

A Data-Centric Empirical Study of Fidelity-Prosody Trade-offs in Data Curation for Zhuang TTS

Fengbo Bai^{1,2}, Qingshi Yang^{1,2}, Xiao Liang^{1*}

¹College of Artificial Intelligence, Guangxi Minzu University, Nanning 530006, China

²University Engineering Research Center of Computational Linguistics and Intelligence, Guangxi, Nanning 530006, China

Keywords: Speech synthesis; Data-centric artificial intelligence; Low-resource languages; Data curation; Prosody modeling

Citation: Bai F.B., Yang Q.S., Liang X.: A Data-Centric Empirical Study of Fidelity-Prosody Trade-offs in Data Curation for Zhuang TTS. Data Intelligence, 2026, Vol. 8, Art. No.: 20260011. DOI: <https://doi.org/10.3724/2096-7004.di.2026.0011>

ABSTRACT

Training text-to-speech (TTS) systems on project-collected, real-world recordings obtained under multiple recording conditions that exhibit both environmental noise and atypical prosody is challenging, especially for tonal languages where prosodic structure directly affects intelligibility. This study adopts a multi-objective perspective on data curation that balances spectral fidelity and prosodic integrity and instantiates this perspective on a 24 hour screened subset drawn from a real-world Zhuang corpus of approximately 100 hours. Seven preprocessing strategies were compared under a controlled setting in which the training roster and model architecture were held fixed. The main empirical finding is that disabling internal silence compression while retaining trimming, filtering, and normalization (no_int_sil) yielded a substantial perceptual gain: Mean Opinion Scores (MOS) for overall naturalness improved from 3.17 for a minimally processed baseline to 3.72 ($p < 0.05$), even though a heavily cleaned configuration (full_method) attained the lowest Fréchet Audio Distance (FAD) among all systems. This pattern indicates that, in this project-collected Wuming Zhuang setting and under a VITS backbone, aggressive internal silence compression can act as a prosody-distorting operation: it improves apparent spectral cleanliness but perturbs boundary-related timing and pitch behavior that listeners rely on for naturalness.

1. INTRODUCTION

The performance of end-to-end text-to-speech (TTS) technology depends strongly on the quality and scale of training data [1]. For low-resource languages such as Zhuang, acquiring large-scale, high-purity,

*Corresponding author: Xiao Liang (E-mail: xiaoliang1223@foxmail.com; ORCID: 0009-0006-5321-4565).

studio-grade corpora is often prohibitively expensive in practice. A common and practical alternative is to augment training data with existing automatic speech recognition corpora or low-cost recordings [2-4]. Such data, often sourced from diverse real-world environments, contain background noise and mouth clicks, and may also exhibit prosodic patterns that deviate from natural conversation, such as frequent within-utterance pauses and a didactic, slow speaking rate; depending on the context, these pauses may reflect disfluency artifacts but may also encode functional prosodic boundary cues.

Traditional TTS data cleaning pipelines often aim to optimize spectral fidelity, for example, by mimicking the distribution of high-quality datasets such as LJSpeech through aggressive silence compression and filtering. High-quality TTS assessment requires evaluating both spectral fidelity and prosodic naturalness [5]. Prosody encompasses suprasegmental information such as speech rate, pauses, and intonation, and its importance is particularly pronounced in tonal languages such as Zhuang, where pitch variations directly carry semantic functions that distinguish word meanings [6-8]. Traditional aggressive cleaning strategies, while improving signal purity, may corrupt prosodic information, particularly when within-utterance pauses are treated as removable artifacts despite their role as acoustic cues for prosodic boundary marking and discourse interpretation [9-10].

Despite extensive research on low-resource TTS [11] and prosody modeling [12], there is no systematic framework in the literature that explicitly quantifies and guides the intrinsic trade-off between spectral fidelity and prosodic naturalness during data curation, especially when the prosody of the source data is atypical. Existing studies typically present multiple evaluation metrics in parallel, such as Mel-Cepstral Distortion (MCD), Fundamental Frequency (F0) root-mean-square error, and Mean Opinion Score (MOS), but they seldom frame them as a unified multi-objective problem amenable to joint optimization. This study addresses that gap.

The main contributions of this study are threefold. First, it presents a fidelity-prosody multi-objective perspective on data curation for low-resource TTS, framing the choice of preprocessing strategies as an explicit trade-off between spectral fidelity and prosodic integrity, and providing a structured way to compare competing pipelines on a common scale. Second, through a controlled ablation study on a real-world Zhuang corpus with noisy and atypical prosody, it is shown that a configuration that removes noise while disabling internal silence compression (`no_int_sil`) achieves the best aggregate ranking and the highest human MOS, with the best values on several prosody-sensitive metrics. In contrast, a traditional heavy-cleaning configuration (`full_method`) that compresses internal pauses improves the Fréchet Audio Distance (FAD) but does not enhance perceived naturalness and, in some cases, worsens it. Third, it offers empirical evidence that, in found-data conditions where internal pauses carry boundary-related prosodic information, over-aggressive silence compression can act as a prosody-distorting operation in a tonal language: by shortening linguistically functional pauses and altering F0 timing, it degrades rhythm and intonation in ways that are not captured by spectral-distance metrics alone, underscoring the need to prioritize prosodic integrity and to rely on prosody-aware metrics when selecting data-curation strategies in similar found-data settings.

2. RELATED WORK AND BACKGROUND

This study addresses data curation challenges in low-resource tonal-language TTS, with theoretical and practical roots in data-centric artificial intelligence, low-resource speech processing, and prosody modeling. Figure 1 provides a high-level overview of the data curation pipeline and the fidelity-prosody trade-off that motivates this work. This section reviews core advances in these areas and focuses on the limitations of existing methods, thereby motivating the research questions and core contributions.

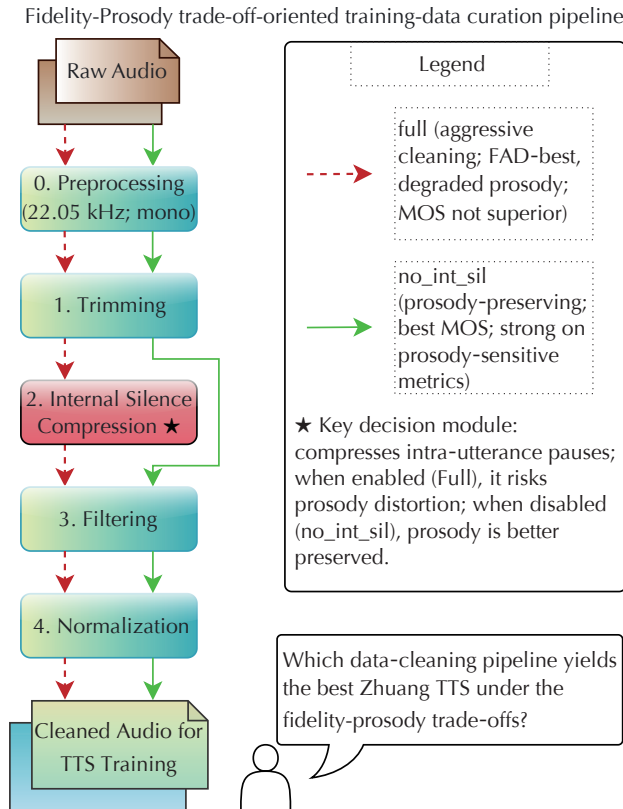


Figure 1. Overview of the data curation pipeline oriented towards the fidelity-prosody trade-off. Raw audio is preprocessed at 22.05 kHz (mono) and passes trimming, internal silence compression, filtering, and normalization. Red denotes the heavy-cleaning configuration `full_method`, which achieves the best FAD but degrades prosody and does not achieve superior MOS. Green denotes `no_int_sil`, which disables internal silence compression, preserves prosody, and achieves the best MOS. Section 4.4.1 details the operators.

2.1 The Data-Centric Paradigm in Low-Resource Speech Synthesis

As the capabilities of deep learning models grow, the research paradigm in artificial intelligence is shifting from model-centric to data-centric. A recent survey [13] argues that the performance bottleneck of many systems is shifting from model architecture to the design, development, and maintenance of training data.

This shift is particularly important for low-resource TTS. The mainstream response to data scarcity has been model-driven. Self-supervised learning pre-training on massive multilingual and multi-speaker corpora, followed by fine-tuning on a small amount of target data, has become a standard approach. A recent study [14] examines the language adaptation of self-supervised TTS models in low-resource scenarios and finds that phonetic similarity between pre-training and target languages, as well as the amount of fine-tuning data, affects final performance. Nevertheless, these methods primarily address how to compensate for data deficiencies with stronger models and more sophisticated fine-tuning strategies.

The data-centric perspective offers a complementary view: instead of continually enhancing models, it is important to systematically improve data quality. For example, a study on TTS for Indian languages [2] reports that low-cost, targeted data augmentation, such as additional recordings for specific phonetic combinations, can significantly improve performance. Specifically, targeted data engineering often yields performance gains that are comparable to architectural modifications in low-resource settings. However, prior work primarily focuses on data augmentation or model adaptation and rarely analyzes how data curation strategies should be balanced when the source data itself has prosodic defects. In this context, the present study adopts a data-centric perspective on low-resource TTS by treating data curation decisions themselves as primary optimization variables rather than fixed preprocessing choices.

2.2 Prosody Modeling and Increasing Dependence on Data

Prosody is a collection of suprasegmental features, including F0, duration, and intensity, that play a decisive role in the naturalness, emotion, and intelligibility of synthesized speech. A systematic review on prosody evaluation [5] notes that F0 is the most studied parameter in prosody research and that objective acoustic metrics, such as F0 root-mean-square error, are standard tools for measuring the closeness of synthesized prosody to natural prosody.

The research community has made substantial progress on the model side to generate expressive prosody. From the prosody encoder in Tacotron [15], to explicit F0 and energy modeling in FastSpeech 2 [16], and more recent approaches such as ProsodyFlow [17] that leverage large speech language models to generate diverse prosody, architectures have become increasingly complex. These models can learn and reproduce very subtle prosodic variations present in the training data.

This modeling power increases dependence on data. If the prosodic patterns in the training data are flawed, even the most powerful models reproduce these flaws with high fidelity. For tonal languages, this problem becomes more severe. In such languages, tone, primarily reflected by the F0 contour, is not merely a stylistic feature but a core linguistic device that distinguishes word meanings. For Zhuang, for example, the syllable *gyaem* can mean accented (Tone 1, *gyaem*), firm (Tone 3, *gyaemj*), moreover (Tone 4, *gyaemx*), purple (Tone 5, *gyaemq*), or repeat (Tone 6, *gyaemh*), depending solely on the tone [6, 18]. In this situation, faithfully reproducing a degraded F0 contour leads directly to semantic errors as well as reduced naturalness.

2.3 The Research Gap: Conflict Between Data Cleaning and Prosody Modeling

Traditional data cleaning pipelines target spectral fidelity. Examples include filtering based on automatic speech recognition confidence [19] or anomaly detection based on acoustic features [20], with the goal of removing noise, reverberation, disfluencies, and abnormal pauses. These operations improve spectral purity but rarely consider the effects on higher-level linguistic features, such as prosody.

A typical conflict arises in the treatment of silence. An algorithm designed to remove hesitation pauses may also eliminate short but grammatically necessary pauses that mark phrase boundaries, thereby destroying the prosodic structure. Overly aggressive denoising may smooth out subtle F0 variations that are crucial for tone perception, resulting in tonal distortion. When training data already exhibit atypical prosodic patterns, for example, read-aloud speech used for language teaching or automatic speech recognition corpus collection [21], the problem becomes even more complex. Blindly encouraging a model to reproduce this atypical prosody can yield speech that sounds unnatural, despite good performance on objective matching metrics.

Reflections on evaluation methods support this view. A TTS evaluation study [5] argues that relying solely on spectral distortion metrics, such as MCD, is insufficient for a comprehensive assessment. Prosodic naturalness, beyond spectral fidelity, has become a key dimension of TTS quality.

In summary, there is a gap in current research: when processing low-resource data that contains both noise and atypical prosody, there is an inherent conflict between improving spectral fidelity and maintaining or correcting prosodic naturalness. Existing research lacks a theoretical framework that systematically models, analyzes, and balances this conflict.

