2.34, 2.76]	100.45	4.40 [4.23, 4.56]	12.62
S	3.15 [2.97, 3.34]	20.45	3.39 [3.22, 3.56]	18.03	4.23 [4.05, 4.40]	11.63
R	1.87 [1.69, 2.06]	68.64	2.12 [1.93, 2.32]	44.94	2.35 [2.15, 2.56]	12.81
R→S	2.72 [2.52, 2.92]	**17.93**	3.27 [3.07, 3.47]	**17.13**	4.32 [4.12, 4.50]	11.65
S→R: 1.0	3.78 [3.55, 4.00]	26.82	3.78 [3.61, 3.94]	23.85	4.35 [4.17, 4.51]	11.69
S→R: 0.5	3.91 [3.71, 4.11]	25.06	3.84 [3.65, 4.02]	23.33	4.42 [4.24, 4.58]	11.41
S→R: cubic	**4.12** [3.93, 4.30]	19.34	**3.97** [3.77, 4.17]	18.53	**4.51** [4.35, 4.66]	**9.83**

3. EXPERIMENTS

3.1. Experimental setup

Data. Burmese S/R: 160,918 Azure Nilar/Thiha and 48,043 DVB pairs (291.148/92.263 h); Lao: 119,246/26,004 pairs (200.004/45.117 h). ASR_1/ASR_2 are Gemini 2.5 Flash/Omnilingual ASR [23]. Synthetic filters: valid audio/text, 0.4–20 s, ASR CER $\leq$ 20%, speech fraction ≥ 0.65, clipping ≤ 0.005, deduplication, evaluation exclusions, reference eligibility, and $\leq$1,024 tokens. Real recordings also undergo duration, single-speaker, audio-quality, and transcript-validity screening in our released video pipeline before agreement weighting. Burmese S: 171,240 candidates; 717 real targets lack references. Real speech: 300 recordings, 1,200 local speaker groups, 286 unverified cross-recording identity clusters.

Training. Burmese S/R use 53,640/16,015 updates at learning rates $2 \times 10^{-6}/10^{-6}$; Lao uses 39,749/8,668 updates. Orders share AdamW [24], 3% warmup and cosine decay. All adaptation/control runs use one device, accumulation/effective batch 3, and training seeds 42/17/73. Tables 1a and 3 report adapted CER/SIM-O as mean ± sample SD over these seeds; Base has no training-seed SD. Dev120 (60 texts × two references) selects $\gamma = 3$, $w_{\min} = 0.10$ by minimum three-seed mean CER with SIM-O ≥ 0.6970, over $\gamma \in \{0.5, 1, 2, 3, 4, 6\}$, $w_{\min} \in \{0, 0.05, 0.10, 0.20\}$. Broadcast R uses 16,015 updates, not test-selected.

Evaluation. Burmese Common400 (100 general/300 cloning conditions; 150 texts) uses Omnilingual ASR Unlimited 7B v2. Clone300 uses 30 Burmese/English/Chinese references; SIM-O averages ECAPA-TDNN/WavLM-large cosines [25, 26]. Lao uses Gemini on 404 FLEURS [27] rows (260 inputs, one reference). CER comparisons are within-language. Common400 reuses FLEURS train/development conditions. Normalized evaluation texts do not overlap archived DVB or Lao S/R transcripts. Burmese hashes match none of the 100 checked target/reference conditions. Audited Clone300 speakers are disjoint from adaptation speakers; audio near-duplicate auditing remains incomplete. FireRedTTS3 decoding uses seed 42, 10 flow steps, guidance 2.0, and stop threshold 0.5; long/empty outputs remain scored. Joint score $H = 200as/(a + s)$ uses $a = \text{clip}(1 - \text{CER}/100, 0, 1)$, $s = \text{clip}(\text{SIM-O}, 0, 1)$, and $H = 0$ if $a + s = 0$.

Listening test. Per language, 20 listeners rated all 30 matched text–reference conditions per system (naturalness 1–5; 600 ratings/system). Burmese backbones share listeners and conditions. Approximate 95% percentile CIs use 10,000 crossed-bootstrap replicates [28], independently resampling listeners and conditions with common indices across systems within each language. We report paired cubic-minus-uniform-0.5 differences. Listening-test audio uses each adapted system's seed-42 checkpoint.

3.2. Main results

Table 1 shows order dependence, also in matched-batch Lao: uniform S→R restores similarity but can worsen CER; R→S favors CER. Cubic has the highest observed H and mean MOS. Against uniform 0.5, paired MOS differences are $+0.210\,[0.025, 0.392]$, $+0.130\,[-0.047, 0.315]$, and $+0.090\,[-0.052, 0.230]$ for FireRedTTS3/Burmese, FireRedTTS3/Lao, and OmniVoice/Burmese. Only the first marginal 95% interval excludes zero; these exploratory comparisons are unadjusted for multiplicity. We do not claim consistent MOS improvement across settings. OmniVoice exceeds R in H by just 0.08/100.

