WenetSpeech-Yue: A Large-scale Cantonese Speech Corpus with Multi-dimensional Annotation

Longhao Li[1*], Zhao Guo[1*], Hongjie Chen[2], Yuhang Dai[1], Ziyu Zhang[1], Hongfei Xue[1], Tianlun Zuo[1], Chengyou Wang[1], Shuiyuan Wang[1], Xin Xu[3], Hui Bu[3], Jie Li[2], Jian Kang[2], Binbin Zhang[4], Ruibin Yuan[5], Ziya Zhou[5], Wei Xue[5], Lei Xie[1†]

[1]Audio, Speech and Language Processing Group (ASLP@NPU), Northwestern Polytechnical University
[2]Institute of Artificial Intelligence (TeleAI), China Telecom
[3]Beijing AISHELL Technology Co., Ltd.
[4]WeNet Open Source Community
[5]Hong Kong University of Science and Technology
lhli@mail.nwpu.edu.cn, gzhao@mail.nwpu.edu.cn, lxie@nwpu.edu.cn

Abstract

The development of speech understanding and generation has been significantly accelerated by the availability of large-scale, high-quality speech datasets. Among these, ASR and TTS are regarded as the most established and fundamental tasks. However, for Cantonese (Yue Chinese), spoken by approximately 84.9 million native speakers worldwide, limited annotated resources have hindered progress and resulted in suboptimal ASR and TTS performance. To address this challenge, we propose WenetSpeech-Pipe, an integrated pipeline for building large-scale speech corpus with multi-dimensional annotation tailored for speech understanding and generation. Based on this pipeline, we release WenetSpeech-Yue, the first large-scale Cantonese speech corpus with multi-dimensional annotation for ASR and TTS, covering 21,800 hours across 10 domains with annotations including ASR transcription, text confidence, speaker identity, age, gender, speech quality scores, among other annotations. We also release WSYue-eval, a comprehensive Cantonese benchmark with two components: WSYue-ASR-eval, a manually annotated set for evaluating ASR on short and long utterances, code-switching, and diverse acoustic conditions, and WSYue-TTS-eval, with base and coverage subsets for standard and generalization testing. Experimental results show that models trained on WenetSpeech-Yue achieve competitive results against state-of-the-art (SOTA) Cantonese ASR and TTS systems, including commercial and LLM-based models, highlighting the value of our dataset and pipeline.

Project — https://github.com/ASLP-lab/WenetSpeech-Yue

Introduction

Recent advances in speech understanding and generation have been largely fueled by the availability of large-scale, diverse, and richly annotated datasets. Core tasks such as ASR and TTS exemplify how model performance across diverse linguistic and acoustic conditions hinges on this very foundation. For instance, advanced ASR and TTS systems, such as FireRedASR (Xu et al. 2025b), Whisper (Radford et al. 2023), and VALL-E (Wang et al. 2023a), have achieved significant performance breakthroughs by leveraging the advantages of large, diverse corpora. To meet the growing demand for high-quality resources, scalable data pipelines like GigaSpeech2 (Yang et al. 2024), WenetSpeech (Zhang et al. 2022), and Emilia-Pipe (He et al. 2025) series have streamlined the construction of large, multilingual, and multi-domain corpora, thereby facilitating the efficient development of ASR and TTS systems.

Despite these advances, Cantonese, as a Chinese dialect with high practical usage and linguistic complexity, remains severely under-resourced. On one hand, Cantonese is spoken by over 84.9 million people across mainland China, Hong Kong, Macau, and global Chinese communities (Xiang et al. 2022). Its rich tone system of nine tones in six categories, coexistence of literary and colloquial forms, and frequent code-switching with English pose unique modeling challenges. On the other hand, its cultural distinctiveness demands speech systems with high robustness, emotional expressiveness, and stylistic diversity, which current resources fail to adequately support.

Publicly available Cantonese corpora are often limited in scale, style, and label diversity. Projects like Common Voice (Ardila et al. 2020) and MDCC (Yu et al. 2022) rely heavily on manual annotation and offer only small datasets. Evaluation sets are typically composed of short utterances and lack coverage of complex linguistic phenomena. Moreover, most corpora provide only speech-text alignment, with little to no speaker attributes or acoustic quality metadata, severely limiting their use in self-supervised learning, style modeling, and multi-task training. As a result, mainstream ASR and TTS systems perform poorly on Cantonese tasks and exhibit weak generalization to real-world scenarios.

To address these gaps, we introduce WenetSpeech-Pipe, a modular and automated pipeline designed for building large-scale Cantonese datasets. It integrates multiple quality-enhancement strategies and metadata enrichment techniques

[*]These authors contributed equally as first authors.
[†]Corresponding author.

Dataset	Task	Duration (hours)	Long Audio	Code-switch	Multi-domain	Multi-label
Guangzhou Daily Use	ASR	4.06	✗	✗	✗	✗
Guangzhou Cabin	ASR	5.00	✗	✗	✗	✗
Mixed Cantonese and English	ASR	34.8	✗	✓	✗	✗
Common-Voice yue	ASR	203	✗	✗	✗	✗
Common-Voice zh-HK	ASR	108	✗	✗	✗	✗
MDCC	ASR	73.6	✗	✗	✓	✗
ZoengJyutGaai-Storytelling	TTS	188.25	✗	✗	✗	✗
WenetSpeech-Yue	ASR/TTS/Others	21,800	✓	✓	✓	✓

Table 1: Comparison with existing Cantonese speech datasets.

to ensure the resulting corpus is both diverse and richly annotated, supporting a wide range of modeling objectives. Built upon this pipeline, we construct WenetSpeech-Yue, the largest and most comprehensive open-source Cantonese speech corpus to date, spanning over 21,800 hours across eleven domains. To facilitate rigorous evaluation of both ASR and TTS systems, we further release WSYue-eval, a dedicated evaluation suite comprising two subsets: WSYue-ASR-eval for ASR and WSYue-TTS-eval for TTS. Each subset is curated to cover a wide range of linguistic and acoustic scenarios, supporting comprehensive and reliable assessment under diverse conditions. Experimental results show that ASR and TTS models trained on WenetSpeech-Yue achieve performance comparable to or exceeding SOTA systems on multiple test sets. These contributions significantly advance Cantonese speech understanding and generation, addressing critical resource gaps and enabling more robust and expressive speech systems.

