NüshuVoice: Reviving the Voice of Endangered Nüshu with Pitch-Aware Text-to-Speech

Hongkun Yang[1*†], Xinhui Yi[2*†], Xiyan Zhao[2*], Yibo Meng[3*], Lionel Z. Wang[2‡], Lixu Wang[4], Yaqi Zhang[5], Ruiqi Chen[6], Xuanyue Zhao[4], Lanxin Zhang[4], Yu Zeng[7], Weijia Chu[2], Yiming Ma[8], Chenyu Liu[2], Jianghao Lin[7], Xin Xu[2‡]

[1]Ocean University of China, [2]The Hong Kong Polytechnic University, [3]Cornell University, [4]Nanyang Technological University, [5]Shanghai Jiao Tong University, [6]University of Michigan–Ann Arbor, [7]University of Science and Technology of China, [8]Harbin Institute of Technology

Abstract

Nüshu is an endangered phonetic script historically used by women in Jiangyong County, southern Hunan, China. While existing computational studies of Nüshu mainly focus on textual digitization and visual recognition, the acoustic reconstruction of its authentic pronunciation remains largely unexplored. Building a Nüshu text-to-speech (TTS) system is particularly challenging because available recordings are extremely limited and mostly consist of isolated syllable-level pronunciations rather than natural sentence-level utterances. In this work, we introduce **NüshuVoice**, the first TTS benchmark for Nüshu. We construct a sentence-level Nüshu text-to-audio dataset that aligns standardized Unicode Nüshu text, phonetic transcriptions, standard Chinese translations, and archival recordings. To synthesize speech under this extreme low-resource setting, we propose **Nüshu-PitchVITS**, an F_0-conditioned VITS framework that leverages Nüshu's five-level pitch notation as an explicit prosodic inductive bias. Experimental results show that Nüshu-PitchVITS outperforms strong TTS baselines in spectral fidelity, pitch reconstruction, and human-rated intelligibility. We publicly release the dataset and code at: https://anonymous.4open.science/r/Nvshu-TTS-2EB6.

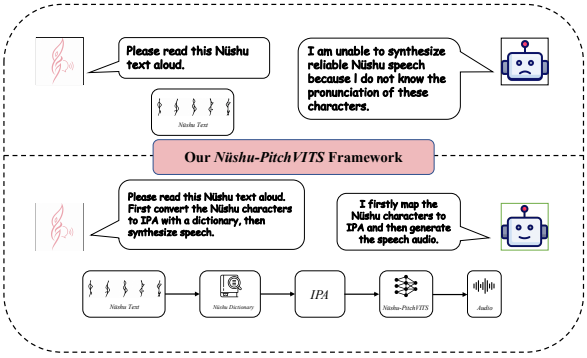


Figure 1: A simplified, stylized illustration of our **Nüshu-PitchVITS**. Nüshu characters are firstly converted to IPA with a dictionary, and then a TTS model generates speech audio

1 Introduction

Nüshu is an endangered script historically used by women in Jiangyong County, southern Hunan, China (Zuo et al., 2024). It is often regarded as the world's only known script created and used primarily by women. Its emergence was shaped by the unequal access to formal education and literacy that women faced in premodern China. In this context, Nüshu represents not only a writing system, but also a form of women's cultural agency under gendered conditions of exclusion. Unlike standard Chinese writing, which is largely morphosyllabic and logographic, Nüshu functions as a phonetic syllabic script. Each character corresponds to a specific syllable in the local Jiangyong dialect, including its initial, final, and tone (Congrong, 2024). This property makes Nüshu not merely a written cultural artifact, but also an acoustic heritage. Its preservation therefore depends in part on recovering historically grounded pronunciations.

Recent work on endangered and Indigenous languages has shown that speech technologies, especially text-to-speech (TTS), can support pronunciation access, language education, and revitalization (Chan and Hammerly, 2025; Hiovain-Asikainen et al., 2025). For Nüshu, such acoustic access is particularly important because the script is phonetic. A TTS system can therefore serve not only as a tool for generating speech, but also as a computational interface between Nüshu glyphs, phonological annotations, and historically grounded pronunciations. However, existing computational studies

[*]Hongkun Yang, Xinhui Yi, Xiyan Zhao, and Yibo Meng contributed equally.

[†]Hongkun Yang and Xinhui Yi are core contributors. The relative order of Hongkun Yang and Xinhui Yi was determined by random draw.

[‡]Corresponding authors.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 2423–2434
November 6-10, 2026 ©2026 Association for Computational Linguistics

of Nüshu have primarily focused on textual digitization, encoding, and visual recognition (Sun et al., 2023; Yang et al., 2024, 2025). These efforts are essential for preserving the written form of Nüshu, but they leave its acoustic modality largely underexplored. As a result, Nüshu can be digitally displayed and recognized, yet its spoken realization, tonal dynamics, and dialectal characteristics remain difficult to access for learners, researchers, and cultural heritage applications.

Building a neural TTS system for Nüshu is technically challenging under such preservation-oriented conditions. Modern TTS models typically rely on large amounts of paired text–speech data (Xu et al., 2020; Park et al., 2023; Basher et al., 2025), whereas available Nüshu recordings are extremely limited, archival, and non-studio in quality [1]. Under this extreme data scarcity, conventional TTS architectures are prone to unstable text-speech alignment and weak prosodic modeling.

In this work, we present the first dedicated study on Nüshu TTS synthesis. We first construct a sentence-level Nüshu TTS dataset by aligning verified Unicode Nüshu text, IPA-based phonetic transcriptions, standard Chinese translations, and sentence-level recordings assembled from archival syllable-level audio. Based on this dataset, we propose an $F0$-conditioned VITS framework tailored to the phonological structure of Nüshu. The key motivation is that Nüshu phonetic annotations use a five-level pitch notation system, which provides explicit tonal information. As shown in **Figure 1**, Nüshu characters are first mapped into phonetic representations based on the International Phonetic Alphabet (IPA) through a dictionary, and the resulting phonetic sequence is then used by the TTS model to generate speech audio.

Experimental results demonstrate that explicit pitch conditioning substantially improves Nüshu speech synthesis under extreme low-resource conditions. Compared with state-of-the-art baselines, our method achieves a 40.1% reduction in MCD, a 43.5% lower F_0 RMSE, and a 27.4% gain in intelligibility MOS over the strongest baseline.. These results suggest that linguistically grounded prosodic modeling is critical for endangered phonetic scripts where acoustic data are scarce but tonal annotations are available.