Independent ASR. Dolphin-small [29] (Burmese) and XLS-R Lao [30] (Lao) rescore the same generated audio and target texts. Neither participated in pseudo-labeling, reliability estimation, filtering, or model selection. Failures, empty transcripts, and overlong outputs are not selectively excluded. Cubic improves on uniform 1.0 in all groups (Table 1c); R→S still yields lower FireRedTTS3 CER, consistent with an accuracy–similarity trade-off.

Adapted OmniVoice leads the shared-text Burmese comparison (Table 2), reducing Base CER by 3.50 points; fixed-voice SIM-O is descriptive. All outputs are scored, including 210/400 mmSpeech Tacotron cases reaching its 12.54 s limit and IMS-Toucan outputs with Burmese phoneme warnings.

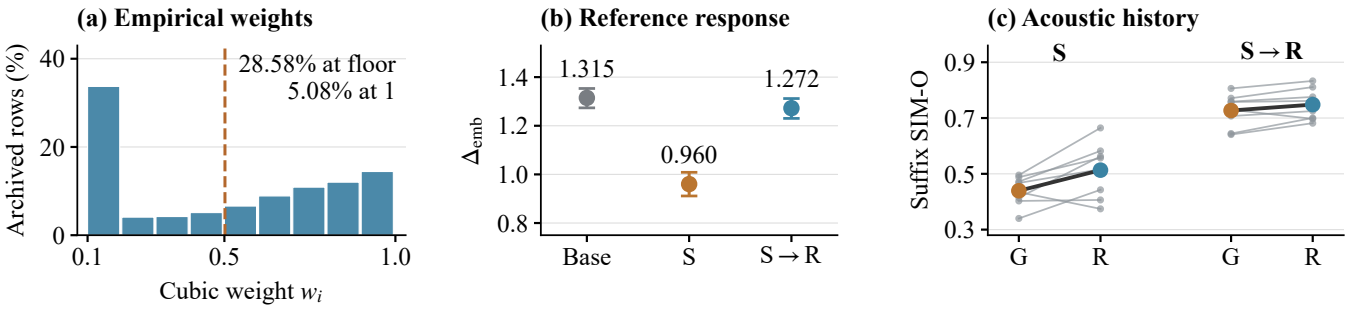

Fig. 2. (a) Cubic weights: 42,785 archived ASR rows. (b) Embedding response: 16 texts, paired-bootstrap 95% intervals. (c) Eight suffix SIM-O pairs: generated (G)/recorded (R) prefixes; thick lines: means. (b,c): historical uniform 0.5 S→R, not cubic.

Table 2. Burmese systems on Common400 / Clone300. *: fixed voices; their SIM-O scores are descriptive.

System	CER (%) ↓	SIM-O ↑
Fish Audio S2-Pro [31]	83.89	0.6376
MMS-TTS-mya* [14]	20.49	0.1029
F5-Myanmar-TTS v2 [32, 33]	35.24	0.5863
IMS-Toucan [34]	95.40	0.3298
mmSpeech Tacotron* [35]	78.70	0.0914
OmniVoice Base [6]	10.35	**0.7278**
FireRedTTS3 cubic	16.90	0.6997
OmniVoice cubic	**6.85**	0.7098

Table 3. Burmese FireRedTTS3 controls from S. Mean ± sample SD over three seeds; shuffles are averaged within seed.

Setting	CER (%) ↓	SIM-O ↑
(a) FLEURS label control; S: Table 1a		
Pseudo-labels, $w = 0.5$	21.17 ± 0.38	0.6345 ± 0.0024
Pseudo-labels, $w = 1.0$	23.67 ± 0.19	$\mathbf{0.6487} \pm 0.0009$
Official text, $w = 1.0$	$\mathbf{15.45} \pm 0.15$	0.6342 ± 0.0012
(b) Broadcast weighting; uniform 1.0: Table 1a		
Uniform 0.5	20.91 ± 0.37	0.6996 ± 0.0005
Linear ($\gamma = 1$)	23.25 ± 0.49	0.6956 ± 0.0030
Global shuffle	21.34 ± 0.52	0.6984 ± 0.0010
Duration-strat. shuffle	21.13 ± 0.22	0.6988 ± 0.0032
Uniform, equal mean	21.09 ± 0.55	0.6993 ± 0.0008
Cubic ($\gamma = 3$)	$\mathbf{16.90} \pm 0.26$	$\mathbf{0.6997} \pm 0.0016$

4. ABLATION AND ANALYSIS

4.1. Supervision order and pseudo-label reliability

Label quality. Table 3a compares labels on identical FLEURS recordings: 3,058 (12.135 h), retaining 2,921 (11.578 h) after overlap/duration exclusions. Official text reduces CER by 8.22 points at $w = 1$, retaining similarity gains; halving pseudo-label weight helps less. Label noise contributes alongside other domain effects.