Our contributions can be summarized as follows:

1. We propose WenetSpeech-Pipe, a large-scale, multi-domain, multi-label data pipeline tailored for both speech understanding and generation in Cantonese.
2. We release WenetSpeech-Yue, a 21,800-hour large-scale Cantonese speech corpus with rich multi-dimensional annotations—currently the largest open-source resource for Cantonese speech research.
3. We release WSYue-eval, a comprehensive Cantonese benchmark comprising WSYue-ASR-eval, a human-annotated test set for ASR, and WSYue-TTS-eval for TTS.
4. We demonstrate that ASR and TTS models trained on WenetSpeech-Yue achieve SOTA performance across multiple benchmarks.

Related Work

Large-scale Speech Corpora and the Scarcity of Cantonese Resources

Recent years have seen growing efforts in constructing large-scale open-source speech datasets across languages and modalities. GigaSpeech2 (Yang et al. 2024) adopts a self-iterative label refinement strategy to develop multilingual corpora for Southeast Asian languages (Thai, Indonesian, and Vietnamese), automating collection, transcription, and alignment. Multilingual LibriSpeech (Panayotov et al. 2015)(MLS), derived from LibriVox audiobooks, offers approximately 451,000 hours of read speech, with 445,000 hours in English and 6,000 in seven other languages. WenetSpeech (Zhang et al. 2022) curates 22,435 hours of Chinese speech audio from YouTube and podcasts, combining with refined transcripts via a well-designed pipeline. WenetSpeech4TTS (Ma et al. 2024) further refines this data for speech synthesis by adjusting segment length, filtering for speaker consistency, and enhancing audio quality, resulting in 12,800 hours of high-quality TTS-ready speech. Emilia introduces a large-scale multilingual speech generation dataset of over 101,000 hours across six languages, supporting both TTS training and synthesis research.

Despite these advances, publicly available Cantonese corpora remain limited in both scale and linguistic diversity. For example, Mozilla's Common Voice (Ardila et al. 2020) provides around 311 hours of validated Cantonese read speech. The MDCC (Yu et al. 2022) corpus contains 73.6 hours of high-quality audiobook recordings from Hong Kong, spanning domains such as philosophy, education, and lifestyle. The ZoengJyutGaai-Storytelling Dataset[1] offers 112.54 hours of expressive single-speaker speech. These datasets are often constrained to read speech, narrow domain coverage, and lack of rich metadata—posing challenges for building robust models for real-world Cantonese speech applications.

Cantonese ASR and TTS Models

Despite recent progress, Cantonese-specific speech modeling remains under-resourced compared to major languages. On the ASR front, SenseVoice (An et al. 2024), trained on 300,000 hours of multilingual speech including 9,600 hours of Cantonese, is regarded as the SOTA Cantonese ASR system. Whisper-large-v3 (Radford et al. 2023), a general-purpose multilingual model trained on over 5 million hours of speech, also demonstrates strong cross-lingual generalization, achieving competitive performance on Cantonese without dialect-specific tuning. More recently, TeleASR (Chen et al. 2024), a dialect-aware model pre-trained on 300,000 hours of unlabeled audio and fine-tuned across 30 Chinese dialects, has demonstrated robust recognition capabilities in regional languages including Cantonese.

On the synthesis front, open-source TTS models have increasingly extended support to Cantonese and other di-

[1]https://huggingface.co/datasets/CanCLID/zoengjyutgaai

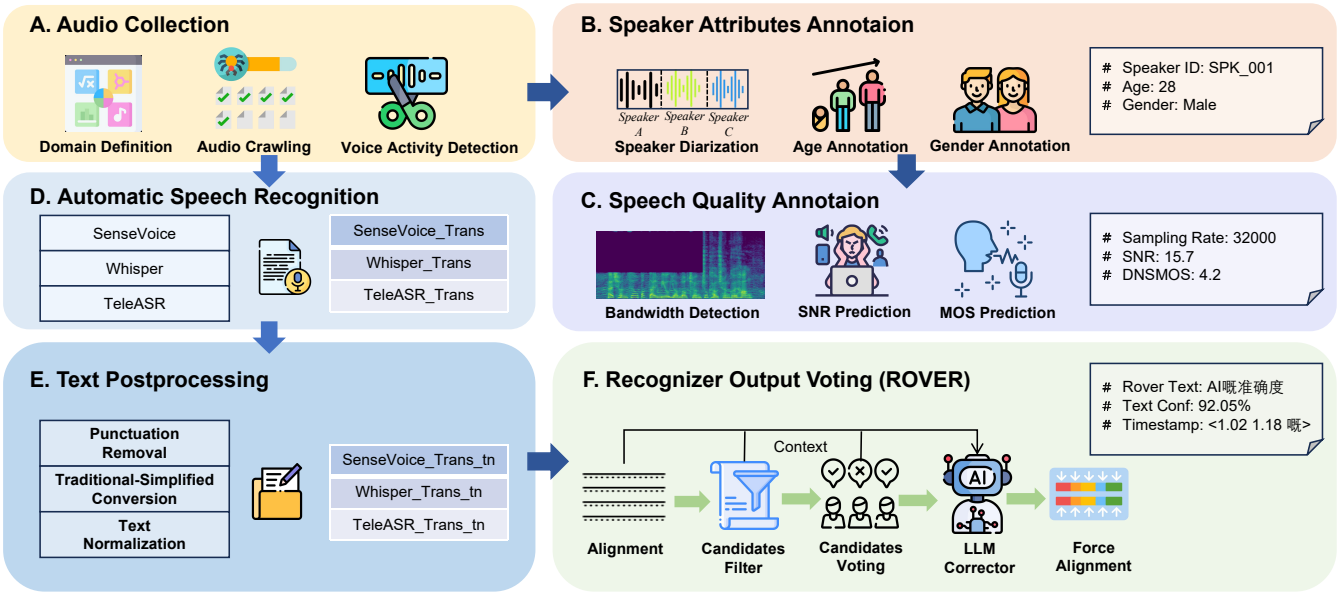


Figure 1: An overview of the WenetSpeech-Pipe processing pipeline.

