# WenetSpeech-Wu: Datasets, Benchmarks, and Models for a Unified Chinese Wu Dialect Speech Processing Ecosystem

**Anonymous ACL submission**

## Abstract

Speech processing for low-resource dialects remains a fundamental challenge in developing inclusive and robust speech technologies. Despite its linguistic significance and large speaker population, the Wu dialect of Chinese has long been hindered by the lack of large-scale speech data, standardized evaluation benchmarks, and publicly available models. In this work, we present **WenetSpeech-Wu**, the first large-scale, multi-dimensionally annotated open-source speech corpus for the Wu dialect, comprising approximately 8,000 hours of diverse speech data. Building upon this dataset, we introduce **WenetSpeech-Wu-Bench**, the first standardized and publicly accessible benchmark for systematic evaluation of Wu dialect speech processing, covering automatic speech recognition (ASR), Wu-to-Mandarin translation, speaker attribute prediction, speech emotion recognition, text-to-speech (TTS) synthesis, and instruction-following TTS (instruct TTS). Furthermore, we release a suite of strong open-source models trained on WenetSpeech-Wu, establishing competitive performance across multiple tasks and empirically validating the effectiveness of the proposed dataset. Together, these contributions lay the foundation for a comprehensive Wu dialect speech processing ecosystem, and we open-source proposed datasets, benchmarks, and models to support future research on dialectal speech intelligence [1].

## 1 Introduction

Speech processing has become a fundamental component of artificial intelligence, enabling natural and efficient human–machine interaction across a wide range of real-world applications (Peng et al., 2025; Xu et al., 2025a). For high-resource languages such as Mandarin Chinese and English, the speech processing ecosystem has reached a high level of maturity, characterized by large-scale and diverse datasets (Zhang et al., 2022; Zen et al., 2019; He et al., 2024), publicly available and multi-dimensional evaluation benchmarks (Sakshi et al., 2025; Wang et al., 2025a,b), and a growing number of powerful open-source models (An et al., 2024; Gao et al., 2022; Radford et al., 2023). Together, these components form a virtuous cycle in which data, benchmarks, and models mutually reinforce one another, continuously driving rapid progress in both academic research and practical deployment. In contrast, the Wu dialect [2], a crucial Chinese dialect that remains relatively underexplored in speech technology community, is associated with a severely underdeveloped speech processing ecosystem. The lack of sufficient datasets, standardized benchmarks, and accessible models has substantially hindered both research advancement and real-world adoption. To bridge this gap and promote inclusive speech technology, this work lays the foundation for a comprehensive speech processing ecosystem for the Wu dialect.


Figure 1: Highlights of WenetSpeech-Wu.

The Wu dialect is a major branch of Chinese spoken by approximately 100 million speakers across Shanghai, Zhejiang, Jiangsu, and overseas communities, and is linguistically characterized by exceptional complexity. It preserves the fully voiced consonant system inherited from Old Chinese. It features a highly intricate tone sandhi system in

[1] https://anonymous.4open.science/r/WenetSpeech-Wu

[2] https://en.wikipedia.org/wiki/Wu_Chinese

Table 1: Comparison of typical low-resource speech processing resources related to WenetSpeech-Wu. ✓, △, and ✗ indicate experimentally verified availability, unverified availability, and absence of dimension, respectively. Abbreviations: Data Comp (Data Composition; CN = China, SEA = Southeast Asia), Hrs (hours), Trans (transcription), Para (paralinguistics), Transl (translation), Q Tier (quality grading tier), Instr TTS (instruct TTS), Sp Und. (speech understanding), Wu-SH (Shanghai sub-dialect of Wu).

| | **Dataset** | | | | | | **Benchmark** | | | | | **Model** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Resource** | Data Comp | Hrs | Trans | Para | Transl | Q Tier | ASR | AST | Para | TTS | Instr TTS | ASR | TTS | Sp Und | Instr TTS |
| KeSpeech | 8 CN Accents | 1.5k | ✓ | △ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ |
| GigaSpeech2 | 3 SEA Langs | 30k | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ |
| WenetSpeech-Yue | CN Yue | 21k | ✓ | △ | ✗ | ✗ | ✓ | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ |
| WenetSpeech-Chuan | CN Sichuan | 10k | ✓ | △ | ✗ | ✗ | ✓ | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ |
| Magicdata-Shanghai | CN Wu-SH | 4 | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| **WenetSpeech-Wu** | CN Wu | 8k | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

connected speech (Rose, 2015), both of which pose substantial challenges for speech modeling. Moreover, substantial variation exists among its sub-dialects, such as Shanghainese, Suzhounese, and Hangzhounese, further complicating the development of robust speech processing systems.

Despite its linguistic significance and broad speaker base, the Wu dialect remains severely under-resourced for speech processing. From a data perspective, existing open-source resources are extremely limited in both scale and coverage: the only publicly available dataset, MagicData-Shanghai[3], provides merely 4.19 hours of annotated Shanghainese speech for automatic speech recognition (ASR), with no open datasets available for other Wu sub-dialects. Moreover, high-quality emotion annotations, speaker attributes and expressive speech–text pairs are entirely absent, which are essential for supporting a wide range of speech processing tasks. From an evaluation perspective, the lack of publicly available benchmarks prevents fair comparison and systematic assessment across different methods. At the model level, even foundational speech processing systems, such as ASR and text-to-speech synthesis (TTS) models, are either unavailable or perform poorly, making both open-source and commercial systems essentially unusable for Wu dialect applications. Collectively, these limitations underscore an urgent need to establish a comprehensive Wu dialect speech processing ecosystem, so as to facilitate dialectal speech research and foster the growth of the open-source community.

To address these challenges, we present a comprehensive open-source effort that lays the foundation for a Wu dialect speech processing ecosystem. Specifically, we construct **WenetSpeech-Wu**, the first large-scale, multi-dimensionally annotated open-source corpus for the Wu dialect, comprising approximately 8,000 hours of eight Wu sub-dialects. Building upon this dataset, we introduce **WenetSpeech-Wu-Bench**, the first standardized and publicly accessible benchmark designed for systematic evaluation of Wu dialect speech processing, covering ASR, Wu-to-Mandarin automatic speech translation (AST), speaker attribute prediction (gender and age), speech emotion recognition, TTS, and instruction-following TTS (instruct TTS). Furthermore, we release a suite of strong open-source models trained on WenetSpeech-Wu, including ASR, TTS, unified speech understanding, and instruct TTS models, which substantially outperform existing open-source and commercial systems, thereby establishing strong models and empirically demonstrating the effectiveness of the proposed dataset. Together, these contributions lay the foundation for a comprehensive, publicly accessible ecosystem for Wu dialect speech processing.