To fill this gap, this study formalizes TTS data curation as a multi-objective optimization problem concerning fidelity and prosody. The proposed framework provides theoretical guidance and a practical path for finding an effective operating point in the fidelity-prosody space, starting from the question of which data-curation strategy maximizes downstream model performance. This direction builds on prior multi-objective work in speech enhancement [22] and voice conversion [23], but shifts the focus from model losses to upstream data-engineering decisions. The present study applies this perspective to a low-resource tonal TTS scenario and uses it to organize the empirical comparison of data-curation strategies.

3. METHODOLOGY: A FORMAL FRAMEWORK FOR FIDELITY-PROSODY TRADE-OFFS

This section formalizes data curation for low-resource TTS as a multi-objective evaluation problem rather than a conventional denoising task. The goal is to compare a discrete set of preprocessing strategies in terms of two quality dimensions: spectral fidelity and prosodic behavior. The framework provides a common scale on which heterogeneous objective metrics are aggregated and used to analyze trade-offs. Section 3.2 describes the strict alignment protocol that fixes the data roster across strategies, and Section 3.4 connects the framework to the concrete objective metrics and composite scores.

3.1 Problem Formalization and Strategy Space

Rather than proposing a new optimization algorithm, this subsection introduces a formal lens for organizing and comparing data-curation strategies along two axes: spectral fidelity and prosodic behavior. We model data-curation decisions as a choice among a discrete set of preprocessing strategies that trade off spectral fidelity and prosodic behavior. Let $D = \{d_i\}_{i=1}^N$ denote the fixed training and validation roster under strict alignment (Section 3.2), with $N = 15,100$. The raw pool of approximately 62,119 utterances and the screened working set of 16,798 utterances are supersets of this roster.

Each strategy corresponds to a parameter setting $\theta \in \Theta$ of a preprocessing pipeline T_θ that applies trimming, internal silence compression, filtering, and normalization before model training (Section 4.4.1). In this study, the search space is restricted to a finite set of seven linguistically motivated strategies $\Theta = \{\theta_1, \dots, \theta_7\}$.

To compare strategies, we introduce two conceptual objectives on D : spectral fidelity $\mathcal{F}(\theta)$ and prosodic performance $\mathcal{P}(\theta)$. For a given configuration θ , let $\psi^*(\theta)$ denote the parameters of the converged TTS model trained on the preprocessed data $T_\theta(D)$, where the superscript $*$ indicates that training has converged to a (local) optimum under the standard VITS loss. Conceptual per-sample contributions $f(d; \theta, \psi^*(\theta))$ and $p(d; \theta, \psi^*(\theta))$ measure fidelity and prosody on item d , aligned such that larger values indicate better performance. The roster-level objectives are defined as expectations.

$$\mathcal{F}(\theta) = \mathbb{E}_{d \in D}[f(d; \theta, \psi^*(\theta))] \quad (1)$$

$$\mathcal{P}(\theta) = \mathbb{E}_{d \in D}[p(d; \theta, \psi^*(\theta))] \quad (2)$$

For a given fidelity-prosody preference weight $\alpha \in [0, 1]$, a scalarized utility

$$S_\alpha(\theta) = \alpha \mathcal{F}(\theta) + (1 - \alpha) \mathcal{P}(\theta), \quad \alpha \in [0, 1] \quad (3)$$

provides an ordering of strategies $\theta \in \Theta$. In this work, Eq. (3) is used as an evaluation criterion on a discrete candidate set rather than as an objective for gradient-based optimization. The true expectations $\mathcal{F}(\theta)$ and $\mathcal{P}(\theta)$ are approximated by the test-set composites introduced below.

3.2 Practical Instantiation: Strict Alignment

To ensure that performance differences arise solely from preprocessing decisions rather than roster variations, a strict alignment protocol is adopted. From the raw pool of approximately 62,119 utterances, a screened working set of 16,798 utterances is constructed using strategy-agnostic criteria such as duration and amplitude bounds. Internal silence compression and other strategy-specific transforms are *not* applied during screening.

A held-out test set of size $U = 1,698$ is reserved. The fixed training and validation roster with $N = 15,100$ items is split once as

$$|D_{\text{train}}| = 13,430, \quad |D_{\text{val}}| = 1,670$$

with a fixed random seed shared across all strategies. No test-item audio is used for training or validation. Early stopping and checkpoint selection, when employed, use identical criteria on the validation set. Under strict alignment, every strategy $\theta \in \Theta$ is trained and validated on the same list of utterance identifiers; only the waveforms differ because of T_θ . This design controls for sample selection effects and ensures that observed performance differences arise from preprocessing choices rather than roster variation. The complete empirical evaluation workflow under this strict-alignment design is summarized in Algorithm 1.

Algorithm 1. Empirical evaluation under strict alignment.

Input: Fixed roster D ($N = 15,100$), one-time split $D_{\text{train}}, D_{\text{val}}$; test transcripts $\mathcal{T}_{\text{test}} = \{t_u\}_{u=1}^U$ and references $A_{\text{ref}} = \{a_u^{\text{ref}}\}_{u=1}^U$ with $U = 1,698$; strategies $\Theta = \{\theta_1, \dots, \theta_7\}$; pipelines $\{T_\theta\}$ (roster-invariant); metric families $\mathcal{M}_{\text{fid}}, \mathcal{M}_{\text{pro}}$; direction signs $s_m \in \{+1, -1\}$; a small constant $\varepsilon = 10^{-8}$ to avoid division by zero during standardization; model: VITS

Output: $(\hat{\mathcal{F}}_{\text{test}}(\theta), \hat{\mathcal{P}}_{\text{test}}(\theta))$ for each θ ; Pareto set Θ_{pareto} ; preference-based choice θ_α^*

for $\theta \in \Theta$ **do**

$$D_{\text{train}}^{(\theta)} \leftarrow T_\theta(D_{\text{train}}), D_{\text{val}}^{(\theta)} \leftarrow T_\theta(D_{\text{val}});$$

$$\text{Ensure } |D_{\text{train}}^{(\theta)}| = |D_{\text{train}}| \text{ and } |D_{\text{val}}^{(\theta)}| = |D_{\text{val}}|;$$

$$\psi^*(\theta) \leftarrow \text{FitVITS}(D_{\text{train}}^{(\theta)}, D_{\text{val}}^{(\theta)})$$

for $u = 1$ **to** U **do**

$$\hat{a}_u(\theta) \leftarrow \text{Synthesize}(\psi^*(\theta), t_u)$$

$$\hat{A}_{\text{test}}(\theta) \leftarrow \{\hat{a}_u(\theta)\}_{u=1}^U$$

$$M_{\text{FAD}}(\theta) \leftarrow \text{FAD}(A_{\text{ref}}, \hat{A}_{\text{test}}(\theta)), M_{\text{FD}}(\theta) \leftarrow \text{FD}(A_{\text{ref}}, \hat{A}_{\text{test}}(\theta));$$

for $u = 1$ **to** U **do**

$$\begin{aligned} & m_u^{\text{set}}(\text{FAD}; \theta), m_u^{\text{set}}(\text{FD}; \theta); \\ & m_u(\text{MCD}; \theta), m_u(\text{DurMAPE}; \theta), (m_u(\text{F0-RMSE}; \theta), m_u(\text{F0-Pearson}; \theta)); \\ & m_u(\text{PauseError}; \theta), m_u(\text{PauseRatio}; \theta) \end{aligned}$$

for $m \in \mathcal{M}_{\text{fid}} \cup \mathcal{M}_{\text{pro}}$ **do**

$$v_{m,u}(\theta) \leftarrow s_m m_u(\theta)$$

$$\mu_m \leftarrow \frac{1}{|\Theta|U} \sum_{\theta \in \Theta} \sum_{u=1}^U v_{m,u}(\theta), \sigma_m^2 \leftarrow \frac{1}{|\Theta|U-1} \sum_{\theta \in \Theta} \sum_{u=1}^U (v_{m,u}(\theta) - \mu_m)^2;$$

$$z_{m,u}(\theta) \leftarrow \frac{v_{m,u}(\theta) - \mu_m}{\sigma_m + \varepsilon}$$

for $\theta \in \Theta$ **do**

Algorithm 1. *Continued.*

$$\begin{cases} \tilde{m}(\theta) \leftarrow \frac{1}{U} \sum_{u=1}^U z_{m,u}(\theta) \text{ for all } m; \\ \hat{\mathcal{F}}_{\text{test}}(\theta) \leftarrow \frac{1}{|\mathcal{M}_{\text{fid}}|} \sum_{m \in \mathcal{M}_{\text{fid}}} \tilde{m}(\theta), \hat{\mathcal{P}}_{\text{test}}(\theta) \leftarrow \frac{1}{|\mathcal{M}_{\text{pro}}|} \sum_{m \in \mathcal{M}_{\text{pro}}} \tilde{m}(\theta) \end{cases}$$

Define $\mathbf{s}(\theta) = (\hat{\mathcal{F}}_{\text{test}}(\theta), \hat{\mathcal{P}}_{\text{test}}(\theta))$;

$\Theta_{\text{Pareto}} \leftarrow \{\theta \in \Theta : \nexists \theta' \in \Theta \text{ s.t. } \mathbf{s}(\theta') \geq \mathbf{s}(\theta), \mathbf{s}(\theta') \neq \mathbf{s}(\theta)\}$

Given $\alpha \in [0, 1]$, let $S_\alpha(\theta) = \alpha \hat{\mathcal{F}}_{\text{test}}(\theta) + (1-\alpha) \hat{\mathcal{P}}_{\text{test}}(\theta)$ and choose $\theta_\alpha^* \in \arg \max_{\theta \in \Theta} S_\alpha(\theta)$;

3.3 Test-Set Evaluation: Pairwise and Distribution-Level Metrics

Evaluation uses the held-out test set with transcripts and references

$$A_{\text{ref}} = \{a_u^{\text{ref}}\}_{u=1}^U, \quad \hat{A}_{\text{test}}(\theta) = \{\hat{a}_u(\theta)\}_{u=1}^U \quad (4)$$

where $\hat{a}_u(\theta)$ is obtained by synthesizing from $\psi^*(\theta)$ on transcript t_u . For convenience, write $\hat{A}_{\text{test}}^{(i)} = \hat{A}_{\text{test}}(\theta_i)$.

Two categories of objective metrics instantiate fidelity and prosody:

Distribution-level metrics. FAD and Fréchet Distance (FD) compare the distributions of deep audio or speech embeddings extracted from A_{ref} and $\hat{A}_{\text{test}}(\theta)$, producing a single scalar $M(\theta)$ for each metric and strategy. To combine them with utterance-level metrics, this scalar is broadcast to all test items as

$$m_u^{\text{set}}(\theta) = M(\theta) \quad \text{for all } u \in \{1, \dots, U\} \quad (5)$$

Utterance-level metrics. For each pair $(a_u^{\text{ref}}, \hat{a}_u(\theta))$, fidelity is quantified by MCD; prosody is quantified by duration mean absolute percentage error, F0 root-mean-square error, F0 Pearson correlation, pause sequence alignment error, and pause ratio error. Section 4.7 provides formal definitions and alignment procedures. These per-utterance quantities can be grouped into fidelity and prosody contributions,

$$f_u(\theta) = f(a_u^{\text{ref}}, \hat{a}_u(\theta); \theta, \psi^*(\theta)) \quad (6)$$

$$p_u(\theta) = p(a_u^{\text{ref}}, \hat{a}_u(\theta); \theta, \psi^*(\theta)) \quad (7)$$

which are aggregated into scalar estimates of $\mathcal{F}(\theta)$ and $\mathcal{P}(\theta)$ via the standardization and compositing procedure in the next subsection.

3.4 Metric Families, Direction Alignment, and Standardization

The test-set metrics are grouped into two families:

$$\mathcal{M}_{\text{fid}} = \{\text{FD}, \text{FAD}, \text{MCD}\} \quad (8)$$

$$\mathcal{M}_{\text{pro}} = \{\text{DurMAPE}, \text{F0-RMSE}, \text{F0-Pearson}, \text{PauseError}, \text{PauseRatio}\} \quad (9)$$

with $u \in \{1, \dots, U\}$ indexing test items.