alects. CosyVoice2 (Du et al. 2024), which enables zero-shot voice cloning and multilingual synthesis, includes Cantonese among supported languages such as Mandarin, English, Japanese, and Korean. Meanwhile, Step-Audio-TTS-3B (Huang et al. 2025), a lightweight model within the Step-Audio framework, leverages knowledge distillation from a larger model to support controllable speech generation, including emotion, speaking rate, dialectal variation (e.g., Cantonese, Sichuanese).

WenetSpeech-Pipe

Our proposed WenetSpeech-Pipe framework, as illustrated in Figure 1, comprises six modules: (A) Audio Collection, (B) Speaker Attributes Annotation, (C) Speech Quality Annotation, (D) Automatic Speech Recognition, (E) Text Postprocessing, and (F) Recognizer Output Voting.

Audio Collection WenetSpeech-Pipe begins with the large-scale collection of in-the-wild speech data spanning diverse domains, such as storytelling, drama, commentary, vlogs, food, entertainment, news, and education. Raw recordings are typically long-form, often tens of minutes to over an hour, which makes them unsuitable for direct model training or alignment. To generate utterance-level data suitable for downstream transcription and quality assessment, the audio is automatically segmented into short clips using a voice activity detection (VAD) module, producing a large and domain-diverse pool of segmented audio clips ready for subsequent processing.

Speaker Attribute Annotation To enrich the dataset with speaker-level metadata for multi-speaker modeling and style-aware synthesis, WenetSpeech-Pipe incorporates a Speaker Attributes Annotation stage. First, speaker diarization is performed using the pyannote toolkit (Bredin 2023), which assigns local speaker labels to short audio segments from the same source, providing intra-recording speaker separation. Second, both age and gender are estimated for each segment using the Vox-Profile (Feng et al. 2025), providing speaker attribute annotations. This process produces utterance-level segments annotated with speaker identity, age, and gender, forming a multi-dimensional metadata that facilitates both supervised and style-controllable speech modeling.

Speech Quality Annotation To ensure that the curated audio supports high-fidelity speech generation tasks such as TTS and voice conversion (VC), WenetSpeech-Pipe incorporates a comprehensive speech quality assessment stage. Each audio segment is assessed through three complementary approaches. First, noise levels are characterized using Brouhaha (Lavechin et al. 2023), which produces segment-level signal-to-noise ratio (SNR) annotations. Second, perceptual quality is measured with DNSMOS (Reddy, Gopal, and Cutler 2022), a non-intrusive model that predicts the mean opinion score (MOS), which reflects human-perceived speech clarity and naturalness.

Finally, spectral characteristics are analyzed via bandwidth detection, which estimates the effective upper frequency and spectral coverage of each recording, complementing nominal sampling-rate metadata. Together, these multi-dimensional assessments generate a structured quality annotation for each segment, providing both quantitative scores (SNR and DNSMOS) and a reliable spectral reference (sampling rate) for downstream high-fidelity speech processing.

Automatic Speech Recognition Single ASR systems often exhibit systematic biases and error patterns due to architectural constraints (Shao et al. 2025), training data limitations, or domain mismatch. To mitigate these issues

and improve transcription reliability, we adopt a multi-system ensemble approach that leverages diverse recognition paradigms. Specifically, each audio segment is independently transcribed using three high-performance Cantonese ASR systems: the open-source models SenseVoice and Whisper, and the commercial solution TeleASR. These systems differ in architecture, training data, and optimization objectives—enabling complementary error profiles and diverse linguistic hypotheses. The output of this module consists of three parallel transcriptions per utterance, forming the foundation for subsequent fusion and refinement. These multi-hypothesis outputs serve as the primary input to the Recognizer Output Voting stage, where consensus-based alignment and voting yield a more accurate and robust final transcription.

Text Postprocessing To ensure reliable cross-system alignment and effective integration of multi-source transcriptions, it is essential to standardize the output formats from different ASR systems. The raw transcriptions from the aforementioned ASR systems exhibit significant variations in character sets (traditional vs. simplified Chinese), inclusion of non-lexical tags (e.g., `[laughter]`), and formatting inconsistencies for numerals and code-switched text. These discrepancies can impede accurate fusion and consensus formation during subsequent processing stages. Therefore, we apply a text post-processing pipeline to all transcription streams. This pipeline converts traditional Chinese characters to simplified form using the `OpenCC`[2] tool, removes all punctuation marks and special symbols, standardizes numerical expressions and date formats through rule-based rewriting, and inserts whitespace between Cantonese and English words to facilitate bilingual modeling. By applying these steps sequentially, we generate cleaned and canonical transcriptions that are consistent across the three systems. These standardized outputs serve as robust input representations for the ROVER module, ensuring that differences in surface form do not interfere with phonetic or lexical alignment during fusion.

Recognizer Output Voting While Text Postprocessing unifies the surface forms of transcriptions across multiple ASR systems, persistent variations remain in lexical selection, word segmentation, and phonetic representation. To generate a unified, high-accuracy reference transcription, we adopt the Recognizer Output Voting Error Reduction (ROVER) framework (Fiscus 1997), a fusion strategy based on multi-system voting to enhance transcription accuracy.