Our contributions are summarized as follows:

- We construct **NüshuVoice**, a sentence-level multimodal Nüshu TTS dataset by aligning standardized Nüshu Unicode text, phonetic transcriptions, Chinese translations, and archival syllable-level recordings.
- We propose **Nüshu-PitchVITS**, an pitch-conditioned VITS framework that leverages the five-level pitch notation of Nüshu as an explicit prosodic inductive bias.
- We provide the first systematic objective and subjective evaluation of neural TTS models for Nüshu, demonstrating the effectiveness of explicit pitch modeling in an extreme low-resource setting.
- We contribute to the computational preservation and acoustic revitalization of Nüshu by making its pronunciation accessible for future research, documentation, and cultural heritage applications.

2 Related Work

2.1 Digital Preservation and Computational Studies of Nüshu

While Nüshu has recently garnered increasing cultural and linguistic interest (Thurnell-Read et al., 2022), computational research on this endangered script remains limited. Foundational preservation efforts have mainly focused on linguistic documentation and digitization (Rao and Zheng, 2022). Notably, Xie Zhimin compiled, translated, and provided phonetic annotations for surviving Nüshu manuscripts [2], while the electronic dictionary developed by Xie Zhimin and Wang Lihua digitized 873 isolated syllable recordings, preserving the authentic pronunciations of fluent speakers [3].

Building upon these linguistic resources, the machine learning and NLP communities have recently initiated exploratory studies on Nüshu. Early interdisciplinary work by Sun et al. (2023) introduced an AI-driven approach to conceptualize a standardized Nüshu encoding system. More recently, computational efforts have shifted toward specific modalities. In the text domain, Yang et al.

[1] Xie, Zhimin, and Wang, Lihua. *Nüshu Fasheng Dianzi Zidian* [*Nüshu Pronunciation Electronic Dictionary*]. Wuhan: Huazhong University of Science and Technology Press, 2002. In Chinese.

[2] Xie, Zhimin. *Jiangyong Nüshu zhi mi* [*The Mystery of Jiangyong Nüshu*]. Zhengzhou: Henan Renmin Chubanshe [Henan People's Publishing House], 1991, pp. 700–709. In Chinese.

[3] Xie, Zhimin, and Wang, Lihua. *Nüshu Fasheng Dianzi Zidian* [*Nüshu Pronunciation Electronic Dictionary*]. Wuhan: Huazhong University of Science and Technology Press, 2002. In Chinese.

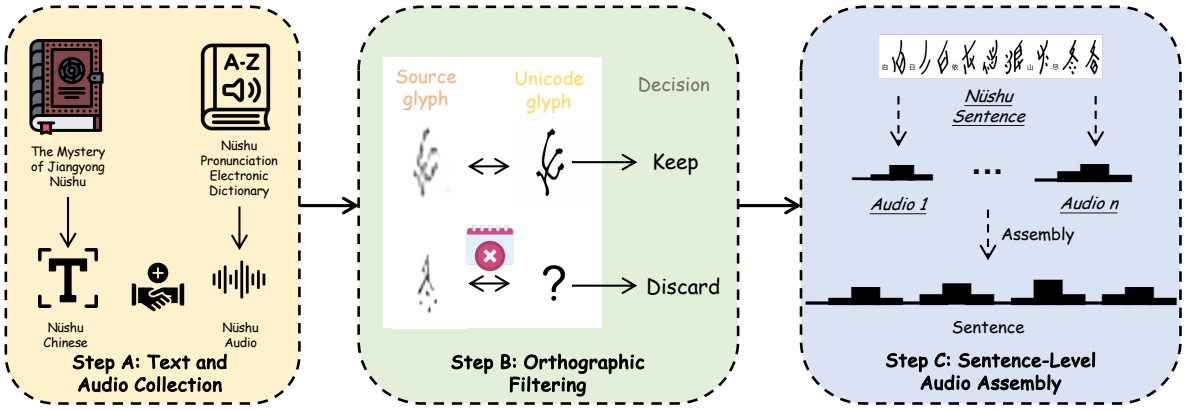


Figure 2: Overview of the construction pipeline of **NüshuVoice**. The pipeline consists of three stages: data collection, rigorous orthographic filtering, and sentence-level audio assembly.

(2024) proposed a framework for leveraging large language models on extremely low-resource endangered languages like Nüshu. In the visual domain, Yang et al. (2025) established foundational optical character recognition (OCR) benchmarks by constructing specialized visual datasets and fine-tuning vision-based recognition models. Despite these emerging advancements in text generation and visual recognition, the acoustic modality of Nüshu remains largely unexplored.

2.2 Text-to-Speech for Low-Resource Languages

Neural text to speech systems usually require substantial paired text and speech data, which are unavailable for many low resource languages and endangered writing traditions (Xu et al., 2020; Park et al., 2023). However, recent studies often define low resource conditions in a less restrictive way than the setting considered in this work. For example, Kwon et al. (2025) studied cross lingual continuous fine tuning with a 12 hour Korean corpus. Basher et al. (2025) achieved few shot adaptation with 4.22 hours of high quality target speech, supported by 3.85k hours of continuous pretraining. These settings still assume several hours of clean target data or large scale external acoustic priors. They are therefore not directly applicable to Nüshu, where the available recordings are extremely limited, archival, and non studio in quality.

This constraint makes the choice of TTS architecture critical. Autoregressive models such as Tacotron 2 (Shen et al., 2018) can produce natural speech, but their attention mechanisms are prone to collapse when trained with limited data. Non autoregressive models such as FastSpeech 2 (Ren et al., 2020) are more stable, but they usually require duration labels or external alignment supervision. Recent flow matching and zero shot models such as F5 TTS (Chen et al., 2025) show strong generalization, but their performance depends on large scale acoustic pretraining. In contrast, VITS (Kim et al., 2021) is a more suitable base model for this setting because it is trained end to end and uses Monotonic Alignment Search to learn text and speech alignment without an external aligner.

Standard VITS is still not sufficient because its unsupervised alignment process can become unstable when the paired data are too limited. We therefore introduce explicit fundamental frequency conditioning. $F0$ provides direct pitch information and acts as a prosodic inductive bias which helps regularize the acoustic latent representation. This design is also consistent with the structure of the Nüshu data, since the corpus uses a five level pitch notation (Zhengzhang and Zheng, 2015). $F0$ conditioning is therefore both an architectural choice for stabilizing VITS and a data driven choice that reflects the available Nüshu annotation.