Step 1: Direction alignment and global standardization. For any metric m and utterance u , let $m_u(\theta)$ denote the raw value for strategy θ , where distribution-level metrics use the broadcast values in Eq. (5). A sign $s_m \in \{+1, -1\}$ is assigned so that

$$v_{m,u}(\theta) = s_m m_u(\theta) \quad (10)$$

is monotonically increasing in performance. Standardization statistics are computed over the Cartesian product $\Theta \times \{1, \dots, U\}$:

$$\mu_m = \frac{1}{|\Theta|U} \sum_{\theta \in \Theta} \sum_{u=1}^U v_{m,u}(\theta) \quad (11)$$

$$\sigma_m^2 = \frac{1}{|\Theta|U-1} \sum_{\theta \in \Theta} \sum_{u=1}^U (v_{m,u}(\theta) - \mu_m)^2 \quad (12)$$

The standardized score (z-score) is

$$z_{m,u}(\theta) = \frac{v_{m,u}(\theta) - \mu_m}{\sigma_m + \varepsilon}, \quad \tilde{m}(\theta) = \frac{1}{U} \sum_{u=1}^U z_{m,u}(\theta) \quad (13)$$

where $\varepsilon = 10^{-8}$ is a small constant to prevent division by zero when a metric has zero variance, and $\tilde{m}(\theta)$ is the per-strategy average of the standardized metric across the test set.

Step 2: Class-level composites. For each strategy θ , composite scores are defined as equal-weight averages within each family:

$$\hat{\mathcal{F}}_{\text{test}}(\theta) = \frac{1}{|\mathcal{M}_{\text{fid}}|} \sum_{m \in \mathcal{M}_{\text{fid}}} \tilde{m}(\theta) \quad (14)$$

$$\hat{\mathcal{P}}_{\text{test}}(\theta) = \frac{1}{|\mathcal{M}_{\text{pro}}|} \sum_{m \in \mathcal{M}_{\text{pro}}} \tilde{m}(\theta) \quad (15)$$

These estimates serve as empirical surrogates for $\mathcal{F}(\theta)$ and $\mathcal{P}(\theta)$ in Eq. (3).

In practice, our composite scores are computed in four steps: (i) align each metric direction so that larger values are better; (ii) standardize metric values across all strategies and test utterances using z-scores to remove scale differences; (iii) average standardized metrics within the fidelity family and within the prosody family to obtain $\hat{\mathcal{F}}_{\text{test}}(\theta)$ and $\hat{\mathcal{P}}_{\text{test}}(\theta)$; and (iv) optionally combine the two families with a preference weight α to obtain a scalar utility for sensitivity analysis.

3.5 Pareto Frontier and Discrete Search

Each strategy θ is summarized by a bi-objective performance vector

$$\mathbf{s}(\theta) = (\hat{\mathcal{F}}_{\text{test}}(\theta), \hat{\mathcal{P}}_{\text{test}}(\theta))$$

Inequalities on $\mathbf{s}(\cdot)$ are understood componentwise. A strategy θ' strictly dominates θ if

$$\theta' \succ \theta \Leftrightarrow \mathbf{s}(\theta') \geq \mathbf{s}(\theta) \text{ and } \mathbf{s}(\theta') \neq \mathbf{s}(\theta) \quad (16)$$

The empirical Pareto set is then

$$\Theta_{\text{Pareto}} = \{\theta \in \Theta : \nexists \theta' \in \Theta \text{ with } \theta' \succ \theta\} \quad (17)$$

In principle, one could pose a bilevel optimization problem in which the lower level trains the TTS model for each pipeline configuration, and the upper level searches over a continuous strategy space. In practice, the relationship between discrete preprocessing choices and downstream metrics is non-differentiable, and the test-set objectives are set-based and non-convex. This study, therefore, adopts a discrete empirical search: the finite set Θ is enumerated, one model $\psi^*(\theta_i)$ is trained per strategy under matched conditions, and the non-dominated frontier Θ_{Pareto} is identified in the $(\hat{\mathcal{F}}_{\text{test}}, \hat{\mathcal{P}}_{\text{test}})$ plane.

To obtain a single recommendation for a given preference weight $\alpha \in [0, 1]$, the post hoc scalarization

$$S_\alpha(\theta) = \alpha \hat{\mathcal{F}}_{\text{test}}(\theta) + (1 - \alpha) \hat{\mathcal{P}}_{\text{test}}(\theta), \quad \alpha \in [0, 1] \quad (18)$$

is used, and

$$\theta_\alpha^* \in \arg \max_{\theta \in \Theta} S_\alpha(\theta) = \arg \max_{\theta \in \Theta} [\alpha \hat{\mathcal{F}}_{\text{test}}(\theta) + (1 - \alpha) \hat{\mathcal{P}}_{\text{test}}(\theta)] \quad (19)$$

This discrete scalarization is consistent with the conceptual objective in Eq. (3) and is used in the analysis to illustrate how the preferred strategy varies with the relative weight placed on fidelity and prosody.

3.6 Curation Pipeline and the Role of Filtering

The data-curation pipeline T_θ comprises four configurable operators: trimming, internal silence compression, filtering, and normalization. Filtering computes quality-control signals, such as duration range, root-mean-square energy bounds, zero-crossing rate thresholds, and short-time Fourier transform (STFT) magnitude statistics. Under strict alignment, filtering acts in a monitor-only fashion: batches are indexed by a single shared 15,100-utterance manifest across all strategies, and roster membership does not change with the filtering switch. When a strategy enables filtering, thresholds generate quality-control flags and, where applicable, amplitude-level adjustments. Waveforms may change, but the set of utterance identifiers used for training and validation does not change. Quality-control flags are logged with stable identifiers. Missing files trigger consistency checks, which are resolved before training; extra files not in the manifest are ignored. The seven strategies differ only in operator activation and parameter settings.

4. EXPERIMENTAL SETUP

Based on the framework in Section 3, this section describes the dataset, preprocessing pipeline, model configuration, and evaluation protocol used to instantiate the seven data-curation strategies.

4.1 Dataset and Experimental Alignment

The experiments use an in-house, project-collected Zhuang automatic speech recognition (ASR) corpus. The original recording pool comprises approximately 100 hours of audio (approximately 62,119 utterances) recorded by multiple volunteer speakers under diverse real-world conditions. Due to extreme environmental noise and recording inconsistency, a multi-stage screening process was applied: strategy-agnostic criteria such as duration and amplitude bounds yielded a screened working set of 16,798 utterances, from which the fixed training and validation roster of 15,100 utterances (approximately 24 hours) was drawn under the strict alignment protocol described in Section 3.2. The remaining 1,698 utterances form the held-out test set. In this low-resource regime, the retained subset should be interpreted as in-the-wild speech rather than studio-recorded speech: compared with standard corpora such as LJSpeech, it exhibits a higher noise floor and a more didactic speaking style, which increases acoustic variability and reduces prosodic consistency for TTS training.

Wuming Zhuang is a Kra-Dai (Tai-Kadai) tonal language. Under the standard Zhuang Latin orthography adopted in this work, a lexical syllable is conventionally analyzed as an onset, a rime, and a tone category (Table 1). Tone categories are only partially encoded in the orthography. For smooth (non-checked) syllables, Tone 1 is zero-marked (no tone letter), whereas Tones 2–6 are marked by the tone letters *z*, *j*, *x*, *q*, and *h*, respectively. For checked-tone syllables, tone categories are indicated by rime-final stop codas *p/t/k*

Table 1. Phonological inventory and tone marking in the standard Zhuang orthography used in this work.

Alphabet	Aa, Bb, Cc, Dd, Ee, Ff, Gg, Hh, Ii, Kk, Ll, Mm, Nn, Oo, Pp, Qq, Rr, Ss, Tt, Uu, Vv, Ww, Xx, Yy, Zz
Initials	b, mb, m, f, v, d, nd, n, s, l, g, gv, ng, h, r, c, y, ny, ngv, by, gy, my
Finals	(Vowels) a, e, i, o, u, w (Diphthongs) ae, ai, au, aw, ei, eu, ie, iu, oe, oi, ou, ui, uw (Nasals) am, an, ang, em, en, eng, im, in, ing, om, on, ong, um, un, ung (Stops) (-p/-t/-k group) ap, at, ak, ep, et, ek, ip, it, ik, op, ot, ok, up, ut, uk; (-b/-d/-g group) ab, ad, ag, eb, ed, eg, ib, id, ig, ob, od, og, ub, ud, ug
Tone marking	Tone 1 on smooth (non-checked) syllables is zero-marked (no tone letter). Tones 2–6 on smooth syllables are marked by the tone letters <i>z</i> , <i>j</i> , <i>x</i> , <i>q</i> , <i>h</i> , respectively. Checked-tone syllables are indicated by rime-final stop codas <i>p/t/k</i> or <i>b/d/g</i> .

Note: The phonological inventory follows the standard Zhuang Latin orthography; International Phonetic Alphabet (IPA) transcriptions are not used in this work.

or *b/d/g*. This writing-system coupling between tone marking, F0 realization, and syllable timing provides a direct rationale for prioritizing prosodic-structure preservation in Zhuang TTS.

Despite the presence of tone information in the text, the speech recordings lack syllable-level tone labels aligned to the waveform, and reliable time-aligned syllable boundaries are not available for this project-collected ASR corpus. Consequently, we do not report tone-category accuracy. Instead, we adopt boundary-conditioned F0 diagnostics in pause-following phrase-initial regions as a tone-relevant proxy that does not require explicit tone annotations.

To ensure fairness in the ablation study, all strategies share a single roster manifest and random seed, as described in Section 3.2. A shared manifest lists 15,100 utterances, and all data loaders index batches by this list. Preprocessing may differ across strategies, but each strategy is trained and validated on the same roster. Extra files produced by preprocessing are ignored. Quality-control checks are used for logging and diagnostics and do not remove items from the roster. If an item in the shared manifest cannot be produced by a given preprocessing configuration, it is treated as an implementation inconsistency and resolved prior to training to preserve strict roster alignment across strategies.

4.2 Analysis of Raw Data Heterogeneity and Atypical Prosody

Before model training, we profile the corpus heterogeneity and atypical prosodic patterns by analyzing both the raw recording pool and the strategy-instantiated datasets. Table 2 shows a highly imbalanced speaker distribution, with some speakers contributing thousands of utterances and others only a few. These very low-sample speakers are intentionally retained to reflect a realistic low-resource scenario and to test robustness under speaker imbalance.

Table 2. Speaker-Level Statistics of the Zhuang Corpus.

ID	Utterances	Dur. (min)	ID	Utterances	Dur. (min)	ID	Utterances	Dur. (min)
15	2224	149.08	0	715	44.48	5	67	8.47
4	1588	172.28	13	636	84.82	18	67	7.43
16	1533	165.85	3	630	60.48	7	25	2.85
20	1471	164.55	9	621	75.66	12	23	1.34
10	1442	142.70	1	610	66.25	11	10	0.78
22	1013	98.69	17	364	40.54	2	2	0.09
21	837	47.22	19	346	28.46	14	2	0.09
8	793	72.78	6	81	8.70			

The dataset contains 23 speakers totaling 15,100 utterances ($\approx$ 24 h). Data is sorted by utterance count to illustrate the long-tail distribution.

Informal expert elicitation with native Zhuang speakers and advanced L2 learners indicated substantial prosodic variability in the raw corpus, motivating pause handling as a key design axis in data curation. Some speakers exhibit a didactic, highly articulated speaking style with a reduced speech rate and frequent within-utterance pauses, which can reduce perceived fluency while maintaining high segmental clarity. The resulting trade-off is whether aggressive pause correction can be applied without sacrificing prosodic structure. Figure 2 shows extreme dispersion in duration- and silence-related characteristics (duration, RMS, silence ratio, and ZCR) across representative strategies, highlighting the corpus as an informative testbed for pause-handling decisions.