Our main contributions are threefold:

- We release **WenetSpeech-Wu**, the first large-scale and multi-dimensionally annotated open-source corpus for the Wu dialect, containing approximately 8,000 hours of speech data.
- We introduce **WenetSpeech-Wu-Bench**, the first standardized benchmark for evaluating Wu dialect speech processing, covering ASR, Wu-to-Mandarin AST, speaker attribute prediction, speech emotion recognition, TTS, and instruct TTS tasks.

[3] https://magichub.com/datasets/shanghai-dialect-conversational-speech-corpus/

- We open-source strong models trained on WenetSpeech-Wu, demonstrating substantial improvements over existing systems and validating the effectiveness of proposed dataset.

## 2 Related Work

### 2.1 Speech Datasets for Low-Resource Languages and Chinese Dialects

Recently, significant efforts have been devoted to constructing speech corpora for low-resource languages and dialects, aiming to mitigate data scarcity and promote inclusive speech technologies (Shao et al., 2025). Representative large-scale corpora such as Common Voice (Ardila et al., 2020), GigaSpeech2 (Yang et al., 2024), and KeSpeech (Tang et al., 2021) have substantially advanced research on speech processing for low-resource or underrepresented languages. Within the Chinese dialect family, several dialect-oriented speech resources have been released recently, including WenetSpeech-Yue (Li et al., 2025) for Cantonese and WenetSpeech-Chuan (Dai et al., 2025) for Sichuanese. However, these datasets primarily focus on ASR and, to a limited extent, TTS: Common Voice and GigaSpeech2 provide only transcriptions, KeSpeech augments transcriptions with dialect and basic speaker labels, and WenetSpeech-Yue and WenetSpeech-Chuan include paralinguistic annotations but evaluate their resources only through ASR and TTS experiments, without validating the effectiveness of the paralinguistic labels. As a result, the usability of such resources for broader speech processing tasks, such as speech emotion recognition, AST, and instruct TTS, remains limited.

Despite its linguistic and practical importance, the Wu dialect remains severely under-resourced in terms of publicly available speech data. To the best of our knowledge, MagicData-Shanghai is the only open-source Wu dialect dataset, providing merely 4.19 hours of annotated Shanghainese speech for ASR. Such extreme data scarcity, together with the lack of diverse annotations has substantially hindered progress in Wu dialect speech processing research.

### 2.2 Wu Dialect Speech Processing: Models and Evaluation

Early studies on Wu dialect speech processing mainly focused on ASR or TTS for individual sub-dialects, most notably Shanghainese, while largely ignoring other major variants such as Suzhounese and Hangzhounese. More recent works incorporate the Wu dialect as a minor component within large-scale multilingual ASR or TTS systems, including Dolphin (Meng et al., 2025), Qwen3-ASR, Qwen3-TTS, Step-Audio2 (Wu et al., 2025), and Qwen3-Omni (Xu et al., 2025a). In these systems, Wu dialect speech is treated as a low-priority subset rather than a primary research target, resulting in foundational ASR and TTS models whose performance remains insufficient for practical Wu dialect applications. Beyond ASR and TTS, speech understanding models and instructing TTS systems tailored to the Wu dialect remain largely unexplored.

More importantly, even for basic ASR and TTS tasks, there is currently no publicly available and standardized benchmark for evaluation. As a consequence, existing studies typically evaluate models on private test sets, severely hindering fair comparison, reproducibility, and systematic progress.

In summary, as illustrated in Tabel 1, these limitations motivate the introduction of WenetSpeech-Wu, which aims to establish a complete Wu dialect speech processing ecosystem by jointly advancing datasets, benchmarks, and models.

## 3 WenetSpeech-Wu

### 3.1 Data Construction Pipeline

We propose an automatic and scalable pipeline for constructing a large-scale Wu dialect speech dataset with multi-dimensional annotations, as illustrated in Figure 2. The pipeline is designed to enable efficient data collection, robust automatic transcription, and diverse downstream annotations.

**Data Collection and Filtering.** We collect large-scale in-the-wild Wu dialect speech from diverse domains and sub-dialects. Non-Wu data are first removed based on metadata filtering, followed by WebRTC VAD-based segmentation. We further apply quality filtering using DNSMOS and signal-to-noise ratio (SNR), resulting in a high-quality speech corpus.

**Annotation Tool Construction.** To support large-scale automatic transcription, we fine-tune two pretrained ASR models using 880 hours of manually annotated Wu dialect speech. Specifically, we fine-tune Tele-CTC-FT[4], a Connectionist Temporal Classification (CTC)-based dialect self-supervised learning (SSL) model, and Step-

[4] https://huggingface.co/Tele-AI/TeleSpeech-ASR1.0

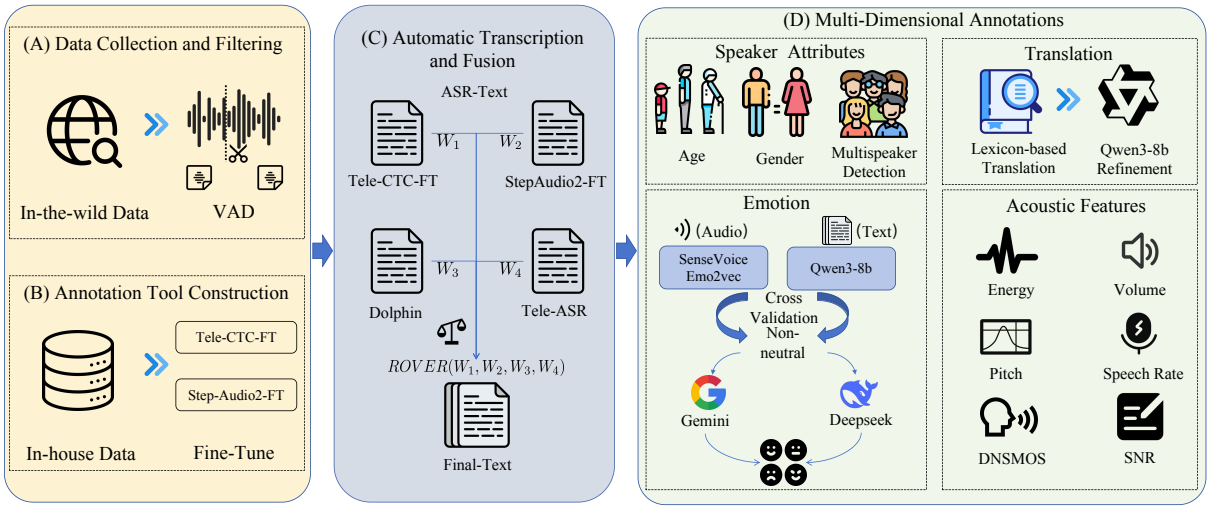


Figure 2: Data construction pipeline for WenetSpeech-Wu.