3 NüshuVoice Construction

As illustrated in **Figure 2**, this section details the construction of our Nüshu text-to-audio dataset, including data collection, rigorous orthographic filtering and audio assembly.

3.1 Text and Audio Collection

The foundation of our dataset relies on two primary historical and linguistic resources. The textual corpus is derived from *The Mystery of Jiangyong*

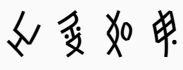
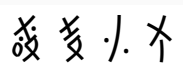
Audio	Chinese	Nüshu	IPA
A0001	教尔聪明		tciou44 ai42 tshang44 nou33
A0002	坐莫移身		tsew21 mai33 y42 cie44
A0003	轻言细语		tchiong44 njing42 siu35 njie33

Table 1: Some examples of **NüshuVoice** showing the multimodal alignment.

Nüshu.[4] This comprehensive compilation contains approximately 63,665 characters across various literary forms, including original letters, narrative poems, and folk songs. Crucially, this resource provides the original Nüshu characters alongside their phonetic transcriptions and standard Chinese translations at the sentence level. For the acoustic modality, we extracted raw audio data from the *Nüshu Pronunciation Electronic Dictionary*.[5] This dictionary contains 873 isolated syllable recordings in WAV format, capturing the pronunciation of individual Nüshu characters directly from fluent speakers.

3.2 Rigorous Orthographic Filtering

Given the highly idiosyncratic and historically non-standardized nature of Nüshu writing, establishing a reliable mapping between archival Nüshu texts and modern Unicode representations is essential. We therefore implemented a strict visual and semantic cross-referencing protocol to map handwritten and printed Nüshu glyphs from the 1991 manuscript to their corresponding standardized Unicode forms.

Annotators manually verified every character within each candidate sentence. To ensure orthographic fidelity, we adopted a conservative filtering strategy: if any character in a sentence lacked a corresponding Unicode equivalent, or exhibited even minor structural discrepancies, such as a missing stroke or a variant component, the entire sentence was discarded. Sentences that passed this filtering stage were stored in a structured metadata table containing the standardized Nüshu Unicode string, phonetic transcription, standard Chinese translation, and sentence identifier.

Data Split	Utterances	Duration (mins)
Train	1,002	79.00
Validation	125	10.00
Test	125	9.80
Total	**1,252**	**98.80**

Table 2: Statistical summary of the **NüshuVoice**.

3.3 Sentence-Level Audio Assembly

Since the primary acoustic resource only offers character-level pronunciations, we developed a deterministic synthesis pipeline to generate continuous, sentence-level audio. Relying on the precise phonetic transcriptions in the original corpus, which utilize a five-level tone marking system[6], we mapped each verified Unicode character to its corresponding isolated WAV file from the electronic dictionary. These character-level audio segments were then temporally concatenated via Audacity to reconstruct the complete utterance. Finally, the synthesized audio files were assigned unique sentence identifiers, establishing a one-to-one multimodal alignment between text and audio. Representative examples of the multimodal alignment are shown in **Table 1**. The statistical summary is showed in **Table 2**.

3.4 Expert Validation

To assess the reliability of NüshuVoice, we randomly sampled 10% of the dataset, i.e., 125 out of

[4]Xie, Zhimin. *Jiangyong Nüshu zhi mi* [*The Mystery of Jiangyong Nüshu*]. Zhengzhou: Henan Renmin Chubanshe [Henan People's Publishing House], 1991, pp. 700–709. In Chinese.

[5]Xie, Zhimin, and Wang, Lihua. *Nüshu Fasheng Dianzi Zidian* [*Nüshu Pronunciation Electronic Dictionary*]. Wuhan: Huazhong University of Science and Technology Press, 2002. In Chinese.

[6]Unlike standard Mandarin Pinyin, Nüshu phonetic notation employs the International Phonetic Alphabet (IPA) alongside a five-level pitch scale to capture the local dialect.

1,252 entries, and invited two Nüshu domain experts to independently validate them. The experts examined glyph–Unicode consistency, phonetic transcriptions, Chinese translations, and audio–syllable alignment. The two experts achieved 90.00% agreement, with a Cohen's κ of 0.7887, indicating substantial agreement. After adjudication, 117 entries were retained as correct, corresponding to a correctness rate of 93.6%, suggesting the high overall quality of NüshuVoice.

4 Proposed Nüshu-PitchVITS

4.1 Overview

As shown in **Figure 3**, Nüshu-PitchVITS extends the original VITS (Kim et al., 2021) architecture by incorporating an explicit fundamental frequency (F_0) predictor. This design is motivated by the phonological characteristics of Nüshu, which employs a five-level pitch transcription system representing tone values from 1 to 5.

4.2 Revisit: VITS

VITS (Kim et al., 2021) can be formulated as a conditional variational autoencoder (VAE). Given a target waveform $\mathbf{y}$ and a condition $\mathbf{c}$, its objective is to maximize $\log p_\theta(\mathbf{y} \mid \mathbf{c})$ via the ELBO:

$$\log p_\theta(\mathbf{y} \mid \mathbf{c}) \geq \mathbb{E}_{q_\phi}\left[\log p_\theta(\mathbf{y} \mid \mathbf{Z})\right] - D_{\mathrm{KL}}\left(q_\phi(\mathbf{Z} \mid \mathbf{X}_{\mathrm{lin}}) \,\|\, p_\theta(\mathbf{Z} \mid \mathbf{c})\right). \tag{1}$$

where $\mathbf{Z}$ is the latent acoustic representation and $\mathbf{X}_{\mathrm{lin}}$ is the linear-scale spectrogram used by the posterior encoder.

4.3 Nüshu-PitchVITS Architecture

The input to Nüshu-PitchVITS is the phonetic transcription of a Nüshu sentence, where each syllabic unit includes its five-level tone value. This tone-aware representation is encoded by the text encoder and serves as the basis for both acoustic alignment and pitch prediction.

To model the tonal characteristics of Nüshu, we augment VITS with a frame-level pitch prediction branch. Let the text encoder output be

$$\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]^\top \in \mathbb{R}^{N \times d}, \tag{2}$$

where N is the phoneme length and d is the hidden dimension.

