WhispSynth: Scaling Multilingual Whisper Corpus through Real Data Curation and A Novel Pitch-free Generative Framework

Anonymous ACL submission

Abstract

Whisper generation is constrained by the difficulty of data collection. Because whispered speech has low acoustic amplitude, high-fidelity recording is challenging. In this paper, we introduce **WhispSynth**, a large-scale multilingual corpus constructed via a novel high-fidelity generative framework. Specifically, we propose a pipeline integrating Differentiable Digital Signal Processing (DDSP)-based pitch-free method with Text-to-Speech (TTS) models. This framework refines a comprehensive collection of resources, including our newly constructed WhispXXX dataset[1], into 118 hours of high-fidelity whispered speech from 479 speakers. Unlike standard synthetic or noisy real data, our data engine faithfully preserves source vocal timbre and linguistic content while ensuring acoustic consistency, providing a robust foundation for text-to-whisper research. Experimental results demonstrate that WhispSynth exhibits significantly higher quality than existing corpora. Moreover, our CosyWhisper, tuned with WhispSynth, achieves speech naturalness on par with ground-truth samples. We will release the implementation code and available resources to enable the reproduction of the WhispSynth generation pipeline.

1 Introduction

Whisper represents an innate and ubiquitous mode of human vocalisation. Primarily employed to attenuate acoustic propagation for secure communication (Tartter, 1989; Andersen, 2015; Rekimoto, 2023; Hiraki and Rekimoto, 2025), this modality also carries significant paralinguistic weight, often signalling intimacy or inducing a soothing effect. Notably, these affective characteristics have driven the recent surge in Autonomous Sensory Meridian Response (ASMR) content across social media.

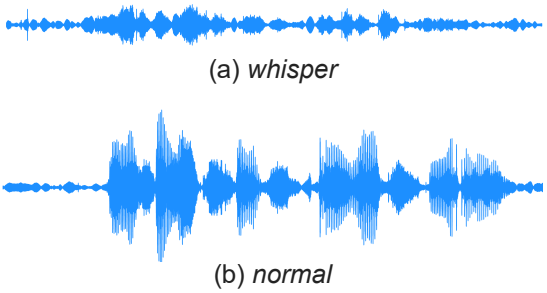


Figure 1: Different Dynamic Range. Whispers exhibit a significantly lower sound pressure level compared to normal speech, even when the linguistic content is identical.

Research interest in whisper generation is rapidly growing, yet a critical data bottleneck fundamentally constrains progress. Publicly available whispered corpora are typically characterized by limited scale, heterogeneous recording quality, and a lack of speaker diversity. Specifically, the inherently low sound pressure level of whispers (see Figure 1) makes high-fidelity capture exceptionally difficult. Furthermore, existing datasets often contain substantial acoustic artifacts that are resistant to standard speech enhancement, because whispered speech is produced with predominantly noise-driven excitation.

While Text-to-Speech (TTS) remains the primary paradigm for synthesising whisper, and general TTS has made remarkable strides in capturing nuanced prosodic features such as rhythm, intonation, and even non-verbal vocalizations (Du et al., 2025; Zhou et al., 2025; Chen et al., 2025), achieving high-fidelity whisper generation remains a significant challenge. Recently, we have observed that commercial services such as Doubao AI Chatbot (ByteDance, 2025), ElevenLabs TTS (ElevenLabs, 2025), and MiniMax TTS (ElevenLabs,

[1]We refer to the dataset as "WhispXXX" to maintain institutional anonymity. The real dataset name and resource will be disclosed upon publication.

2025) can produce natural-sounding whispered speech. However, these systems rely on proprietary datasets and closed-source pipelines. Consequently, the lack of high-quality, open-access resources remains a critical bottleneck that directly impedes advances in general whisper synthesis.

To further investigate the limitations of current open-source TTS models in whisper generation, we reveal that TTS models remain severely flawed at zero-shot whisper generation. Even when provided with explicit whisper prompts or textual instructions, these models fail to bridge the domain gap, highlighting a significant deficiency in serving this special modality. This fragility can be attributed to two complementary biases: (*i*) corpora and vocoders are predominantly optimized for modal phonation, establishing implicit priors that favor harmonic-rich signals with clear fundamental frequencies (F0); and (*ii*) speaker-embedding extractors trained on such data frequently misconstrue whisper prompts as hoarseness or environmental noise rather than authentic breathy voice.

Paradoxically, these TTS models trained on normal speech possess a latent capacity for whisper generation, derived from sporadic devoiced segments that share similar acoustic energy distributions. The overlap is anatomically grounded: whispered production suppresses vocal-fold vibration while retaining turbulent airflow at the same supralaryngeal constriction sites exploited for voiceless obstruent in normal speech (Ito et al., 2005). Consequently, the acoustic realisations of whisper and devoiced segments are governed by nearly identical aerodynamic–articulatory constraints, endowing standard models with a dormant yet exploitable reservoir of whisper-like priors.

To address these limitations, we introduce **WhispSynth**, a large-scale, multilingual corpus constructed via a novel high-fidelity generative framework. Specifically, we propose a whisper generation pipeline integrating Differentiable Digital Signal Processing (DDSP)-based pitch-free with TTS models. This framework refines a comprehensive collection of resources into 118 hours of high-fidelity 24 kHz whispers over 479 speakers. The main contributions are summarized as follows:

1. **WhispReal:** We construct the first union of realistic whisper by consolidating 6 publicly available corpora and releasing our own newly recorded Mandarin dataset, **WhispXXX** (∼45 h). The resulting collection, termed **WhispReal** (∼118 h), provides standardized splits and metadata, establishing a robust data foundation for the field.

2. **WhispSynth:** We propose a scalable data engine that synergizes state-of-the-art TTS (e.g., CosyVoice 3) with a novel pitch-free method. This pipeline effectively disentangles pitch from wav signal, transforming the noisy WhispReal source into **WhispSynth**—a studio-grade synthetic corpus of approximately 118 hours. The generated whisper preserves linguistic content, reducing CER and WER by 11%, and also improves audio quality, increasing DNSMOS by 3%.

3. **CosyWhisper:** We fine-tune CosyVoice 3 on our WhispSynth to develop **CosyWhisper**. Experimental results demonstrate that training on WhispSynth yields significantly better performance than training on realistic recordings, achieving speech naturalness on par with ground-truth samples. We will release the fine-tuned model as the first open-source and reproducible *text-to-whisper* system.