In our implementation, we extend the standard ROVER pipeline to handle the linguistic characteristics of Cantonese better. First, transcriptions after text normalization from the aforementioned ASR systems, are aligned using dynamic programming. To ensure robustness against outlier hypotheses, we introduce a candidate filtering module that computes the edit distance between each system's output and the average transcription of the other two. Outputs exceeding a predefined threshold are excluded from voting. At each aligned position, the most frequently occurring word is selected, and the average voting frequency across all positions is recorded as the utterance-level text confidence score. We extend the voting paradigm to Cantonese pinyin by implementing a pronunciation-specific confidence measure, operating in parallel with the character-level voting to reinforce phoneme consistency.

To further improve transcription accuracy, we employ an LLM, Qwen3-4B (Yang et al. 2025), to perform minimal, context-aware refinements on the consensus output. The LLM considers all original ASR hypotheses as contextual references and applies only necessary corrections to grammar, word choice, or named entities, preserving the integrity of the spoken content.

Finally, we perform character-level forced alignment between the refined transcription and the original audio using a pre-trained acoustic model. This yields precise timestamps for each character, enabling fine-grained speech processing and supporting downstream tasks.

[2]https://github.com/BYVoid/OpenCC

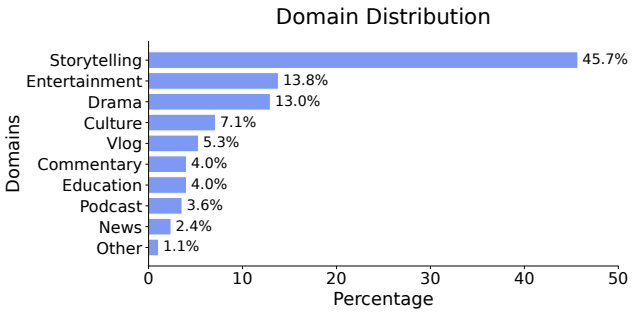


Figure 2: Domain distribution of WenetSpeech-Yue.

WenetSpeech-Yue

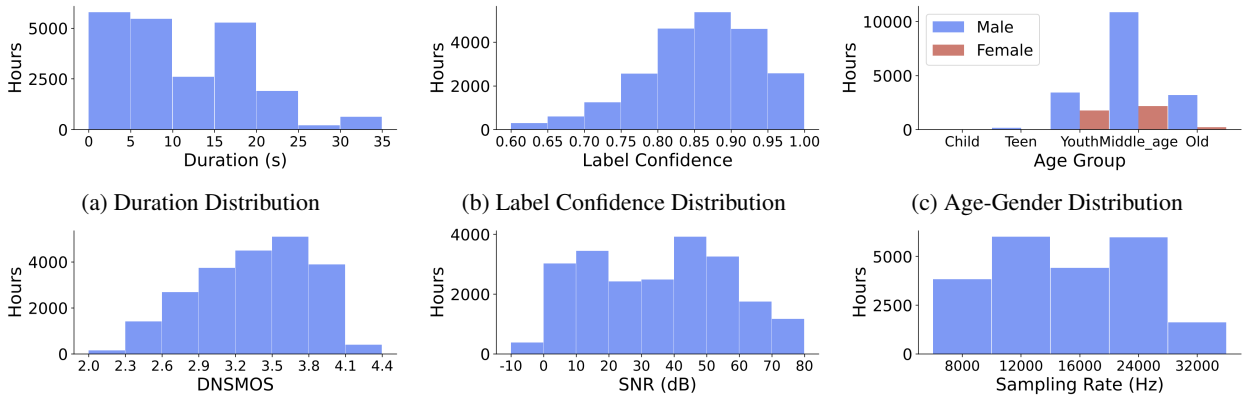
Dataset

Metadata Metadata is stored in a single JSON file. The metadata fields include *audio path*, *duration*, *text confidence* , *speaker identity*, *SNR*, *DNSMOS*, *age*, *gender*, and *character-level timestamps*. These fields are extensible, and additional metadata tags may be incorporated in the future.

Domains The domain of the WenetSpeech-Yue corpus is classified into ten categories: *Storytelling*, *Entertainment*, *Drama*, *Culture*, *Vlog*, *Commentary*, *Education*, *Podcast*, *News*, and *Others*. The distribution of these domains is illustrated in Figure 2.

Duration WenetSpeech-Yue contains a total of 21,800 hours of audio, including both short and long recordings, with an average duration of 11.40 seconds per audio segment. The major distribution is shown in Figure 3a.

Confidence In this work, we retain only labels with a text confidence score above 0.6. Based on the confidence score, we partition the data into three subsets: *strong labels* (confidence > 0.9, 6,771.43 hours), *moderate labels* ($0.8 <$ confidence ≤ 0.9, 10,615.02 hours), and *weak labels* ($0.6 <$ confidence ≤ 0.8, 4,488.13 hours). Detailed distribution is shown in Figure 3b.



(a) Duration Distribution (b) Label Confidence Distribution (c) Age-Gender Distribution

(d) DNSMOS Distribution (e) SNR Distribution (f) Sampling Rate Distribution

Figure 3: Visualization of statistical analysis for the WenetSpeech-Yue corpus.

Speech Quality As illustrated in Figure 3, we evaluated the audio quality of the WenetSpeech-Yue dataset. Specifically, Figure 3d presents *DNSMOS* scores spanning 2.0 to 4.4; Figure 3e depicts *SNR* values varying from -5 to 80 dB; and Figure 3f shows the distribution of *sampling rate*, which range from 8,000 to 32,000 Hz. To ensure suitability for generative tasks, we filtered the dataset by retaining samples with *DNSMOS* greater than 2.5 and *SNR* above 25 dB, resulting in a subset of 12,000 hours high-quality utterances for TTS applications. This filtering strategy can be further adapted for downstream tasks such as vocoder, codec, and voice conversion.

Speaker Attributes As shown in Figure 3c, we present the distribution of age and gender in the WenetSpeech-Yue corpus. The corpus is predominantly composed of male speakers, especially in the *Middle_age* (50.6%) group, while female speakers are relatively underrepresented in all age groups.