Audio2-FT (Wu et al., 2025). These models are used as complementary automatic annotators for Wu dialect transcription.

**Automatic Transcription and Fusion.** We adopt Recognizer Output Voting Error Reduction (ROVER) (Fiscus, 1997) to fuse transcription from multiple ASR systems. Specifically, we combine the outputs from the two fine-tuned Wu dialect ASR models, Dolphin (Meng et al., 2025) and TeleASR (Chen et al., 2024), and determine the model weights via grid search. The fused results provide final transcriptions along with confidence scores.

**Multi-Dimensional Annotations.** Each speech sample is annotated across multiple dimensions. Speaker attributes, including gender and age, are inferred using the VoxProfile (Feng et al., 2025), while multi-speaker presence is detected using Pyannote (Bredin, 2023). Wu-to-Mandarin translations are generated through lexicon-based mapping and further refined using Qwen3-8B (Yang et al., 2025), a large language model, aiming to provide fluent standard Mandarin references. Emotion annotations are obtained through a multi-stage, cross-modal procedure. Initial predictions are produced using SenseVoice (An et al., 2024) and Emo2Vec (Wang et al., 2020) for acoustic signals, and Qwen3-8B for textual content. Samples jointly identified as non-neutral are further analyzed using DeepSeek-R1 based on text and Gemini-2.5-Pro based on acoustic information, with the final label determined by the intersection of the two. In addition, prosodic acoustic features, including speech rate, loudness, energy, and pitch, are extracted by Dataspeech (Lyth and King, 2024) to support speech generation tasks.

### 3.2 Datasets

WenetSpeech-Wu is the first large-scale Wu dialect speech corpus with multidimensional annotations. It contains rich metadata and annotations, including transcriptions with confidence scores, Wu-to-Mandarin translations, domain and sub-dialect labels, speaker attributes, emotion annotations, and audio quality measures. The dataset comprises approximately 8,000 hours of speech collected from diverse domains and covers eight Wu sub-dialects. To support a wide range of speech processing tasks with heterogeneous quality requirements, we further adopt a task-specific data quality grading strategy. In the following, we describe the dataset statistics, domain and dialect coverage, annotation distributions, and quality control mechanisms in detail.

**Duration and Confidence Distribution.** WenetSpeech-Wu contains 8,000 hours of speech with 3.86M utterances, with utterance durations up to 30 seconds and an average duration of 7.45 seconds. We use the transcription confidence produced by the weighted ROVER as a measure of annotation quality, and retain utterances with confidence scores above 0.55. The detailed distributions of utterance duration and transcription confidence are shown in Figure 3b and Figure 3c.

**Domain and Sub-Dialect Coverage.** WenetSpeech-Wu covers a wide range of speech domains and Wu sub-dialects. The domains include *News, Culture, Vlog, Entertainment, Education, Podcast, Commentary, Interview, Radio*

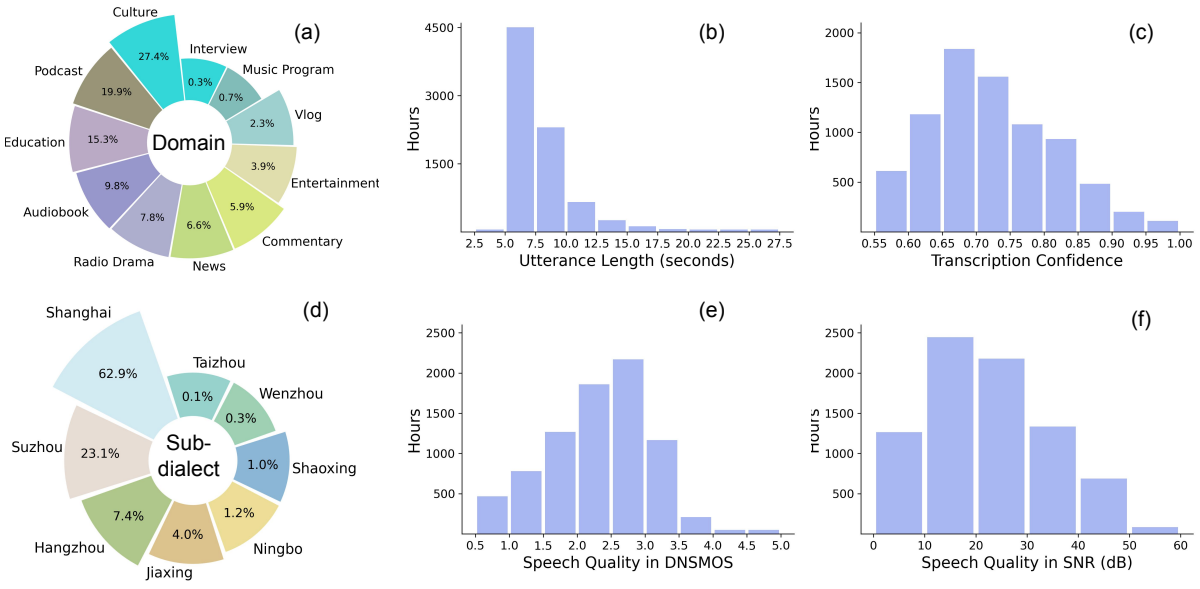


Figure 3: Statistical overview of WenetSpeech-Wu.

Table 2: Statistics of speaker attributes and emotion annotations in WenetSpeech-Wu.

| Task | Category | Hours |
|---|---|---|
| Gender | Male | 4,135 |
| | Female | 1,331 |
| Age | Teenagers | 372 |
| | Youth | 1,673 |
| | Middle-aged | 2,003 |
| | Elderly | 1,418 |
| Emotion | Neutral | 5,102 |
| | Happy | 73 |
| | Sad | 81 |
| | Angry | 109 |
| | Surprised | 101 |

*Drama, Music Program, and Audiobook*, with their distribution shown in Figure 3a. In terms of dialectal coverage, approximately 37% of the recordings cannot be reliably assigned to a specific Wu sub-dialect and are therefore labeled as Unknown. The remaining recordings span multiple identified Wu sub-dialects, including *Shanghainese, Suzhounese, Shaoxingnese, Ningbonese, Hangzhounese, Jiaxingnese, Taizhounese, and Wenzhounese*, whose distribution is illustrated in Figure 3d.