2 WhispReal

WhispReal dataset is a curated collection of existing publicly available whispered speech corpora combined with our proprietary **WhispXXX (anonymized)** dataset. As summarized in Table 1, the number of open-source whisper datasets is limited, and they exhibit significant variations in scale, linguistic diversity, and accessibility. The licensing terms (CC Type) for these datasets are also heterogeneous, including Creative Commons Attribution-NonCommercial-ShareAlike (NC-SA), Attribution-NoDerivatives (ND), Attribution-NonCommercial (NC), among others, while the license information for some datasets remains unspecified (N.A.). This heterogeneity directly impacts data availability, redistribution rights, and the freedom to utilize the data for academic and commercial research. In constructing WhispReal, we meticulously respected and adhered to the licensing agreements of all source datasets, ensuring the full legal compliance of the resulting collection.

Existing whispered speech corpora documented in the literature can be broadly categorized into three primary types, with a subset being publicly available as highlighted in gray in Table 1:

Dataset	Source	Sample Rate	Size (*h*)	Avg (*s*)	Language	Speaker	P?	CC Type
UTVE-I	(Zhang and Hansen, 2010)	44100	<1	N.A.	en	12	N	N.A.
UTVE-II	(Ghaffarzadegan et al., 2014)	44100	1	N.A.	en	112	N	N.A.
AVWD	(Zhou et al., 2019)	44100	<2.44	N.A.	zh	10	Y	N.A.
Whi-spe	(Grozdić et al., 2012)	22050	<5	N.A.	sr	10	Y	N.A.
AV-Whisper	(Tran et al., 2013)	48000	<10	N.A.	en	11	Y	N.A.
CIAIR	(Kawaguchi et al., 2002)	16000	15	N.A.	ja	123	Y	N.A.
iWhisper-Mandarin	(Lee et al., 2014a)	16000	15	N.A.	zh	80	Y	N.A.
wSPIRE	(Singhal et al., 2021)	44100	18	N.A.	en	88	Y	N.A.
AISHELL6-Whisper	(Li et al., 2025)	48000	29.75	9.01	zh	167	Y	BY-NC-SA
CHAINs (subset)	(Cummins et al., 2006)	44100	2.55	4.12	en	36	Y	BY-ND
EARs (subset)	(Richter et al., 2024)	48000	3.22	16.58	en	107	Y	BY-NC
Expresso (subset)	(Nguyen et al., 2023)	48000	1.32	3.13	en	4	Y	BY-NC
Whisper40	(Yang and Zhou, 2024)	16000	6.10	12.25	zh	40	Y	N.A.
wTIMIT	(Li-Li et al., 2005)	44100	29.44	5.04	en	48	Y	N.A.
WhispXXX	Our new curation	44100	45.24	9.01	zh	77	Y	BY-NC
WhispReal	Our aggregated corpus	Mixed	117.62	9.91	en, zh	479	Y	Mixed
WhispSynth	Synthesized from WhispReal	24000	≈118	≈10	en, zh	479	N	MIT

Table 1: Overview of open-source and newly introduced datasets used in this work, with the highlighted (gray) rows indicating all obtainable real whisper corpora included in our study. Entries include data duration (Size), average utterance length (Avg), language, speaker count, whether the paired normal segments are available (P?), and license type (CC Type), "Mixed" indicates multiple license terms or sample rate, "N.A." denotes unknown or unspecified.

Pure Whisper Datasets These collections are specifically dedicated to whispered vocalizations, designed for targeted acoustic analysis (Li-Li et al., 2005; Jou et al., 2005; Zhang and Hansen, 2010; Lim, 2011; Ghaffarzadegan et al., 2014; Lee et al., 2014b; Singhal et al., 2021; Hiraki and Rekimoto, 2022; Yang and Zhou, 2024). Among these, wTIMIT (Li-Li et al., 2005) is the most notable and widely adopted. It is a classic, relatively large-scale English whispered speech corpus that also includes parallel normal speech. The linguistic content of these collections is largely confined to the word or phrase level.

Subsets within Expressive Speech Databases Several speech corpora designed for emotional or expressive analysis incorporate whispered speech as one vocalization style (Cummins et al., 2006; Nguyen et al., 2023; Richter et al., 2024). These datasets often contain rich emotional annotations, facilitating research on affective whispering, although the whisper portions themselves are typically limited in size.

Audiovisual Whisper Collections This is an emerging category that integrates whispered speech with synchronized visual recordings (Kawaguchi et al., 2002; Tran et al., 2013; Petridis et al., 2018; Zhou et al., 2019; Li et al., 2025). Within this domain, AISHELL6-Whisper (Li et al., 2025) represents one of the largest Chinese datasets, characterized by a substantial number of speakers, which makes it well-suited for multimodal analysis and robustness research. It is worth noting that the upcoming dataset, WhispXXX (anonymized), also falls into this category. However, due to the heightened complexity of the recording process, the scalability of such audiovisual collections is constrained.

In the following sections, we present information on the existing available whispered data sources and a unique processing required for each. This is followed by a comprehensive account of the data collection procedures for our dataset.

2.1 Existing Available Resources

AISHELL6-Whisper (Li et al., 2025): it provides audio-visual whisper speech dataset, featuring 30 hours each of whisper speech and parallel normal speech, with synchronized frontal facial videos. We used only the audio and followed the train/test/-valid splits. It should be noted that some utterances within the retained set still lack paired counterparts.

wTIMIT (Lim, 2011): it is constructed after the TIMIT dataset (Garofolo et al., 1993). The first phase of the project collected the speech from 20 Singaporeans, and the second phase collected the speech from 28 North Americans. After excluding files with evident sampling errors, we retained all remaining audio files. It should be noted that some utterances within the retained set still lack whispered counterparts. Subsequently, we reorganized the dataset by speaker into a 60%–20%–20% split for training, validation, and testing, respectively.

Whisper40 (Yang and Zhou, 2024): it is a Mandarin Chinese corpus comprising whispered and normal speech recordings from 40 speakers without corresponding transcripts. However, since the normal and whispered speech pairs share identi-

cal linguistic content, the normal speech is clearly articulated, and the text is consistent across speakers, the authors performed manual transcription and verification for all utterances. We followed the train/test/valid splits of the original dataset.

CHAINS (Cummins et al., 2006): **CHA**racterizing **IN**dividual **S**peakers contains recordings from 36 speakers (8 from the UK/US and 28 from Ireland), each producing 37 utterances across six different speaking conditions: solo speech, retelling, synchronous speech, repetitive synchronous imitation, fast speech, and whispered speech. We extracted the solo speech and whispered speech from CHAINs and segmented its four short fables (Cinderella, Rainbow text, North Wind and the Sun, and Members of the Body) into sentences to ensure that each audio file is under 30 seconds in duration, and split the data by speaker into 60% training, 20% validation, and 20% test sets.