Benchmark

To address the unique linguistic characteristics of Cantonese, we propose **WSYue-eval**, a comprehensive benchmark encompassing both ASR and TTS tasks. This integrated evaluation framework is specifically tailored to assess model performance across critical dimensions of Cantonese language processing.

ASR Benchmark As a representative task of speech understanding, we developed the **WSYue-ASR-eval** test set for the ASR task. The WSYue-ASR-eval set is annotated through multiple rounds of manual labeling and includes tags such as text transcription, emotion, age, and gender. Based on differences in audio duration, WSYue-ASR-eval is divided into two subsets, *Short* and *Long*, as shown in Table 2, to enable comprehensive evaluation of Cantonese speech of varying lengths. In addition, WSYue-ASR-eval covers a wide range of Cantonese usage scenarios, including Cantonese-English code-switching and multi-domain conditions.

Set	Duration	Speakers	Hours
Short	0–10 s	2861	9.46
Long	10–30 s	838	1.97

Table 2: WSYue-ASR-eval Subsets

TTS Benchmark We introduce **WSYue-TTS-eval**, a benchmark specifically designed for zero-shot Cantonese TTS evaluation. WSYue-TTS-eval comprises two subsets: *Base* and *Coverage*. The *Base* subset contains 1,000 prompt-text pairs sampled from the CommonVoice dataset, enabling assessment of model performance on real-world data distributions. However, as CommonVoice primarily consists of daily conversational data, its coverage of diverse domains and linguistic phenomena is limited. To address this, the *Coverage* subset combines manually curated and LLM-generated texts. This subset spans a wide range of domains, including daily life, news, entertainment, and poetry, and incorporates various Cantonese linguistic phenomena such as polyphonic characters, tone sandhi, code-switching, proper nouns, numerals, and other challenging cases. This comprehensive design enables rigorous evaluation of model generalization and robustness across diverse and complex scenarios.

Experiments

ASR Task

Experimental Setup To assess the effectiveness of the WenetSpeech-Yue corpus and quantify its contribution to model performance, we conducted ASR experiments with two model categories, traditional architectures without large language models and LLM-augmented hybrids, denoted as `w/o LLM` and `w/ LLM`, respectively. The `w/o LLM` group includes `U2pp-Conformer-Yue`, a U2pp-Conformer (Wu et al. 2021; Li et al. 2025) trained from scratch to maximize the dataset's supervision signal, `Whisper-medium-Yue`, a Whisper-medium model

Model	#Params	In-House		Open-Source					WSYue-ASR-eval	
	(M)	Dialogue	Reading	yue	zh-HK	MDCC	Daily_Use	Commands	Short	Long
w/o LLM										
Paraformer (Gao et al. 2022)	220	83.22	51.97	70.16	68.49	47.67	79.31	69.32	73.64	89.00
SenseVoice-small (An et al. 2024)	234	21.08	6.52	8.05	**7.34**	6.34	5.74	6.65	6.69	9.95
SenseVoice-small-Yue	234	19.19	6.71	6.87	8.68	5.43	5.24	6.93	5.23	8.63
Dolphin-small (Meng et al. 2025)	372	59.2	7.38	39.69	51.29	26.39	7.21	9.68	32.32	58.20
U2pp-Conformer-Yue	130	**16.57**	7.82	7.72	11.42	5.73	5.73	8.97	5.05	8.89
TeleASR (Chen et al. 2024)	700	37.18	7.27	7.02	7.88	6.25	8.02	**5.98**	6.23	11.33
Whisper-medium (Radford et al. 2023)	769	75.50	68.69	59.44	62.50	62.31	64.41	80.41	80.82	50.96
Whisper-medium-Yue	769	18.69	6.86	6.86	11.03	5.49	4.70	8.51	5.05	8.05
FireRedASR-AED-L (Xu et al. 2025b)	1100	73.70	18.72	43.93	43.33	34.53	48.05	49.99	55.37	50.26
Whisper-large-v3 (Radford et al. 2023)	1550	45.09	15.46	12.85	16.36	14.63	17.84	20.70	12.95	26.86
w/ LLM										
Qwen2.5-Omni-3B(Xu et al. 2025a)	3000	72.01	7.49	12.59	11.75	38.91	10.59	25.78	67.95	88.46
Kimi-Audio (KimiTeam et al. 2025)	7000	68.65	24.34	40.90	38.72	30.72	44.29	45.54	50.86	33.49
FireRedASR-LLM-L (Xu et al. 2025b)	8300	73.70	18.72	43.93	43.33	34.53	48.05	49.99	49.87	45.92
U2pp-Conformer-LLM-Yue	4200	17.22	**6.21**	**6.23**	9.52	**4.35**	**4.57**	6.98	**4.73**	**7.91**

Table 3: ASR Results (MER%) on various Cantonese test sets for our models and comparison models. The gray bars represent the baseline models, while the green bars represent our proposed models. The dashed lines are used to separate models of different sizes. The best results are highlighted in bold, while the second-best results are underlined.

fine-tuned with a low learning rate for efficient Cantonese adaptation, and `SenseVoice-small-Yue`, a fine-tuned SenseVoice-small (An et al. 2024) variant serving as a strong small-scale Cantonese baseline. The `w/ LLM` is showcased by `U2pp-Conformer-LLM-Yue`, a Conformer–LLM hybrid with the U2pp-Conformer encoder connected to the Qwen3-4B via a lightweight adapter module.

All models adopt a two-stage training strategy: an initial stage using mixed medium- and high-confidence labels for rapid convergence, followed by a fine-tuning stage on high-confidence labels to maximize transcription accuracy. This setup both reduces training cost and directly reflects the dataset's quality impact.

Baselines To demonstrate the effectiveness of our training on the WenetSpeech-Yue dataset, we benchmarked against several prominent and competitive Cantonese ASR systems as shown in Table 3.

