arXiv:2601.11027v1 [cs.SD] 16 Jan 2026

WenetSpeech-Wu：面向统一汉语吴方言语音处理生态系统的数据集、基准与模型

Chengyou Wang[1*]、Mingchen Shao[1*]、Jingbin Hu[1*]、Zeyu Zhu[1*]、Hongfei Xue[1],Bingshen Mu[1], Xin Xu[2], Xingyi Duan[6],Binbin Zhang[3], Pengcheng Zhu[3],Chuang Ding[4], Xiaojun Zhang[5],Hui Bu[2],Lei Xie[1†], [1]西北工业大学音频、语音与语言处理组（ASLP@NPU）[2]北京希尔贝壳科技有限公司， [3] WeNet开源社区[4]Moonstep AI， [5]西交利物浦大学， [6]包玉刚实验学校

摘要

低资源方言的语音处理仍是开发包容、稳健语音技术的一项根本挑战。尽管吴语在语言学上意义重大，且拥有庞大的说话人群体，但长期以来，由于缺少大规模语音数据、标准化评测基准和公开可用的模型，汉语吴方言的发展受到制约。在本研究中，我们推出**WenetSpeech-Wu**，这是首个面向吴语的大规模、多维度标注开源语音语料库，包含约8,000小时的多样化语音数据。在此数据集基础上，我们推出**WenetSpeech-Wu-Bench**，这是首个标准化且公开可用的基准测试，可系统评估吴语语音处理，涵盖自动语音识别（ASR）、吴语到普通话翻译、说话人属性预测、语音情感识别、文本转语音（TTS）合成以及指令跟随式文本转语音（指令式TTS）。此外，我们还发布了一套基于WenetSpeech-Wu训练的高性能开源模型，在多项任务上取得了具有竞争力的表现，并通过实证验证了所提数据集的有效性。这些成果共同为构建全面的吴方言语音处理生态系统奠定了基础。我们将所提出的数据集、基准和模型开源，以支持未来方言语音智能研究 [1]。



图1：WenetSpeech-Wu 的亮点。

1 引言

语音处理已成为人工智能的基础组成部分，使人们能够在广泛的现实应用中实现自然、高效的人机交互（Peng 等人，2025年；Xu 等人，2025a）。对于普通话和英语等资源丰富的语言，语音处理生态系统已发展得相当成熟，其特点包括大规模、多样化的数据集（Zhang 等人，2022年；Zen 等人，2019年；He 等人，2024年）、公开可用且多维度的评测基准（Sakshi 等人，2025年；Wang 等人，2025a，b），以及不断增多的强大开源模型（An 等人，2024年；Gao 等人，2022年；Radford 等人，2023）。这些组成部分共同形成了良性循环，数据、基准和模型相互促进，持续推动学术研究和实际应用的快速发展。相比之下，吴语[2]是重要的汉语方言，但在语音技术领域仍未得到充分探索，其语音处理生态系统也严重滞后。数据集、标准化基准和易于获取的模型不足，极大阻碍了研究进展和现实应用。为弥合这一差距并推动包容性语音技术的发展，本研究为构建全面的吴语语音处理生态系统奠定了基础。

吴语是汉语的一大分支，约有一亿人在上海、浙江、江苏及海外社区使用，其语言特征极为复杂。吴语保留了从上古汉语继承而来的全浊辅音系统。它

*贡献相同。 †通讯作者。[1] https://github.com/ASLP-lab/WenetSpeech-Wu-Repo

[2]https://en.wikipedia.org/wiki/Wu_汉语

表1：与 WenetSpeech-Wu 相关的典型低资源语音处理资源对比。 ✓、 △ 和 分别表示经实验验证可用、未经验证可用以及不具备该维度。缩写：Data Comp（数据构成；CN =中国，SEA = 东南亚）、Hrs（小时）、Trans（转录）、Para（副语言学）、Transl（翻译）、Q Tier（质量分级层级）、Instr TTS（指令式文本转语音）、Sp Und.（语音理解）、Wu-SH（吴语上海次方言）。

	数据集						基准测试					模型			
资源	数据构成	小时	语音转写	段落翻译	Q Tier	自动语音识别	自动语音翻译 Para	文本转语音		指令 TTS	自动语音识别	文本转语音	Sp Und	指令 TTS
KeSpeech	8 种中国口音	1.5k	✓	△	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗
GigaSpeech2	3SEA 语言	30k	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗
WenetSpeech-Yue	中国粤语	21k	✓	△	✗	✗	✓	✗	✗	✓	✗	✓	✓	✗	✗
WenetSpeech-Chuan	中国四川话	10k	✓	△	✗	✗	✓	✗	✗	✓	✗	✓	✓	✗	✗
Magicdata-Shanghai	中国上海吴语	4	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
WenetSpeech-Wu	中国吴语	8k	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

在连续语流中具有高度复杂的连读变调系统(Rose，2015年)，这两者都给语音建模带来了巨大挑战。此外，其各个次方言之间也存在显著差异，例如上海话、苏州话和杭州话，这进一步增加了开发稳健语音处理系统的难度。

尽管吴语具有重要的语言学意义，且使用者众多，但其语音处理资源仍严重匮乏。从数据角度来看，现有开源资源在规模和覆盖范围上都极为有限：唯一公开可用的数据集 MagicData- Shanghai[3]，仅提供 4.19 小时带标注的上海话语音，用于自动语音识别（自动语音识别）；其他吴语次方言均无公开数据集。此外，高质量的情感标注、说话人属性以及表现力语音—文本对也完全缺失，而这些资源对于支持各类语音处理任务至关重要。从评测角度来看，缺乏公开可用的基准，导致无法对不同方法进行公平比较和系统评估。在模型层面，即使是自动语音识别和文本转语音合成（文本转语音）模型等基础语音处理系统，也要么无法获取，要么表现不佳，使得开源系统和商用系统基本都无法用于吴语应用。总而言之，这些局限凸显了建立全面的吴语语音处理生态系统的迫切需求，以促进方言语音研究并推动开源社区发展。