After Monotonic Alignment Search (MAS), we obtain a hard monotonic alignment matrix

$$\mathbf{A}^* \in \{0,1\}^{T \times N}, \qquad \sum_{n=1}^{N} A^*_{t,n} = 1, \tag{3}$$

where T is the number of acoustic frames. The phoneme-level hidden states are expanded to the frame level by

$$\tilde{\mathbf{H}} = \mathbf{A}^* \mathbf{H} \in \mathbb{R}^{T \times d}. \tag{4}$$

The pitch predictor estimates the frame-level log-F_0 sequence:

$$\hat{\mathbf{f}} = g_\psi(\tilde{\mathbf{H}}), \qquad \hat{\mathbf{f}} \in \mathbb{R}^T, \tag{5}$$

where $g_\psi(\cdot)$ is a lightweight two-layer 1D convolutional network with layer normalization and ReLU activation.

The predicted log-F_0 sequence is projected to the hidden dimension and added to the latent representation:

$$\mathbf{E}_{F_0} = W_{F_0}(\hat{\mathbf{f}}) \in \mathbb{R}^{T \times d}, \qquad \mathbf{Z}' = \mathbf{Z} + \mathbf{E}_{F_0}. \tag{6}$$

Here, W_{F_0} denotes a frame-wise linear projection from $\mathbb{R}$ to $\mathbb{R}^d$. The waveform decoder then generates

$$\hat{\mathbf{y}} = G_\theta(\mathbf{Z}'). \tag{7}$$

To avoid disturbing alignment estimation in the low-resource setting, the pitch branch is excluded from MAS. The alignment $\mathbf{A}^*$ is computed only from the prior statistics and the flow-transformed posterior latent variables. Thus, the pitch predictor affects acoustic realization only after the alignment has been determined.

For supervision, the ground-truth fundamental frequency sequence $\mathbf{F}_0 \in \mathbb{R}^T$ is extracted from the target waveform using the DIO algorithm in WORLD and converted to the logarithmic domain:

$$\mathbf{f} = \log(\mathbf{F}_0 + \epsilon), \qquad \mathbf{f} \in \mathbb{R}^T, \tag{8}$$

where ϵ is a small positive constant. Since both $\mathbf{f}$ and $\hat{\mathbf{f}}$ are log-F_0 sequences, the pitch loss is

$$\mathcal{L}_{F_0} = \frac{1}{T}\left\|\mathbf{f} - \hat{\mathbf{f}}\right\|_2^2. \tag{9}$$

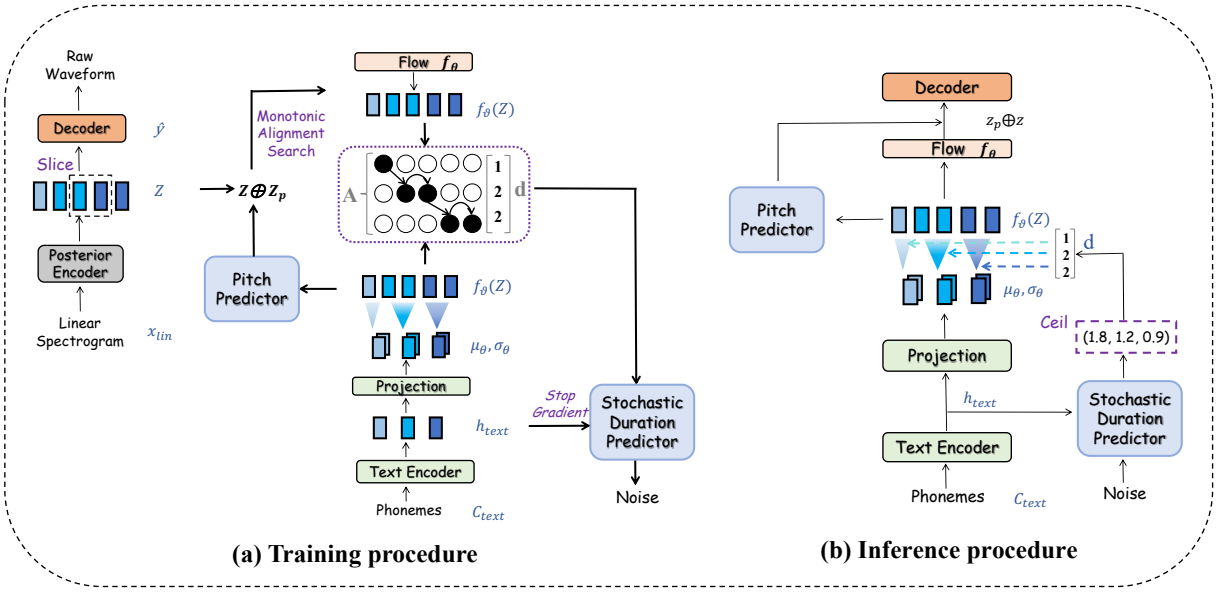


Figure 3: Overview of the proposed **Nüshu-PitchVITS** framework. (a) During training, the text encoder produces phoneme-level hidden states, which are expanded to the frame level by MAS. A pitch predictor estimates the frame-level log-F_0 sequence and injects the projected pitch representation into the latent acoustic representation before waveform decoding. The model is optimized with reconstruction, KL-divergence, duration, pitch prediction, adversarial, and feature matching losses. (b) During inference, the predicted log-F_0 is used together with the text-derived latent representation to generate the final waveform.

4.4 Two-stage Training

To accelerate convergence and ensure stable optimization, we employ a two-stage training strategy initialized from a pre-trained English VITS model. In the first stage, we freeze the waveform decoder and posterior encoder, and train only the text encoder, pitch predictor, duration predictor, and normalizing flow. This stage is optimized using $\mathcal{L}_{\text{kl}}$, $\mathcal{L}_{\text{dur}}$, and $\mathcal{L}_{F_0}$. In the second stage, we unfreeze all modules and perform end-to-end joint training.

4.5 Training Objective

The total generator loss is

$$\begin{aligned}\mathcal{L}_G =& \mathcal{L}_{\text{adv}}(G) + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}}(G) + \lambda_{\text{mel}}\mathcal{L}_{\text{recon}} \\ &+ \lambda_{\text{dur}}\mathcal{L}_{\text{dur}} + \lambda_{\text{kl}}\mathcal{L}_{\text{kl}} + \lambda_{F_0}\mathcal{L}_{F_0},\end{aligned} \tag{10}$$

where the λ terms are loss weights.