ASR Test Sets To comprehensively assess the generalization capabilities of various ASR models, we employ a diverse array of Cantonese test sets categorized into three groups: (1) **In-house collections**, including *Dialogue* for conversational speech analysis, and *Reading* for read speech evaluation; (2) **Open-source resources**, consisting of the *Common Voice* series (yue and zh-HK) (Ardila et al. 2020), the multi-domain *MDCC*[3] dataset, *Daily_Use*[4] (containing 4.06 hours of transcripted daily conversations), and *Commands*[5] (featuring 5 hours of in-vehicle command recordings); and (3) Our proposed ***WSYue-ASR-eval*** **benchmark** including *Short* and *Long* for short and long-sentence evaluation respectively.

ASR Results and Analysis We use the Mixed Error Rate (MER), which calculates errors at the character level for Chinese and the word level for English, as the evaluation metric to compare models trained on WenetSpeech-Yue and baselines.

As shown in Table 3, the experimental results reveal several consistent observations: (1) Across all model scales—including the small, medium, and w/LLM configurations—our models achieve the best performance on most evaluation sets; (2) Within the small-scale group, both `SenseVoice-small-Yue` and `U2pp-Conformer-Yue` achieve competitive results, with `SenseVoice-small-Yue` outperforming all baselines despite having the smallest size, indicating that our corpus substantially enhances efficiency in low-capacity settings; (3) Within the `w/o LLM` category, both `U2pp-Conformer-Yue` and `Whisper-medium-Yue` surpass the large scale baselines; (4) Within the `w/ LLM` group, `U2pp-Conformer-LLM-Yue`, consistently attains the-state-of-the-art (SOTA) accuracy. Collectively, these observations highlight that our propose WenetSpeech-Yue not only improves overall performance but also maximizes model potential across different parameter regimes, validating its utility for both traditional and LLM-enhanced ASR paradigms.

Table 5 reports the impact of our two-stage training strategy. Stage 1, trained on the mixed-confidence dataset, already achieves very competitive Cantonese ASR performance, while Stage 2 fine-tuning on high-confidence data yields significant gains across both test sets of WSYue-ASR-eval. These observations confirm that high-confidence labels are the primary driver of performance improvements. We believe that retaining confidence information is essen-

[3]https://github.com/HLTCHKUST/cantonese-asr

[4]https://huggingface.co/datasets/AlienKevin/guangzhou-daily-use-speech

[5]https://huggingface.co/datasets/AlienKevin/guangzhou-cabin-speech

Model	Base		Coverage		UTMOSv2 ↑	I-MOS ↑	S-MOS ↑	A-MOS ↑
	MER (%) ↓	SIM ↑	MER (%) ↓	SIM ↑				
Llasa-1B	53.31	0.732	43.68	0.754	2.360	2.60 ± 1.01	3.05 ± 0.87	2.32 ± 0.98
Step-Audio-TTS-3B	27.79	0.762	24.25	0.781	2.496	3.22 ± 0.70	3.14 ± 0.58	2.82 ± 0.69
CosyVoice2	14.38	0.812	13.74	0.826	2.989	3.72 ± 0.50	3.52 ± 0.36	3.22 ± 0.60
Edge-TTS†	**8.30**	-	**9.27**	-	2.997	4.12 ± 0.28	-	3.48 ± 0.56
Llasa-1B-Yue	10.89	0.762	12.78	0.772	2.696	4.30 ± 0.23	$\mathbf{4.11 \pm 0.37}$	$\mathbf{4.34 \pm 0.34}$
Cosyvoice2-Yue	**10.33**	**0.821**	**9.49**	**0.834**	**3.021**	$\mathbf{4.45 \pm 0.16}$	3.78 ± 0.53	4.21 ± 0.27

† Commercial system with single fixed speaker, and speaker similarity is not considered.

Table 4: TTS performance comparison on WSYue-TTS-eval using both objective and subjective metrics.

Model	Subset	WSYue-ASR-eval	
		Short	Long
Whisper-medium-Yue	Stage 1	7.27	11.19
	Stage 2	5.05	8.05
U2pp-Conformer-Yue	Stage 1	7.62	12.01
	Stage 2	5.05	8.89
U2pp-Conformer-LLM-Yue	Stage 1	6.81	10.75
	Stage 2	4.73	7.91

Table 5: Model performance on WSYue-ASR-eval subsets (MER%)

tial because it facilitates flexible training strategies, allowing high-confidence subsets to drive fine-tuning, while carefully leveraging low-confidence segments can improve model robustness in semi-supervised or domain-adaptive scenarios.

TTS Task

Experimental Setup To evaluate the effectiveness of WenetSpeech-Yue for speech synthesis, we adopt a transfer learning approach on two pretrained TTS models: Llasa-1B and CosyVoice2. Llasa-1B is a zero-shot TTS model pretrained on 250,000 hours of Mandarin and English speech. Both models are further fine-tuned on the TTS subset of WenetSpeech-Yue to adapt them to the linguistic and acoustic characteristics of Cantonese.

Baselines We compare our proposed methods, `Llasa-1B-Yue` and `CosyVoice2-Yue`, with several zero-shot baselines: (1) `Llasa-1B` (without fine-tuning on our dataset); (2) `CosyVoice2` (without fine-tuning on our dataset); and (3) `Step-Audio-TTS-3B`. In addition, we include `Edge-TTS`, a commercial system with single fixed speaker, as a reference for high-quality synthesis.

Evaluation Metrics We employ both objective and subjective metrics to comprehensively assess model performance. For objective evaluation, we use the WSYue-TTS-eval benchmark, which consists of two test subsets: *Base* and *Coverage*. We report MER to measure intelligibility and speaker similarity (SIM) to evaluate voice consistency on both subsets. For MER, we transcribe the generated audio using `U2pp-Conformer-Yue`, our state-of-the-art Cantonese ASR model, and compute MER against the reference text. SIM is measured using Wespeaker (Wang et al. 2023b), which calculates the similarity between speaker embeddings of the synthetic and reference audio. Additionally, we adopt UTMOSv2 (Baba et al. 2024) to assess the overall quality of the synthesized audio.