为应对这些挑战，我们推出了一项全面的开源计划，为吴语语音处理生态系统奠定基础。具体而言，我们构建了**WenetSpeech-Wu**，这是首个面向吴语的大规模、多维度标注开源语料库，涵盖八种吴语次方言，时长约 8,000小时。在此数据集基础上，我们推出 **WenetSpeech-Wu-Bench**，这是首个标准化且公开可用的基准，旨在系统评估吴语语音处理，涵盖自动语音识别、吴语到普通话自动语音翻译（自动语音翻译）、说话人属性预测（性别和年龄）、语音情感识别、文本转语音以及指令跟随文本转语音（指令式文本转语音）。此外，我们还发布了一系列基于 WenetSpeech-Wu 训练的高性能开源模型，包括自动语音识别、文本转语音、统一语音理解和指令式文本转语音模型。这些模型大幅超越现有开源系统和商用系统，从而确立了强大的模型基线，并通过实验证明了所提出数据集的有效性。总之，这些贡献为建立全面且公开可用的吴语语音处理生态系统奠定了基础。

我们的主要贡献有三点：

- 我们发布了**WenetSpeech-Wu**，这是首个面向吴语的大规模、多维标注开源语料库，包含约 8,000小时语音数据。

- 我们推出了 **WenetSpeech-Wu-Bench**，这是首个用于评估吴语语音处理的标准化基准测试，涵盖自动语音识别、吴语到普通话自动语音翻译、说话人属性预测、语音情感识别、文本转语音以及指令式文本转语音任务。

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

•我们开源了基于 WenetSpeech-Wu 训练的强大模型，其性能较现有系统取得了显著提升，验证了所提出数据集的有效性。

2 相关工作

2.1 低资源语言和汉语方言语音数据集

近年来，为缓解数据匮乏并推动包容性语音技术的发展，人们投入了大量精力，为低资源语言和方言构建语音语料库（Shao 等人，2025年）。Common Voice（Ardila 等人，2020年）、GigaSpeech2（Yang 等人，2024年）和 KeSpeech（Tang 等人，2021年）等具有代表性的大规模语料库，极大地推动了低资源或代表性不足语言的语音处理研究。在汉语方言中，近年来也发布了若干面向方言的语音资源，包括针对粤语的 WenetSpeech-Yue（Li 等人，2025年）和针对四川话的 WenetSpeech-Chuan（Dai 等人，2025年）。然而，这些数据集主要聚焦于自动语音识别，在一定程度上也涉及文本转语音：Common Voice 和 GigaSpeech2 仅提供转写文本；KeSpeech 在转写文本的基础上增添了方言和基本说话人标签；WenetSpeech-Yue 和 WenetSpeech-Chuan 虽然包含副语言标注，但仅通过自动语音识别和文本转语音实验评估其资源，未能验证这些副语言标签的有效性。因此，这类资源在更广泛的语音处理任务中的可用性仍然有限，例如语音情感识别、自动语音翻译和指令式文本转语音。

尽管吴语在语言学和实际应用方面都具有重要意义，但公开可用的语音数据资源仍严重匮乏。据我们所知，MagicData-Shanghai 是唯一开源的吴语数据集，仅提供了 4.19 小时带标注的上海话语音，用于自动语音识别。如此严重的数据短缺，加上标注类型单一，极大地阻碍了吴语语音处理研究的进展。

2.2 吴语语音处理：模型与评估

早期吴语语音处理研究主要聚焦于单一方言的自动语音识别或文本转语音，尤其是上海话，而在很大程度上忽略了苏州话和杭州话等其他主要变体。近期研究将吴语作为较小的组成部分纳入大规模多语言自动语音识别或文本转语音系统，包括 Dolphin（Meng 等人，2025年）、Qwen3-ASR、Qwen3-TTS、Step-Audio2（Wu 等人，2025年）以及 Qwen3-Omni（Xu 等人，2025a）。在这些系统中，吴语语音被视为优先级较低的子集，而非主要研究目标，因此基础自动语音识别和文本转语音模型的性能仍不足以满足吴语实际应用的需求。除自动语音识别和文本转语音外，针对吴语的语音理解模型和指令式文本转语音系统也基本未得到探索。

更重要的是，即使对于基础的自动语音识别和文本转语音任务，目前也没有公开且标准化的评测基准。因此，现有研究通常在私有测试集上评估模型，严重阻碍了公平比较、结果复现和系统性进展。

总之，如表1所示，这些局限促使我们推出 WenetSpeech-Wu，旨在通过协同推进数据集、基准和模型，建立完整的吴方言语音处理生态系统。

3 WenetSpeech-Wu

3.1 数据构建流程

我们提出了一套自动化、可扩展的流程，用于构建带有多维标注的大规模吴语语音数据集，如图2所示。该流程旨在实现高效的数据收集、稳健的自动转录以及多样化的下游标注。

数据收集与筛选。我们从不同领域和次方言中收集大规模、自然场景下的吴语语音。首先基于元数据过滤剔除非吴语数据，随后使用 WebRTC 语音活动检测进行分段。我们还利用 DNSMOS 和信噪比（SNR）进行质量筛选，最终得到高质量的语音语料库。

标注工具构建。 为支持大规模自动转录，我们使用 880 小时的人工标注吴语语音对两个预训练自动语音识别模型进行微调。具体而言，我们对基于连接主义时序分类（CTC）的方言自监督学习（SSL）模型 Tele-CTC-FT[4]，a，以及 Step-