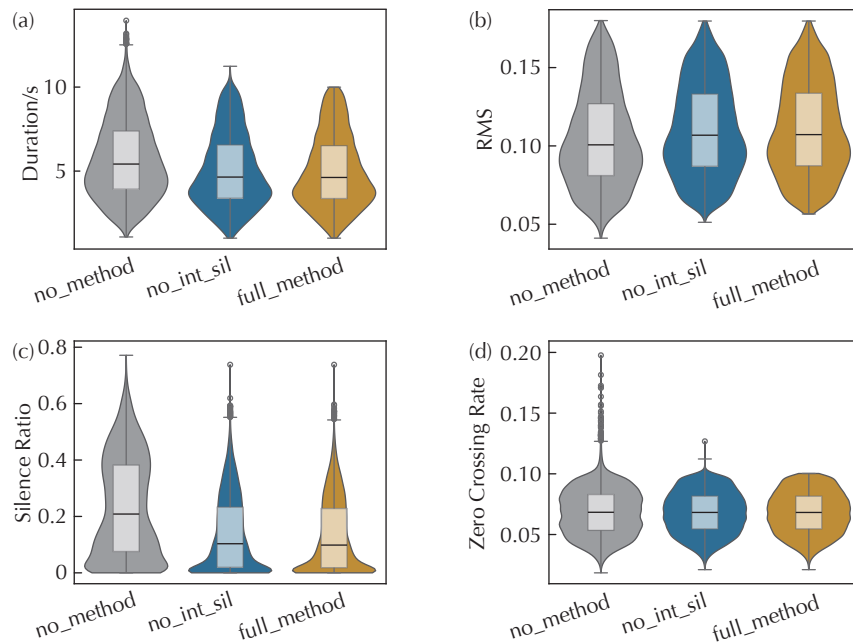


Figure 2. Distributions of four acoustic features on the training data after strategy-specific preprocessing for three core strategies: the baseline system no_method, the heavily cleaned configuration full_method, and the proposed system no_int_sil. Panels (a)–(d) show duration (seconds), root-mean-square (RMS) energy, silence ratio, and zero-crossing rate, respectively. Each violin indicates the estimated distribution density; the overlaid boxplot shows the interquartile range and median; the overlaid boxplot shows the interquartile range and median.

For raw corpus profiling (e.g., the benchmark statistics in Table 3), we scan recordings in multiple audio formats and decode them to a unified 22.05 kHz mono representation. Feature extraction uses top_db = 40 (as in librosa.effects.split) and clips analysis to a maximum duration of 12 s per recording. Decoding failures are rare (<0.05%), and no threshold-based filtering is applied at this stage; therefore, these raw statistics reflect the near-complete pool prior to strategy-specific screening. This profiling clip is used only for raw-pool analysis and is independent of the training-time duration filter reported in Table 4. Strategy-level feature distributions are reported separately in Figure 2.

Table 3. Illustrative silence statistics of the raw Zhuang corpus versus common benchmarks.

Dataset	Utterances	Silence ratio (mean)	Duration (mean)
LJSpeech	13,100	0.0215	6.57 s
VCTK [24] (22 kHz)	44,242	0.0115	3.56 s
Zhuang (Raw)	62,119	0.1675	7.03 s

Note: Silence ratio is computed from non-silent intervals estimated with $\text{top_db} = 40$ (librosa) as in Eq. (25). Statistics are reported on the raw pool (prior to strategyspecific screening and strict-alignment roster fixing).

Table 4. Key threshold settings for preprocessing.

Module	Parameter	Value
Resampling	Target sampling rate f_s	22.05 kHz
Trimming/split	Non-speech threshold (top_db)	40 dB
Internal silence compression	Max retained internal pause $\tau_{\max}$	0.8 s
Filtering	Duration range T	[1, 10] s
Filtering	Peak cap $A_{\max}$	0.9
Filtering	RMS bounds (pre-norm)	[0.06, 0.18]
Filtering	ZCR upper bound $ZCR_{\max}$	0.1
Filtering	STFT magnitude mean $\mu s $	≥ 0.25
Filtering	STFT magnitude std $\sigma s $	≥ 1.0
Normalization	Target RMS RMS_{tgt}	0.12

Note: Thresholds are fixed a priori to avoid test-set-driven tuning. STFT statistics are computed with $n_{\text{fft}} = 1024$ and hop length 256 (see Section 4.4.1). Internal silence compression truncates each within-utterance silence gap to at most $\tau_{\max}$ seconds when enabled.

4.3 Rationale for Threshold Selection

Internal silence compression in `full_method` truncates within-utterance silence gaps to a maximum retained duration $\tau_{\max}$. We fixed $\tau_{\max} = 0.8$ s to limit long internal pauses that are prevalent in the raw corpus, while avoiding overly aggressive temporal editing in the end-to-end ablation comparisons. As an illustrative benchmark, Table 3 reports the mean silence ratio of the raw Zhuang pool alongside common TTS datasets; the substantially higher silence ratio motivates introducing an explicit pause cap in `full_method`.

To make this hyperparameter choice transparent, we conduct a preprocessing-level sweep of $\tau_{\max}$ from 0.4 s to 1.2 s in steps of 0.2 s under the same trimming procedure, internal silence compression

procedure, and threshold settings used by full_method. All sweep statistics are computed on the same fixed roster (train + val, $N = 15,100$) to isolate the effect of $\tau_{\max}$ from sample selection. The sweep results are reported later in Section 5.3. Based on this analysis, we adopt $\tau_{\max} = 0.8$ s as a fixed operating point for the controlled end-to-end comparisons in the ablation study.

As detailed in Section 5.3, decreasing $\tau_{\max}$ monotonically increases the amount of temporal editing ($\Delta\bar{T}$) on the same fixed roster ($N = 15,100$), whereas coarse low-level statistics (RMS and ZCR) change minimally. We therefore selected $\tau_{\max} = 0.8$ s as a conservative mid-range operating point with diminishing marginal changes beyond 0.8 s.

Within the family of fixed-cap internal silence compression heuristics (as in full_method), a larger $\tau_{\max}$ corresponds to weaker compression and thus reduces the risk of prosody distortion in boundary-adjacent regions. However, any global cap remains a coarse rule that does not distinguish linguistically functional boundary pauses from non-functional disfluency pauses. Designing pause-type-aware and context-adaptive compression policies is left for future work.

Accordingly, our main comparison focuses on whether applying a global fixed-cap pause truncation is preferable to preserving internal pauses (no_int_sil), rather than exhaustively optimizing $\tau_{\max}$.

4.4 Implementation Details

4.4.1 Data Preprocessing

All raw audio files first undergo a uniform preprocessing step involving resampling to 22.05 kHz and mixing to mono. As visualized in Figure 1, the subsequent data-curation pipeline consists of four optional modules: trimming, internal silence compression, filtering, and normalization. The seven strategies listed in Table 5 represent discrete configurations of this pipeline and serve as instantiations of the parameter θ within the theoretical framework.

Table 5. Configuration of Data Curation Strategies and Operator Settings.

Strategy	Trimming	Internal Silence Compression	Filtering	Normalization
no_method	×	×	×	×
no_norm	✓	✓	✓	×
no_filter	✓	✓	×	✓
trim + norm	✓	×	×	✓
no_silence	×	×	✓	✓
full_method	✓	✓	✓	✓
no_int_sil	✓	×	✓	✓

Note: ✓ = enabled; × = disabled. The shaded row highlights the proposed strategy.

In the trimming stage, leading and trailing non-speech intervals are removed based on an energy threshold. This operation, implemented via *librosa* [25], improves boundary alignment and stabilizes subsequent duration-based metrics without altering linguistic content.

Internal silence compression targets within-utterance pauses. Non-silent intervals are identified, and each silence gap between consecutive non-silent regions is truncated such that at most $\tau_{\max}$ seconds of silence are retained. This suppresses prolonged hesitations while preserving short, potentially functional pauses.

Filtering computes quality-control signals by thresholding duration, RMS amplitude, zero-crossing rate, and STFT magnitude statistics; violating utterances are flagged (and, where applicable, amplitude-adjusted) but, under strict alignment, are never removed from the roster. When enabled, utterances are flagged if they violate criteria based on duration, RMS amplitude, zero-crossing rate, or STFT magnitude statistics. Finally, a two-stage level normalization applies peak limiting to prevent clipping and RMS-based gain adjustment with a gain cap to avoid amplifying background noise.

Unless otherwise stated, preprocessing uses a fixed set of thresholds shared across strategies to ensure reproducibility and avoid test-set-driven tuning. Trimming and non-silent interval detection use $\text{top_db} = 40$. In *full_method*, internal silence compression truncates each detected silence gap to at most $\tau_{\max} = 0.8$ s. Quality-control checks apply to the post-trimmed (and, when enabled, post-compressed) waveform and include duration bounds of 1 s to 10 s, a peak cap of 0.9, RMS bounds of [0.06, 0.18], a zero-crossing rate upper bound of 0.1, and STFT-magnitude constraints $\text{mean}(|\text{STFT}|) \geq 0.25$ and $\text{std}(|\text{STFT}|) \geq 1.0$ (computed with $n_{\text{fft}} = 1024$ and a hop length of 256). RMS-based gain adjustment uses $\text{RMS}_{\text{tgt}} = 0.12$ as a reference level and is applied conservatively with gain limits. Non-silent interval detection follows the settings in Table 4.

Under strict alignment, the roster of utterance identifiers is fixed across strategies. Items that fail to meet a criterion under a given strategy are logged and resolved by consistency checks against the shared manifest prior to training (Section 3.2).

4.4.2 Model Architecture and Training Parameters

To ensure the identifiability of data-curation effects, we hold the acoustic-model backbone and the training protocol constant across strategies and vary only the preprocessing configuration.

A multi-speaker version of VITS [26] is used as the backbone for all seven strategies. The acoustic front-end follows a STFT configuration with an FFT size of 1024, a window length of 1024, a hop length of 256, and 80-channel mel-spectrograms. Core architectural hyperparameters are $\text{hidden_channels} = 192$, $\text{gin_channels} = 256$, and $n_{\text{layers}} = 6$. All models are trained independently on four NVIDIA RTX 3090 GPUs (24GB) under a unified schedule with a learning rate of 2×10^{-4} , a batch size of 32, and a fixed random seed (seed = 1234). Each strategy trains for approximately 200,000 steps (about 3.5 days), and losses converge to the 15.x–16.x range, supporting direct comparison under matched conditions across

curation strategies. All preprocessing, training, and evaluation code, including configuration files, will be released upon acceptance at <https://github.com/GUANGXI-MINZU-UNIVERSITY/yqs>.

4.5 Subjective Listener Panel

The subjective evaluation uses ratings from 20 listeners with substantial experience in the target Zhuang variety (Table 6). Seventeen listeners are undergraduates majoring in the Zhuang language at Guangxi Minzu University, who have received approximately one year of formal instruction and thus represent advanced late L2 learners. One listener is a third-year graduate student in Zhuang linguistics, and two listeners are Zhuang-language instructors with graduate-level training, including one with a doctoral degree; these three listeners are native speakers of the target Wuming Zhuang dialect. All listeners have normal hearing by self-report and are familiar with the orthographic conventions used in the transcripts. Each listener evaluated 18 synthesized speech samples, comprising six source utterances rendered under three curation strategies. Each sample was rated separately for clarity and overall naturalness, yielding 36 MOS ratings per listener. In the CMOS test, each listener evaluated 12 pairs of synthesized speech samples on the same two dimensions, yielding 24 CMOS ratings per listener.

Table 6. Listener composition for the MOS and CMOS tests (N = 20).

Group	Count	Description
Advanced L2 learners	17	Zhuang-major undergraduates (1 year of instruction)
Native/expert L1	3	1 graduate student + 2 instructors (Wuming Zhuang)
Total	20	All familiar with the target variety

4.6 Evaluation Overview

To evaluate the impact of different curation strategies on synthesized speech, the study uses a protocol covering three dimensions: acoustic fidelity, prosodic behavior, and human subjective perception. Objective metrics quantify spectral-envelope and suprasegmental phenomena, and controlled listening tests provide perceptual ground truth with statistical inference.

4.7 Objective and Subjective Metrics

For exposition, objective metrics are organized into two families that are used consistently throughout the paper. Fidelity metrics quantify spectral or distributional similarity to reference audio, whereas prosody metrics quantify temporal structure, pause behavior, and pitch-contour consistency.