For subjective evaluation, we randomly select Thirty samples from the *base* test set and Twenty samples from the *coverage* test set. Ten native Cantonese speakers are recruited to evaluate the five TTS systems. We conduct Mean Opinion Score (MOS) evaluations using a 5-point scale and report results with 95% confidence intervals. The MOS evaluation covers three aspects: intelligibility (I-MOS), speaker similarity (S-MOS), and accent nativeness (A-MOS). Specifically, I-MOS measures the intelligibility and consistency of the synthesized speech with respect to the input text, S-MOS evaluates the similarity between the synthesized and reference speakers, and A-MOS assesses the nativeness of the pronunciation.

TTS Results and Analysis For objective metrics, both `Llasa-1B-Yue` and `CosyVoice2-Yue` achieve substantial improvements over their pretrained counterparts. In particular, `CosyVoice2-Yue` attains the lowest MER among all systems, reducing the error rate to 10.33% on the *base* set and 9.49% on the *coverage* set, while also achieving the highest SIM scores (0.821 and 0.834), indicating superior speaker similarity. `Llasa-1B-Yue` also significantly reduces MER (10.89% and 12.78%) compared to `Llasa-1B`, and maintains competitive SIM values. In addition, UTMOSv2 scores show that both fine-tuned models generate more natural-sounding speech than zero-shot baselines.

In terms of subjective evaluation, `CosyVoice2-Yue` achieves the highest intelligibility (I-MOS: $\mathbf{4.45 \pm 0.16}$), while `Llasa-1B-Yue` outperforms all other systems in speaker similarity (S-MOS: $\mathbf{4.11 \pm 0.37}$) and accent nativeness (A-MOS: $\mathbf{4.34 \pm 0.34}$). Notably, although `CosyVoice2-Yue` obtains higher objective SIM scores, its perceived speaker similarity is lower than that of `Llasa-1B-Yue`. This may be attributed to the in-context learning inference approach in `Llasa-1B-Yue`, which leads to more natural prosody and speaking style, thus improving perceived speaker similarity. Compared to zero-shot baselines, both fine-tuned models significantly improve all MOS metrics, and even though `Edge-TTS` achieves a lower

MER, its I-MOS is lower than that of `CosyVoice2-Yue` or `Llasa-1B-Yue`, likely because it produces more mechanical and less natural-sounding speech. Overall, these results confirm the effectiveness of WenetSpeech-Yue for improving Cantonese speech synthesis.

Conclusion

In this work, we present WenetSpeech-Yue, the largest and most comprehensive open-source Cantonese speech corpus to date, together with WenetSpeech-Pipe, a scalable and modular data processing pipeline designed for building high-quality speech corpus with multi-dimensional annotations. We also release the WSYue-eval benchmark to support rigorous ASR and TTS evaluation. Built upon WenetSpeech-Yue, we obtain several ASR/TTS models with SOTA performance consistently. We believe that WenetSpeech-Yue and WenetSpeech-Pipe will serve as valuable resources for the community, facilitating future research on multi-domain Cantonese speech understanding and generation.

References

An, K.; Chen, Q.; Deng, C.; Du, Z.; Gao, C.; Gao, Z.; Gu, Y.; He, T.; Hu, H.; Hu, K.; Ji, S.; Li, Y.; Li, Z.; Lu, H.; Luo, H.; Lv, X.; Ma, B.; Ma, Z.; Ni, C.; Song, C.; Shi, J.; Shi, X.; Wang, H.; Wang, W.; Wang, Y.; Xiao, Z.; Yan, Z.; Yang, Y.; Zhang, B.; Zhang, Q.; Zhang, S.; Zhao, N.; and Zheng, S. 2024. FunAudioLLM: Voice Understanding and Generation Foundation Models for Natural Interaction Between Humans and LLMs. *CoRR*, abs/2407.04051.

Ardila, R.; Branson, M.; Davis, K.; Kohler, M.; Meyer, J.; Henretty, M.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *LREC*, 4218–4222. European Language Resources Association.

Baba, K.; Nakata, W.; Saito, Y.; and Saruwatari, H. 2024. The T05 System for the voicemos challenge 2024: Transfer Learning from Deep Image Classifier to Naturalness MOS Prediction of High-Quality Synthetic Speech. In *SLT*, 818–824. IEEE.

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

Chen, H.; Li, Z.; Xia, G.; Liu, B.; Yang, Y.; Kang, J.; and Li, J. 2024. *TeleSpeechPT: Large-Scale Chinese Multi-dialect and Multi-accent Speech Pre-training*, 183–190. Springer. ISBN 978-981-96-1044-0.

Du, Z.; Wang, Y.; Chen, Q.; Shi, X.; Lv, X.; Zhao, T.; Gao, Z.; Yang, Y.; Gao, C.; Wang, H.; Yu, F.; Liu, H.; Sheng, Z.; Gu, Y.; Deng, C.; Wang, W.; Zhang, S.; Yan, Z.; and Zhou, J. 2024. CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models. *CoRR*, abs/2412.10117.

Feng, T.; Lee, J.; Xu, A.; Lee, Y.; Lertpetchpun, T.; Shi, X.; Wang, H.; Thebaud, T.; Moro-Velázquez, L.; Byrd, D.; Dehak, N.; and Narayanan, S. 2025. Vox-Profile: A Speech Foundation Model Benchmark for Characterizing Diverse Speaker and Speech Traits. *CoRR*, abs/2505.14648.

Fiscus, J. 1997. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*.

Gao, Z.; Zhang, S.; McLoughlin, I.; and Yan, Z. 2022. Paraformer: Fast and Accurate Parallel Transformer for Non-autoregressive End-to-End Speech Recognition. In *INTERSPEECH*, 2063–2067. ISCA.