Expresso (Nguyen et al., 2023): it is a high-quality (48 kHz) expressive speech dataset containing expressive read speech in 8 styles and improvised dialogues in 26 styles, recorded from 4 speakers. We selected default style and whisper style of it. Following a consistent partitioning scheme, we split the data by speaker into 60% training, 20% validation, and 20% test sets.

EARs (Richter et al., 2024) : **E**xpressive **A**nechoic **R**ecordings of **S**peech is a high-quality expressive speech collection, recorded at 48 kHz and comprising 107 speakers from diverse backgrounds, totaling approximately 100 hours of clean, anechoic speech. EARs covers the full dynamic range of vocal expression, from whispering to yelling and screaming. For this study, we extracted the whispered and regular speech subsets (split: 60% train / 20% valid / 20% test by speaker).

2.2 The WhispXXX corpus

The WhispXXX corpus consists of 85 hours of paired whispered and normal speech from 77 speakers of students recruited at X University. WhispXXX is designed to facilitate research and development in Mandarin whispered ASR. It is constructed after the THCHS-30 dataset (Wang and Zhang, 2015), which is usually used for the study of Mandarin speech recognition.

Speaker Recruitment The dataset was recorded by 77 native Mandarin speakers (37 male and 40 female), aged 22-26, recruited from X University. All participants met the following selection criteria essential for speech production research:

- Native proficiency in Mandarin Chinese.
- No diagnosed speech or hearing disorders.
- No acute oral/auditory health conditions.

These criteria ensured consistent recording quality while complying with ethical research standards. Participants received compensation and provided informed consent.

Record Settings The participants were instructed to produce each sentence twice: once in normal speech and once in whispered speech. Due to the time-consuming nature of the recordings, sessions could be divided into multiple sittings, and submissions were only accepted after all recordings were completed. To ensure universal applicability, no restrictions were placed on the speakers' regional origin or accent. During recording, no obstructions were allowed between the speaker and the recording device to avoid interfering with airflow pickup by the microphone.

Data preparation Following the partitioning scheme of the THCHS-30 dataset (Wang and Zhang, 2015), our dataset is divided into four groups according to the recording text: A (sentence ID from 1 to 250), B (sentence ID from 251 to 500), C (sentence ID form 501 to 750), and D(sentence ID from 751 to 1000). Each participant recorded 250 sentences (approximately 50 words per sentence), totaling approximately one hour of recording per speaker. Speakers were randomly assigned unique identification numbers and evenly distributed across the four groups.

3 WhispSynth

3.1 Generation Pipeline

Although recent advances in TTS synthesis have enabled the generation of highly natural and intelligible speech, when synthesizing whispered speech many state-of-the-art systems tend to produce residual pitch or fundamental frequency (F0) components. To address this issue, we propose a post-processing pipeline that detects and removes pitch components from synthesized whispers while preserving the desired noise-like whisper quality. Figure 4 outlines the proposed procedure.

The core idea is to leverage a DDSP decomposition to isolate and suppress harmonic content in pitch-contaminated segments, followed by an overlap-add (OLA) reconstruction to ensure smooth transitions. The full algorithm is provided in Appendix A.

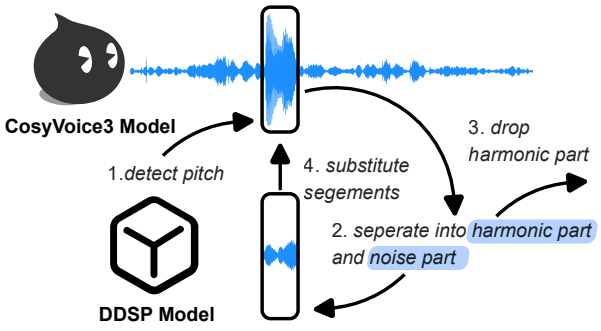


Figure 2: Visualization of the WhispSynth's generation pipeline. We apply CosyVoice3 and DDSP Model to generate pitch-free voice.

3.2 CosyVoice3: State-of-the-Art Speech Synthesis

CosyVoice3 (Du et al., 2025), developed by Alibaba Group, represents a next-generation speech synthesis model designed for high-quality and natural speech in open-environment scenarios. It integrates a multi-task speech tokenizer, scalable Differentiable Reward Optimization for post-training, and is trained on an ultra-large-scale multilingual corpus with instruction-following controllability.

The model exhibits particular strength in whisper synthesis. Its multi-task tokenizer, trained jointly using speech emotion recognition and audio event detection tasks, effectively encodes paralinguistic features such as softness and breathiness. Moreover, its million-hour-scale training corpus, drawn from diverse real-world recordings, contains abundant whisper and low-volume speech samples, enabling robust learning of such vocal qualities. Nonetheless, during whisper synthesis, the model occasionally exhibits fundamental frequency artifacts. This issue stems from the fact that the neural modules are predominantly trained on phonated speech, leading them to inadvertently introduce subtle periodicity even when generating whisper-like outputs.

3.3 Pitch-free with DDSP: An Overview

The DDSP model (Engel et al., 2020), proposed by researchers from Google, adopts the harmonic-plus-noise model (Serra and Smith, 1990) to decompose a monophonic sound into a harmonic component H and a noise component N:

$$\mathbf{S}[i] = \mathbf{H}[i] + \mathbf{N}[i]. \tag{1}$$

Given the upsampled estimated $\widetilde{F}_0[i]$, in this work we approximate $\mathbf{H}$ by a sawtooth signal (Wu et al., 2022), which contains an equal number of even and odd harmonics with decaying magnitudes:

$$\widetilde{\mathbf{H}}[i] = \sum_{k=1}^{K} \frac{1}{k} \sin(\phi_k[i]). \tag{2}$$

The instantaneous phase $\phi_k(t)$ is obtained by cumulative summation of the instantaneous frequency samples $k\widetilde{F}_0[i]$:

$$\phi_k[i] = 2\pi \sum_{n=0}^{N} k\widetilde{F}_0[i], \tag{3}$$

where $\widetilde{\mathbf{H}}$ is treated as the "excitation signal" and shaped into the desirable $\mathbf{y}_\mathrm{h}$ by means of an linear time-varying finite impulse response (LTV-FIR) filter $\psi_\mathrm{h}(i) \in R^{L_\mathrm{h}}$:

$$\bar{\mathbf{H}}[i] = \widetilde{\mathbf{H}}[i] * \psi_\mathrm{h}[i], \tag{4}$$

where $\mathbf{N}$ is approximated by convolving a uniform distributed noise signal ζ ranging from -1 to 1 (with the same length as a frame) with an LTV-FIR filter $\psi_\mathrm{n}[i] \in R^{L_\mathrm{n}}$ estimated per frame:

$$\bar{\mathbf{N}}[i] = \zeta[i] * \psi_\mathrm{n}[i]. \tag{5}$$

Jointly, the parameters $\Phi := \{\widetilde{F_0}[i], \psi_\mathrm{h}[i], \psi_\mathrm{n}[i]\}_{i=1}^{N}$ are estimated from an input mel-spectrogram $\mathbf{X}$ per frame by an NN mapping function f_NN (Gulati et al., 2020):

$$\Phi = f_\mathrm{NN}(\mathbf{X}). \tag{6}$$

DDSP entails a source-filter model (Wang et al., 2019) that can successfully separate the noise component from normal speech.

