SYNPARASPEECH：用于语音生成和理解的副语言数据集自动化合成

Bingsong Bai[1⋆] Qihang Lu[1⋆] Wenbing Yang[1⋆] Zihan Sun[2]
YueRan Hou[2] Peilei Jia[2] Songbai Pu[2] Ruibo Fu[3]
Yingming Gao[1] Ya Li[1†] Jun Gao[2†]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China
[2]Hello Group Inc., China [3]Institute of Automation, Chinese Academy of Sciences, China

ABSTRACT

副语言声音，如笑声和叹息，在合成更真实、更具吸引力的语音中至关重要。然而，现有的方法通常依赖于专有的数据集，而公开可用的资源往往存在语音不完整、时间戳不准确或缺失以及现实世界相关性有限的问题。为了解决这些问题，我们提出了一种用于生成大规模副语言数据的自动化框架，并将其应用于构建 SynParaSpeech 数据集。该数据集包含 6 类副语言声音，共计 118.75 小时的数据和精确的时间戳，全部源自自然对话语音。我们的贡献在于引入了首个用于构建大规模副语言数据集的自动化方法，并发布了 SynParaSpeech 语料库，这通过更自然的副语言合成推进了语音生成，并通过改进副语言事件检测增强了语音理解。该数据集和音频样本可在 https://github.com/ShawnPi233/SynParaSpeech 获取。

Index Terms— 副语言，语音合成，语音理解，数据集

1. 介绍

随着深度学习的迅速发展，文本到语音（TTS）和语音语言模型（SLMs）等领域已经实现了高质量的语音合成。大多数以前的方法侧重于语义内容，经常忽略副语言声音。然而，在自然对话中，笑声和叹息声是常见的。最近，更多的研究者致力于改进副语言语音合成以增强交互性和逼真度 [1, 2, 3, 4]。

尽管提到的方法可以合成副语言语音，但它们依赖于具有副语言注释的专有数据集，这些数据集不公开可用，限制了可扩展的研究。因此，开放的副语言数据集至关重要。如表 1 所示，当前开源数据集可分为两类。第一类包括仅含音频的副语言事件数据集，例如 AudioSet[5]、ESC-50[6]、VocalSound[7] 和 Nonspeech7k[8]。这些数据集旨在用于声音事件识别，涵盖了各种非语义事件，但缺乏语音、文本和精确的时间戳，因此不适合副语言语音合成或事件定位。第二种类型是包含语音和文本的副语言演讲数据集，包括 Switchboard[9]、Fisher Speech[10]、MagicData-RAMC[11]、NVS[12] 和 NVSpeech[13]。虽然这些提供了转录和副语言标签，但它们有一些限制。Switchboard 和 Fisher Speech 采样率较低，Fisher Speech 和 MagicData-RAMC 涵盖的事件类别较少，而 Switchboard 和 NVSpeech 缺乏精确的时间戳。NVS 主要由动画、电影和节目组成，其中非言语表达被夸张并过度表示，因此它偏离了自然发生的对话。同样，NVS 和 NVSpeech 通过训练带有手动标注副语言事件的 ASR 模型来构建副语言数据集，并随后使用这些模型为新录音打标签。然而，这种基于 ASR 的方法存在两大限制：副语言类别的不平衡性引入了偏差到训练数据中，以及副语言语音数据的稀缺性和高成本限制了可扩展性。

为了解决这些挑战，我们提出了一种用于合成大规模副语言语音数据的自动化方法，并构建了 SynParaSpeech 数据集。与之前的基于 ASR 的扩展方法 [12, 13] 不同，我们的方法使用副语言音频事件和非表演性语音录音自动生成一个带有精确时间戳标注的大规模数据集。该方法可以轻松扩展到各种副语言类别和语言，为构建大规模多语种副语言数据集提供了见解。我们的贡献如下：

1. 据我们所知，我们提出了第一个用于合成大规模副语言语音数据集的自动化方法。
2. 我们介绍了 SynParaSpeech，这是一个包含 6 个副语言类别的中文语音数据集，具有精确时间戳的转录和总计 118.75 小时的时长。
3. 我们通过微调 CosyVoice2 和 F5-TTS 展示了 SynParaSpeech 的有效性，在副语言语音生成方面取得了显著改进。
4. 我们将提示调优应用于如 Qwen 2.5 Omni 和 Kimi Audio 等模型，增强对副语言事件的检测，并探索不同提示上下文长度的影响。

2. 方法

Below, we describe how we built our dataset and the paralinguistic speech generation and understanding systems based on it. The construction of the dataset includes creating paralinguistic labeled text, synthesizing paralinguistic speech audio, and manual checks, as shown in figure 1.