The mel-spectrogram reconstruction loss is

$$\mathcal{L}_{\text{recon}} = \left\| \mathbf{X}_{\text{mel}} - \hat{\mathbf{X}}_{\text{mel}} \right\|_1. \tag{11}$$

The KL-divergence loss is defined as

$$\mathcal{L}_{\text{kl}} = D_{\text{KL}}\big(q_\phi(\mathbf{Z} \mid \mathbf{X}_{\text{lin}}) \| p_\theta(\mathbf{Z} \mid \mathbf{c}_{\text{text}}, \mathbf{A}^*)\big), \tag{12}$$

where $\mathbf{c}_{\text{text}}$ is the text condition and $q_\phi(\mathbf{Z} \mid \mathbf{X}_{\text{lin}})$ is a diagonal Gaussian posterior.

The adversarial losses follow the least-squares GAN objective in VITS:

$$\mathcal{L}_{\text{adv}}(D) = \sum_{k=1}^{K} \mathbb{E}\left[(D_k(\mathbf{y}) - 1)^2 + D_k(\hat{\mathbf{y}})^2\right], \tag{13}$$

$$\mathcal{L}_{\text{adv}}(G) = \sum_{k=1}^{K} \mathbb{E}\left[(D_k(\hat{\mathbf{y}}) - 1)^2\right], \tag{14}$$

where $\{D_k\}_{k=1}^{K}$ denotes the set of discriminators.

The feature matching loss is

$$\mathcal{L}_{\text{fm}}(G) = \sum_{k,\ell} \frac{1}{N_{k,\ell}} \mathbb{E}\left[\|D_{k,\ell}(\mathbf{y}) - D_{k,\ell}(\hat{\mathbf{y}})\|_1\right], \tag{15}$$

where $D_{k,\ell}(\cdot)$ is the ℓ-th feature map of the k-th discriminator, and $N_{k,\ell}$ is the number of elements in that feature map.

5 Experiments

5.1 Implementation Details

We initialized Nüshu-PitchVITS using the weights of a pre-trained English VITS model (Kim et al., 2021), rather than training all parameters from scratch. The model was then trained with our two-stage optimization strategy on **NüshuVoice**. All training procedures were conducted on a single

Algorithm 1. Nüshu-PitchVITS: pitch-aware training and inference

Input: Tone-aware phonetic sequence $\mathbf{c}$; target waveform $\mathbf{y}$; training stage $s \in \{1, 2\}$
Output: Synthesized waveform $\hat{\mathbf{y}}$
Initialize: Pre-trained VITS parameters Θ; pitch predictor g_ψ; pitch projection W_{F_0}

Training phase
Encode $\mathbf{c}$ into phoneme-level hidden states: $\mathbf{H} \leftarrow \mathrm{TextEnc}_\Theta(\mathbf{c})$
Extract posterior latent representation: $\mathbf{Z} \leftarrow \mathrm{PostEnc}_\Theta(\mathbf{y})$
Estimate monotonic alignment path using MAS: $\mathbf{A}^* \leftarrow \mathrm{MAS}(\mathbf{Z}, \mathbf{H})$
Expand hidden states to the acoustic frame level: $\widetilde{\mathbf{H}} \leftarrow \mathbf{A}^*\mathbf{H}$
Predict frame-level log-pitch: $\hat{\mathbf{f}} \leftarrow g_\psi(\widetilde{\mathbf{H}})$
Inject pitch information into the latent representation: $\mathbf{Z}' \leftarrow \mathbf{Z} + W_{F_0}(\hat{\mathbf{f}})$
Generate waveform: $\hat{\mathbf{y}} \leftarrow G_\Theta(\mathbf{Z}')$
Extract ground-truth log-F_0 from $\mathbf{y}$ using WORLD-DIO: $\mathbf{f} \leftarrow \log(F_0(\mathbf{y}) + \epsilon)$
if $s = 1$ **then**
    Freeze posterior encoder and waveform decoder
    Update text encoder, duration predictor, flow, and pitch predictor using $\mathcal{L}_{\mathrm{kl}} + \mathcal{L}_{\mathrm{dur}} + \mathcal{L}_{F_0}$
else if $s = 2$ **then**
    Unfreeze all modules
    Update the full model using $\mathcal{L}_G = \mathcal{L}_{\mathrm{adv}} + \lambda_{\mathrm{fm}}\mathcal{L}_{\mathrm{fm}} + \lambda_{\mathrm{mel}}\mathcal{L}_{\mathrm{recon}} + \lambda_{\mathrm{dur}}\mathcal{L}_{\mathrm{dur}} + \lambda_{\mathrm{kl}}\mathcal{L}_{\mathrm{kl}} + \lambda_{F_0}\mathcal{L}_{F_0}$
Inference phase
Predict duration and alignment from $\mathbf{c}$
Predict $\hat{\mathbf{f}}$ from the expanded text representation
Construct pitch-conditioned latent representation $\mathbf{Z}'$
Decode $\mathbf{Z}'$ to synthesize $\hat{\mathbf{y}}$

NVIDIA RTX 5090 GPU. To improve computational efficiency, we used mixed-precision training (FP16) and a data loading pipeline with four worker processes. We used AdamW optimizer with an initial learning rate of 1×10^{-4} in the first training stage and reduced it to 5×10^{-5} during second.

5.2 Baselines

We compare Nüshu-PitchVITS with five representative TTS baselines: Tacotron 2 (Shen et al., 2018), FastSpeech 2 (Ren et al., 2020), Glow-TTS (Kim et al., 2020), F5-TTS (Chen et al., 2025), and standard VITS (Kim et al., 2021). These models cover major neural TTS paradigms, including autoregressive sequence-to-sequence synthesis, non-autoregressive duration-based synthesis, flow-based synthesis, flow-matching-based synthesis, and end-to-end variational adversarial synthesis. For a fair comparison, all baseline models were trained and evaluated on the same Nüshu-Voice train, validation, and test splits.

5.3 Evaluation Metrics

To comprehensively evaluate the quality of the synthesized Nüshu speech, we employ a combination of objective and subjective metrics. Objective assessments include Mel-Cepstral Distortion (MCD) (Kubichek, 1993) to measure spectral fidelity. Furthermore, given the critical role of pitch guidance in our proposed architecture, we compute the Root Mean Square Error (RMSE) and the Pearson correlation coefficient of the fundamental frequency (F_0) to evaluate prosodic reconstruction accuracy. For subjective evaluation, we conduct Mean Opinion Score (MOS) tests to assess the generated audio across two dimensions: naturalness and intelligibility (Sector, 1996).