3.4 Pitch-free with DDSP: Training Pipeline

Based on the prior work of DDSP in singing vocoders (Wu et al., 2022), the authors identified a critical limitation: the model produces buzzing artifacts in unvoiced and semi-voiced segments when used as a subtractive synthesizer[2]. To overcome this limitation and adapt DDSP for high-quality whispered speech synthesis, we propose the following three training strategies:

(i) **Adversarial Training with Normal Speech:** We integrate the DDSP generator into the BigV-GAN (Lee et al., 2022) framework, applying a multi-resolution discriminator (MRD) and a

[2] https://github.com/YatingMusic/ddsp-singing-vocoders/tree/main/postprocessing

multi-period discriminator (SPD) with least-square GAN loss $\mathcal{L}_{\text{adv}}$ (Mao et al., 2017). This adversarial training is essential because GAN-based discrimination has been shown to effectively improve the modeling of harmonic structures and suppress artifact-prone synthesis patterns.

(ii) **Continued Training with Whispered Speech:** After initially training the DDSP vocoder on normal speech data, we continue training it on the WhispReal dataset without adversarial objectives for faster training speed. This two-stage procedure allows the model to first learn robust pitch-conditioned synthesis, then specialize in whispered speech generation while avoiding instability that might arise from direct training on limited whispered data.

(iii) **Semi-supervised Dual-focus Training:** We observe that the WhispReal dataset contains samples with perceptible pitch contours resembling normal speech, due to speaker errors or temporary vocal fatigue. Therefore the training scheme always treats the output as a sum of harmonic and stochastic components, but shifts the learning emphasis depending on the input type: whispered/normal speech share loss, but the gradient reweights harmonic/noise terms so each half teaches the other. This dual-focus approach enables mutually reinforcing improvements in both harmonic and stochastic modeling.

4 CosyWhisper

We fine-tune CosyVoice3 with Whisper, following the exact procedure outlined in the official script. The CosyVoice3 architecture comprises three primary components: the text-to-speech large language model (LLM), the Conditional Flow Matching (CFM) model, and HiFi-GAN vocoder. In our scenario, the distinction between whispered and normal speech lies primarily in acoustic modeling rather than semantic content. Therefore, we selectively fine-tune only the CFM model, which is responsible for converting semantic tokens into high-fidelity acoustic features, while keeping the LM and HiFi-GAN unchanged.

Although the official script currently supports only LLM training (line 65: "Run train. We only support LLM training for now"), we successfully extend it to CFM fine-tuning. Key adaptations include adding a token projection layer, replacing the encoder with a direct embedding lookup, and revising the conditioning mechanism. These changes improve token-acoustic alignment and model efficiency. Implementation specifics are provided in the Appendix B. We refer to the resulting fine-tuned model as **CosyWhisper**, a name that reflects its origin in the CosyVoice3 architecture and its specialized capability in Whisper-based speech synthesis.

5 Experiments

5.1 Experimental setups

Data Settings We used the English and Mandarin Chinese subsets of the WhispReal dataset for training and evaluation. And the WhispSynth dataset is partitioned accordingly. For each source, data were divided as described in Section 2.1. The validation set was used for model checkpoint selection, while the final sound quality evaluation was performed only on the test set. Detailed statistics of the dataset distribution are provided in Table 2.

	en		zh	
	# Train	**# Test**	**# Train**	**# Test**
File Count	20694	4770	32704	7256
Size (h)	29.62	6.91	65.86	15.25
Avg (s)	5.15	5.21	7.25	7.57
Speaker Count	78M 87F	15M 15F	206M 194F	26M 25F

Table 2: The statistics of the WhispReal dataset train/test distribution.

Training Settings We transformed the normalized speech in all the datasets into mel-spectrograms with a frame length of 1280 and a hop length of 320. Adversarial training was conducted using a batch size of 8, while standard (non-adversarial) training employed a batch size of 32. All experiments were performed on eight V100-32GB GPUs.

Evaluation Settings The generation performance is evaluated using both subjective and objective metrics. In terms of subjective evaluation, the proposed and baseline systems were evaluated using a 5-point Whisper-likeness Mean Opinion Score (W-MOS). It measured the perceptual similarity between the synthesized whisper and the listener's mental prototype of a whisper, ranging from 0 ("Totally Not similar") to 5 ("Extremely similar"). We have developed an automated subjective evaluation interface, as shown in Appendix C. For objective assessment, we adopt standard TTS metrics, as no widely used metrics are specifically designed for whisper. For the naturalness, we measure spectral differences with Mel Cepstral Distortion (MCD) (Chen et al., 2022), DNSMOS (Reddy et al., 2021), and UTMOS (Saeki et al., 2022)

Dataset	Source	Language	Naturalness ↑		Intelligibility ↓	F0 ↓
			DNSMOS	UTMOS	CER / WER (%)	VTR(%)
wTIMIT	(Li-Li et al., 2005)	en	2.76	1.31	50.99	0.69
CHAINs (*subset*)	(Cummins et al., 2006)	en	2.76	1.48	11.55	0.95
Expresso (*subset*)	(Nguyen et al., 2023)	en	3.24	1.50	7.77	0.58
EARs (*subset*)	(Richter et al., 2024)	en	3.43	2.01	6.31	0.30
Whisper40	(Yang and Zhou, 2024)	zh	2.61	1.27	72.80	0.99
AISHELL6-Whisper	(Li et al., 2025)	zh	2.75	1.68	38.26	0.94
WhispXXX	Ours	zh	2.90	1.33	36.74	0.98
WhispReal	Ours	en,zh	2.80	1.44	39.30/37.58	0.88
WhispSynth	Ours	en,zh	**2.89**	**1.46**	**31.16/20.98**	0.87

Table 3: Objective evaluation results on open-sourced and our proposed datasets. Comprehensive evaluation on whisper speech conversion and synthesis. **Quality**: Subjective naturalness (DNSMOS/UTMOS ↑). **Intelligibility**: Character/Word Error Rate (↓). **Pitch**: Voiced Time Ratio (VTR) (↓).

to estimate perceptual audio quality. We likewise calculate the Word Error Rate (WER) (Wang et al., 2018) for English and Character Error Rate (CER) (Xu et al., 2025) for Chinese to gauge intelligibility using SOTA multilingual ASR model `Fun-ASR-nano-2512` (An et al., 2025). For the timbre, we calculate cosine similarity metrics based on `ECAPA-TDNN` (Desplanques et al., 2020) to obtain speaker identity similarity (SpkSim). Additionally, we evaluate Voiced Time Ratio (VTR) for pitch-free assessment.