Reliability validation. On these 2,921 utterances, ASR_1 label CER e_i against official text correlates with d_i (Spearman $\rho = 0.560$; Pearson $r = 0.546$). Mean e_i is 5.94% ($d_i < 0.1$, $n = 971$) versus 32.30% ($d_i \geq 0.6$, $n = 286$). On the full Burmese FLEURS training split, both ASRs produce identical but incorrect normalized transcripts for 3.2% of utterances.

Weight assignment. Table 3b fixes initialization, targets, references, order, and updates. Global shuffles of cubic weights (101/202/303 crossed with three training seeds) break assignment; duration-decile shuffles control length correlations. Both preserve weights yet, like equal-mean uniform weighting, underperform cubic: assignment matters beyond loss scale. The 42,785-row archive (Fig. 2a) has mean/SD 0.50166/0.33337 and quartiles 0.10000/0.54771/0.81090. Its 588 later-excluded rows and missing later transcripts distinguish it from the 48,043-target pool (mean 0.500946). The pool's duration-weighted mean 0.58715 becomes 0.50110/≈0.5865 with global/duration-stratified shuffling.

Mixtures. A 0/10/25/50/75/90/100% real-example sweep from S uses 16,015 updates with unit S/cubic R weights; synthetic replay restores less similarity than real-only continuation. Joint training from Base matches 160,920 S/48,045 R exposures and the learning-rate schedule (restart at 53,640), but yields 20.160% CER/0.65120 SIM-O, supporting progressive adaptation under matched exposure.

4.2. Reference conditioning and acoustic history

Reference response. Base/S/historical uniform 0.5 S→R yield Δ_{emb} of 1.3147/0.9602/1.2723 (Fig. 2b); matched-reference SIM-O is 0.7804/0.5672/0.7625. Conflicts follow embeddings; R restores S-weakened response; cubic is untested.

History dependence. Recorded versus generated prefixes improve suffix SIM-O by 0.0739/0.0211 for S/uniform 0.5 S→R (Fig. 2c), in 7/8 cases each. Paired 95% intervals $[0.0225, 0.1204]$ and $[0.0025, 0.0374]$ condition on cases, not seeds. This supports history dependence, not naturally accumulating voice drift.

5. CONCLUSION

We presented trust-aware progressive adaptation for low-resource TTS. It combines a synthetic-to-real adaptation schedule with transcript-agreement weighting. Supervision order matters: synthetic speech improves content accuracy but can reduce speaker similarity; real speech restores reference-speaker control at the risk of transcription noise. Transcript-agreement weighting helps preserve content accuracy, with improvements over uniform weighting persisting under independent ASR evaluation. Ground-truth label analysis supports transcript agreement as an informative, imperfect reliability proxy, while conditioning interventions reveal changes in reference response and sensitivity to acoustic history. Together, these findings suggest that weak supervision should be organized around the capabilities it strengthens or preserves, rather than treated as interchangeable additional data. By combining the pronunciation supervision available from existing TTS systems with the acoustic diversity of real speech, this approach offers a practical path toward zero-shot voice cloning in low-resource languages with reduced dependence on manual transcription of target-language speech.

6. ETHICAL STATEMENT

We use synthetic speech and existing real recordings for research, without linking speakers to personal identities. Voice cloning can enable impersonation; released models and examples should be used only with speaker authorization. Users should respect source terms and applicable privacy and copyright requirements when handling recordings, transcripts, and generated speech.

7. REFERENCES

[1] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Proc. ICML*, 2022, vol. 162 of *PMLR*, pp. 2709–2720.

[2] E. Casanova et al., "XTTS: a massively multilingual zero-shot text-to-speech model," in *Proc. Interspeech*, 2024, pp. 4978–4982.

[3] M. Le et al., "Voicebox: Text-guided multilingual universal speech generation at scale," in *Adv. Neural Inf. Process. Syst.*, 2023, vol. 36.

[4] Z. Zhang et al., "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *arXiv preprint arXiv:2303.03926*, 2023.

[5] Z. Du et al., "CosyVoice 3: Towards in-the-wild speech generation via scaling-up and post-training," *arXiv preprint arXiv:2505.17589*, 2025.