**Audio Quality.** As shown in Figure 3e and Figure 3f, most utterances have SNRs between 10 and 40 dB, with a peak at 20–30 dB. MOS scores are primarily in the range 2.0–3.5.

**Speaker Attributes and Emotion Annotations.** We annotate gender, age, and emotion labels for single-speaker segments. Gender is categorized into *Male* and *Female*, age into four groups including 0–17 as *Teenagers*, 18–35 as *Youth*, 36–59 as *Middle-aged*, and 60+ as *Elderly*, and emotion into five classes: *Neutral, Happy, Sad, Surprised, and Angry*. The distribution of each category is shown in Table 2.

**Task-Specific Data Quality Grading.** To support practical training across heterogeneous speech tasks, we adopt a task-specific data-quality grading strategy aligned with the task-specific quality requirements. For ASR and TTS, we construct two quality tiers. The normal-quality subset is designed for large-scale pretraining and prioritizes data coverage and diversity, requiring only moderate transcription confidence. In contrast, the high-quality subset targets supervised fine-tuning (SFT) and applies stricter filtering criteria, including high transcription confidence, clean acoustic conditions, and reliable speaker segmentation, to provide strong and stable supervision. For tasks that are more sensitive to annotation noise and semantic ambiguity, including Wu-to-Mandarin AST, speaker attribute prediction, speech emotion recognition, TTS, and instruct TTS, we adopt stringent selection criteria, such as single-speaker recordings, high MOS, SNR, and pitch standard deviation, and verified annotation consistency, as shown in Table 3.

## 4 WenetSpeech-Wu-Bench

We introduce **WenetSpeech-Wu-Bench**, the first publicly available, manually curated benchmark

Table 3: Task-specific data selection and quality tiers in WenetSpeech-Wu. Abbreviations: Q Tier (quality grading tier), Text Conf (transcription confidence), Spk (speaker type, Mono = single-speaker), Pitch Std. (pitch standard deviation), Expr (expressive data obtained via emotion label), Spk Attr (speaker attributes), Emo (emotion), Inst Pro (instruct TTS prosodic control), Inst Emo (instruct TTS emotional control).

| Task | Hrs | Q Tier | Text Conf | Spk | SNR | Pitch Std. | Expr |
|---|---|---|---|---|---|---|---|
| ASR | 7388 | Mid | $>0.6$ | – | – | – | – |
| | 795 | High | $>0.85$ | – | – | – | – |
| AST | 795 | High | $>0.85$ | – | – | – | – |
| Spk Attr | 2986 | High | – | Mono | – | – | – |
| Emo | 500 | High | – | Mono | >10 | >50 | ✓ |
| TTS | 7388 | Mid | $>0.6$ | – | – | – | – |
| | 1500 | High | $>0.65$ | Mono | >10 | >50 | – |
| Inst Pro | 679 | High | $>0.7$ | Mono | >30 | – | – |
| Inst Emo | 161 | High | $>0.7$ | Mono | >10 | >50 | ✓ |

for Wu dialect speech processing, covering ASR, Wu-to-Mandarin AST, speaker attributes, emotion recognition, TTS, and instruct TTS, and providing a unified platform for fair evaluation.

**Automatic Speech Recognition.** The ASR test set of WenetSpeech-Wu-Bench comprises 9.75 hours of Shanghainese, Suzhounese, and Mandarin code-mixed speech, including single-speaker and multi-speaker scenarios. Performance is evaluated using character error rate (CER), with character-level errors reported for detailed analysis.

**Wu-to-Mandarin Speech Translation.** We construct a Wu-to-Mandarin AST test set. The test set contains 3000 Wu dialect utterances totaling 4.4 hours with manually verified standard Mandarin translations, covering multiple domains. Translation quality is evaluated using the BLEU score [5].

**Speaker Attribute Prediction and Speech Emotion Recognition.** This test set evaluates the prediction of age, gender, and emotion for Wu dialect speech. For speaker attributes, gender is coded as male or female, with 1,500 samples per class. And age into four groups: teenagers aged 17 and under, youth aged 18 to 35, middle-aged adults aged 36 to 59, and elderly aged 60 and above, with 500 samples in each group. For emotion, there are 300 neutral samples, 200 happy samples, 100 sad samples, 200 angry samples, and 200 surprised samples, for a total of 1,000 samples. Performance is measured by category-wise and overall classification accuracy.

**Text to Speech.** We constructed a TTS testset as part of WenetSpeech-Wu-Bench, comprising 144 easy and 98 hard test sentences. The texts were further reviewed and refined by professional Wu dialect experts. Prompt audio samples were selected from the open-source Magicdata-Shanghai, and 12 speakers of the Wu dialect were selected using strict filtering criteria. For evaluation, speaker similarity is measured using WeSpeaker [6] based speaker embedding similarity, while intelligibility is assessed by computing CER with our proposed Step-Audio2-Wu-ASR model. Additionally, subjective listening tests are conducted, including intelligibility MOS (IMOS), similarity MOS (SMOS), and accent MOS (AMOS). The subjective evaluation was conducted with 23 listeners, each rating 20 selected samples.

**Instruct TTS.** WenetSpeech-Wu-Bench includes two evaluation sets for instruct TTS. The prosodic control test set consists of five speech prompts, each spoken at a moderate speaking rate and normal fundamental frequency, and synthesized into 20 sentences with controlled variations in speaking rate and pitch. For evaluation, two experiments are conducted: one with a fast speaking rate and high pitch, and another with a slow rate and low pitch. These samples are automatically annotated using Dataspeech (Lyth and King, 2024). Each pair is scored as one if the relative speaking rate and pitch match the intended instructions, and zero otherwise. The prosodic classification metric is the average score across all pairs, reflecting the model's ability to follow prosodic instructions (Kuan et al., 2023). The emotional control test set evaluates a model's ability to follow emotion-related instructions. We select 10 reference prompts that contain no explicit emotional expression. Based on each prompt, 50 sentences are synthesized for each of the four target emotions: anger, sadness, happiness, and surprise. The samples are evaluated using our Step-Audio2-Wu-Und model. A sample is considered correct if the predicted emotion matches the intended target emotion, and the mean classification accuracy is reported as the evaluation metric (Gao et al., 2025). Additionally, subjective listening tests are conducted to assess the quality of instruction-