5.2 Evaluating on Benchmarks

Based on the objective evaluation results in Table 3, several key observations can be made regarding the quality of whispered corpora. Among the publicly available English datasets, the EARs subset demonstrates superior performance, achieving the highest naturalness scores (DNSMOS: 3.43, UTMOS: 2.01) and the lowest CER: 6.31%, indicating its high-quality whisper characteristics. For Chinese datasets, our proposed WhispXXX corpus shows a favorable balance, attaining a higher DNSMOS score (2.90) compared to other Chinese whisper collections like Whisper40 (2.61) and AISHELL6-Whisper (2.75). The primary advantages of our newly introduced datasets, WhispReal and WhispSynth, are threefold. First, they provide multilingual coverage for both English and Chinese, addressing a gap in existing resources that are predominantly monolingual. Second, WhispSynth exhibits exceptional intelligibility, achieving the lowest error rates (CER: 31.16%, WER: 20.98%) among all evaluated datasets, which is beneficial for speech recognition and synthesis tasks requiring high clarity. Third, it maintains a balanced naturalness, with competitive DNSMOS (2.89) and UTMOS (1.46) scores while demonstrating a reasonable pitch profile (VTR: 0.87) comparable to high-quality datasets like Expresso.

Overall, WhispSynth presents an optimal trade-off between intelligibility and naturalness, making it a suitable resource for whisper-based speech processing applications. In addition, the results also reveal that existing objective metrics are poorly aligned with whisper characteristics; ultimately, human judgment remains indispensable.

5.3 Experiments for generated samples

5.3.1 Comparative Systems

Whisper-Effect (Roh et al., 2025) is a designed acoustic perturbation that simulates whisper by applying a high-pass filter combined with white noise.
toWhisper (zeta, 2017) is an Linear Predictive Coding (LPC)-based tool that synthesizes whispered speech by processing white noise through a vocal tract filter estimated via LPC vocoding.
Normal2Whisper (Lin et al., 2023) converts normal speech to whisper through a two-stage process (Glottal source removal and Spectral modification) by the WORLD vocoder (Morise et al., 2016).
SeedVC (Liu, 2024) is a zero-shot voice conversion framework that disentangles representations of content and speaker information.
CosyVoice3 (Du et al., 2025) is a multilingual TTS model for high-quality and efficient generation.

5.4 Experiments for CosyWhisper

5.4.1 Quantitative Evaluation

The objective results in Table 4 highlight two key findings. First, in *Normal-to-Whisper Conversion*, voice conversion models like SeedVC yield high speaker similarity and intelligibility, but their whisperness remains limited (W-MOS: 2.18). Although traditional baselines (e.g., toWhisper) are non-trainable, lack adaptability, and yield lower W-MOS despite decent objective scores. Our pitch-

Method	Source	Whisperiness	Naturalness			Intelligibility	Timbre	Pitch
		W-MOS ↑	DNSMOS ↑	UTMOS ↑	MCD ↓	CER/WER (%) ↓	SpkSim ↑	VTR (%) ↓
Ground Truth	Test Set of WhispReal	4.33±0.33	2.80	1.44	0.00	39.30/37.58	1.00	0.88
Normal-to-Whisper Conversion								
Whisper-Effect	(Roh et al., 2025)	-	3.14	1.30	70.23	59.46/80.35	0.52	0.12
toWhisper	(zeta, 2017)	1.02±0.29	2.94	1.35	55.73	12.07/ 9.27	0.58	0.92
Normal2Whisper	(Lin et al., 2023)	-	2.89	1.28	60.00	11.92/28.28	0.59	0.94
SeedVC	(Liu, 2024)	2.18±0.47	3.01	1.61	54.85	58.65/19.97	0.66	0.71
Pitch-free Model	Ours	1.30±0.29	2.90	1.36	63.05	21.31/45.18	0.63	0.98
Text-to-Whisper Synthesis								
CosyVoice3	(Du et al., 2025)	3.40±0.51	3.00	1.47	56.03	12.51/9.81	0.83	0.86
CosyWhisper	Ours	**4.53±0.20**	3.08	1.48	59.31	12.76/29.22	0.80	0.88

Table 4: Evaluation on whisper speech conversion and synthesis. **Quality**: Objective whisperiness (W-MOS ↑), Subjective naturalness (DNSMOS/UTMOS ↑) and objective distortion (MCD ↓). **Intelligibility**: Character/Word Error Rate (↓). **Timbre**: Speaker similarity cosine similarity (↑). **Pitch**: Voiced Time Ratio (VTR) (↓).

free model is trainable and better captures natural whisper variability, confirming that not fixed transformations for authentic whisper synthesis. Second, in *Text-to-Whisper Synthesis*, even strong TTS systems like CosyVoice3 produce speech that is intelligible and natural, yet still perceptibly non-whisper (W-MOS: 3.40). In contrast, our CosyWhisper achieves a W-MOS of 4.53—the highest among all synthesized approaches and even better than the ground-truth (4.33), demonstrating our superiority to generate authentic whisper from text.

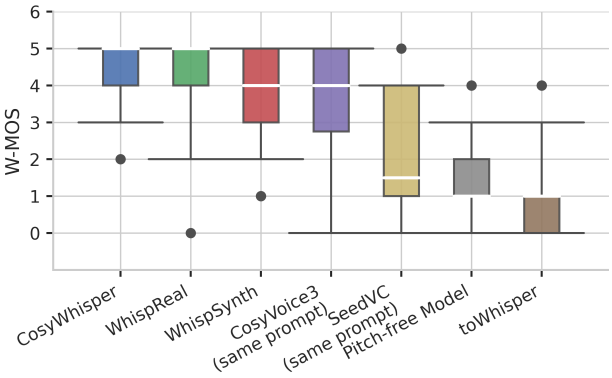


Figure 3: Results of the listening test. Whisper-likeness Mean Opinion Score (W-MOS) based on 20 participants visualized in a standard box plot.

For subjective evaluation, the W-MOS box plot from the listening test statistically validates that listeners perceive the output of CosyWhisper as similar to natural whisper. The interquartile range (IQR) for CosyWhisper is expected to be compact and positioned high on the scale (near 4.0), with a median close to 4.1, indicating consistent high-quality ratings from participants with low disagreement. In contrast, the boxes for other methods (e.g., CosyVoice3, SeedVC) would be located lower, with potentially larger IQRs, reflecting greater variability in listener perception and lower consensus on their whisper authenticity. CosyWhisper successfully bridges the gap between high intelligibility and authentic whisper perceptual quality.