This study evaluates curation strategies using complementary objective metrics for acoustic fidelity and prosodic behavior, together with controlled listening tests. For acoustic fidelity, it reports MCD and FAD and includes Fréchet Distance (FD), computed on Wav2Vec2-Base embeddings, as an additional distribution-level fidelity metric. MCD measures spectral-envelope distortion via the Euclidean distance

between mel-cepstral coefficients of synthesized and reference speech, with dynamic time warping alignment computed using pymcd [27]. FAD computes the Fréchet distance between distributions of deep audio embeddings (VGGish) extracted from two sets of waveforms [28], providing a distribution-level surrogate for perceived quality. FD applies the same computation to Wav2Vec2-Base embeddings [29]. Unless otherwise stated, FD is reported alongside FAD and is included in the fidelity composite used in the scalarized objective.

Prosody is quantified using the mean absolute percentage error of duration, F0 root-mean-square error, Pearson correlation of aligned F0 contours, pause error, and pause ratio error. The duration mean absolute percentage error captures the overall speaking-rate deviation. F0 root-mean-square error and F0 Pearson correlation compare intonation trajectories after alignment, with F0 extracted using WORLD [30-31]. Pause detection identifies internal silences using librosa; the pause error then matches pause sequences between synthesized and reference audio using the Hungarian algorithm, penalizing extraneous, mismatched, or missing pauses [25]. The pause ratio error compares the proportion of silence to the total duration between synthesized and reference audio. Together, these metrics cover rhythm and intonation cues that are critical to perceived naturalness and intelligibility; for a tonal language, they serve as proxies for tone realization in context.

Since human perception remains the definitive standard, controlled listening tests are conducted with 20 raters in a quiet environment, using high-fidelity headphones and a randomized presentation order. For the three core strategies (no_method, full_method, and no_int_sil), the tests collect MOS ratings on a five-point scale for clarity and overall naturalness, together with Comparative MOS (CMOS) ratings on a seven-point scale from -3 to $+3$ through pairwise comparisons of matched sentences. The primary inferential analysis of the MOS ratings uses a linear mixed-effects model [32] fitted to the unaggregated ratings, with strategy as a fixed effect and rater as a random intercept. This model accounts for the repeated ratings contributed by each listener. As a complementary nonparametric analysis, paired two-sided Wilcoxon signed-rank tests [33] are applied to listener-level condition means obtained by equally averaging the six synthesized speech samples within each strategy and perceptual dimension. Within each inferential method, Holm–Bonferroni correction [34] is applied jointly to all six pairwise MOS contrasts formed by the three strategy comparisons across the two perceptual dimensions. The mixed-effects model constitutes the primary repeated-measures analysis, whereas the Wilcoxon analysis is reported as a distribution-free robustness check. Effect estimates and Holm-adjusted p-values are reported in Section 5.1.

4.7.1 Formal Definitions

Let U denote the fixed held-out test set, indexed by $u \in \{1, \dots, |U|\}$. For each test item u , let a_u^{ref} be the reference waveform and $\hat{a}_u^{(s)}$ the waveform synthesized by strategy $s \in \Theta$.

MCD is computed per utterance using dynamically time-warped aligned mel-cepstral coefficient sequences. Let $\{\mathbf{c}_t^{\text{ref}}\}_{t=1}^T$ and $\{\mathbf{c}_t^{\text{syn}}\}_{t=1}^T$ be K -dimensional mel-cepstral coefficients extracted from a_u^{ref} and $\hat{a}_u^{(s)}$, respectively. Dynamic time warping yields an alignment path $\pi = \{(i_k, j_k)\}_{k=1}^L$. The utterance-level MCD is

$$\text{MCD}_u^{(s)} = \frac{10}{\ln 10} \cdot \frac{1}{L} \sum_{k=1}^L \sqrt{2 \sum_{d=1}^K (c_{i_k, d}^{\text{ref}} - c_{j_k, d}^{\text{syn}})^2} \quad (20)$$

where $c_{i_k, d}^{\text{ref}}$ and $c_{j_k, d}^{\text{syn}}$ are the d -th mel-cepstral coefficients ($d = 1, \dots, K$) at the aligned frame pair (i_k, j_k) determined by the dynamic time warping path π , and L is the total number of aligned frame pairs. The system-level MCD is the arithmetic mean $\text{MCD}^{(s)} = \frac{1}{|U|} \sum_u \text{MCD}_u^{(s)}$.

FD and FAD are defined via Fréchet distances between deep audio embedding distributions. For a fixed embedding function $\phi(\cdot)$ (Wav2Vec2-Base for FD and VGGish for FAD), define

$$\mathcal{X}^{\text{ref}} = \{\phi(a_u^{\text{ref}})\}_{u \in U}, \quad \mathcal{X}^{(s)} = \{\phi(\hat{a}_u^{(s)})\}_{u \in U}$$

Approximating each set by a multivariate Gaussian with mean and covariance (μ_r, Σ_r) and (μ_s, Σ_s) yields the Fréchet distance

$$\text{FD}^{(s)} = \|\mu_r - \mu_s\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_s - 2(\Sigma_r \Sigma_s)^{1/2}) \quad (21)$$

with FAD defined identically by changing ϕ .

The mean absolute percentage error of duration is computed from utterance durations. Let T_u^{ref} and $T_u^{(s)}$ denote the total durations in seconds of a_u^{ref} and $\hat{a}_u^{(s)}$, obtained from file metadata. The utterance-level duration mean absolute percentage error is

$$\text{DurMAPE}_u^{(s)} = \left| \frac{T_u^{(s)} - T_u^{\text{ref}}}{T_u^{\text{ref}}} \right| \quad (22)$$

and the system-level value is $\text{DurMAPE}^{(s)} = \frac{1}{|U|} \sum_u \text{DurMAPE}_u^{(s)}$.

Pause-based metrics operate on internal, within-utterance pauses derived from non-silent intervals. Let voiced segments be detected on a waveform y by `split()` (as implemented in `librosa.effects.split`) [25], yielding non-silent intervals $\{[s_i, e_i]\}_{i=1}^M$. Here and in Eq. (25), “voiced” refers to the energy-based non-silent segments returned by `split()`, rather than to phonetically voiced speech; the phonetic voiced/unvoiced distinction is used only for the F0 metrics in Eqs. (27)–(28). Internal pauses are the gaps between consecutive voiced segments,

$$\mathcal{P}(y) = \{(e_i, s_{i+1}) : s_{i+1} - e_i \geq \tau_{\min}\} \quad (23)$$

where $\tau_{\min}$ is a minimum pause duration. Denote by $\mathbf{p}_u^{\text{ref}} \in \mathbb{R}^m$ and $\mathbf{p}_u^{(s)} \in \mathbb{R}^n$ the resulting pause-length vectors in seconds for a_u^{ref} and $\hat{a}_u^{(s)}$. Pause error is defined via a minimum-cost assignment:

$$\text{PauseError}_u^{(s)} = \frac{1}{\max(m, n)} \left(\sum_{(i, j) \in \mathcal{M}} |p_{u, i}^{\text{ref}} - p_{u, j}^{(s)}| + \sum_{i \notin \mathcal{M}} p_{u, i}^{\text{ref}} + \sum_{j \notin \mathcal{M}} p_{u, j}^{(s)} \right) \quad (24)$$

where $\mathcal{M} \subseteq \{1, \dots, m\} \times \{1, \dots, n\}$ denotes the optimal one-to-one matching obtained by solving the linear assignment problem with cost matrix $C_{ij} = |p_{u,i}^{\text{ref}} - p_{u,j}^{(s)}|$ via the Hungarian algorithm; the notation $i \notin \mathcal{M}$ (resp. $j \notin \mathcal{M}$) refers to reference (resp. synthesized) pauses that remain unmatched, and their durations are added as penalties. The system-level pause error is the average over u .

A pause ratio metric operates on the fraction of silence. Let $\text{PR}(y)$ denote the pause ratio of y , defined as

$$\text{PR}(y) = 1 - \frac{\text{voiced_samples}(y)}{\text{total_samples}(y)} \quad (25)$$

using voiced intervals from $\text{split}(\cdot)$. The utterance-level pause ratio error is

$$\text{PauseRatio}_u^{(s)} = |\text{PR}(a_u^{\text{ref}}) - \text{PR}(\hat{a}_u^{(s)})| \quad (26)$$

F0-based metrics operate on frame-synchronous F0 contours and voiced/unvoiced masks extracted by WORLD at a fixed frame period. For each pair $(a_u^{\text{ref}}, \hat{a}_u^{(s)})$, let $f_{0,u}^{\text{ref}}[t]$, $f_{0,u}^{(s)}[t]$, and the voiced/unvoiced masks $v_u^{\text{ref}}[t]$, $v_u^{(s)}[t]$ be time-synchronized by frame index and define the jointly voiced mask $M_u = \{t : v_u^{\text{ref}}[t] = v_u^{(s)}[t] = 1\}$. The synchronized F0 root-mean-square error and Pearson correlation are

$$\text{F0-RMSE}_u^{(s)} = \sqrt{\frac{1}{|M_u|} \sum_{t \in M_u} (f_{0,u}^{\text{ref}}[t] - f_{0,u}^{(s)}[t])^2} \quad (27)$$

$$\text{F0-Pearson}_u^{(s)} = \text{corr}(f_{0,u}^{\text{ref}}[M_u], f_{0,u}^{(s)}[M_u]) \quad (28)$$

System-level values are obtained by averaging across utterances, excluding items with $|M_u|$ that are below a small threshold.

For duration error and correlation diagnostics, let d_u^{ref} and $d_u^{(s)}$ denote utterance durations. The absolute and relative errors are

$$\Delta d_u^{(s)} = |d_u^{(s)} - d_u^{\text{ref}}|, \quad \epsilon_u^{(s)} = \frac{d_u^{(s)} - d_u^{\text{ref}}}{d_u^{\text{ref}}} \quad (29)$$

with system-level summaries provided by mean absolute error $\text{MAE}^{(s)} = \mathbb{E}_u[\Delta d_u^{(s)}]$, mean absolute percentage error $\mathbb{E}_u[|\epsilon_u^{(s)}|]$, and Pearson correlation $\text{corr}(d_u^{\text{ref}}, d_u^{(s)})$.

Subjective MOS and CMOS metrics are derived from listener ratings. Let $r \in \mathcal{R}$ index raters and q index MOS questions. For MOS, each response $y_{rq}^{(d,s)} \in \{1, \dots, 5\}$ is recorded for dimension $d \in \{\text{clarity}, \text{overall}\}$ and strategy s . The strategy- and dimension-specific MOS is

$$\text{MOS}^{(d,s)} = \frac{1}{|\mathcal{R}_s^{(d)}|} \sum_{r \in \mathcal{R}_s^{(d)}} \bar{y}_r^{(d,s)}, \quad \bar{y}_r^{(d,s)} = \frac{1}{Q_r^{(d,s)}} \sum_{q \in Q_r^{(d,s)}} y_{rq}^{(d,s)} \quad (30)$$

where $Q_r^{(d,s)}$ is the set of questions answered by rater r for (d, s) , and $\mathcal{R}_s^{(d)}$ is the set of raters with at least one valid response.

For CMOS, each paired comparison between systems s_i and s_j yields a score $c_{r,p}^{(s_i, s_j)} \in \{-3, \dots, +3\}$ for rater r and pair index p . Symmetric net preference is computed by combining both presentation orders and normalizing by the total count:

$$\Delta_{\text{CMOS}}(s_i, s_j) = \frac{1}{N_{ij}^{\text{tot}}} \left(\sum_{r, p: s_i \succ s_j} c_{r,p}^{(s_i, s_j)} - \sum_{r, p: s_j \succ s_i} c_{r,p}^{(s_j, s_i)} \right) \quad (31)$$

with N_{ij}^{tot} equal to the total number of paired ratings for the unordered pair $\{s_i, s_j\}$. Bootstrap confidence intervals and permutation tests on $\Delta_{\text{CMOS}}(s_i, s_j)$ are used for inference.