2.1. 合成带标签的文本语音参数

首先我们设计了一个处理流程来生成准确的转录、对齐的时间戳和副语言标签。音频在两个并行步骤中进行处理：(1) 应用了三个 ASR 模型——Whisper Large V3[14]、SenseVoice[15] 和 Paraformer[16]，通过多数投票获得了句子级别的转录。(2) 语音活动检测（VAD）并分割音频为较短的片段。

由于 ASR 在非常短的片段上表现不佳，我们通过以下方式验证每个由 VAD 确定的分割点：在这些点处将音频分为左、右子片段，使用相同的 ASR 模型转录每个子片段，并计算这些子片段转录与完整句子转录之间的编辑距离，以确定子片段文本在完整句子中的对齐情况。对于每个分割点，我们采用了较长子片段的转录结果，将其与完整句子转录对应部分进行字符错误率（CER）计算，并将 CER 低于 0.1 的情况判定为分割准确（生成带时间戳的文本片段）。最终对齐则利用这些 VAD 输出和 ASR 结果，并通过 Stable Whisper[17] 进行验证，以生成精确带时间戳的文本段落。

为了添加副语言标签，我们使用了一个主流的大语言模型（LLM）Deepseek Chat V3[18] 来构建一个带标签的文本数据集。我们将前述转录输入到 LLM 中，提示其从 [laugh],[sigh],[gasp],[throat clearing],[pause],[tsk] 中选择最优标签而不改变原始文本。该 LLM 还在每个 VAD 分割的片

⋆Equal contribution. †Corresponding authors.

Table 1: 副语言数据集比较。

数据集 类型 时间戳	小时 语言。 演讲	剪辑 SR（千赫兹） 可用的
AudioSet [5]	72.3	26,088
18	-	-
×	×	✓
ESC-50 [6]	0.33	240
10	-	22.05
×	×	✓
VocalSound [7]	20.46	21,024
6	-	16/44.1
×	×	✓
Nonspeech7k [8]	6.75	7,014
7	-	32
×	×	✓
Switchboard [9]	260	11,699
42	En	8
×	✓	✓
Fisher Speech [10]	984	5,850
2	En	8
✓	✓	✓
MagicData-RAMC [11]	180	219,325
3	Zh	16
✓	✓	✓
NVS[12]	131	38,718
10	Zh/En	24
✓	✓	✓
NVSpeech [13]	573.4	174,179
18	Zh	-
×	✓	✓
SynParaSpeech (Ours)	118.75	79,986
6	Zh	24
✓	✓	✓

段（例如，句子开始/剪辑结束）中找到了最佳插入位置。最后，我们获得了带有副语言标签的文本数据。

2.2. 合成语音音频参数同步

音频合成阶段（第二阶段）的目标是通过结合已建立的副语言音频和语音音频合成方法，生成包含符合第一阶段产生的副语言注释文本的副语言线索的语音。

首先，使用第一阶段的副语言标签（例如，如图 1 所示的[笑]），我们从匹配的副语言音频语料库中随机挑选一个音频片段。为了保持副语言音频和语音之间的音色一致，我们进行声音转换（VC）：副语言音频是源，语音音频是目标。我们使用 ASR 模型 Whisper Large V3[14] 来编码源音频的语义内容，并使用说话人编码器 CAM++[19] 提取目标说话人的嵌入。这些特征被输入到零样本 VC 模型 SeedVC[20] 中，以获得音色调整后的副语言音频。

此外，语音音频基于第一阶段的时间戳文本进行切分。然后，音色转换后的副语言音频被插入到相应的语音段落中。最后，所有处理过的语音段落按时间顺序合并以获得最终的语音。

2.3. 手动辅助验证

尽管所提出的流程能够高效地实现副语言语音的自动化合成，但仍进行了额外的手动验证以确保与人类感知在自然度、音质和副语言类别方面的一致性。邀请了语音专业人士来评估并优化合成的 SynParaSpeech。评估考虑了四个方面：音频的自然度，重点在于语音与副语言线索之间的流畅过渡；副语言表达的质量，包括音色一致性和准确的标签匹配；整体音质，确保没有噪声、削波或失真；以及音频和文本之间的时间对齐精度，避免遗漏、错误或冗余字符。优化后，仅保留符合标准的音频片段。