[4]https://huggingface.co/Tele-AI/TeleSpeech-自动语音识别1.0

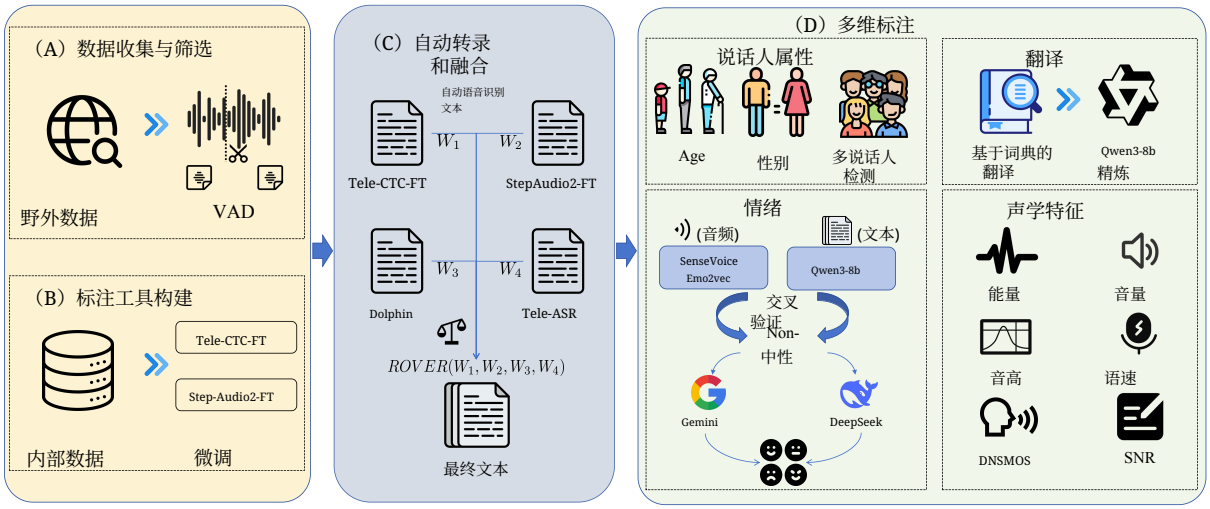

图 2：WenetSpeech-Wu 的数据构建流水线。

Audio2-FT（Wu 等人，2025年）进行微调。这些模型被用作吴语方言转录的互补自动标注器。

自动转录与融合。我们采用识别器输出投票错误消减（ROVER）（Fiscus，1997）来融合多个自动语音识别系统的转录结果。具体而言，我们结合两个经过微调的吴语方言自动语音识别模型 Dolphin（Meng 等人，2025年）和 TeleASR（Chen 等人，2024年）的输出，并通过网格搜索确定模型权重。融合结果提供带置信度分数的最终转写文本。

多维度标注。每个语音样本均在多个维度上进行标注。说话人属性（包括性别和年龄）由 VoxProfile（Feng 等人，2025年）推断，多说话人存在情况则使用 Pyannote（Bredin，2023）检测。吴语到普通话的翻译通过基于词典的映射生成，并使用 Qwen3-8B（Yang 等人，2025年）进一步润色。Qwen3-8B 是一种大语言模型，旨在提供流畅的标准普通话参考文本。情感标注通过多阶段跨模态流程获得。首先，使用 SenseVoice（An 等人，2024年）和 Emo2Vec（Wang 等人，2020年）对声学信号进行预测，并使用 Qwen3-8B 分析文本内容。对于经共同判定为非中性的样本，进一步使用 DeepSeek-R1 根据文本进行分析，并使用 Gemini-2.5-Pro 根据声学信息进行分析，最终标签由两者的交集确定。此外，使用 Dataspeech（Lyth 和 King，2024年）提取语速、响度、能量和音高等韵律声学特征，以便支持语音生成任务。

3.2 数据集

WenetSpeech-Wu 是首个带有多维标注的大规模吴语语音语料库。它包含丰富的元数据和标注，包括带置信度分数的转写文本、吴语到普通话的翻译、领域和次方言标签、说话人属性、情感标注以及音频质量指标。该数据集包含约 8,000小时来自不同领域的语音，覆盖八种吴语次方言。为支持对质量要求各异的多种语音处理任务，我们进一步采用任务特定的数据质量分级策略。下文将详细介绍数据集统计信息、领域和方言覆盖范围、标注分布以及质量控制机制。

时长和置信度分布。
WenetSpeech-Wu包含8,000小时的语音，共有386万条话语，单条话语时长最长可达30秒，平均时长为7.45秒。我们使用加权ROVER生成的转录置信度作为标注质量的衡量指标，并保留置信度分数高于0.55的话语。话语时长和转录置信度的详细分布如图3b和图3c所示。

领域和次方言覆盖范围。
WenetSpeech-Wu 覆盖广泛的语音领域和吴语次方言。领域包括新闻、文化、视频博客、娱乐、教育、播客、评论、访谈、广播剧、音乐节目和有声读物，其分布情况如下

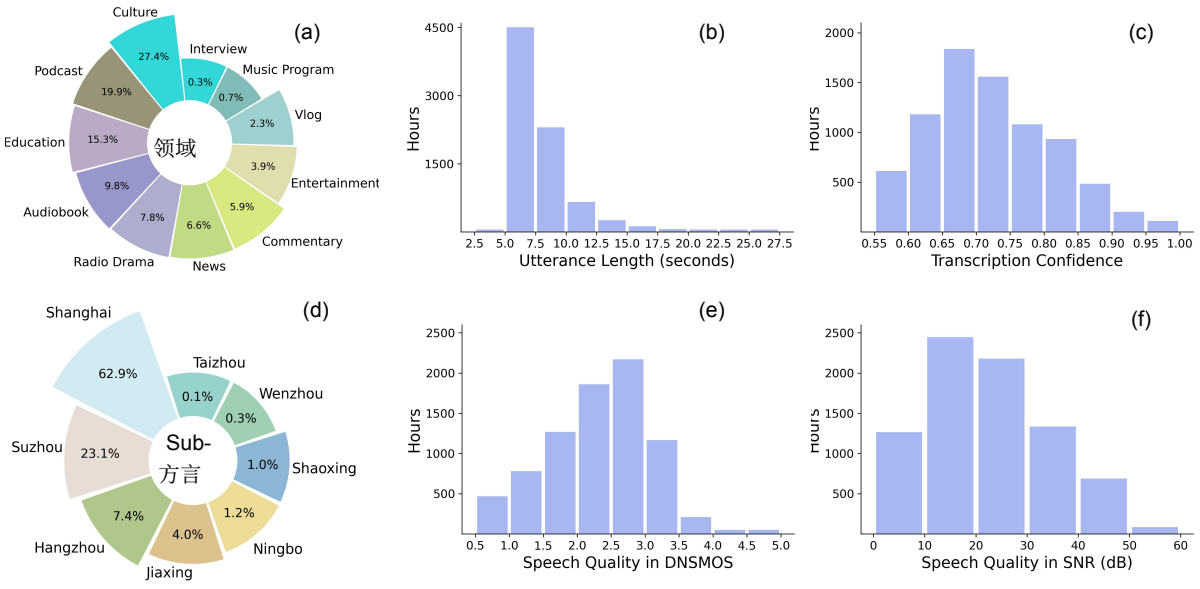

图3：WenetSpeech-Wu 的统计概览。

表2：WenetSpeech-Wu 中说话人属性和情感标注的统计数据。

Task	类别	小时
性别	Male	4,135
	女性	1,331
Age	青少年	372
	青年	1,673
	中年	2,003
	老年	1,418
情绪	中性	5,102
	开心	73
	Sad	81
	愤怒	109
	惊讶	101