Notably, while Word Error Rate (WER) computed via an off-the-shelf Automatic Speech Recognition (ASR) model is the standard objective proxy for intelligibility in conventional TTS studies, the complete absence of Nüshu ASR systems precludes this approach. Consequently, we rely on human-evaluated intelligibility MOS as a necessary alternative to assess the linguistic clarity and pronunciation correctness of the synthesized outputs. Detailed metric definitions and the subjective evaluation guidelines are provided in Appendix B.

6 Results

6.1 Objective TTS Evaluation

The objective evaluation results, detailed in **Table 3**, clearly demonstrate the fragility of conventional TTS architectures under extreme low-resource constraints. Traditional models, including Tacotron 2, FastSpeech 2, and Glow-TTS, suffer catastrophic degradation, exhibiting severe spectral distortion (MCD > 37) and virtually no pitch correlation. While the standard VITS model emerges as the strongest baseline, our proposed method achieves substantial improvements across all metrics. By explicitly integrating an F_0 predictor, our framework minimizes the MCD to 3.11 and drastically reduces the F_0 RMSE to 13.23. Furthermore, it achieves a highly accurate F_0 correlation of 0.92, confirming that explicit prosodic guidance effectively stabilizes acoustic generation and prevents spectral collapse.

Model	Objective Metrics			Subjective MOS	
	MCD ↓	F_0 RMSE ↓	F_0 Corr. ↑	Naturalness ↑	Intelligibility ↑
Tacotron 2	44.26 ± 4.34	47.15 ± 16.01	0.32 ± 0.38	1.66 ± 0.84	1.03 ± 0.18
FastSpeech 2	37.53 ± 5.91	29.51 ± 13.06	0.66 ± 0.27	1.90 ± 0.71	1.10 ± 0.30
F5-TTS	9.96 ± 1.92	37.03 ± 18.70	0.47 ± 0.26	3.10 ± 0.92	3.27 ± 0.44
Glow-TTS	39.96 ± 3.91	28.44 ± 13.77	0.72 ± 0.22	2.61 ± 0.82	1.03 ± 0.18
VITS	5.19 ± 1.28	23.42 ± 13.13	0.73 ± 0.31	3.03 ± 0.56	3.90 ± 0.55
Nüshu-PitchVITS	**3.11** ± 1.34	**13.23** ± 9.63	**0.92** ± 0.12	**3.31** ± 0.53	**4.97** ± 0.18
Ground Truth	–	–	–	3.34 ± 0.60	N/A†

Table 3: Objective and subjective evaluation results on **NüshuVoice**. Objective metrics include Mel-Cepstral Distortion (MCD), F_0 RMSE, and F_0 correlation. Subjective metrics include naturalness and intelligibility MOS. The best model results are shown in **bold**. †Ground-truth intelligibility MOS was not reported.

6.2 Subjective Human Evaluation

The subjective Mean Opinion Scores (**Table 3**) further corroborate our objective findings. Standard baselines fail entirely to produce comprehensible Nüshu speech, yielding Intelligibility MOS scores near 1.0 (indicating unintelligible babbling). Although recent models like F5 and VITS manage to generate partially recognizable audio, our method significantly surpasses them, achieving a near-perfect Intelligibility MOS of 4.97. Crucially, our model attains a Naturalness MOS of 3.31, performing highly competitively with the ground-truth recordings (3.34). These results conclusively demonstrate that our approach not only preserves the strict phonetic integrity of the Nüshu script but also reconstructs its authentic tonal dynamics.

6.3 Spectral Analysis

As shown in **Figure 4**, both VITS and our Nüshu-PitchVITS generate recognizable harmonic structures and clear formant trajectories, suggesting that VITS-style end-to-end alignment is relatively robust in the Nüshu low-resource setting. However, standard VITS still exhibits noticeable over-smoothing, weakened harmonic contrast, and less stable energy distribution, especially around syllable transitions. In contrast, Nüshu-PitchVITS produces spectrograms that are closer to the ground truth, with sharper harmonic bands, smoother temporal continuity, and more consistent pitch-related structures. These differences indicate that the explicit F_0 branch helps the model preserve tonal dynamics and improves prosodic stability beyond what can be achieved by standard latent acoustic modeling alone.

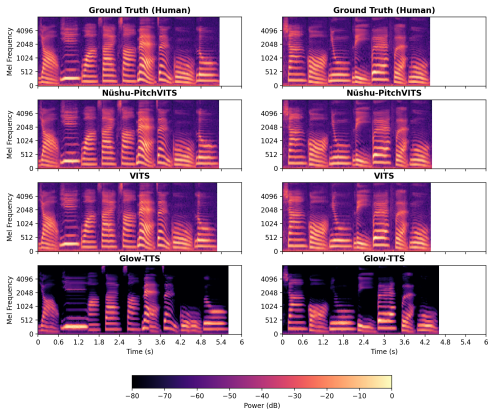


Figure 4: Mel-spectrogram comparison of two Nüshu utterances. The left column corresponds to an utterance with IPA transcription [iou13 i5 va33 si44 suow44 mai42 tseng42 ku21 lu21], and the right column corresponds to [ciang13 kou21 njyu33 tie42 pwe5 fwe13 ngu13].

7 Conclusion

This paper presents the first dedicated study on text-to-audio modeling for Nüshu, an endangered phonetic script with unique linguistic and cultural significance. To address the lack of acoustic resources, we constructed a sentence-level Nüshu TTS dataset by aligning verified Unicode Nüshu text, phonetic transcriptions, standard Chinese translations, and archival recordings.

Building on this dataset, we proposed Nüshu-PitchVITS tailored to the extreme low-resource characteristics of Nüshu. By introducing an explicit pitch prediction branch, the model incorporates tonal information as a prosodic inductive bias. Experimental results show that the proposed method consistently outperforms strong TTS baselines in spectral fidelity, pitch reconstruction, and human

perceptual evaluation.

Beyond model performance, this work highlights the importance of acoustic preservation for endangered writing traditions. Since Nüshu is inherently phonetic and closely tied to the Jiangyong dialect, recovering its pronunciation is essential for preserving not only its written form but also its oral and cultural heritage. We hope that the proposed dataset and modeling framework can serve as a foundation for future research on low-resource TTS, endangered language revitalization, and computational approaches to intangible cultural heritage.