5.5 Ablation on Tuning Dataset

To evaluate the efficacy of our synthetic data for text-to-whisper synthesis, we fine-tune our CosyWhisper on two variants of the corpus: WhispReal and our synthetic WhispSynth. As shown in Table 5, WhispSynth consistently outperforms across all key metrics: it reduces CER/WER by 46%, lowers pitch distortion (VTR) by 9%, and improves naturalness by 8%, despite being fully synthetic. This confirms that WhispSynth not only mitigates the noise and inconsistency inherent in real whisper data but also provides high-fidelity data that better supports stable whisper generation from text.

Dataset	Naturalness ↑		Intelligibility ↓	Timbre ↑	F0 ↓
	DNSMOS	UTMOS	CER / WER (%)	Cosine	VTR(%)
WhispReal	2.93	1.33	28.3/46.5	0.75	0.77
WhispSynth	**3.08**	**1.48**	**12.8/29.2**	**0.80**	**0.70**

Table 5: Performance comparison of our CosyWhisper fine-tuned on WhispReal versus WhispSynth.

6 Conclusion

In this paper, we address the fundamental data scarcity challenge in whisper research by introducing WhispSynth, a large-scale, high-fidelity multilingual whisper corpus. Our core contribution is a novel generative framework that integrates a DDSP-based pitch-free method with advanced TTS models, transforming diverse and noisy real whispered recordings into a clean, studio-grade synthetic dataset. This pipeline ensures the faithful preservation of vocal timbre and linguistic content while significantly enhancing acoustic quality.

Limitations

The impact of non-linguistic variations (e.g., the use of different microphones) on model performance was not systematically assessed. Although existing research suggests that CosyVoice3 are relatively robust to such variations, the influence of hardware differences in real-world deployment scenarios on specific task performance requires further verification. Besides, to address security concerns, the CosyWhisper model used in this study will be released with an automatically embedded real-time audio watermark. While this measure is crucial for responsible usage tracking, it may have a potential impact on the acoustic properties of the synthesized audio and subsequent analyses.

Ethics Statement

Our paper evaluated various methods that could make developing text-to-whisper synthesize systems more viable for languages where paired whisper and transcriptions are difficult to obtain. In our experiments, we only used already publicly available data (CHAINs, EARs, Expresso, Whisper40) or data for which we have obtained informed consent for public release from the data custodians (AISHELL6-Whisper, wTIMIT). To make our findings as relevant as possible for other language projects, we minimized the amount of computing time used.

References

Keyu An, Yanni Chen, Chong Deng, Changfeng Gao, Zhifu Gao, Bo Gong, Xiangang Li, Yabin Li, Xiang Lv, Yunjie Ji, and 1 others. 2025. Fun-asr technical report. *arXiv preprint arXiv:2509.12508*.

Joceline Andersen. 2015. Now you've got the shiveries: Affect, intimacy, and the asmr whisper community. *Television & New Media*, 16(8):683–700.

ByteDance. 2025. Doubao ai chatbot. https://www.doubao.com/chat. Accessed: [2025-07-01].

Qi Chen, Mingkui Tan, Yuankai Qi, Jiaqiu Zhou, Yuanqing Li, and Qi Wu. 2022. V2C: Visual voice cloning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21210–21219.

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.

Fred Cummins, Marco Grimaldi, Thomas Leonard, and Juraj Simko. 2006. The chains corpus: Characterizing individual speakers. In *Proc of SPECOM*, pages 1–6.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Annu. Conf. Int. Speech Commun. Assoc.*, pages 3830–3834.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, and 1 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

ElevenLabs. 2025. Elevenlabs: Free text to speech & ai voice generator. https://elevenlabs.io. Accessed: [2025-07-01].

Jesse Engel, Chenjie Gu, Adam Roberts, and 1 others. 2020. Ddsp: Differentiable digital signal processing. In *International Conference on Learning Representations (ICLR)*.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon Technical Report N*, 93:27403.

Shabnam Ghaffarzadegan, Hynek Bořil, and John HL Hansen. 2014. Ut-vocal effort ii: Analysis and constrained-lexicon recognition of whispered speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2544–2548. IEEE.

Đorđe T Grozdić, Branko Marković, Jovan Galić, and Slobodan T Jovičić. 2012. Application of neural networks in whispered speech recognition. In *2012 20th Telecommunications Forum (TELFOR)*, pages 728–731. IEEE.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and 1 others. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Hirotaka Hiraki and Jun Rekimoto. 2022. Silentwhisper: Faint whisper speech detection using wearable microphones. In *Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–3.

Hirotaka Hiraki and Jun Rekimoto. 2025. Silentwhisper: inaudible faint whisper speech input for silent speech interaction. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Taisuke Ito, Kazuya Takeda, and Fumitada Itakura. 2005. Analysis and recognition of whispered speech. *Speech Communication*.

Szu-Chen Jou, Tanja Schultz, and Alex Waibel. 2005. Whispery speech recognition using adapted articulatory features. In *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, volume 1, pages 1009–1012. IEEE.

N Kawaguchi, K Takeda, S Matsubara, I Yokoo, T Ito, K Tatara, T Shinde, and F Itakura. 2002. Ciair speech corpus for real world applications. In *The International conference Committee for the Coordination and Standardization of Speech Databases and Assesment Techniques*.

Pei Xuan Lee, Darren Wee, Hilary Si Yin Toh, Boon Pang Lim, Nancy F Chen, and Bin Ma. 2014a. A whispered mandarin corpus for speech technology applications. In *INTERSPEECH*, pages 1598–1602.

Pei Xuan Lee, Darren Wee, Hilary Si Yin Toh, Boon Pang Lim, Nancy F Chen, and Bin Ma. 2014b. A whispered mandarin corpus for speech technology applications. In *Proc. Interspeech 2014*, pages 1598–1602.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2022. Bigvgan: A universal neural vocoder with large-scale training. *International Conference on Learning Representations (ICLR)*.

Cancan Li, Fei Su, Juan Liu, Hui Bu, Yulong Wan, Hongbin Suo, and Ming Li. 2025. Aishell6-whisper: A chinese mandarin audio-visual whisper speech dataset with speech recognition baselines. *arXiv preprint arXiv:2509.23833*.

Yang Li-Li, Li Yan, and Xu Bo-Ling. 2005. The establishment of a chinese whisper database and perceptual experiment. *Journal of Nanjing University (Natural Sciences Edition)*, 41(3):311.