分布情况见图3a。就方言覆盖范围而言，约 37% 的录音无法可靠地归属到某一种吴语次方言，因此标记为未知。其余录音涵盖多种已识别的吴语次方言，包括上海话、苏州话、绍兴话、宁波话、杭州话、嘉兴话、台州话和温州话，其分布情况如图3d 所示。

音频质量。如图3e 和图 3f 所示，大多数话语的信噪比介于 10 至 40 dB 之间，峰值为 20–30 dB。平均意见分主要在 2.0–3.5 之间。

说话人属性和情感标注。我们为单说话人片段标注性别、年龄和情绪标签。性别分为两类：男性和女性；年龄分为四组：0–17 岁为青少年，18–35 岁为青年，36–59 岁为中年，以及 60+为老年；情绪分为五类：中性、开心、悲伤、惊讶和愤怒。各类别的分布如表2所示。

任务特定数据质量分级。为支持异构语音任务的实际训练，我们采用与任务特定质量要求相匹配的任务特定数据质量分级策略。对于自动语音识别和文本转语音，我们构建了两个质量等级。普通质量子集面向大规模预训练，优先考虑数据覆盖范围和多样性，只要求转录置信度达到中等水平。相比之下，高质量子集面向监督微调（SFT），采用更严格的筛选标准，包括高转录置信度、纯净的声学条件以及可靠的说话人分段，以提供强而稳定的监督信号。对于对标注噪声和语义歧义更敏感的任务，包括吴语到普通话自动语音翻译、说话人属性预测、语音情感识别、文本转语音和指令式文本转语音，我们采用严格的筛选标准，例如单说话人录音、高平均意见得分、高信噪比、较大的音高标准差以及经核实的标注一致性，如表3所示。

4 WenetSpeech-Wu-Bench

我们推出**WenetSpeech-Wu-Bench**，这是首个公开发布、经人工整理的吴语语音处理基准测试，涵盖自动语音识别，

表3：WenetSpeech-Wu 中的任务特定数据选择和质量等级。缩写：Q Tier（质量分级层级）、Text Conf（转录置信度）、Spk（说话人类型，Mono = 单说话人）、Pitch Std.（音高标准差）、Expr（通过情感标签获取的表现力数据）、Spk Attr（说话人属性）、Emo（情感）、Inst Pro（指令式TTS韵律控制）、Inst Emo（指令式TTS情感控制）。

任务	时长	Q Tier	Text Conf	说话人	信噪比	音高 Std.	Expr
ASR	7388	Mid	> 0.6	–	–	–	–
	795	High	> 0.85	–	–	–	–
AST	795	High	> 0.85	–	–	–	–
Spk Attr	2986	高	–	单声道	–	–	–
Emo	500	High	–	单声道	>10	>50	✓
TTS	7388	Mid	> 0.6	–	–	–	–
	1500	高	> 0.65	单声道	>10	>50	–
Inst Pro	679	High	> 0.7	单声道	>30	–	–
Inst Emo	161	High	> 0.7	单声道	>10	>50	✓

吴语到普通话自动语音翻译、说话人属性、情感识别、文本转语音和指令式文本转语音，并提供统一平台以进行公平评估。

自动语音识别。 WenetSpeech-Wu-Bench 的 ASR 测试集包含 9.75 小时的上海话、苏州话和普通话语码混合语音，涵盖单说话人和多说话人场景。评估采用字符错误率（CER），并报告字符级错误以供详细分析。

吴语到普通话语音翻译。 我们构建了一个吴语到普通话自动语音翻译测试集。该测试集包含 3000 条吴语话语，总计 4.4 小时，并配有经过人工核验的标准普通话翻译，涵盖多个领域。翻译质量使用 BLEU分数进行评估 [5]。

说话人属性预测和语音情感识别。 该测试集用于评估吴语语音的年龄、性别和情绪预测。说话人属性方面，性别分为男性和女性，每个类别有 1,500 个样本。年龄分为四组：17 岁及以下的青少年、18 至 35 岁的青年、36 至 59 岁的中年人，以及 60 岁及以上的老年人，每组各有 500 个样本。情绪方面，包括 300 个中性样本、200 个开心样本、100 个悲伤样本、200 个愤怒样本和 200 个惊讶样本，总计 1,000 个样本。性能通过各类别准确率和总体分类准确率进行衡量。

文本转语音。 作为 WenetSpeech-Wu-Bench 的一部分，我们构建了一个 TTS 测试集，其中包含 144 条简单测试句和 98 条困难测试句。随后，专业吴语专家对文本进行了进一步审核和润色。提示音频样本选自开源数据集 Magicdata-Shanghai，并依据严格的筛选标准选取了 12 名吴语说话人。在评估中，我们使用基于 WeSpeaker [6] 的说话人嵌入相似度来测量说话人相似度，同时使用我们提出的 Step-Audio2-Wu-ASR 模型，通过计算字符错误率来评估可懂度。此外，我们还开展了主观听评测试，包括可懂度平均意见得分（IMOS）、相似度平均意见得分（SMOS）和口音平均意见得分（AMOS）。主观评估由 23 名听评员完成，每人对 20 个选定样本进行评分。