Limitations

This work has several limitations. First, our sentence-level audio is assembled from isolated syllable recordings rather than recorded as natural continuous speech. As a result, the dataset cannot fully capture coarticulation, pauses, rhythm, and discourse-level prosody that would appear in spontaneous or naturally read utterances. Second, the scale of the dataset is inherently constrained by the scarcity of surviving Nüshu acoustic resources. Although NüshuVoice contains 1,252 utterances and 98.8 minutes of audio, this amount necessarily limits the diversity of speakers, speaking styles, and phonetic contexts available for model training and evaluation. Third, because sentence-level audio is assembled from a finite inventory of archival syllable recordings, train and test splits may share underlying acoustic units. Therefore, objective results should be interpreted within this controlled archival reconstruction setting. Finally, our strict Unicode-based filtering improves annotation reliability, but it also excludes historical glyph variants without exact digital equivalents.

Ethics Considerations

Nüshu is an endangered script and an important form of women's cultural heritage. This work aims to support preservation and linguistic documentation, not to replace community knowledge or living cultural practice. Since the audio comes from archival pronunciation resources, any release of derived data, models, or generated samples should respect the licensing conditions and rights associated with the original recordings. Synthetic speech should be clearly labeled as machine-generated to avoid cultural misrepresentation or confusion with authentic historical recordings. Future work should involve Nüshu scholars, local cultural practitioners, and community stakeholders.

The archival recordings used in this work originate from the *Nüshu Pronunciation Electronic Dictionary*. We use these materials solely for academic research and cultural preservation. The release of NüshuVoice follows the access and redistribution conditions of the original resources. Where redistribution of the original recordings is not permitted, we release only the derived metadata, code, and reproducible processing pipeline rather than the original audio files. The dataset and models are intended for non-commercial research, documentation, and educational use related to Nüshu and low-resource speech technology.

Declaration of AI Usage

The authors used AI-assisted writing tools to improve the clarity, grammar, and readability of the manuscript. All AI-generated suggestions were carefully reviewed, edited, and verified by the authors. The authors take full responsibility for the content of the paper, including the proposed method, dataset construction, experimental design, results, and conclusions.

References

Mohammad Jahid Ibna Basher, Md Kowsher, Md Saiful Islam, Rabindra Nath Nandi, Nusrat Jahan Prottasha, Mehadi Hasan Menon, Tareq Al Muntasir, Shammur Absar Chowdhury, Firoj Alam, Niloofar Yousefi, and 1 others. 2025. Bntts: Few-shot speaker adaptation in low-resource setting. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4956–4968.

Viann Sum Yat Chan and Christopher Hammerly. 2025. Evaluating indigenous language speech synthesis for education: A participatory design workshop on ojibwe text-to-speech. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 47–64.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271.

Li Congrong. 2024. History, characteristics, and modern vitality of nüshu: A cultural anthropology perspective. *Anthropological Explorations of Gender, Identity, and Economics*, 85.

Katri Hiovain-Asikainen, Thomas B Kjærstad, Maja Lisa Kappfjell, and Sjur Moshagen. 2025. The

world's first south sámi tts–a revitalisation effort of an endangered language by reviving a legacy voice. In *Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages*, pages 12–21.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International conference on machine learning*, pages 5530–5540. PMLR.

Robert Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, volume 1, pages 125–128. IEEE.

Ki-Joong Kwon, Jun-Ho So, and Sang-Hoon Lee. 2025. Parameter-efficient fine-tuning for low-resource text-to-speech via cross-lingual continual learning. In *Proc. Interspeech*, volume 2025, pages 1613–1617.

Seongyeon Park, Myungseo Song, Bohyung Kim, and Tae-Hyun Oh. 2023. Unsupervised pre-training for data-efficient text-to-speech on low resource languages. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jian Rao and Xinyue Zheng. 2022. Intangible cultural heritage protection based on mar algorithm. In *International Conference on Multi-modal Information Analytics*, pages 136–144. Springer.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

International Telecommunication Union. Telecommunication Standardization Sector. 1996. *Methods for subjective determination of transmission quality*. International Telecommunication Union.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, and 1 others. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Yuqian Sun, Yuying Tang, Ze Gao, Zhijun Pan, Chuyan Xu, Yurou Chen, Kejiang Qian, Zhigang Wang, Tristan Braud, Chang Hee Lee, and 1 others. 2023. Ai nüshu: An exploration of language emergence in sisterhood through the lens of computational linguistics. In *SIGGRAPH Asia 2023 Art Papers*, pages 1–7.

Thomas Thurnell-Read, Michael Skey, and Marie Heřmanová. 2022. Introduction: Cultures of authenticity.

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2802–2812.

Ivory Yang, Xiaobo Guo, Yuxin Wang, Hefan Zhang, Yaning Jia, William Dinauer, and Soroush Vosoughi. 2025. Recontextualizing revitalization: A mixed media approach to reviving the nüshu language. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12430–12439.

Ivory Yang, Weicheng Ma, and Soroush Vosoughi. 2024. Nushurescue: Revitalization of the endangered nushu language with ai. *arXiv preprint arXiv:2412.00218*.

Shangfang Zhengzhang and Wei Zheng. 2015. Wu dialect. *The Oxford handbook of Chinese linguistics*, pages 189–199.

Wen Zuo, Metta Sirisuk, and 1 others. 2024. *Jiangyong Nvshu in China: the Gender, Signifier, Signified, and Sustainable Inheritance Path of the Local Culture*. Ph.D. thesis, Mahasarakham University.

A Additional Dataset Construction Details

This appendix provides additional details on the construction of NüshuVoice, including archival resource recovery, Unicode-based orthographic filtering, sentence-level audio assembly, and dataset organization. These details complement the main dataset construction section and are intended to improve reproducibility.

A.1 Archival Resource Recovery

The construction of NüshuVoice relies on two complementary archival resources. The textual source is *The Mystery of Jiangyong Nüshu*, which provides Nüshu sentences together with standard Chinese translations and phonetic transcriptions. The acoustic source is the *Nüshu Pronunciation Electronic Dictionary*, which contains isolated syllable-level pronunciation recordings of Nüshu characters.