Boon Pang Lim. 2011. *Computational Differences Between Whispered and Non-whispered Speech*. University of Illinois at Urbana-Champaign.

Zhaofeng Lin, Tanvina Patel, and Odette Scharenborg. 2023. Improving whispered speech recognition performance using pseudo-whisper based data augmentation. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Songting Liu. 2024. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802.

Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. 2016. World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, 99(7):1877–1884.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarandi, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, and 1 others. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. In *Proc. Interspeech 2023*, pages 4823–4827.

Stavros Petridis, Jie Shen, Doruk Cetin, and Maja Pantic. 2018. Visual-only recognition of normal, whispered and silent speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6219–6223. IEEE.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. 2021. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *IEEE Conf. Acoust. Speech Signal Process.*, pages 6493–6497.

Jun Rekimoto. 2023. Wesper: Zero-shot and real-time whisper to normal voice conversion for whisper-based speech interactions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–12.

Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann. 2024. Ears: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *Proc. Interspeech 2024*, pages 4873–4877.

Jaechul Roh, Virat Shejwalkar, and Amir Houmansadr. 2025. Multilingual and multi-accent jailbreaking of audio llms. *arXiv preprint arXiv:2504.01094*.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Xavier Serra and Julius Smith. 1990. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24.

Bhavuk Singhal, Abinay Reddy Naini, and Prasanta Kumar Ghosh. 2021. wspire: A parallel multi-device corpus in neutral and whispered speech. In *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 146–151. IEEE.

VC Tartter. 1989. What's in a whisper? *The Journal of the Acoustical Society of America*, 86(5):1678–1683.

Tam Tran, Soroosh Mariooryad, and Carlos Busso. 2013. Audiovisual corpus to analyze whisper speech. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8101–8105. IEEE.

Dong Wang and Xuewei Zhang. 2015. Thchs-30: A free chinese speech corpus. *arXiv preprint arXiv:1512.01882*.

Xin Wang, Shinji Takaki, and Junichi Yamagishi. 2019. Neural source-filter waveform models for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:402–415.

Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Int. Conf. on Mach. Learn.*, volume 80, pages 5167–5176.

Da-Yi Wu, Wen-Yi Hsiao, Fu-Rong Yang, Oscar Friedman, Warren Jackson, Scott Bruzenak, Yi-Wen Liu, and Yi-Hsuan Yang. 2022. Ddsp-based singing vocoders: A new subtractive-based synthesizer and comprehensive evaluation. In *International Society for Music Information Retrieval Conference (ISMIR)*.

Kai-Tuo Xu, Feng-Long Xie, Xu Tang, and Yao Hu. 2025. Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration. *arXiv preprint arXiv:2501.14350*.

Jingwen Yang and Ruohua Zhou. 2024. Whisper40: A multi-person chinese whisper speaker recognition dataset containing same-text neutral speech. *Information*, 15(4):184.

zeta. 2017. towhisper. https://github.com/zeta-chicken/toWhisper. Accessed: [2025-07-01].

Chi Zhang and John HL Hansen. 2010. Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):883–894.

Jian Zhou, Yuting Hu, Hailun Lian, Cong Pang, Huabin Wang, and Liang Tao. 2019. An audio-visual whisper database in chinese. In *Journal of Physics: Conference Series*, volume 1237, page 022106. IOP Publishing.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*.

Appendix

A Algorithm

Algorithm 1 outlines the proposed procedure.

Algorithm 1: Whisper Synthesis with Pitch-Aware Segment Replacement

Input: Prompt audio $\mathbf{A}_p$, Corresponding text T_p, Instruction I
Output: Synthesized whisper audio $\mathbf{A}_w$ Initialize $\mathbf{A}_w \leftarrow \emptyset$ `// Final whisper audio`
$\mathbf{A}_w^{\text{init}} \leftarrow \text{CosyVoice3}(\mathbf{A}_p, T_p, I)$
$\mathcal{P} \leftarrow \text{DDSP\_PitchDetection}(\mathbf{A}_w^{\text{init}})$
if $\mathcal{P} \neq \emptyset$ **then**
  foreach *pitch segment* $\mathbf{S}_i \in \mathcal{P}$ **do**
    $(\mathbf{H}_i, \mathbf{N}_i) \leftarrow \text{DDSP\_Decompose}(\mathbf{S}_i)$
    `//` $\mathbf{H}_i$`: harmonic segment,` $\mathbf{N}_i$`: noise segment`
    $\mathbf{S}_i^{\text{noise}} \leftarrow \mathbf{N}_i$
    $\mathbf{S}_i^{\text{replaced}} \leftarrow \text{ApplyWindow}(\mathbf{S}_i^{\text{noise}})$
    $\mathbf{A}_w^{\text{init}} \leftarrow \text{ReplaceSegment}(\mathbf{A}_w^{\text{init}}, \mathbf{S}_i, \mathbf{S}_i^{\text{replaced}})$
  $\mathbf{A}_w \leftarrow \text{OverlapAdd}(\mathbf{A}_w^{\text{init}})$
else
  $\mathbf{A}_w \leftarrow \mathbf{A}_w^{\text{init}}$ `// No pitch detected, use original`
return $\mathbf{A}_w$

First, an initial whisper is generated using CosyVoice3, conditioned on a prompt audio, its transcript; note that a instruction is optional. The pitch detection module in DDSP then identifies segments containing residual F0. For each such segment, DDSP decomposition separates the harmonic and noise components. The harmonic part is dropped, while the noise component, which carries the whisper's spectrnal envelope and aperiodic excitation, is retained. Finally, an OLA operation is applied to reconstruct a continuous, pitch-free whisper waveform.

B Code

Below are two versions of the code: the original implementation (Original) and the revised version (Modified). After implementing these modifications in the file[3], we were able to successfully fine-tune the Flow model within CosyVoice3.

```python
def forward(
        self,
        batch: dict,
        device: torch.device,
) -> Dict[str, Optional[torch.Tensor]]:

    # Data extraction
```