指令式文本转语音。 WenetSpeech-Wu-Bench 包含两个用于指令式文本转语音的评测集。韵律控制测试集由五条语音提示组成，每条提示均以适中的语速和正常的基频说出，并合成为 20 个语句，且语速和音高各不相同。评测分为两项实验：一项采用较快语速和较高音高，另一项采用较慢语速和较低音高。这些样本由 Dataspeech 自动标注（Lyth 和 King，2024年）。如果一对样本的相对语速和音高符合预期指令，则记为 1 分，否则记为 0 分。韵律分类指标为所有样本对的平均得分，用于反映模型遵循韵律指令的能力（Kuan 等人，2023）。情感控制测试集用于评估模型遵循情绪相关指令的能力。我们选取 10 条不含明确情感表达的参考提示。针对每条提示，为四种目标情绪——愤怒、悲伤、快乐和惊讶——分别合成 50 个语句。我们使用 Step-Audio2-Wu-Und 模型对样本进行评估。如果模型预测的情绪与预期目标情绪一致，则样本判定为正确，并将平均分类准确率作为评测指标（Gao 等人，2025年）。此外，我们还开展主观听音测试，以评估指令遵循质量。听评员使用两个

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

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

表 4：各测试集上的自动语音识别结果（字符错误率 [illegible] 和深 [illegible] 表示开源基线、商业模型、在 WenetSpeech-Wu 上训练的自动语音识别模型，以及在内部数据上训练的标注模型。**粗体**数字表示最佳结果；带下划线的数字表示次佳结果。

模型	内部数据		WS-Wu-Bench
	对话朗读		ASR
自动语音识别模型			
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**
标注模型			
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

指标：韵律平均意见得分（PMOS）（Chan 和 Kuang，2025年）和情感平均意见分（EMOS）（Cho 等人，2025年）。评测由 23 名听评员参与，每人对 15 个样本进行评分，以判断合成语音在多大程度上遵循了预期的情感指令。

5 模型与实验

为解决吴语语音处理模型匮乏的问题，我们基于 WenetSpeech-Wu，针对语音处理的两个核心方面——语音理解和语音生成——开发了相应的模型。其中，语音理解方面包括自动语音识别模型和统一语音理解模型；语音生成方面则包括文本转语音模型和指令式文本转语音模型。

5.1 语音理解

自动语音识别模型。为适应不同的应用场景，我们开发了三种不同规模的吴语方言自动语音识别模型：Conformer-U2pp-Wu（Wu 等人，2021年）模型、Whisper-Medium-Wu（Radford 等人，2023）模型，以及 Step-Audio2-Wu-ASR 模型。Conformer-U2pp-Wu 和 Whisper-Medium-Wu 模型均基于 Wenet 框架（Yao 等人，2021年）训练，其中 Conformer-U2pp-Wu 从头开始训练，Whisper-Medium-Wu 则基于公开的预训练权重进行微调。Step-Audio2-Wu-ASR 模型是在 MS-Swift 框架（Zhao 等人，2025年）中使用 LoRA（Hu 等人，2021年）进行参数高效微调的 Step-Audio2-mini。所有模型均在 ASR-中子集上进行预训练，并在 ASR-高子集上进行监督微调。评估采用 WenetSpeech-Wu-Bench 的 ASR 测试集，以及两个内部人工标注的测试集，涵盖对话和朗读场景，从而能够全面评估模型在不同说话条件下的表现。

表5：WenetSpeech-Wu-Bench 上的语音理解性能。

模型	自动语音识别	自动语音翻译	性别	年龄	情绪
Qwen3-Omni	44.27				0.667
Step-Audio2-mini	26.72				0.460
Step-Audio2-Wu-Un					**0.712**

作为对比，我们纳入了开源和商业基线，包括 Dolphin、Sen- seVoice、Paraformer（Gao 等人，2022年）、Fir- eRedASR（Xu 等人，2025b）作为开源基线，以及 Qwen3-ASR、Gemini-2.5-Pro、Tele- ASR（Chen 等人，2024年）和 Tencent-Cloud- ASR 作为商业基线。对于 Conformer-U2pp-Wu、Whisper-medium-Wu 和 Step-Audio2-Wu-ASR，两个内部测试集均属于严格的集外评测。反过来，WS-Wu- Bench-ASR 测试集则是 Step-Audio2-FT 和 Tele-CTC-FT 的集外评测集，这两个模型仅使用内部数据训练。

如表4所示，所有现有的开源和商业自动语音识别系统在全部三个测试集上的表现都很差，表明它们不适用于吴语识别。相比之下，在 WenetSpeech-Wu 上训练的模型在所有规模下均达到最先进的性能，其中最小的 Conformer-U2pp-Wu 也显著优于此前的所有系统。

语音理解模型。对于语音理解，我们使用任务感知且按质量分级的数据，在 WenetSpeech-Wu 语料库上训练 Step-Audio2-Wu-Und 模型。该模型首先使用 ASR-Mid、高质量自动语音翻译以及年龄、性别和情绪标注进行预训练，随后结合 ASR-high子集和相同的非自动语音识别任务进行微调。

在基准测试中，我们将其与包括 Step-Audio2-mini 和 Qwen3- Omni 在内的基线模型进行比较，用于语音理解任务。如表5所示，经过多任务微调后，Step-Audio2-Wu-Und 的自动语音识别性能较 Step-Audio2-Wu-ASR 略有

表 6：WenetSpeech-Wu-Bench 上的文本转语音结果。**粗体**和带下划线的数值分别表示最佳和次佳结果；浅绿色行表示在 WenetSpeech-Wu 上训练或在内部高质量数据集上进一步微调的模型。

模型	WS-Wu-Eval-TTS-easy					WS-Wu-Eval-TTS-hard				
	字符错误率(%)↓	相似度↑	IMOS↑	相似度平均意见得分↑	平均意见得分↑	字符错误率(%)↓	相似度↑	IMOS↑	相似度平均意见得分↑	平均意见得分↑
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

† 采用单个固定说话人的商用系统，不考虑说话人相似度。* 单说话人微调模型，不评估说话人相似度。

下降，但仍取得了第二优的结果。在吴语到普通话 AST 任务上，其性能显著优于基线模型。与 Step-Audio2-mini 的比较说明，在性别、年龄和情绪预测方面，吴语与普通话之间存在不匹配，而我们的数据有效解决了这一问题。与 Qwen3-Omni 相比，我们的模型在年龄和情绪预测方面取得了显著提升，但在性别分类方面的表现略低。