5. RESULTS AND ANALYSIS

This section analyzes the impact of different data-curation strategies on noisy, atypically prosodic data by combining subjective listening tests with objective diagnostics. Section 5.1 reports MOS-based perceptual results. Section 5.2 summarizes objective fidelity and prosody metrics and highlights their disconnect. We then assess the robustness of the internal pause-cap choice in full_method via a preprocessing-level τ_{max} sweep (Section 5.3). Section 5.4 provides mechanistic evidence based on boundary-conditioned F0 behavior, and Section 5.5 instantiates the utility function and evaluates sensitivity to preference weights.

5.1 Subjective Evaluation Results: Superiority of the no_int_sil Strategy

The subjective listening tests demonstrated a consistent preference for no_int_sil over both baselines in clarity and overall naturalness. As shown in Figure 3, no_int_sil attained the highest MOS for both clarity and overall naturalness. The error bars represent 95% bootstrap confidence intervals of the mean across the 20 listeners. The improvements of no_int_sil over full_method and no_method range roughly from 0.3 to 0.5 on a five-point scale, representing a perceptible enhancement in synthesis quality.

The statistical robustness of this preference ordering was tested using two complementary inference methods, as detailed in Table 7. The Wilcoxon signed-rank test confirmed a significant advantage of no_int_sil over both baselines in clarity ($p_{\text{Holm}} < 0.05$). For naturalness against full_method, the same effect ($\Delta\text{MOS} = +0.317$) did not reach significance under the conservative Wilcoxon test ($p_{\text{Holm}} = 0.102$), whereas the primary, higher-powered linear mixed-effects model identified it as significant ($p_{\text{Holm}} = 0.019$). Because both analyses estimate the same effect on this balanced design, the divergence reflects a difference in statistical power rather than in the magnitude or direction of the effect. As a descriptive consistency check, we computed the mean MOS within the native/expert subgroup ($N = 3$) and obtained the same ordering as in the full panel, with no_int_sil rated highest, followed by full_method and no_method. Because the subgroup size is small, we do not perform separate statistical inference and report this result for qualitative completeness only.

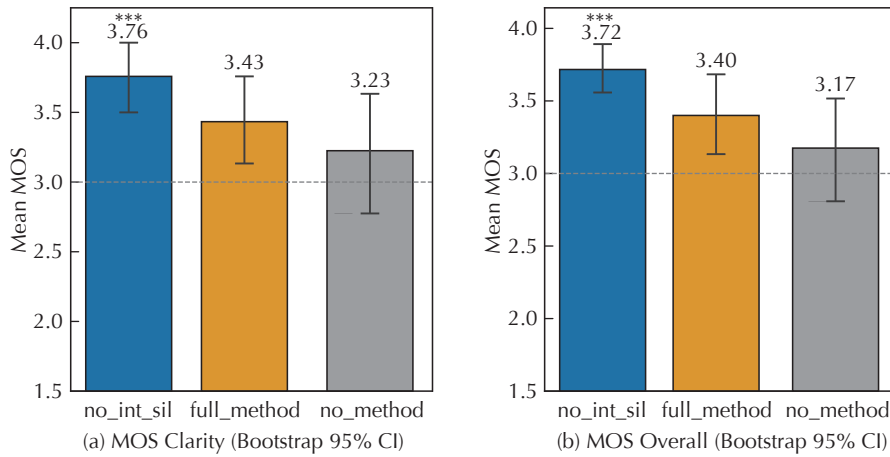


Figure 3. Mean MOS and 95% bootstrapped confidence intervals for clarity (a) and overall naturalness (b). Asterisks denote the significance of no_int_sil versus no_method under the primary linear mixed-effects models with Holm–Bonferroni correction (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$); full pairwise results are reported in Table 7. All scores are averaged over 20 listeners (17 advanced L2 learners and 3 native and expert speakers of Wuming Zhuang).

Table 7. Statistical Analysis of Pairwise Subjective Preference.

Dim.	Contrast	ΔMOS	Wilcoxon p_{Holm}	MixedLM p_{Holm}
Clarity	no_int_sil vs. full_method	+0.325	0.017	0.015
Naturalness	no_int_sil vs. full_method	+0.317	0.102	0.019
Clarity	no_int_sil vs. no_method	+0.533	0.017	<0.001
Naturalness	no_int_sil vs. no_method	+0.542	0.026	<0.001

Note: ΔMOS is the mean paired difference in MOS units (no_int_sil minus the comparison strategy), obtained after averaging the six per-listener ratings within each strategy and dimension. Because the design is complete and balanced ($N = 20$, six items per cell), the mixed-effects fixed-effect estimate coincides with ΔMOS ; the two analyses therefore share the same effect estimate and differ only in their p -values. The **MixedLM** column reports the primary repeated-measures test, fitted to the unaggregated ratings with a random rater intercept, which retains within-listener information and thus has higher power. The **Wilcoxon** column reports a conservative distribution-free test on the 20 listener-level means. This accounts for the single divergence: for naturalness (no_int_sil vs. full_method) the same effect (+0.317) is significant under the primary model ($p_{\text{Holm}} = 0.019$) but not under the lower-powered Wilcoxon test ($p_{\text{Holm}} = 0.102$). Holm–Bonferroni correction was applied within each method across all six pairwise MOS contrasts; the two full_method vs. no_method contrasts (both non-significant) are omitted. Bold indicates $p_{\text{Holm}} < 0.05$.

The analysis of Comparative MOS (CMOS) yielded confidence intervals straddling zero (e.g., $[-0.255, 0.245]$ for clarity), indicating that pairwise forced choices were less discriminative than absolute ratings in this experimental design. Consequently, the conclusions primarily rely on the stable MOS differences and mixed-effects modeling.

5.2 Objective Evaluation: The Fidelity-Prosody Disconnect

Table 8 summarizes the objective evaluation and reveals a fidelity-prosody disconnect in this project-collected, in-the-wild setting. In particular, the heavily cleaned configuration `full_method` achieved the lowest FAD (0.4639), indicating strong spectral-fidelity and noise-reduction effects. However, this advantage did not translate into the best prosody-sensitive metrics nor the highest MOS, whereas `no_int_sil` ranked best overall by preserving prosodic structure while retaining measured denoising.

Table 8. Objective Evaluation Results Comparing Fidelity and Prosodic Integrity.

Strategy	Rank	Prosody Metrics					Fidelity Metrics		
		Dur↓	F0-RMSE↓	F0-r↑	Pause↓	P.Ratio↓	FD↓	FAD↓	MCD↓
<code>no_method</code>	6	0.1882	2.3088	0.1040	0.1219	0.1179	1.8381	0.9775	6.0988
<code>no_norm</code>	2	0.1609	2.3038	0.1762	0.1302	0.0568	0.9708	0.5285	6.6866
<code>no_filter</code>	4	0.1649	2.3479	0.1669	0.1376	0.0600	0.9572	0.6265	6.6798
<code>trim + norm</code>	5	0.1643	2.2794	0.1685	0.1339	0.0575	0.9936	1.8124	6.7107
<code>no_silence</code>	7	0.1832	2.4503	0.0980	0.1345	0.1173	1.6383	0.9799	6.6025
<code>full_method</code>	3	0.1637	2.3249	0.1606	0.1387	0.0625	0.9188	0.4639	6.7195
<code>no_int_sil</code>	1	0.1522	2.2521	0.1793	0.1308	0.0618	0.8386	0.7192	6.6090

Note: Arrows indicate performance direction. **Dur**: Duration Mean Absolute Percentage Error. **F0-RMSE/r**: RMSE and Pearson correlation of aligned F0 contours. **Pause**: Internal pause alignment error. **P.Ratio**: Silence-to-speech ratio error. **FD**: Fréchet Distance on Wav2Vec2-Base embeddings. **MCD**: Mel Cepstral Distortion. **Rank**: Aggregated ranking obtained by computing the equal-weighted average of all metrics after direction alignment (so that higher values indicate better performance) and z-score standardization across strategies, then ordering strategies by this composite score (1 = best).

These results suggest that internal pause handling can perturb timing cues that are critical for prosodic organization; we therefore provide a robustness analysis of the pause-cap choice (Section 5.3) and then examine boundary-conditioned F0 behavior to explain the perceptual gap (Section 5.4).

In contrast, `no_int_sil` ranks highest in the aggregated evaluation. It preserves the dynamic range of fundamental frequency (highest F0 correlation, $r = 0.1793$) and achieves a competitive pause-ratio error (P.Ratio), while also achieving the best FD among all strategies. Figure 4 visualizes this trade-off as a Pareto front between spectral fidelity (FD) and prosodic timing (Duration MAPE). The proposed strategy occupies an optimal region, balancing noise reduction with the preservation of critical linguistic information, whereas `full_method` sacrifices prosodic accuracy for spectral purity.

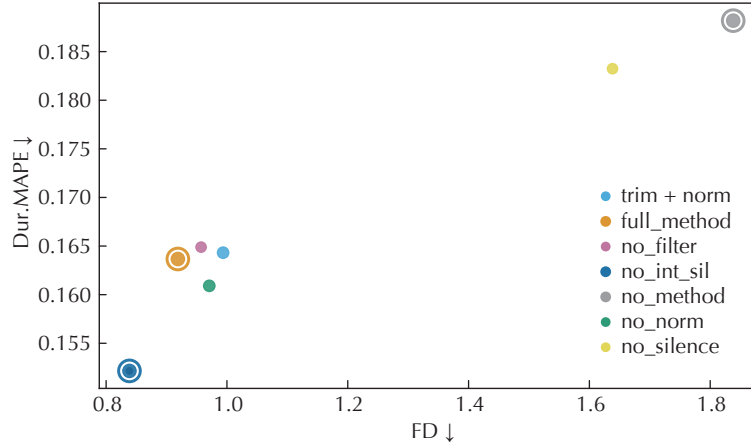


Figure 4. Pareto front analysis comparing Fidelity (FD) and Prosodic Timing (Duration MAPE). Lower values indicate better performance for both metrics. The hollow rings highlight the three principal strategies.

5.3 Robustness of the Pause-Cap Choice in full_method

Internal silence compression in full_method depends on the hyperparameter $\tau_{\max'}$, which may affect how aggressively within-utterance pauses are truncated. To make this design choice transparent and to address potential sensitivity concerns, we conduct a preprocessing-level sweep of $\tau_{\max}$ on the same fixed train+val roster ($N = 15,100$) under identical trimming and threshold settings.

Table 9 indicates that $\tau_{\max}$ primarily controls temporal editing strength ($\Delta\bar{T}$), while leaving RMS and ZCR essentially unchanged, suggesting that the effect of $\tau_{\max}$ is predominantly temporal rather than reflected in these low-level waveform statistics. This sweep is intended to calibrate corpus-level temporal editing induced by $\tau_{\max'}$, rather than to optimize $\tau_{\max}$ or diagnose fine-grained segmental truncation.

Table 9. Preprocessing-level sensitivity of internal silence compression in full_method to the pause cap $\tau_{\max}$ on the fixed train + val roster ($N = 15,100$; same trimming and QC settings).

$\tau_{\max}$ (s)	N	$\bar{T}_{\text{orig}}$	$\bar{T}_{\text{proc}}$	$\Delta\bar{T}$	$\overline{\text{RMS}}$	$\overline{\text{ZCR}}$
0.4	15100	5.734	4.941	0.793	0.11126	0.06818
0.6	15100	5.734	4.993	0.741	0.11069	0.06818
0.8	15100	5.734	5.017	0.718	0.11045	0.06817
1.0	15100	5.734	5.028	0.706	0.11034	0.06817
1.2	15100	5.734	5.034	0.700	0.11029	0.06818

Note: T_{orig} is the original utterance duration before any preprocessing, and T_{proc} is the duration after trimming and internal silence compression under the same QC settings. All statistics are computed on the same fixed roster (train + val), hence $\bar{T}_{\text{orig}}$ is identical across the sweep. $\Delta\bar{T} = \bar{T}_{\text{orig}} - \bar{T}_{\text{proc}}$ quantifies the average amount of temporal editing; its monotonic decrease indicates that smaller $\tau_{\max}$ corresponds to stronger temporal editing, whereas RMS and ZCR remain nearly invariant.