3. SYNPARASPEECH 数据集

如表 2 所示，SynParaSpeech 数据集包括六种副语言事件的语音数据：叹息、清嗓子、笑声、停顿、啧舌和喘气。其统计指标涵盖每种事件的总时长（小时）、片段数量（片段数）、平均片段时长（平均秒）以及在完整数据集中所占的比例（占比）。

由于语义场景的自然分布影响了副语言事件的频率，我们在构建数据集时并未强制要求每个类别都有相等的比例。即便如此，统计数据显示所有类别的比例仍然保持良好的平衡：最高比例（叹气，23.76%）和最低比例（喘息，9.36%）之间的差异在合理范围内，并且没有一个类别过于主导或极其稀缺（样本量从 8,846 到 18,827 不等）。因此，该数据集可以为副语言事件的分析和建模提供平衡的多类别支持。

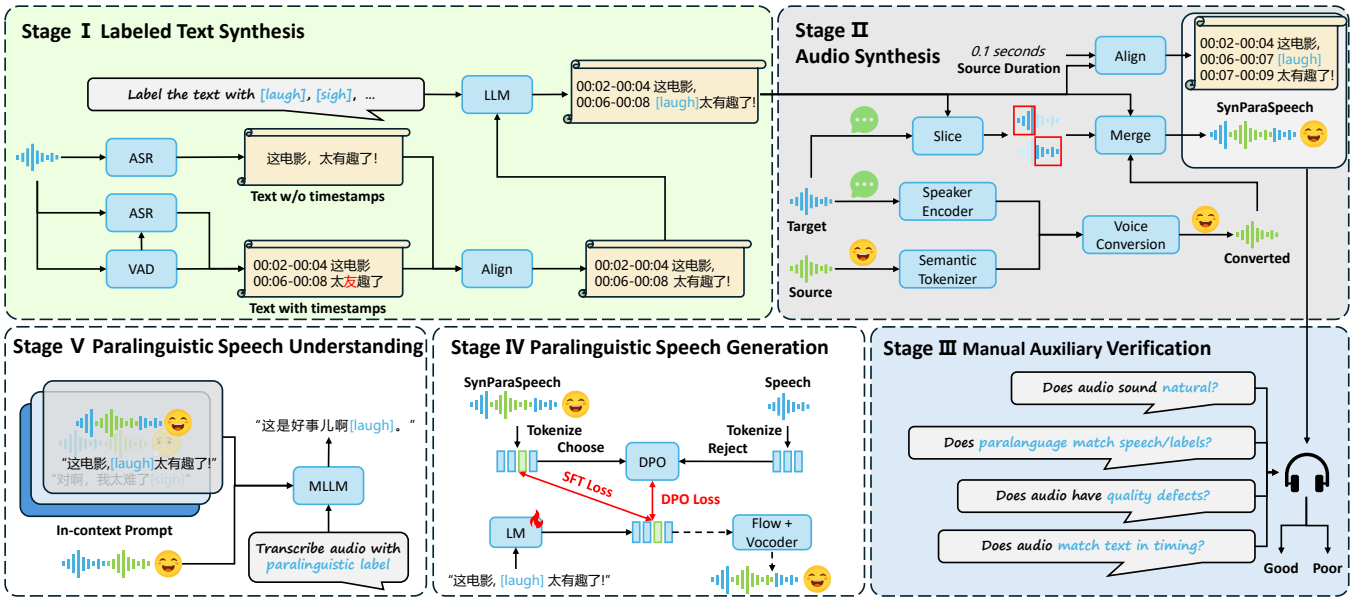

Fig. 1: 概述 SynParaSpeech。(1) 合成带有副语言时间戳的文本。(2) 合成音频并与其副语言信息对齐。(3) 从多个维度验证合成的音频。(4) 通过优化建模生成副语言语音。(5) 通过上下文提示实现副语言语音的理解。

Table 2: SynParaSpeech 数据集的统计。

类别	小时	剪辑	平均值。(秒)	分享
Sigh	28.22	17,706	5.74	23.76 %
Throat Clearing	25.45	18,827	4.87	21.43 %
Laugh	20.84	13,023	5.76	17.55 %
Pause	18.30	9,643	6.83	15.41 %
Tsk	14.82	11,941	4.47	12.48 %
Gasp	11.11	8,846	4.52	9.36 %
Total	118.75	79,986	5.34	100.00 %