5.2 语音生成

文本转语音模型。 在持续预训练（CPT）阶段，我们从开源 CosyVoice2（Du 等人，2024年）检查点继续训练，使用 TTS-Mid 数据集训练十轮。在监督微调阶段，我们使用 TTS-High 数据集训练三轮。最后，我们使用一个 10 小时的内部高质量数据集进行说话人专属微调。

如表 6所示，实验结果表明，分阶段训练策略显著提升了 CosyVoice2 在吴语能力方面的语音合成表现。CPT 阶段利用大规模流水线处理数据，增强了模型的基础能力和鲁棒性，尤其是在具有挑战性的样本上。监督微调阶段进一步提升了语音自然度和表现力。最后，单说话人监督微调（SS-SFT）阶段在字符错误率、IMOS 和 AMOS 等指标上取得了最佳表现。总体而言，CosyVoice2-Wu-SS 在大多数指标上接近或超越基线模型 Qwen3-TTS、DiaMoE-TTS（Chen 等人，2025年）和 CosyVoice2，尤其是在困难样本合成方面。

指令式文本转语音模型。 指令训练数据来自 Inst Pro 和 Inst Emo 数据集，如表 3所示。训练时使用指令文本和标签，以较小的学习率训练五轮。与指令微调前的模型相比，结果显示，模型在 WenetSpeech-Wu-Bench 的所有可控性指标上均有明显提升，如表 7所示；主观听音测试也证实了显著的感知控制效果，验证了所提数据的有效性。

表 7：指令式文本转语音模型的性能。

Type	指标	CosyVoice2-Wu-SFT	CosyVoice2-Wu-指令
情绪	开心↑	0.87	**0.94**
	愤怒↑	0.83	**0.87**
	Sad↑	0.84	**0.88**
	惊讶↑	0.67	**0.73**
	EMOS↑	3.66	**3.83**
韵律	音高↑	0.24	**0.74**
	语速↑	0.26	**0.82**
	PMOS↑	2.13	**3.68**

6 结论

在本研究中，我们构建了一个涵盖数据集、基准测试和模型的吴语语音处理生态系统。我们推出了 WenetSpeech-Wu，这是一个包含 8,000 小时语音、覆盖多个吴语次方言且具有丰富多维标注的大规模语料库，并发布了首个公开的吴语基准，涵盖自动语音识别、吴语到普通话自动语音翻译、语音属性与情绪识别、文本转语音以及指令式文本转语音。基于这一数据集，我们发布了一系列自动语音识别、文本转语音、统一语音理解模型和指令式文本转语音模型，以支持社区研究，并展示了 WenetSpeech-Wu 的有效性。

局限性

尽管 WenetSpeech-Wu 为方言语音处理社区作出了贡献，但仍有若干局限有待未来工作解决。尽管 WenetSpeech-Wu 规模庞大且覆盖多个吴语次方言，但不同方言和领域的数据分布并不完全均衡，这可能会影响模型对代表性较低变体的泛化能力。此外，许多标注是通过自动化或半自动化流程生成的。尽管我们实施了严格的质量控制和过滤策略，与完全人工标注相比，这些标注仍可能含有噪声。最后，我们的基线模型旨在提供强大且可复现的参考基准，而非针对特定任务的最优解决方案；未来工作可通过更具针对性的建模和训练策略进一步提升性能。

参考文献

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等14人。2024年。Funau-diollm：用于实现人与大语言模型自然交互的语音理解与生成基础模型。*CoRR*，abs/2407.04051。

Rosana Ardila、Megan Branson、Kelly Davis、Michael Kohler、Josh Meyer、Michael Henretty、Reuben Morais、Lindsay Saunders、Francis M. Tyers 和 Gregor Weber。2020年。Common Voice：一个大规模多语言语音语料库。发表于*Proc.LREC*，第4218–4222页。

Hervé Bredin。2023。pyannote.audio 2.1 说话人分离流程：原理、基准测试与实践指南。发表于*Proc. INTERSPEECH*，第1983–1987页。

Cedric Chan 和 Jianjing Kuang。2025年。《迈向客观且可解释的文本转语音韵律评估：一种基于语言学的方法》。*arXiv*，2511.02104。

Hongjie Chen、ZehanLi、Guangmin Xia、Boqing Liu、Yan Yang、Jian Kang 和 Jie Li。2024年。*Tele-SpeechPT*：大规模中文多方言、多口音语音预训练，第183–190页。施普林格。

Ziqi Chen、Gongyu Chen、Yihua Wang、Chaofan Ding、Zihao Chen 和 Wei-Qiang Zhang。2025年。Diamoe-tts：一种基于统一国际音标的方言文本转语音框架，结合专家混合与参数高效的零样本适配。*CoRR*，abs/2509.22727。

Deok-Hyeon Cho、Hyung-Seok Oh、Seung-Bin Kim 和 Seong-Whan Lee。2025年。Emosphere++：通过情感自适应球面向量实现可控情绪的零样本文本转语音。*IEEETrans. Af-fect. Comput.*，16(3)：2365–2380。

Yuhang Dai、Ziyu Zhang、王帅、Longhao Li、Zhao Guo、Tianlun Zuo、Shuiyuan Wang、Hongfei Xue、Chengyou Wang、Qing Wang、Xin Xu、Hui Bu、Jie Li、Jian Kang、Binbin Zhang 和 Lei Xie。2025年。WenetSpeech-Chuan：一个带有丰富标注、用于方言语音处理的大规模四川话语料库。*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、Shil- iang Zhang、Zhijie Yan 和 Jingren Zhou。2024年。Cosyvoice 2：基于大语言模型的可扩展流式语音合成。*CoRR*，abs/2412.10117。

Tiantian Feng、Jihwan Lee、Anfeng Xu、Yoonjeong Lee、ThanathaiLertpetchpun、XuanShi、Helin Wang、Thomas Thebaud、Laureano Moro-Velázquez、Dani Byrd、Najim Dehak 和 Shrikanth Narayanan。2025年。Vox-profile：用于表征多样说话人及语音特征的语音基础模型基准。*CoRR*，abs/2505.14648。