Based on this analysis, we adopt $\tau_{\max} = 0.8$ s as a conservative mid-range operating point for the controlled end-to-end comparisons. More linguistically informed, pause-type-aware compression policies that selectively reduce disfluencies while preserving boundary cues are beyond the scope of the present study and are left for future work.

Accordingly, our end-to-end comparisons emphasize the strategy-level choice between fixed-cap truncation and preserving internal pauses, rather than hyperparameter tuning of $\tau_{\max}$.

5.4 Mechanistic Analysis: Prosodic Integrity and Tone-Relevant Pitch Behavior

To connect temporal editing in pause truncation to perceptual outcomes, we examine boundary-conditioned fundamental-frequency (F0) trajectories, where timing perturbations are expected to be particularly salient in a tonal language. Figure 5 compares the reconstructed F0 contours under the three principal strategies for representative utterances. In each case, full_method exhibits abrupt truncations or

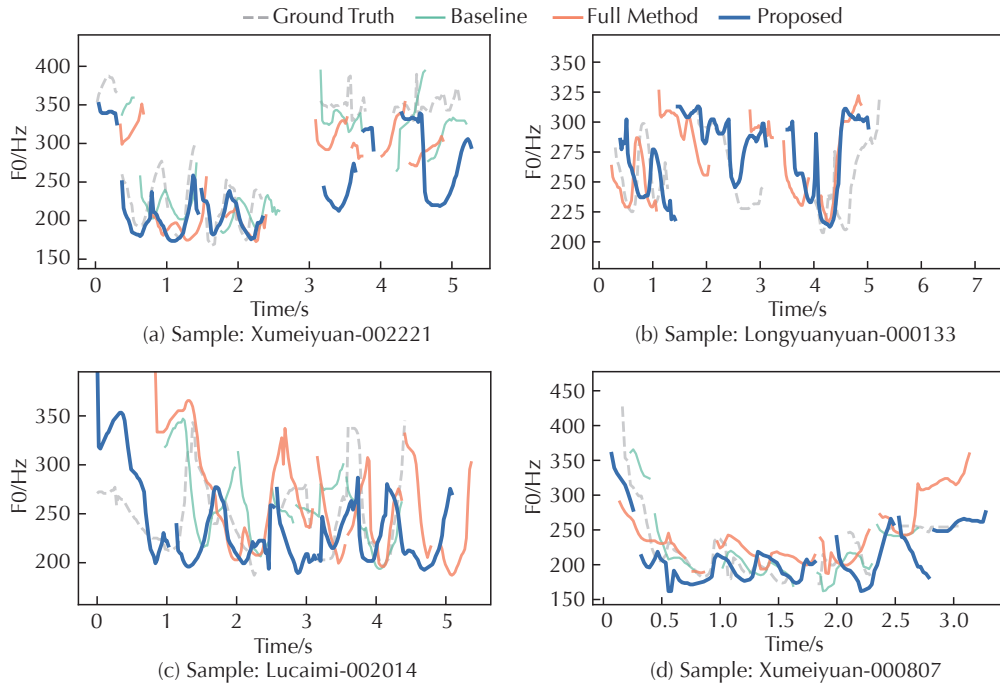


Figure 5. Comparative analysis of F0 contour reconstruction across four representative scenarios. The gray dashed lines represent the Ground Truth (GT). (a) Medium-duration segment with rich prosodic variations, highlighting the frequency drift in the Baseline. (b) Long-duration sequence requiring high temporal synchronization, showing boundary artifacts in the Baseline. (c) High-dynamic-range speech segment, demonstrating the temporal misalignment caused by aggressive silence compression in the Full Method. (d) Short utterance containing micro-prosodic details, where the Proposed Method (no_int_sil; blue line) exhibits superior fine-grained fidelity. Note: These examples illustrate that no_int_sil can achieve a favorable trade-off between pitch fidelity and temporal alignment.

temporal shifts of phrase-level F0 movements under aggressive internal silence compression: pauses at syntactic or prosodic boundaries are shortened, causing adjacent tonal movements to be squeezed together. This behavior is consistent with the higher Duration and Pause errors observed in Table 8. In contrast, `no_int_sil` more consistently preserves the timing and continuity of phrase-level F0 patterns while also reducing nuisance noise, aligning with its superior MOS.

For a tonal language such as Wuming Zhuang, tone realization depends jointly on F0 contour, segmental duration, and boundary-related pauses. The boundary collapse and F0 truncation introduced by `full_method` therefore remove cues that listeners rely on for naturalness and intelligibility, providing a mechanistic explanation for its inferior perceptual performance despite cleaner spectra. This supports our interpretation of aggressive internal silence compression as a prosody corruptor: improving apparent cleanliness at the cost of disrupting functional pauses and boundary-conditioned tonal coarticulation cues.

Prior work on Mandarin corpora further reports that realized tones can systematically deviate from lexical tone labels due to coarticulation and prosodic-structural position [35]. Motivated by this observation, we conduct an additional tone-relevant diagnostic analysis focusing on pause-following phrase-initial regions (Table 10), where boundary cues are known to interact with phrase-initial pitch scaling and boundary-related pitch adjustments in tonal prosody [36-37].

Table 10. Boundary-conditioned pitch diagnostics on pause-following phrase-initial regions.

Strategy	Onset $ \Delta _{st}$ (med.)↓	Post RMSE _{st} (med.)↓	Post r (mean)↑	N_{onset}	N_{post}
<code>no_method</code>	3.29	4.41	0.025	441	638
<code>full_method</code>	3.39	4.75	−0.006	427	615
<code>no_int_sil</code>	3.66	4.54	0.015	445	639

Note: Long internal pauses in the reference waveform (detected with `librosa`, `top_db` = 40, silence gap ≥ 0.30 s) are treated as boundary proxies. Pitch is extracted with `WORLD` (harvest + `stonemask`, frame period 5 ms). **Onset** measures the absolute deviation of median pitch (semitones) in the onset window immediately after each boundary (60 ms in our setting). **Post** metrics are computed on a longer post-boundary window capturing early phrase-initial pitch behavior (300 ms). Metrics are computed on jointly voiced frames only; N_{onset} and N_{post} denote the number of boundary instances with sufficient jointly voiced frames for each metric.

Long internal pauses in the reference audio are treated as proxies for prosodic boundaries, and we quantify pitch deviations in (i) a short onset window immediately after each boundary and (ii) a longer post-boundary window capturing the early phrase-initial trajectory. The onset window is a challenging region for all systems (with limited jointly voiced frames), and all strategies exhibit substantial onset pitch deviations. However, `full_method` shows systematically worse post-boundary pitch behavior—higher (Table 10) pitch error and near-zero (sometimes negative) pitch correlation—consistent with boundary disruption under aggressive silence compression. In contrast, `no_int_sil` avoids the most severe boundary-related degradation while still benefiting from denoising, supporting its overall perceptual advantage.

The joint analysis concludes that for low-resource data with atypical prosody, the objective should not be to imitate studio-quality spectral characteristics at the expense of temporal structure. The integrity of prosodic information, preserved by `no_int_sil`, proves to be the dominant factor in perceptual quality, particularly in a tonal setting.

The present results are obtained from a single Zhuang corpus and a fixed VITS backbone; therefore, they do not establish cross-lingual generalization. We hypothesize that the observed failure mode, namely temporal-structure perturbation induced by indiscriminate truncation of within-utterance pauses, may arise in other in-the-wild TTS settings when internal pauses function as boundary cues. The tonal setting may amplify the perceptual impact of such perturbations due to interactions between boundary timing and F0-based lexical and post-lexical patterns. Importantly, the boundary-conditioned pitch diagnostics reported in this work constitute tone-relevant evidence rather than tone-category evaluation, and the quantitative magnitudes of these effects may depend on language, speaking style, and corpus characteristics.

To contextualize this observed failure mode, we must revisit the assumptions within our data-curation pipeline. As noted in the Introduction, within-utterance pauses in found data may simultaneously reflect disfluency artifacts and encode functional prosodic boundary cues. The data-curation pipeline is motivated by an initial hypothesis that the raw corpus contains unusually slow, classroom-style speech with frequent within-sentence pauses, which may harm fluency when used for TTS training. This hypothesis is supported qualitatively by inspection and by feedback from native and expert listeners, who report that many recordings are markedly slower than conversational speech. A natural first attempt is, therefore, to compress long internal silences in order to approximate the temporal rhythm of mainstream TTS corpora. However, the ablation results revise this hypothesis: aggressive internal silence compression improves apparent cleanliness but introduces temporal discontinuities and distorts boundary-related timing cues, which are critical for prosodic organization and, in a tonal language, for tone realization in context. Consequently, the empirical evidence suggests that a substantial portion of within-utterance pauses in the target domain is not purely detrimental noise but carries prosodic and linguistic functions; removing them indiscriminately acts as a prosody corruptor. This finding speaks to the ambiguity raised in the Introduction: although within-utterance pauses in found data may in principle reflect either disfluency artifacts or functional prosodic boundary cues, in the present Wuming Zhuang corpus, the results are consistent with many within-utterance pauses carrying boundary-related information in this corpus, such that indiscriminate truncation is counterproductive. This revision motivates the proposed `no_int_sil` strategy, which retains internal pause structure while still benefiting from trimming, filtering, and normalization.

The present results should not be interpreted as prescribing unconditional preservation of atypical or didactic prosody. Rather, they indicate that uniform truncation of within-utterance pauses can be detrimental when such pauses participate in boundary marking and interact with phrase-initial pitch dynamics and local timing structure. In applications that explicitly target style normalization, or in regimes where long pauses predominantly reflect non-functional disfluencies, pause modification may still be warranted. These considerations suggest that pause handling should ideally condition temporal editing on prosodic context, rather than applying a uniform global cap to all internal gaps.

5.5 Sensitivity Analysis of Preference Weights

To operationalize the multi-objective perspective introduced in Section 3, we instantiate the post hoc scalarized utility function $S_\alpha(\theta)$ in Eq. (18) and evaluate its sensitivity to the preference weight $\alpha \in [0, 1]$. Specifically, we compute $S_\alpha(\theta)$ across the full spectrum of α using composite scores derived from the system-level summaries in Table 8 by direction alignment (higher is better), per-metric min–max rescaling to $[0,1]$ across the seven strategies, and equal-weight averaging within each metric family. This system-level rescaling differs from the per-utterance standardization defined in Section 3 and is used here only to place the two families on a common $[0, 1]$ scale; the resulting ordering of strategies is unchanged if z-score standardization is applied to the same summaries. Equal weighting within each metric family is adopted as a conservative default in the absence of externally validated, task-specific preference specifications. The sensitivity results in Figure 6 show that no_int_sil remains the top-ranked strategy over a wide range of fidelity-prosody trade-off weights.

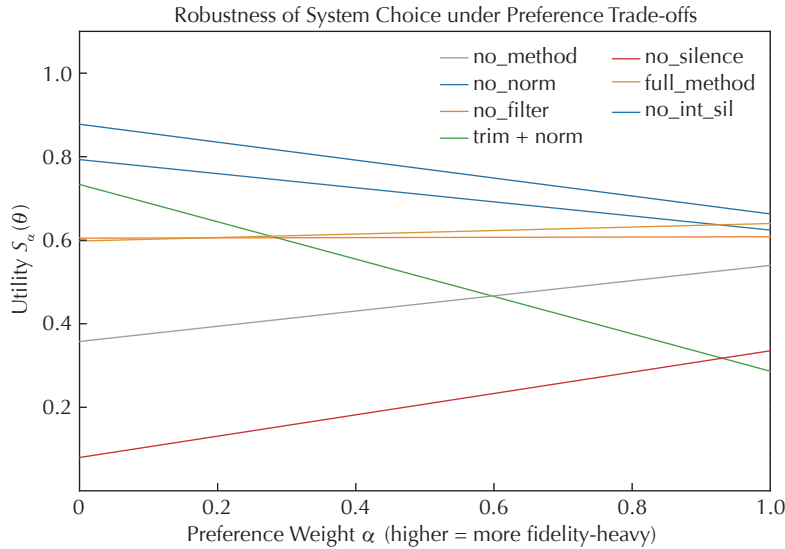


Figure 6. Utility score S_α as a function of fidelity preference α . The proposed no_int_sil (blue line) achieves the highest utility across the entire interval $\alpha \in [0, 1]$, indicating robustness to different prosody-fidelity trade-offs. Here $\hat{\mathcal{F}}_{\text{test}}(\theta)$ and $\hat{\mathcal{P}}_{\text{test}}(\theta)$ are instantiated from the system-level summaries in Table 8 by direction alignment (higher is better), per-metric min–max rescaling to $[0, 1]$ across the seven strategies, and equal-weight averaging within each family.