[5] https://github.com/mjpost/sacrebleu

[6] https://github.com/wenet-e2e/wespeaker

Table 4: ASR results (CER%) on various test sets. Gray , red , light green , and dark green rows denote open-source baselines, commercial models, ASR models trained on WenetSpeech-Wu, and annotation models trained on in-house data. **Bold** numbers indicate best results; underlined numbers indicate second-best results.

| Model | In-House | | WS-Wu-Bench |
|---|---|---|---|
| | Dialogue | Reading | ASR |
| **ASR Models** | | | |
| Paraformer | 63.13 | 66.85 | 64.92 |
| SenseVoice-small | 29.20 | 31.00 | 46.85 |
| Whisper-medium | 79.31 | 83.94 | 78.24 |
| FireRedASR-AED-L | 51.34 | 59.92 | 56.69 |
| Step-Audio2-mini | 24.27 | 24.01 | 26.72 |
| Qwen3-ASR | 23.96 | 24.13 | 29.31 |
| Tencent-Cloud-ASR | 23.25 | 25.26 | 29.48 |
| Gemini-2.5-pro | 85.50 | 84.67 | 89.99 |
| Conformer-U2pp-Wu | 15.20 | 12.24 | 15.14 |
| Whisper-medium-Wu | 14.19 | 11.09 | 14.33 |
| Step-Audio2-Wu-ASR | 8.68 | 7.86 | **12.85** |
| **Annotation Models** | | | |
| Dolphin-small | 24.78 | 27.29 | 26.93 |
| TeleASR | 29.07 | 21.18 | 30.81 |
| Step-Audio2-FT | **8.02** | **6.14** | 15.64 |
| Tele-CTC-FT | 11.90 | 7.23 | 23.85 |

Table 5: Speech understanding performance on WenetSpeech-Wu-Bench.

| Model | ASR | AST | Gender | Age | Emotion |
|---|---|---|---|---|---|
| Qwen3-Omni | 44.27 | 33.31 | **0.977** | 0.541 | 0.667 |
| Step-Audio2-mini | 26.72 | 37.81 | 0.855 | 0.370 | 0.460 |
| Step-Audio2-Wu-Und | **13.23** | **53.13** | 0.956 | **0.729** | **0.712** |

following. Listeners rate the samples using two scores: the prosodic MOS (PMOS) (Chan and Kuang, 2025) and emotional Mean Opinion Score (EMOS) (Cho et al., 2025). The evaluation involves 23 listeners, each rating 15 samples, to provide perceptual judgments on how well the synthesized speech follows the intended emotional instructions.

## 5 Models & Experiments

To address the lack of Wu dialect speech processing models, we develop models for the two core aspects of speech processing, speech understanding and speech generation trained on WenetSpeech-Wu. These include ASR models and unified speech understanding models for understaning, as well as TTS and instruct TTS models for generation.

### 5.1 Speech Understanding

**ASR models.** To accommodate different application scenarios, we develop three Wu dialect ASR models at three scales: a Conformer-U2pp-Wu (Wu et al., 2021) model, a Whisper-Medium-Wu (Radford et al., 2023) model, and a Step-Audio2-Wu-ASR model. The Conformer-U2pp-Wu and Whisper-Medium-Wu models are trained within the Wenet framework (Yao et al., 2021), with Conformer-U2pp-Wu trained from scratch, and Whisper-Medium-Wu fine-tuned from publicly available pre-trained weights. The Step-Audio2-Wu-ASR model is a Step-Audio2-mini fine-tuned in a parameter-efficient manner using LoRA (Hu et al., 2021) within the MS-Swift framework (Zhao et al., 2025). All models are pretrained on the ASR-Mid subset and conduct SFT on the ASR-High subset. Evaluation is performed on the ASR test set of WenetSpeech-Wu-Bench as well as two in-house manually annotated test sets covering dialogue and reading scenarios, enabling comprehensive assessment across diverse speaking conditions.

For comparison, we include both open-source and commercial baselines, such as Dolphin, SenseVoice, Paraformer (Gao et al., 2022), FireRedASR (Xu et al., 2025b) as an open-source baseline, Qwen3-ASR, Gemini-2.5-Pro, TeleASR (Chen et al., 2024), and Tencent-Cloud-ASR as a commercial baseline. The two in-house test sets are strictly out-of-set evaluations for Conformer-U2pp-Wu, Whisper-medium-Wu, and Step-Audio2-Wu-ASR. Conversely, the WS-Wu-Bench-ASR test set serves as an out-of-set evaluation for Step-Audio2-FT and Tele-CTC-FT, which are trained exclusively on in-house data.

As shown in Table 4, all existing open-source and commercial ASR systems perform poorly across all three test sets, indicating that they are not viable for Wu dialect recognition. In contrast, models trained on WenetSpeech-Wu achieve state-of-the-art performance across all model scales, with even the smallest Conformer-U2pp-Wu substantially outperforming all prior systems.

**Speech understanding models.** For speech understanding, we train the Step-Audio2-Wu-Und model on the WenetSpeech-Wu corpus using task-aware, quality-graded data. The model is first pre-trained on ASR-Mid and High-quality AST and age, gender, and emotion annotations, and then fine-tuned with the ASR-high subset alongside the same non-ASR tasks.

For benchmarking, we compare against baseline models, including Step-Audio2-mini and Qwen3-Omni, for speech understanding tasks. As shown in Table 5, after multi-task fine-tuning, the ASR per-

Table 6: TTS results on WenetSpeech-Wu-Bench. **Bold** and underlined values denote the best and second-best results, respectively; light green rows indicate models trained on WenetSpeech-Wu or further fine-tuned on an internal high-quality dataset.