4. 实验

4.1. 实验设置

为了评估 SynParaSpeech 在副语言语音合成和语音理解任务中的有效性，我们在副语言 TTS 和事件检测上进行了实验。

4.1.1. 副语言 TTS

在副语言 TTS 实验中，我们使用 SynParaSpeech 对最先进的开源 TTS 模型进行了有监督微调（SFT），即自回归 CosyVoice2[1] 和非自回归 F5-TTS[21]。为了进行比较，我们也考虑了 NVS 数据集 [12]，这是一个从现实世界媒体收集的大型且多样的副语言资源，并进行了自动标注。为了确保可比性，我们选择了由 SynParaSpeech 和 NVS 都支持的四个常见副语言标签—[笑]、[叹息]、[喘息] 和 [清喉咙]。我们随机保留了 2%的数据作为验证集，其余用于训练。为了评估，使用了 LLM DeepSeek V3[18] 重写种子文本（插入副语言标签），每个类别生成 100 条推理文本。

我们比较了 CosyVoice2 和 F5-TTS 的检查点及其在 SynParaSpeech 和 NVS 上训练得到的微调衍生模型。为了克服 SFT 仅关注模型的整体输出损失而缺乏对副语言合成的关注的问题，我们在两种配置下将直接偏好优化（DPO）[22] 应用于 CosyVoice2：分阶段 DPO（在 SFT 之后进行 DPO）和联合 DPO（同时优化两个目标）。利用 SynParaSpeech 的干净源头，每个语音自然地提供了带有副语言事件的正样本对和不带副语言事件的负样本对。在 DPO 训练过程中，原始语音作为被拒绝的样本 μ_o，而其 SynParaSpeech 对应版本则是被选择的样本 μ_p。因此，偏好学习目标被定义为

$$\Delta(\mu_p, \mu_o; y) = \log \frac{\pi_{\text{ref}}(\mu_p \mid y)}{\pi_\theta(\mu_p \mid y)} - \log \frac{\pi_{\text{ref}}(\mu_o \mid y)}{\pi_\theta(\mu_o \mid y)},$$
$$\mathcal{L}_{\text{dpo}} = -\mathbb{E}_y \left[\log \sigma \left(\beta \cdot \Delta(\mu_p, \mu_o; y)\right)\right], \quad (1)$$

其中，π_θ 是目标模型，π_{ref} 是参考模型，$\beta = 0.01$ 是温度，$\sigma(\cdot)$ 是 sigmoid 函数。

对于训练，CosyVoice2 的语言模型使用 Adam 优化器进行优化，学习率为 1×10^{-5}，共训练 50 个周期，采用提前停止策略（耐心值为 10），恒定的学习率以及 2500 步的线性预热。

通过梯度裁剪（阈值为 5）、两步累积梯度和最大帧数为 2000 的动态批处理来维持训练稳定性。

根据 NVS 设置，F5-TTS 训练了 400 个周期，学习率为 1×10^{-4}，使用带有 1000 次预热更新的余弦退火调度器，并且每张 GPU 基于帧的批处理大小为 30,000。

为了初始化新的副语言标记，我们使用来自 CLAP[24] 中基于 RoBERTa 的 [23] 文本编码器的嵌入，并通过插值对齐维度。

模型性能使用客观和主观指标进行了评估。客观测量包括用于可懂度的字符错误率（CER），用于语音质量的 UTMOSv2[25]，以及用于音色相似性的 SECS。主观评价采用了 5 分制评分：SMOS 用于说话人相似性，NMOS 用于自然度，QMOS 用于整体音频质量，PMOS 用于副语言质量。对于 MOS 评分，21 名志愿者参与了双盲评估。

4.1.2. 副语言事件检测

在副语言事件检测实验中，我们应用了基于 SynParaSpeech 的提示调整到多模态大语言模型（MLLMs），并分析了上下文大小的影响。参考 MMSU[26]，由于其强大的副语言推理能

Table 3: 副语言合成语音与 SynParaSpeech 数据集。MOS 分数的置信区间是 95%。

模型	PMOS ↑	NMOS ↑	**土壤湿度产品** ↑	QMOS ↑	CER(%) ↓	**秒数** ↑	UTMOSv2 ↑
F5-TTS (Baseline)	1.16 ± 0.01	4.08 ± 0.02	4.52 ± 0.02	3.95 ± 0.03	6.01	0.76	3.01
+ NVS SFT	1.49 ± 0.03	3.83 ± 0.03	4.03 ± 0.02	3.75 ± 0.03	12.56	0.74	3.01
+ SynParaSpeech SFT	3.10 ± 0.04	4.16 ± 0.02	4.41 ± 0.02	4.08 ± 0.02	7.26	0.74	2.83
CosyVoice2 (Baseline)	1.88 ± 0.04	4.24 ± 0.02	3.71 ± 0.03	4.00 ± 0.03	6.58	0.70	3.13
+ NVS SFT	2.35 ± 0.05	4.06 ± 0.02	3.47 ± 0.03	3.95 ± 0.03	9.50	0.69	3.02
+ SynParaSpeech SFT	3.31 ± 0.04	4.11 ± 0.02	3.74 ± 0.03	4.01 ± 0.02	11.00	0.71	2.78
+ DPO-Staged	3.40 ± 0.04	4.15 ± 0.02	3.84 ± 0.02	4.09 ± 0.02	10.91	0.70	2.87
+ DPO-Joint	3.46 ± 0.04	4.17 ± 0.02	4.03 ± 0.03	4.12 ± 0.02	11.78	0.71	2.83