He, H.; Shang, Z.; Wang, C.; Li, X.; Gu, Y.; Hua, H.; Liu, L.; Yang, C.; Li, J.; Shi, P.; Wang, Y.; Chen, K.; Zhang, P.; and Wu, Z. 2025. Emilia: A Large-Scale, Extensive, Multilingual, and Diverse Dataset for Speech Generation. *CoRR*, abs/2501.15907.

Huang, A.; Wu, B.; Wang, B.; Yan, C.; Hu, C.; Feng, C.; Tian, F.; Shen, F.; Li, J.; and et al. 2025. Step-Audio: Unified Understanding and Generation in Intelligent Speech Interaction. *CoRR*, abs/2502.11946.

KimiTeam; Ding, D.; Ju, Z.; Leng, Y.; Liu, S.; Liu, T.; Shang, Z.; Shen, K.; Song, W.; Tan, X.; Tang, H.; Wang, Z.; Wei, C.; Xin, Y.; Xu, X.; Yu, J.; Zhang, Y.; Zhou, X.; Charles, Y.; Chen, J.; Chen, Y.; Du, Y.; He, W.; Hu, Z.; Lai, G.; Li, Q.; Liu, Y.; Sun, W.; Wang, J.; Wang, Y.; Wu, Y.; Wu, Y.; Yang, D.; Yang, H.; Yang, Y.; Yang, Z.; Yin, A.; Yuan, R.; Zhang, Y.; and Zhou, Z. 2025. Kimi-Audio Technical Report. *CoRR*, abs/2504.18425.

Lavechin, M.; Métais, M.; Titeux, H.; Boissonnet, A.; Copet, J.; Rivière, M.; Bergelson, E.; Cristià, A.; Dupoux, E.; and Bredin, H. 2023. Brouhaha: Multi-Task Training for Voice Activity Detection, Speech-to-Noise Ratio, and C50 Room Acoustics Estimation. In *ASRU*, 1–7. IEEE.

Li, L.; Li, Y.; Xue, H.; Liu, J.; Fang, S.; Wang, K.; and Xie, L. 2025. Delayed-KD: Delayed Knowledge Distillation based CTC for Low-Latency Streaming ASR. *CoRR*, abs/2505.22069.

Ma, L.; Guo, D.; Song, K.; Jiang, Y.; Wang, S.; Xue, L.; Xu, W.; Zhao, H.; Zhang, B.; and Xie, L. 2024. WenetSpeech4TTS: A 12, 800-hour Mandarin TTS Corpus for Large Speech Generation Model Benchmark. In *INTERSPEECH*. ISCA.

Meng, Y.; Li, J.; Lin, G.; Pu, Y.; Wang, G.; Du, H.; Shao, Z.; Huang, Y.; Li, K.; and Zhang, W. 2025. Dolphin: A Large-Scale Automatic Speech Recognition Model for Eastern Languages. *CoRR*, abs/2503.20212.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP*, 5206–5210. IEEE.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, 28492–28518. PMLR.

Reddy, C. K. A.; Gopal, V.; and Cutler, R. 2022. Dnsmos P.835: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors. In *ICASSP*, 886–890. IEEE.

Shao, M.; Zhu, X.; Wang, C.; Mu, B.; Li, H.; Yan, Y.; Liu, J.; Xie, D.; and Xie, L. 2025. Weakly Supervised Data Refinement and Flexible Sequence Compression for Efficient Thai LLM-based ASR. *CoRR*, abs/2505.22063.

Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; He, L.; Zhao, S.; and Wei, F. 2023a. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *CoRR*, abs/2301.02111.

Wang, H.; Liang, C.; Wang, S.; Chen, Z.; Zhang, B.; Xiang, X.; Deng, Y.; and Qian, Y. 2023b. Wespeaker: A Research and Production Oriented Speaker Embedding Learning Toolkit. In *ICASSP*, 1–5. IEEE.

Wu, D.; Zhang, B.; Yang, C.; Peng, Z.; Xia, W.; Chen, X.; and Lei, X. 2021. U2++: Unified Two-pass Bidirectional End-to-end Model for Speech Recognition. *CoRR*, abs/2106.05642.

Xiang, R.; Tan, H.; Li, J.; Wan, M.; and Wong, K.-F. 2022. When Cantonese NLP Meets Pre-training: Progress and Challenges. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; Zhang, B.; Wang, X.; Chu, Y.; and Lin, J. 2025a. Qwen2.5-Omni Technical Report. *CoRR*, abs/2503.20215.

Xu, K.; Xie, F.; Tang, X.; and Hu, Y. 2025b. FireRedASR: Open-Source Industrial-Grade Mandarin Speech Recognition Models from Encoder-Decoder to LLM Integration. *CoRR*, abs/2501.14350.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025. Qwen3 Technical Report. *CoRR*, abs/2505.09388.

Yang, Y.; Song, Z.; Zhuo, J.; Cui, M.; Li, J.; Yang, B.; Du, Y.; Ma, Z.; Liu, X.; Wang, Z.; Li, K.; Fan, S.; Yu, K.; Zhang, W.; Chen, G.; and Chen, X. 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.

Yu, T.; Frieske, R.; Xu, P.; Cahyawijaya, S.; Yiu, C. T. S.; Lovenia, H.; Dai, W.; Barezi, E. J.; Chen, Q.; Ma, X.; Shi, B. E.; and Fung, P. 2022. Automatic Speech Recognition Datasets in Cantonese: A Survey and New Dataset. In *LREC*, 6487–6494. European Language Resources Association.

Zhang, B.; Lv, H.; Guo, P.; Shao, Q.; Yang, C.; Xie, L.; Xu, X.; Bu, H.; Chen, X.; Zeng, C.; Wu, D.; and Peng, Z. 2022. WENETSPEECH: A 10000+ Hours Multi-Domain Mandarin Corpus for Speech Recognition. In *ICASSP*, 6182–6186. IEEE.