| Model | WS-Wu-Eval-TTS-easy | | | | | WS-Wu-Eval-TTS-hard | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CER(%)↓ | SIM↑ | IMOS↑ | SMOS↑ | AMOS↑ | CER(%)↓ | SIM↑ | IMOS↑ | SMOS↑ | AMOS↑ |
| Qwen3-TTS† | 5.95 | – | 4.35 | – | 4.19 | 16.45 | – | 4.03 | – | **3.91** |
| DiaMoE-TTS | 57.05 | 0.702 | 3.11 | 3.43 | 3.52 | 82.52 | 0.587 | 2.83 | 3.14 | 3.22 |
| CosyVoice2 | 10.33 | 0.713 | 3.83 | 3.71 | 3.84 | 82.49 | 0.618 | 3.24 | 3.42 | 3.37 |
| CosyVoice2-Wu-CPT | 6.35 | **0.727** | 4.01 | **3.84** | 3.92 | 32.97 | **0.620** | 3.72 | **3.55** | 3.63 |
| CosyVoice2-Wu-SFT | 6.19 | 0.726 | 4.32 | 3.78 | 4.11 | 25.00 | 0.601 | 3.96 | 3.48 | 3.76 |
| CosyVoice2-Wu-SS* | **5.42** | – | **4.37** | – | **4.21** | **15.45** | – | **4.04** | – | 3.88 |

† Commercial system with a single fixed speaker, and speaker similarity is not considered.
* Single-speaker finetuned model, and speaker similarity is not evaluated.

formance of Step-Audio2-Wu-Und shows a slight drop than Step-Audio2-Wu-ASR but still achieves the second-best results. For the Wu-to-Mandarin AST task, it significantly outperforms baseline models. Comparison with Step-Audio2-mini illustrates the mismatch between Wu dialect and Mandarin in gender, age, and emotion prediction, which is effectively addressed by our data. Compared with Qwen3-Omni, our model shows notable improvements in age and emotion prediction, while performing slightly lower on gender classification.

## 5.2 Speech Generation

**TTS models.** In the continual pre-training (CPT) stage, we continue training from the open-source CosyVoice2 (Du et al., 2024) checkpoint using the TTS-Mid dataset for ten epochs. In the SFT stage, we use TTS-High dataset trained for three epochs. Finally, we perform speaker-specific fine-tuning on a 10-hour internal high-quality dataset.

As shown in Table 6, the experimental results demonstrate that the staged training strategy significantly improves CosyVoice2's speech synthesis performance in terms of Wu dialect capability. The CPT stage, leveraging large-scale pipeline-processed data, enhances the model's fundamental capabilities and robustness, particularly on challenging samples. The SFT stage, further improves speech naturalness and expressiveness. Finally, the Single-Speaker Supervised Fine-Tuning (SS-SFT) stage achieves the best performance across CER, IMOS, and AMOS metrics. Overall, CosyVoice2-Wu-SS approaches or surpasses the baseline Qwen3-TTS, DiaMoE-TTS (Chen et al., 2025), and CosyVoice2 in most metrics, especially in difficult sample synthesis.

**Instruct TTS models.** The instruction training data are from the Inst Pro and Inst Emo datasets as introduced in Table 3. Training was performed on the instruction text and tags using a small learning rate for five epochs. Compared with the model before instruction fine-tuning, the results showed clear improvements across all controllability metrics on WenetSpeech-Wu-Bench, as shown in Table 7, and subjective listening tests confirmed strong perceptual control effects, validating the effectiveness of proposed data.

Table 7: Performance of instruct TTS model.

| Type | Metric | CosyVoice2-Wu-SFT | CosyVoice2-Wu-instruct |
|---|---|---|---|
| Emotion | Happy↑ | 0.87 | **0.94** |
| | Angry↑ | 0.83 | **0.87** |
| | Sad↑ | 0.84 | **0.88** |
| | Surprised↑ | 0.67 | **0.73** |
| | EMOS↑ | 3.66 | **3.83** |
| Prosody | Pitch↑ | 0.24 | **0.74** |
| | Speech Rate↑ | 0.26 | **0.82** |
| | PMOS↑ | 2.13 | **3.68** |

## 6 Conclusion

In this work, we establish a Wu dialect speech processing ecosystem encompassing datasets, benchmarks, and models. We introduce WenetSpeech-Wu, an 8,000-hour large-scale corpus with rich multi-dimensional annotations across multiple Wu sub-dialects, and present the first public Wu dialect benchmark covering ASR, Wu-to-Mandarin AST, speech attribute and emotion recognition, TTS, and instruct TTS. Building on this dataset, we release a suite of ASR, TTS, unified speech understanding models, and instruct TTS models, enabling community research and demonstrating the effectiveness of WenetSpeech-Wu.

## Limitations

Despite the contributions of WenetSpeech-Wu to the dialectal speech processing community, several limitations remain to be addressed in future work. Although WenetSpeech-Wu is large-scale and covers multiple Wu sub-dialects, the data distribution across dialects and domains is not perfectly balanced, which may affect model generalization to less-represented varieties. In addition, many annotations are produced through automated or semi-automated pipelines. While we apply stringent quality control and filtering strategies, these annotations may still contain noise compared to fully manual labeling. Finally, our baseline models are intended to provide strong and reproducible reference points rather than task-optimal solutions, and future work may further improve performance through more specialized modeling and training strategies.

## Ethical Considerations

The Wu dialect speech data used in this work are collected exclusively from publicly accessible, non-restricted online sources intended for open dissemination. No access control mechanisms were circumvented during data collection. To ensure legal and ethical compliance, we only include audio that is explicitly available for research use.

To further mitigate potential ethical and privacy risks, we collaborate with a professional data service company to conduct independent review and filtering of the collected data. This process focuses on identifying and removing content that may contain personally identifiable information, sensitive attributes, or offensive material. In particular, audio segments involving explicit personal identifiers, such as names, phone numbers, or addresses, are excluded. As a result, the released dataset is anonymized and does not enable the identification of individual speakers.

All benchmark annotations are produced by the same professional data service company through manual labeling, following established internal compliance and quality control procedures. No personally identifiable information is disclosed to annotators, and annotation tasks are strictly limited to speech content relevant to linguistic analysis and benchmarking. Due to commercial confidentiality, detailed information regarding annotator compensation and internal annotation protocols cannot be publicly disclosed; however, the service provider confirms that all annotation activities comply with applicable local labor regulations and professional annotation standards.

The dataset and benchmark are constructed solely to support academic research on Wu dialect speech processing. Based on our data collection and filtering procedures, we think that the intended use of the dataset and benchmark does not introduce ethical or societal risks. To promote transparency and responsible research practices, we provide detailed documentation describing dataset composition, scale, splits, and evaluation protocols, and we clearly report experimental settings and results. The dataset is released with explicit documentation specifying its intended non-commercial research use.

## References


Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, and 14 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *CoRR*, abs/2407.04051.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proc. LREC*, pages 4218–4222.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH*, pages 1983–1987.