[6] H. Zhu et al., "OmniVoice: Towards omnilingual zero-shot text-to-speech with diffusion language models," *arXiv preprint arXiv:2604.00688*, 2026.

[7] Q. Liu et al., "Cross-lingual F5-TTS: Towards language-agnostic voice cloning and speech synthesis," *arXiv preprint arXiv:2509.14579*, 2025.

[8] F. Lux, J. Koch, and N. T. Vu, "Low-resource multilingual and zero-shot multispeaker TTS," in *Proc. AACL-IJCNLP*, 2022, pp. 741–751.

[9] Y. Geng, J. Xu, Z. Liang, J. Yang, X. Shi, and X. Shen, "Scaling under-resourced TTS: A data-optimized framework with advanced acoustic modeling for Thai," in *Proc. ACL Industry Track*, 2025, pp. 593–604.

[10] Y.-A. Chung, Y. Wang, W.-N. Hsu, Y. Zhang, and R. J. Skerry-Ryan, "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," in *Proc. ICASSP*, 2019, pp. 6940–6944.

[11] N. Makishima, S. Suzuki, A. Ando, and R. Masumura, "Speaker consistency loss and step-wise optimization for semi-supervised joint training of TTS and ASR using unpaired text data," in *Proc. Interspeech*, 2022, pp. 526–530.

[12] R. Joshi and N. Garera, "Rapid speaker adaptation in low resource text to speech systems using synthetic data and transfer learning," in *Proc. PACLIC*, 2023, pp. 267–273.

[13] Y. Geng et al., "Bridging the stability-expressivity gap: Synthetic data scaling and preference alignment for low-resource spoken language models," *arXiv preprint arXiv:2605.27383*, 2026.

[14] V. Pratap et al., "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.

[15] Z. Du et al., "CosyVoice 2: Scalable streaming speech synthesis with large language models," *arXiv preprint arXiv:2412.10117*, 2024.

[16] Y. Choi, J. Oh, H. Kim, and H. Kim, "ZeSTA: Zero-shot TTS augmentation with domain-conditioned training for data-efficient personalized speech synthesis," *arXiv preprint arXiv:2603.04219*, 2026.

[17] F. Shen et al., "FireRedTTS3: Unified speech generation and editing with semantically enriched speech representations," *arXiv preprint arXiv:2608.17492*, 2026.

[18] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *Proc. ICML*, 2018, vol. 80 of *PMLR*, pp. 4334–4343.

[19] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "CAM++: A fast and efficient network for speaker verification using context-aware masking," *arXiv preprint arXiv:2303.00332*, 2023.

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

[21] D. Jia et al., "DiTAR: Diffusion transformer autoregressive modeling for speech generation," in *Proc. ICML*, 2025, vol. 267 of *PMLR*, pp. 27255–27270.

[22] Y. Geng et al., "CLASVS: Continuous-latent autoregression for melody-preserving lyric editing in singing voice synthesis," *arXiv preprint arXiv:2608.03253*, 2026.

[23] Omnilingual ASR Team et al., "Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages," *arXiv preprint arXiv:2511.09690*, 2025.

[24] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[25] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[26] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, 2022.

[27] A. Conneau et al., "FLEURS: Few-shot learning evaluation of universal representations of speech," *arXiv preprint arXiv:2205.12446*, 2022.

[28] A. B. Owen, "The pigeonhole bootstrap," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 386–411, 2007.

[29] Dataocean AI, "Dolphin-small," Hugging Face model, https://huggingface.co/DataoceanAI/dolphin-small, Accessed: Sep. 15, 2026.

[30] SiangLao, "XLS-R Lao ASR," Hugging Face model, https://huggingface.co/SiangLao/xls-r-lao-asr, Accessed: Sep. 15, 2026.

[31] Fish Audio, "S2-Pro: Model release," https://huggingface.co/fishaudio/s2-pro, 2026, Accessed September 14, 2026.

[32] freococo, "F5-Myanmar-TTS: Burmese v2 model release," https://huggingface.co/freococo/F5-Myanmar-TTS, 2026, Evaluated with f5-myanmar-tts 0.2.1; accessed September 14, 2026.

[33] Y. Chen et al., "F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching," in *Proc. ACL*, 2025, pp. 6255–6271.

[34] DigitalPhonetics, "IMS-Toucan: MassiveScaleToucan implementation," https://github.com/DigitalPhonetics/IMS-Toucan, MassiveScaleToucan branch; accessed September 14, 2026.

[35] hpbyte, "Myanmar end-to-end text-to-speech," https://github.com/hpbyte/myanmar-tts, mmSpeech Tacotron checkpoint; accessed September 14, 2026.