The utility trajectories further demonstrate that no_int_sil attains the highest utility throughout the entire interval $\alpha \in [0, 1]$. Although the utility of the aggressive full_method approaches that of no_int_sil as the weight on spectral fidelity increases ($\alpha \rightarrow 1$), it never surpasses the proposed strategy. This indicates that the marginal gains in spectral purity offered by aggressive cleaning are insufficient to offset the penalties incurred in prosodic integrity, even under strongly fidelity-oriented preferences.

Taken together with the MOS and objective results, this sensitivity analysis shows that the preference for `no_int_sil` is structurally robust rather than tied to a single operating point. Within the class of rule-based preprocessing schemes considered in this work, preserving internal silence emerges as a consistently preferable data-curation strategy for low-resource Zhuang TTS under noisy and atypically prosodic conditions.

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion: Seeking an Effective Balance in Flawed Data

This study revisits data curation for low-resource Zhuang TTS from a data-centric perspective. Using a real-world corpus with atypical, didactic-style prosody, it demonstrates that maximizing spectral purity does not necessarily yield the best perceived quality in this setting. A configuration that disables internal silence compression while retaining trimming, filtering, and normalization (`no_int_sil`) achieves a higher MOS and stronger prosody-sensitive objective results, such as a lower duration mean absolute percentage error and F0 root-mean-square error, as well as a higher F0 correlation compared to a heavy-cleaning baseline (`full_method`), despite the latter's advantage in FAD.

These results suggest that preserving sentence-level rhythm and pause structure is a first-order requirement for tonal-language TTS when the source prosody is imperfect. In the present setting, aggressive silence compression tends to act as a prosody corruptor; it modestly improves distributional fidelity metrics while degrading cues that listeners rely on for naturalness and intelligibility. More broadly, when curating found data, prosody-sensitive metrics that better reflect human perception should be weighted more heavily than a single spectral-distance measure. Data curation is therefore framed as balancing noise reduction with the protection of linguistically functional prosodic information, rather than pursuing uniform cleanliness.

Overall, the evidence from this corpus and the VITS backbone supports a practical guideline for curating similar found-data corpora with atypical prosody: prioritize the integrity of prosodic structure while applying measured denoising and validating choices with prosody-aware diagnostics in addition to spectral-distance metrics. This conclusion is further supported by the post hoc utility sensitivity analysis, which shows that `no_int_sil` remains top-ranked across a wide range of fidelity-prosody preference weights.

This conclusion should not be construed as endorsing the unconditional preservation of atypical speaking styles; rather, it highlights the risk of indiscriminate pause truncation in found-data settings.

6.2 Limitations and Future Work

This work has several limitations and defines its scope accordingly. All experiments are conducted with a multi-speaker VITS backbone under the strict-alignment protocol. We adopt VITS as a controlled and widely used open-source acoustic model whose training dynamics are strongly coupled to data quality, making it suitable for isolating data-curation effects in a data-centric study. While newer TTS systems

increasingly incorporate additional front-end components, many open-source pipelines still rely on VITS-style generators and flow-based decoders for waveform reconstruction. We hypothesize that the observed trend may generalize to VITS-family variants with similar acoustic decoding mechanisms; however, the present work empirically validates it only for the VITS backbone considered here. Cross-architecture validation remains an important direction for future work.

We do not report tone-category accuracy because the recordings lack syllable-level tone labels and reliable time-aligned syllable boundaries. Instead, our tone-relevant analysis focuses on boundary-conditioned pitch behavior (e.g., pause-following phrase-initial regions), which provides an empirically grounded proxy for tonal prosody without requiring explicit tone annotations.

First, the subjective evaluation includes 20 raters, which is sufficient for statistical testing but modest in size; larger and more diverse listener populations would strengthen generalizability. Second, the conclusions are drawn from one language and a specific noise and prosody profile; cross-lingual validation across tonal and non-tonal languages, as well as evaluation under different noise conditions, is needed to delineate the scope. Third, the curation decisions are rule-based. In addition, we do not report end-to-end performance as a function of multiple $\tau_{\max}$ values; the $\tau_{\max}$ sweep in Section 5.3 is provided as a preprocessing-level calibration of temporal editing strength.

AUTHOR CONTRIBUTIONS

Fengbo Bai: Data curation, Project administration, Conceptualization, Writing-review;

Qingshi Yang: Original writing draft, Data curation, Method implementation;

Xiao Liang: Formal analysis, Methodology, Theoretical validation.

ACKNOWLEDGEMENTS

This work is supported by the Key Research and Development Program of Guangxi (No. FN2504240031, No. AB25069456) and the Central Government Guides for Local Science and Technology Development Fund of Guangxi (No. ZY24212045).

DATA AVAILABILITY

The speech corpus used in this study was collected in-house under a Guangxi Zhuang Autonomous Region funded language documentation and protection project. The recordings were originally collected for ASR research and contain participant speech. Due to privacy considerations and project-related data sharing restrictions, the dataset cannot be publicly released at this time. Access for academic research may be considered on a case-by-case basis upon reasonable request, subject to approval and a data use agreement; interested researchers may contact the corresponding author.

REFERENCES

- [1] K. Ito and L. Johnson, "The LJ speech dataset," [Online]. Available: <https://keithito.com/LJ-Speech-Dataset/>, 2017, accessed: Mar. 7, 2026.
- [2] S. Anand, P. Srinivasa Varadhan, A. Sankar, G. Raju, and M. M. Khapra, "Enhancing out-of-vocabulary performance of Indian TTS systems for practical applications through low-effort data strategies," in *Proc. Interspeech*, 2024, pp. 1200–1204.
- [3] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proc. Interspeech*, 2019, pp. 1526–1530.
- [4] J. Jung, W. Zhang, S. Maiti, Y. Wu, X. Wang, J.-H. Kim, Y. Matsunaga, S. Um, J. Tian, H. Shim, N. Evans, J. S. Chung, S. Takamichi, and S. Watanabe, "Text-to-speech synthesis in the wild," arXiv:2409.08711 [Online]. Available: <https://arxiv.org/abs/2409.08711>, 2025, accessed: Mar. 7, 2026.
- [5] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "A review on subjective and objective evaluation of synthetic speech," *Acoust. Sci. Technol.*, vol. 45, no. 4, pp. 161–183, 2024.
- [6] M. Yip, *Tone*, ser. Cambridge Textbooks in Linguistics. Cambridge Univ. Press, 2002.
- [7] 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.
- [8] P. N. Phuong, C. T. Quang, Q. T. Do, and M. C. Luong, "A study on neural-network-based text-to-speech adaptation techniques for Vietnamese," in *Proc. O-COCOSDA*, 2021, pp. 199–205.
- [9] P. J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *J. Acoust. Soc. Amer.*, vol. 90, no. 6, pp. 2956–2970, 1991.
- [10] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Proc. 5th ISCA Speech Synth. Workshop (SSW5)*, 2004, pp. 223–224.
- [11] P. Do, M. Coler, J. Dijkstra, and E. Klabbers, "A systematic review and analysis of multilingual data strategies in text-to-speech for low-resource languages," in *Proc. Interspeech*, 2021, pp. 16–20.
- [12] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, vol. 80, 2018, pp. 4693–4702. [Online]. Available: <https://proceedings.mlr.press/v80/skerry-ryan18a.html>.
- [13] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, Z. Jiang, S. Zhong, and X. Hu, "Data-centric artificial intelligence: A survey," *ACM Comput. Surv.*, vol. 57, no. 5, pp. 129:1–129:42, 2025.
- [14] C. Gong, E. Cooper, X. Wang, C. Qiang, M. Geng, D. Wells, L. Wang, J. Dang, M. Tessier, A. Pine, K. Richmond, and J. Yamagishi, "An initial investigation of language adaptation for TTS systems under low-resource scenarios," in *Proc. Interspeech*, 2024, pp. 4963–4967.
- [15] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.
- [16] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," arXiv:2006.04558 [Online]. Available: <https://arxiv.org/abs/2006.04558>, 2022, accessed: Mar. 7, 2026.
- [17] H. Wang, S. Shan, Y. Guo, and Y. Wang, "ProsodyFlow: High-fidelity text-to-speech through conditional flow matching and prosody modeling with large speech language models," in *Proc. 31st Int. Conf. Comput. Linguistics (COLING)*, 2025, pp. 7748–7753.
- [18] L. Luo, *Zhuang-Chinese-English Dictionary*. Beijing, China: Minzu Chubanshe, 2004.

- [19] A. Afshan, K. Kumar, and J. Wu, "Sequence-level confidence classifier for ASR utterance accuracy and application to acoustic models," in *Proc. Interspeech*, 2021, pp. 4084–4088.
- [20] K.-Z. Lee, E. Cooper, and J. Hirschberg, "A comparison of speaker-based and utterance-based data selection for text-to-speech synthesis," in *Proc. Interspeech*, 2018, pp. 2873–2877.
- [21] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Commun.*, vol. 32, no. 1–2, pp. 127–154, 2000.
- [22] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 54–70, 2023.
- [23] Y. Cao, Z. Liu, M. Chen, J. Ma, S. Wang, and J. Xiao, "Nonparallel emotional speech conversion using VAE-GAN," in *Proc. Interspeech*, 2020, pp. 3406–3410.
- [24] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," [Online]. Available: <https://datashare.ed.ac.uk/handle/10283/3443>, 2019, accessed: Mar. 7, 2026.
- [25] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf. (SciPy)*, 2015, pp. 18–24.
- [26] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 5530–5540. [Online]. Available: <https://proceedings.mlr.press/v139/kim21f.html>.
- [27] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Commun. Comput. Signal Process.*, vol. 1, 1993, pp. 125–128.
- [28] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms," in *Proc. Interspeech*, 2019, pp. 2350–2354.
- [29] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 12449–12460.
- [30] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. E99-D, no. 7, pp. 1877–1884, 2016.
- [31] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Commun.*, vol. 84, pp. 57–65, 2016.
- [32] S. G. Luke, "Evaluating significance in linear mixed-effects models in R," *Behav. Res. Methods*, vol. 49, no. 4, pp. 1494–1502, 2017.
- [33] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.
- [34] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.
- [35] C.-Y. Chiang, W.-H. Li, and T.-H. Liu, "Tone labeling for Mandarin speech corpus: A deep learning iterative approach," in *Handbook of Chinese Language Learning and Technology*, S.-H. Teng, L.-P. Chang, and T.-H. Liu, Eds. Springer, 2025, pp. 283–335.
- [36] B. Wang, Y. Xu, and Q. Ding, "Interactive prosodic marking of focus, boundary and newness in Mandarin," *Phonetica*, vol. 75, no. 1, pp. 24–56, 2018.
- [37] J. Holzgrefe, C. Wellmann, C. Petrone, H. Truckenbrodt, B. Höhle, and I. Wartenburger, "Brain response to prosodic boundary cues depends on boundary position," *Frontiers Psychol.*, vol. 4, p. 421, 2013.

AUTHOR BIOGRAPHY

Fengbo Bai is a Ph.D. and Master Supervisor at the College of Artificial Intelligence, Guangxi Minzu University, China, with research interests in Natural Language Processing and digital evidence technology. ORCID: 0009-0009-1534-6868; E-mail: baif@gxmzu.edu.cn

Qingshi Yang is a master's student at the College of Artificial Intelligence, Guangxi Minzu University, China, with research interests in speech synthesis. ORCID: 0009-0000-2848-8990; E-mail: yqs441598707@foxmail.com

Xiao Liang received the M.S. degree from Leibniz Universität Hannover, Germany, and is currently pursuing the Ph.D. degree at Universiti Tunku Abdul Rahman, Malaysia, with research interests in Natural Language Processing and Machine Translation. ORCID: 0009-0006-5321-4565; E-mail: xiaoliang1223@foxmail.com