Cedric Chan and Jianjing Kuang. 2025. Toward Objective and Interpretable Prosody Evaluation in Text-to-Speech: A Linguistically Motivated Approach. *arXiv*, 2511.02104.

Hongjie Chen, Zehan Li, Guangmin Xia, Boqing Liu, Yan Yang, Jian Kang, and Jie Li. 2024. *TeleSpeechPT: Large-Scale Chinese Multi-dialect and Multi-accent Speech Pre-training*, pages 183–190. Springer.

Ziqi Chen, Gongyu Chen, Yihua Wang, Chaofan Ding, Zihao Chen, and Wei-Qiang Zhang. 2025. Diamoe-tts: A unified ipa-based dialect TTS framework with mixture-of-experts and parameter-efficient zero-shot adaptation. *CoRR*, abs/2509.22727.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, and Seong-Whan Lee. 2025. Emosphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector. *IEEE Trans. Affect. Comput.*, 16(3):2365–2380.

Yuhang Dai, Ziyu Zhang, Shuai Wang, Longhao Li, Zhao Guo, Tianlun Zuo, Shuiyuan Wang, Hongfei Xue, Chengyou Wang, Qing Wang, Xin Xu, Hui Bu, Jie Li, Jian Kang, Binbin Zhang, and Lei Xie. 2025. Wenetspeech-chuan: A large-scale sichuanese corpus with rich annotation for dialectal speech processing. *CoRR*, abs/2509.18004.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *CoRR*, abs/2412.10117.

Tiantian Feng, Jihwan Lee, Anfeng Xu, Yoonjeong Lee, Thanathai Lertpetchpun, Xuan Shi, Helin Wang, Thomas Thebaud, Laureano Moro-Velázquez, Dani Byrd, Najim Dehak, and Shrikanth Narayanan. 2025. Vox-profile: A speech foundation model benchmark for characterizing diverse speaker and speech traits. *CoRR*, abs/2505.14648.

J.G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Proc. ASRU*.

Changfeng Gao, Zhihao Du, and Shiliang Zhang. 2025. Differentiable reward optimization for LLM based TTS system. In *Proc. Interspeech*.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Proc. Interspeech*, pages 2063–2067.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2024. Emilia: An Extensive, Multilingual, and Diverse Speech Dataset For Large-Scale Speech Generation. In *Proc. SLT*, pages 885–890.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Chun-Yi Kuan, Chen-An Li, Tsu-Yuan Hsu, Tse-Yang Lin, Ho-Lam Chung, Kai-Wei Chang, Shuo-Yiin Chang, and Hung-Yi Lee. 2023. Towards general-purpose text-instruction-guided voice conversion. In *Proc. ASRU*, pages 1–8.

Longhao Li, Zhao Guo, Hongjie Chen, Yuhang Dai, Ziyu Zhang, Hongfei Xue, Tianlun Zuo, Chengyou Wang, Shuiyuan Wang, Jie Li, Jian Kang, Xin Xu, Hui Bu, Binbin Zhang, Ruibin Yuan, Ziya Zhou, Wei Xue, and Lei Xie. 2025. Wenetspeech-yue: A large-scale cantonese speech corpus with multi-dimensional annotation. *CoRR*, abs/2509.03959.

Daniel Lyth and Simon King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *CoRR*, abs/2402.01912.

Yangyang Meng, Jinpeng Li, Guodong Lin, Yu Pu, Guanbo Wang, Hu Du, Zhiming Shao, Yukai Huang, Ke Li, and Wei-Qiang Zhang. 2025. Dolphin: A large-scale automatic speech recognition model for eastern languages. *CoRR*, abs/2503.20212.

Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, YanGui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, Shuai Wang, and Kai Yu. 2025. A Survey on Speech Large Language Models for Understanding. *IEEE Journal of Selected Topics in Signal Processing*, page 1–32.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, volume 202, pages 28492–28518.

Philip Rose. 2015. Tonation in three chinese wu dialects. In *Proc. ICPhS*.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *Proc. ICLR*.

Mingchen Shao, Bingshen Mu, Chengyou Wang, Haizhou Li, Ying Yan, Zhonghua Fu, and Lei Xie. 2025. Towards Building Speech Large Language Models for Multitask Understanding in Low-Resource Languages. *CoRR*, abs/2509.14804.

Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xingjun Tan, Chuandong Xie, Shuran Zhou, Rui Yan, Chenjia Lv, Yang Han, Wei Zou, and Xiangang Li. 2021. KeSpeech: An Open Source Speech Dataset of Mandarin and Its Eight Subdialects. In *Proc. NeurIPS*.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025a. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. *CoRR*, abs/2506.04779.

Shuai Wang, Zhaokai Sun, Zhennan Lin, Chengyou Wang, Zhou Pan, and Lei Xie. 2025b. MSU-Bench: Towards Understanding the Conversational Multi-talker Scenarios. *CoRR*, abs/2508.08155.

Shuo Wang, Aishan Maoliniyazi, Xinle Wu, and Xiaofeng Meng. 2020. Emo2vec: Learning emotional embeddings via multi-emotion category. *ACM Trans. Internet Techn.*, 20(2):13:1–13:17.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 81 others. 2025. Step-audio 2 technical report. *CoRR*, abs/2507.16632.

Di Wu, Binbin Zhang, Chao Yang, Zhendong Peng, Wenjing Xia, Xiaoyu Chen, and Xin Lei. 2021. U2++: unified two-pass bidirectional end-to-end model for speech recognition. *CoRR*, abs/2106.05642.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025a. Qwen3-Omni Technical Report. *CoRR*, abs/2509.17765.

Kaituo Xu, Feng-Long Xie, Xu Tang, and Yao Hu. 2025b. Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to LLM integration. *CoRR*, abs/2501.14350.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Yifan Yang, Zheshu Song, Jianheng Zhuo, Mingyu Cui, Jinpeng Li, Bo Yang, Yexing Du, Ziyang Ma, Xunying Liu, Ziyuan Wang, Ke Li, Shuai Fan, Kai Yu, Wei-Qiang Zhang, Guoguo Chen, and Xie Chen. 2024. Gigaspeech 2: An evolving, large-scale and multi-domain ASR corpus for low-resource languages with automated crawling, transcription and refinement. *CoRR*, abs/2406.11546.

Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. 2021. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. In *Proc. Interspeech*, pages 4054–4058.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proc. Interspeech*, pages 1526–1530.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022. WENETSPEECH: A 10000+ Hours Multi-Domain Mandarin Corpus for Speech Recognition. In *Proc. ICASSP*, pages 6182–6186.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2025. SWIFT: A scalable lightweight infrastructure for fine-tuning. In *Proc. AAAI*, pages 29733–29735.