J.G. Fiscus。1997年。一种可降低词错误率的后处理系统：识别器输出投票错误消减（rover）。发表于*ASRU* 会议论文集。

Changfeng Gao、Zhihao Du 和 Shiliang Zhang。2025年。面向基于 LLM 的文本转语音系统的可微奖励优化。发表于*Interspeech* 会议论文集。

Zhifu Gao、Shiliang Zhang、Ian McLoughlin 和 Zhijie Yan。2022年。Paraformer：用于非自回归端到端语音识别的快速精准并行变换器。发表于*Interspeech* 会议论文集，第2063–2067 页。

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 和 Zhizheng Wu。2024年。Emilia：面向大规模语音生成的广泛、多语言且多样化的语音数据集。发表于*SLT* 会议论文集，第 885–890 页。

Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang 和 Weizhu Chen。2021年。LoRA：大语言模型的低秩适配。*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 和 Hung-Yi Lee。2023。迈向通用文本指令引导的语音转换。发表于*ASRU* 论文集，第 1–8 页。

Longhao Li、Zhao Guo、Hongjie Chen、Yuhang Dai、Ziyu Zhang、HongfeiXue、TianlunZuo、Chengyou Wang、Shuiyuan Wang、Jie Li、Jian Kang、Xin Xu、Hui Bu、Binbin Zhang、Ruibin Yuan、Ziya Zhou,

薛伟和Lei Xie。2025年。WenetSpeech-Yue：一个带有多维标注的大规模粤语语音语料库。*CoRR*，abs/2509.03959。

Daniel Lyth 和 Simon King。2024年。利用合成标注对高保真文本转语音进行自然语言引导。*CoRR*，abs/2402.01912。

孟洋洋、Jinpeng Li、林国栋、蒲宇、王冠博、杜虎、邵志明、黄宇凯、Ke Li 和 Wei-Qiang Zhang。2025年。Dolphin：面向东方语言的大规模自动语音识别模型。*CoRR*，abs/2503.20212。

彭静、王宇成、李博涵、郭一伟、Han-kun Wang、方彦贵、席宇、李浩宇、李旭、张可、王帅和 Kai Yu。2025年。语音大语言模型理解能力综述。*IEEE信号处理选题期刊*，第1–32页。

Alec Radford、Jong Wook Kim、徐涛、Greg Brock-man、Christine McLeavey 和 Ilya Sutskever。2023。通过大规模弱监督实现鲁棒语音识别。载于国际机器学习大会论文集，第202卷，第28492–28518页。

Philip Rose。2015年。三个汉语吴方言中的声调化。载于国际语音科学大会论文集。

S. Sakshi、Utkarsh Tyagi、Sonal Kumar、Ashish Seth、Ramaneswaran Selvakumar、Oriol Nieto、Ramani Du- raiswami、Sreyan Ghosh 和 Dinesh Manocha。2025年。MMAU：大规模多任务音频理解与推理基准测试。载于国际学习表征会议。Mingchen Shao、Bingshen Mu、Chengyou Wang、李海洲、闫颖、傅中华和Lei Xie。2025年。面向低资源语言多任务理解的语音大语言模型构建。*CoRR*，abs/2509.14804。唐志远、王栋、徐阳光、孙建伟、雷晓宁、赵帅江、温成、谭兴军、谢传东、周舒然、闫瑞、Chen- jia Lv、韩洋、邹伟和李向刚。2021年。KeSpeech：一个普通话及其八种次方言的开源语音数据集。载于神经信息处理系统大会。王丁东、吴锦岑、李俊安、杨东超、陈雪媛、张天华和孟海伦。2025a。MMSU：大规模多任务口语理解与推理基准测试。*CoRR*，abs/2506.04779。王帅、孙兆凯、林振南、Chengyou Wang、潘舟和Lei Xie。2025b。MSU-Bench：迈向对话式多说话人场景理解。*CoRR*，abs/2508.08155。王硕、Aishan Maoliniyazi、吴欣乐和Xi- aofeng Meng。2020年。Emo2vec：通过多情感类别学习情感嵌入。*ACM互联网技术汇刊*，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，以及其他81位作者。2025年。Step-audio 2 技术报告。*CoRR*，abs/2507.16632。

Di Wu、Binbin Zhang、Chao Yang、Zhendong Peng、Wenjing Xia、Xiaoyu Chen 和 Xin Lei。2021年。U2++：用于语音识别的统一双遍双向端到端模型。*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 等另外 19 人。2025a。Qwen3-Omni 技术报告。*CoRR*，abs/2509.17765。

Kaituo Xu、Feng-Long Xie、Xu Tang 和 Yao Hu。2025b。FireRedASR：从编码器-解码器到大语言模型集成的开源工业级普通话语音识别模型。*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、Day- iheng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang 等另外 40 人。2025年。Qwen3 技术报告。*CoRR*，abs/2505.09388。Yifan Yang、Zheshu Song、Jianheng Zhuo、Mingyu Cui、Jinpeng Li、Bo Yang、Yexing Du、Ziyang Ma、Xuny- ing Liu、Ziyuan Wang、Ke Li、Shuai Fan、Kai Yu、Wei-Qiang Zhang、Guoguo Chen 和 Xie Chen。2024年。Gigaspeech 2：一个不断发展的、大规模的多领域低资源语言自动语音识别语料库，采用自动爬取、转录和精炼。*CoRR*，abs/2406.11546。Zhuoyuan Yao、Di Wu、Xiong Wang、Binbin Zhang、Fan Yu、Chao Yang、Zhendong Peng、Xiaoyu Chen、Lei Xie 和 Xin Lei。2021年。Wenet：面向生产的流式和非流式端到端语音识别工具包。收录于*Interspeech* 会议论文集，第 4054–4058 页。Heiga Zen、Viet Dang、Rob Clark、Yu Zhang、Ron J. Weiss、Ye Jia、Zhifeng Chen 和 Yonghui Wu。2019年。LibriTTS：一个源自 LibriSpeech 的文本转语音语料库。收录于 *Interspeech* 会议论文集，第 1526–1530 页。Binbin Zhang、Hang Lv、Pengcheng Guo、Qijie Shao、Chao Yang、Lei Xie、Xin Xu、Hui Bu、Xiaoyu Chen、Chenchen Zeng、Di Wu 和 Zhendong Peng。2022年。WENETSPEECH：一个用于语音识别的 10000+ 小时多领域普通话语料库。收录于*ICASSP* 会议论文集，第 6182–6186 页。Yuze Zhao、Jintao Huang、Jinghan Hu、Xingjun Wang、Yunlin Mao、Daoze Zhang、Zeyinzi Jiang、Zhikai Wu,