Because the electronic dictionary was originally released as a CD-ROM resource, its software interface is not reliably executable on modern operating systems. We therefore recovered the raw acoustic data directly from the original CD-ROM contents. Specifically, we extracted the isolated WAV recordings from the `voice` directory and used only these

Case	Decision
The source glyph has an exact Unicode counterpart.	Retain the character.
All characters in the sentence have visually consistent Unicode forms.	Retain the sentence.
A source glyph has no corresponding Unicode character.	Discard the entire sentence.
A source glyph differs from the Unicode form by a missing or additional stroke.	Discard the entire sentence.
A source glyph contains a variant component or structurally different radical.	Discard the entire sentence.
The glyph–Unicode mapping is visually ambiguous.	Discard the entire sentence.

Table 4: Unicode-based orthographic filtering criteria used in the construction of NüshuVoice.

original pronunciation files for sentence-level audio construction. This step was necessary because no modern large-scale Nüshu speech corpus is available, and the recovered recordings constitute a rare acoustic resource for preserving Nüshu.

A.2 Digital Environment and Lookup Tools

Dataset construction was conducted in a Windows environment with Nüshu font and input support. We used the Rime Nüshu input method to input and render standardized Unicode Nüshu characters. For candidate character lookup and Unicode conversion, we referred to the online converter associated with the *Nüshu Standard Character Calligraphy Copybook*. Audio editing, concatenation, and waveform export were performed using Audacity.

These tools were used only for rendering, lookup, and audio editing. All final decisions about character validity were made through manual verification against the archival text rather than automatic conversion alone.

A.3 Unicode-Based Orthographic Filtering

Historical Nüshu glyphs are not fully standardized, and the same syllable may appear in different handwritten or printed variants. To build a reliable Unicode-based TTS dataset, we adopted a conservative sentence-level filtering strategy. Each candidate sentence was manually inspected character by character. A sentence was retained only when every source glyph had a visually and structurally consistent Unicode counterpart.

If any character in a sentence lacked a Unicode equivalent, showed an ambiguous glyph shape, or differed from the corresponding Unicode form in stroke structure or component composition, the entire sentence was discarded. This strict filtering strategy reduces coverage but improves the orthographic reliability of the final dataset.

A.4 Sentence-Level Audio Assembly

The acoustic dictionary provides isolated syllable-level recordings rather than continuous sentence-level speech. Therefore, we constructed sentence-level audio using a deterministic concatenation procedure. For each retained sentence, we first used the phonetic transcription in the textual source to identify the pronunciation of each Nüshu character. We then matched each phonetic unit to its corresponding isolated WAV file in the recovered electronic dictionary.

The selected syllable recordings were arranged in sentence order and concatenated using Audacity to form a complete utterance. Each generated waveform was exported as a sentence-level audio file and assigned the same unique identifier as its corresponding metadata entry. This procedure creates a one-to-one alignment between standardized Nüshu text, phonetic transcription, Chinese translation, and sentence-level audio.

A.5 Dataset Organization

The final dataset consists of a metadata table and a set of waveform files. Each metadata entry contains a unique sentence identifier, standardized Unicode Nüshu text, phonetic transcription, and standard Chinese translation. The waveform folder contains the corresponding sentence-level audio files. The shared sentence identifier links each textual entry to its audio file.

This organization enables both text-to-speech training and multimodal analysis. It also makes the construction process transparent: each sentence-level audio file can be traced back to its Unicode Nüshu text, phonetic transcription, and the sequence of isolated archival syllable recordings used to assemble it.

B Evaluation Details

This appendix provides additional details on the objective metric computation and subjective evaluation protocol used in our experiments.

B.1 Objective Metrics

We evaluate spectral fidelity using Mel-Cepstral Distortion (MCD). Given the ground-truth mel-cepstral coefficients $\mathbf{c}_t$ and the generated coefficients $\hat{\mathbf{c}}_{\pi(t)}$, where $\pi(\cdot)$ denotes the frame alignment obtained by dynamic time warping (DTW), MCD is computed as

$$\mathrm{MCD} = \frac{1}{T}\sum_{t=1}^{T}\frac{10}{\ln 10}\sqrt{2\sum_{m=1}^{M}\left(c_{t,m}-\hat{c}_{\pi(t),m}\right)^2}, \tag{16}$$

where T is the number of aligned frames and M is the number of mel-cepstral coefficients. Lower MCD indicates better spectral fidelity.

For pitch evaluation, we extract the fundamental frequency using the DIO algorithm in WORLD. Unvoiced frames are excluded from pitch-based evaluation. Let f_t and $\hat{f}_t$ denote the ground-truth and generated F_0 values on the aligned voiced frames. The F_0 RMSE is defined as

$$F_0\text{-RMSE} = \sqrt{\frac{1}{|\mathcal{V}|}\sum_{t\in\mathcal{V}}\left(f_t-\hat{f}_t\right)^2}, \tag{17}$$

where $\mathcal{V}$ denotes the set of voiced frames. Lower F_0 RMSE indicates more accurate pitch reconstruction.

We also compute the Pearson correlation coefficient between the ground-truth and generated F_0 contours:

$$\rho_{F_0} = \frac{\sum_{t\in\mathcal{V}}(f_t-\bar{f})(\hat{f}_t-\bar{\hat{f}})}{\sqrt{\sum_{t\in\mathcal{V}}(f_t-\bar{f})^2}\sqrt{\sum_{t\in\mathcal{V}}(\hat{f}_t-\bar{\hat{f}})^2}}, \tag{18}$$

where $\bar{f}$ and $\bar{\hat{f}}$ are the mean ground-truth and generated F_0 values over voiced frames. Higher correlation indicates better preservation of pitch contour dynamics.

B.2 Subjective Evaluation Guidelines

We conduct MOS evaluation along two dimensions: naturalness and intelligibility. Annotators are asked to evaluate each audio sample independently without knowing which model generated it. The two dimensions are defined as follows.

Naturalness. Naturalness measures how human-like and fluent the audio sounds, including smoothness, absence of artifacts, and overall acoustic quality. Annotators are instructed not to focus on whether the content is correct when assigning this score.

Intelligibility. Intelligibility measures whether the pronunciation is clear and whether the intended Nüshu syllables can be recognized. Annotators are instructed to focus on linguistic clarity and pronunciation correctness rather than overall audio pleasantness.

For both dimensions, annotators use a five-point scale:

Score	Description
5	Excellent: very natural or clearly intelligible
4	Good: mostly natural or mostly intelligible
3	Fair: understandable but with noticeable issues
2	Poor: difficult to understand or clearly unnatural
1	Bad: unintelligible or severely degraded