# A Appendix

## A.1 Details of Datasets

**Example of an Annotated Data.** Each speech sample is represented using a unified JSON format. An entry includes a unique key, audio duration, Wu dialect transcription with its Mandarin translation, and an associated confidence score. Contextual metadata, such as domain and sub-dialect, audio quality indicators (e.g., DNSMOS and SNR), and paralinguistic annotations (emotion, age, and gender) are organized into structured fields. In addition, prosodic acoustic features, including speech rate, loudness, fundamental frequency statistics, and energy statistics, are provided to support fine-grained acoustic analysis. Figure 4 illustrates an example of the annotated JSON entry.


```
{
  "key": "mx9ok0979_qtmvx_8765690",
  "duration": "2.3",
  "transcription": "侬好今朝天气蛮好
个",
  "translation": "你好今天天气挺好的",
  "confidence": "0.89",
  "meta-info": {
    "domain": "Interview",
    "sub-dialect": "Shanghai",
  },
  "audio-quality": {
    "DNSMOS": "3.98",
    "SNR": "58.74",
  },
  "paralinguistic": {
    "emotion": "Happy",
    "age": "Youth",
    "gender": "Male",
  },
 "low-level-feature": {
    "speech_rate": 15.094,
    "loudness": -17.72,
    "f0_mean": 167.17,
    "f0_std": 40.74,
    "energy_mean": 0.0684,
    "energy_std": 0.078,
  }
}
```


Figure 4: Example of an Annotated Data Entry in JSON format

## A.2 Statistics of the WenetSpeech-Wu-Bench

As summarized in Table 9, the WenetSpeech-Wu Bench consists of diverse test sets tailored for multiple speech-related tasks. For each task, we report the number of utterances and the corresponding total duration in hours.

## A.3 Details of Experiments

**Optimization and Training Hyperparameters.** The training hyperparameters for different models

Table 8: Optimization and training hyperparameters for proposed ASR and TTS models.

| Model | Para. (M) | LR | LR sched. | Warmup | Grad. Acc. | Batch size |
|---|---|---|---|---|---|---|
| Whisper-Medium-Wu | 769 | $8 \times 10^{-5}$ | WarmupLR | 4000 steps | 4 | Dynamic (24k frames) |
| Conformer-U2pp-Wu | 123 | $1 \times 10^{-3}$ | WarmupLR | 25000 steps | 4 | Dynamic (60k frames) |
| Step-Audio2-Wu-ASR | 7000 | $1 \times 10^{-5}$ | WarmupLR | 0.05 ratio | 8 | 8 |
| CosyVoice2-Wu-CPT | 500 | $1 \times 10^{-4}$ | WarmupLR | 25000 steps | 2 | Dynamic (10k frames) |
| CosyVoice2-Wu-SFT | 500 | $1 \times 10^{-5}$ | ConstantLR | 0 | 2 | Dynamic (2k frames) |
| CosyVoice2-Wu-SS | 500 | $5 \times 10^{-6}$ | ConstantLR | 0 | 2 | Dynamic (1k frames) |
| CosyVoice2-Wu-instruct | 500 | $5 \times 10^{-6}$ | ConstantLR | 0 | 2 | Dynamic (1k frames) |

Table 9: Statistics of the WenetSpeech-Wu-Bench.

| Task | Category | Utterances | Duration (h) |
|---|---|---|---|
| ASR | - | 4851 | 9.75 |
| AST | - | 3000 | 4.4 |
| Gender | Male | 1500 | 2.10 |
| | Female | 1500 | 2.29 |
| | Total | 3000 | 4.39 |
| Age | Youth | 500 | 0.69 |
| | Middle-aged | 500 | 0.79 |
| | Elderly | 500 | 0.69 |
| | Total | 1500 | 2.22 |
| Emotion | Neutral | 300 | 0.50 |
| | Happy | 200 | 0.23 |
| | Sad | 100 | 0.26 |
| | Angry | 200 | 0.15 |
| | Surprised | 200 | 0.27 |
| | Total | 1000 | 1.41 |

are summarized in Table 8.

**Supplementary Experimental Results.** To provide a more detailed analysis of model performance on WS-Wu-Bench, we report category-wise classification accuracy for gender, age, and emotion across WS-Wu-Und and baseline models. As shown in Figure 5, our model achieves more balanced and consistently higher performance across different paralinguistic categories, demonstrating stronger paralinguistic recognition capability.

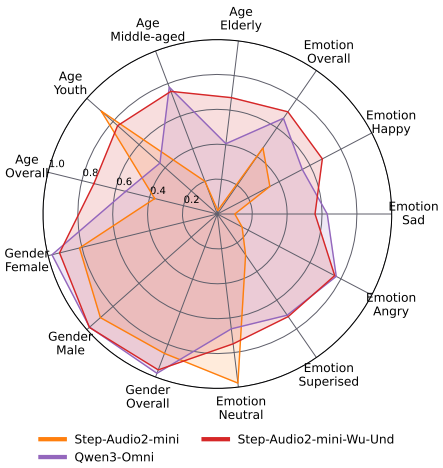


Figure 5: Supplementary comparison results on age, gender, and emotion recognition tasks.

**Prompt Templates Used in Experiments.** For all speech understanding tasks, the same prompt templates are used during both training and inference to avoid potential distribution mismatch.For instruct TTS tasks, instruction-based templates are likewise employed in both training and inference to enable explicit instruction control. Table 10 lists the prompt formulations for each task. The English prompts are shown for ease of presentation, while equivalent instructions are employed in other languages when applicable.

Table 10: Details of prompt templates used in speech understanding and instruct TTS experiments.

| Task Type | Prompt Content |
|---|---|
| ASR | Please transcribe the speech. |
| AST | Please listen to this speech carefully and translate its content into Mandarin. |
| Age | Based on the acoustic features of the speech, determine the speaker's age. Choose one label from: youth, middle-aged, or elderly. |
| Gender | Based on the acoustic features of the speech, determine the speaker's gender. Choose one label from: male or female. |
| Emotion | Based on the acoustic features and semantics of the speech, determine the emotion. Choose one label from: neutral, happy, sad, surprised, or angry. |
| Inst Emo | You must speak with anger, sadness, happiness, or surprise. |
| Inst Pro | This is a man or woman speaking in a low or high voice and at a slow or fast pace. |