Baole Ai、Ang Wang、Wenmeng Zhou 和 Yingda Chen。2025年。SWIFT：一种可扩展的轻量级微调基础设施。收录于*AAAI* 会议论文集，第 29733–29735 页。

A 附录

A.1 数据集详情

标注数据示例。 每个语音样本均采用统一的 JSON 格式表示。一条记录包含唯一键、音频时长、吴语方言转录及其普通话翻译，以及相应的置信度分数。领域和次方言、音频质量指标（例如 DNSMOS 和 SNR）以及副语言标注（情绪、年龄和性别）等上下文元数据，均整理在结构化字段中。此外，还提供包括语速、响度、基频统计和能量统计在内的韵律声学特征，以支持细粒度的声学分析。图 4展示了标注 JSON 条目的示例。


```
{
  "key": "mx9ok0979_qtmvx_8765690",
  "duration": "2.3",
  "转录": "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": "访谈",
    "sub-dialect": "上海",
  },
  "音频质量    ": {
    "DNSMOS": "3.98",
    "信噪比": "58.74",
  },
  "副语言": {
    "emotion": "开心",
    "age": "青年",
    "gender": "男性",
  },
 "低级特征": {
    "语音_rate": 15.094,
    "响度": -17.72,
    "f0_mean": 167.17,
    "f0_std": 40.74,
    "能量_mean": 0.0684,
    "能量_std": 0.078,
  }
}
```


图 4：JSON 格式标注数据条目示例

A.2 WenetSpeech-Wu-Bench 统计信息

如表 8所总结，WenetSpeech-Wu 基准包含多个多样化的测试集，针对多种语音相关任务量身设计。对于每项任务，我们报告话语数量及相应的总时长（小时）。

A.3 实验详情

优化与训练超参数。不同模型的训练超参数汇总于表 9。

补充实验结果。 为了更详细地分析模型在 WS-Wu-Bench 上的性能，我们报告了 WS-Wu-Und 和基线模型在性别、年龄和情绪方面按类别划分的分类准确率。如图 5所示，我们的模型在不同副语言类别上的表现更加均衡且始终更高，展现出更强的副语言识别能力。

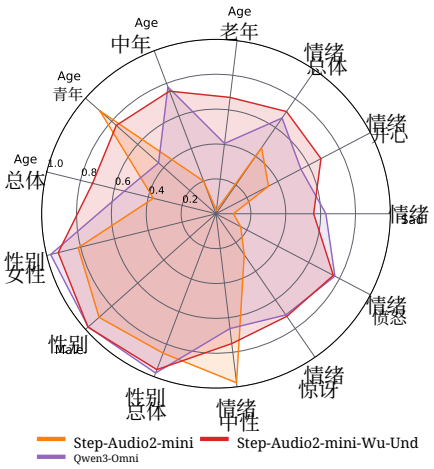

图 5：年龄、性别和情感识别任务的补充对比结果。

实验中使用的提示模板。 对于所有语音理解任务，训练和推理阶段均使用相同的提示模板，以避免潜在的分布不匹配。对于指令式文本转语音任务，训练和推理阶段同样采用基于指令的模板

表 8：WenetSpeech-Wu-Bench 统计信息。

Task	类别	话语	持续时间（小时）
ASR	-	4851	9.75
AST	-	3000	4.4
性别	Male	1500	2.10
	女性	1500	2.29
	总计	3000	4.39
Age	青年	500	0.69
	中年	500	0.79
	老年	500	0.69
	总计	1500	2.22
情绪	中性	300	0.50
	开心	200	0.23
	Sad	100	0.26
	愤怒	200	0.15
	惊讶	200	0.27
	总计	1000	1.41

表 9：所提出的自动语音识别和文本转语音模型的优化与训练超参数。

模型	参数。(M)	LR	学习率调度器	预热	梯度。Acc.	批次大小
Whisper-Medium-Wu	769	8×10^{-5}	预热学习率	4000 步	4	动态（24k 帧）
Conformer-U2pp-Wu	123	1×10^{-3}	预热学习率	25000 步	4	动态（60000 帧）
Step-Audio2-Wu-ASR	7000	1×10^{-5}	预热学习率	0.05 比例	8	8
CosyVoice2-Wu-CPT	500	1×10^{-4}	预热学习率	25000 步	2	动态（10000 帧）
CosyVoice2-Wu-SFT	500	1×10^{-5}	恒定学习率	0	2	动态（2k 帧）
CosyVoice2-Wu-SS	500	5×10^{-6}	恒定学习率	0	2	动态（1k 帧）
CosyVoice2-Wu-instruct	500	5×10^{-6}	恒定学习率	0	2	动态（1k 帧）

以实现显式指令控制。表 10列出了各项任务的提示语表述。为便于展示，此处提供英文提示；在适用情况下，其他语言也会采用相应的指令。

表10：语音理解和指令式文本转语音实验所用提示模板的详细信息。

Task Type	提示内容
ASR	请转写这段语音。
AST	请仔细聆听这段语音，并 将其内容翻译成普通话。
Age	根据语音的声学特征， 判断说话人的年龄。选择一个 标签：青年、中年或老年。
性别	根据语音的声学特征， 判断说话人的性别。选择一个 标签来自：男性或女性。
情绪	根据语音的声学特征和语义 确定情绪。 从以下标签中选择一个：中性、开心、悲伤、 惊讶或愤怒。
Inst Emo	你必须用愤怒、悲伤、 快乐或惊讶的语气说话。
Inst Pro	说话者是男性或女性，嗓音低沉或 高音调，并且语速较慢或较快。