[3] https://github.com/FunAudioLLM/CosyVoice/blob/652132ebaa3133428bf4db7e092a3cd1e073ca80/examples/libritts/cosyvoice3/cosyvoice/flow/flow.py

```
    token = batch['speech_token'].to(device)
    token_len = batch['speech_token_len'].to(device)
    feat = batch['speech_feat'].to(device)
    feat_len = batch['speech_feat_len'].to(device)
    embedding = batch['embedding'].to(device)

    # NOTE unified training, static_chunk_size > 0 or = 0
    streaming = True if random.random() < 0.5 else False

    # xvec projection
    embedding = F.normalize(embedding, dim=1)
    embedding = self.spk_embed_affine_layer(embedding)

    # concat text and prompt_text
    mask = (
        (~make_pad_mask(token_len))
        .float()
        .unsqueeze(-1)
        .to(device)
    )

    token = (
        self.input_embedding(torch.clamp(token, min=0))
        * mask
    )

    # text encode
    h, h_lengths = self.encoder(
        token,
        token_len,
        streaming=streaming
    )
    h = self.encoder_proj(h)

    # get conditions
    conds = torch.zeros(feat.shape, device=token.device)
    for i, j in enumerate(feat_len):
        if random.random() < 0.5:
            continue
        index = random.randint(0, int(0.3 * j))
        conds[i, :index] = feat[i, :index]
    conds = conds.transpose(1, 2)

    mask = (
        ~make_pad_mask(
            h_lengths.sum(dim=-1).squeeze(dim=1)
        )
        .to(h)
    )
    loss, _ = self.decoder.compute_loss(
        feat.transpose(1, 2).contiguous(),
        mask.unsqueeze(1),
        h.transpose(1, 2).contiguous(),
        embedding,
        cond=conds,
        streaming=streaming,
    )
    return {'loss': loss}
```

Listing 1: Original

```
self.token_proj = nn.Linear(input_size, input_size)

def forward(
    self,
    batch: dict,
    device: torch.device,
) -> Dict[str, Optional[torch.Tensor]]:

    # 1. Data extraction
    token = batch['speech_token'].to(device)
    token_len = batch['speech_token_len'].to(device)
    feat = batch['speech_feat'].to(device)
    feat_len = batch['speech_feat_len'].to(device)
    embedding = batch['embedding'].to(device)


    streaming = random.random() < 0.5

    # 2. Speaker embedding processing
    embedding = F.normalize(embedding, dim=-1)
    embedding = self.spk_embed_affine_layer(embedding)

    # 3. Token to embedding mapping
    # Ensure token shape is [B, T]
    if token.dim() == 3:
        token = token.squeeze(-1)

    token = token.long()

    # Token mask: [B, T, 1]
    token_mask = (
        (~make_pad_mask(token_len))
        .unsqueeze(-1)
        .to(device)
    )

    # Embedding lookup
    h = self.input_embedding(token) * token_mask
    h_lengths = token_len

    # 4. Token-to-mel alignment
    h = h.repeat_interleave(self.token_mel_ratio, dim=1)
    token_mask = token_mask.repeat_interleave(
        self.token_mel_ratio,
        dim=1
    )
    h_lengths = h_lengths * self.token_mel_ratio

    # 5. Conditional mel (cond) preparation
    conds = torch.zeros_like(feat)
    for i, j in enumerate(feat_len):
        if random.random() < 0.5:
            continue
        index = random.randint(0, int(0.3 * j))
        conds[i, index] = feat[i, index]

    conds = conds.transpose(1, 2)  # [B, mel_dim, T_mel]

    # 6. Flow mask preparation
    # Flow expects shape [B, T]
    flow_mask = (~make_pad_mask(h_lengths)).to(device)

    # 7. Flow loss computation
    loss, _ = self.decoder.compute_loss(
        feat.transpose(1, 2).contiguous(),
        flow_mask.unsqueeze(1),
        h.transpose(1, 2).contiguous(),
        embedding,
        cond=conds,
        streaming=streaming,
    )

    return {'loss': loss}
```

Listing 2: Modified

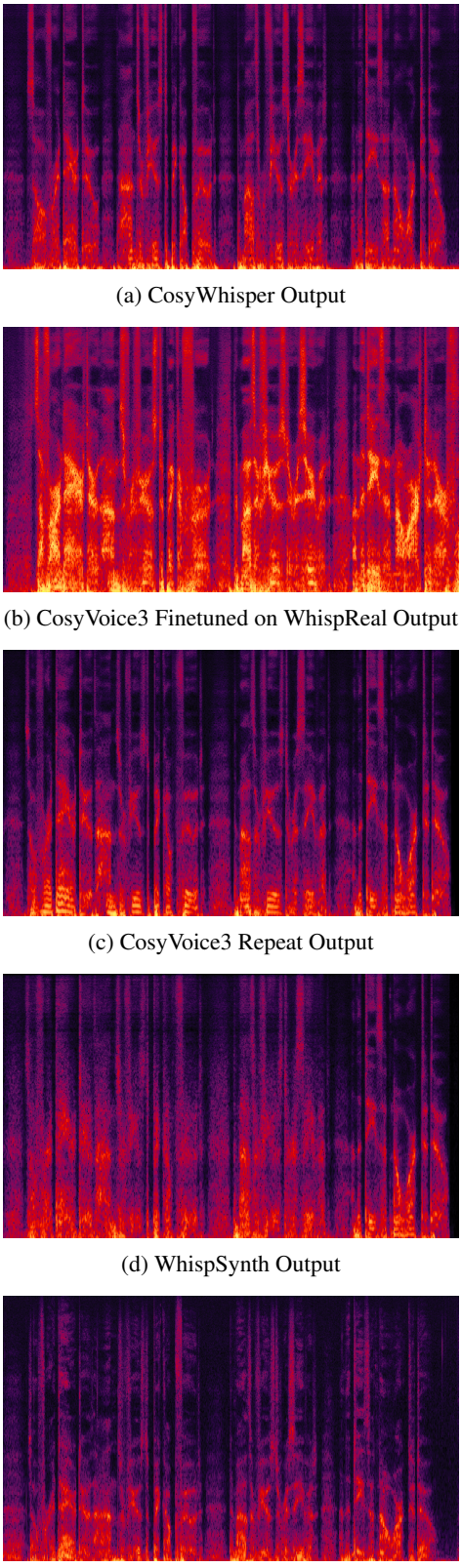
C Subjective Evaluation

We developed a web-based automated subjective evaluation interface (Figure 3) to collect real-time listening ratings. Twenty audio stimuli (four per proposed method) were pre-loaded into four Latin-square sequences to minimise order and carry-over effects; the sequence was automatically selected according to the participant ID entered at the start of the session. Each trial started with the automatic playback of one stimulus (24 kHz, 16-bit WAV; approximately 12 s). Immediately after playback, a five-point Likert scale (0 = very bad, 5 = excellent) appeared on the screen and the participant had 5 s to tap the desired score on a touch device or click with a mouse. Responses were timestamped and written to a CSV file that was automatically downloaded at the end of the session.

D Case Study

Sample spectrograms are provided above; additional examples and the full demo are available in the supplementary material.

Figure 4: User interface of the listening test.



(a) CosyWhisper Output

(b) CosyVoice3 Finetuned on WhispReal Output

(c) CosyVoice3 Repeat Output

(d) WhispSynth Output

(e) WhispReal Output

Figure 5: Sample spectrograms. (We invite reviewers to see the supplementary demo for more examples.)