Table 4: 副语言事件检测结果。

模型	**上下文**	**准确率** ↑	F1 **分数** ↑	CER (%)↓
金米音频	–	0.320	0.294	17.79
	1	0.314	0.312	11.30
	3	0.354	0.336	10.61
	5	0.382	0.340	11.11
	7	0.371	0.331	11.01
Qwen 2.5 多功能	–	0.215	0.189	23.52
	1	0.337	0.357	21.18
	3	0.460	0.447	20.60
	5	0.473	0.471	19.48
	7	0.423	0.362	20.07

力选择了 Kimi Audio[27]，并且由于其优越的副语言感知能力选择了 Qwen 2.5 Omni[28]。两个模型都在有和没有 SynParaSpeech 提示调整的情况下进行了评估。对于每个副语言类别，使用了 100 对音频-文本进行测试，而上下文提示则分别采样了 1、3、5 和 7 对。性能通过准确率、宏 F1 分数和 CER 来衡量。

4.2. 实验结果与分析

4.2.1. 副语言 TTS 实验结果

副语言 TTS 的实验结果如表 3 所示。比较两个数据集上的 F5-TTS，经过 SynParaSpeech 微调的模型在副语言合成、自然度和整体质量方面有所提升。这种增益在副语言合成中最为明显，PMOS 显著增加，与基于 NVS 进行微调的模型相比。SynParaSpeech 还提供了在自然度、发音人相似性、整体质量和清晰度方面的优势。同样，经过 SynParaSpeech 微调的 CosyVoice2 在副语言合成方面取得了大幅改进，在发音人相似性方面有中等程度的增长，并且在副语言质量、自然度、发音人相似性和整体质量方面优于基于 NVS 进行微调的模型。这些结果证实了 SynParaSpeech 可以增强副语言合成，同时保持自然度、发音人相似性和整体语音质量。

此外，F5-TTS 使用非自回归流匹配技术，它实现了高自然度和说话人相似性。相比之下，CosyVoice2 采用其自回归多阶段设计，更好地建模了副语言特征，并产生了更高的音频质量。在此基础上，我们研究了 DPO 如何使 CosyVoice2 更好地捕捉副语言细节并进一步增强其副语言能力。最后三行显示 DPO 提升了副语言质量、自然度和说话人相似性。对于两种类型的 DPO，我们的联合 SFT 训练优于分阶段训练，提高了所有主观评分，并实现了最佳的 PMOS 和 QMOS。

我们还观察到，在副语言数据集上训练的模型在目标指标 CER、SECS 和 UTMOSv2 上表现出轻微下降，因为这些指标是为标准语音设计的。例如，笑声可能会被转录为 “ha ha”，从而提高 CER。然而，在 SynParaSpeech 上训练的模型仍然保持高 NMOS 和 QMOS，并且在 PMOS 上有明显的提升。

4.2.2. 副语言事件检测实验结果

表 4 强调了 SynParaSpeech 促进调整和上下文数量对副语言事件检测的影响。首先，结果证实了 SynParaSpeech 的实用性：对于 Kimi Audio 和 Qwen 2.5 Omni，引入 SynParaSpeech 提示相比无上下文基准显著提高了性能，特别是在准确率和宏 F1 分数方面。这验证了 SynParaSpeech 数据集为副语言感知和推理提供了有效的监督信号。

其次，上下文示例的数量产生关键影响。性能通常随着上下文数量的增加而提升，并且在 5-shot 设置下两种模型都观察到了峰值性能增益。然而，过多的上下文（例如 7-shot）未能带来进一步改进，甚至可能导致性能下降。这表明了提供信息提示与避免由冗余输入导致的模型过载之间的权衡。

总体而言，这些发现表明 SynParaSpeech 增强了模型在副语言事件检测中的能力，而适量的上下文示例对于充分发挥其优势至关重要。

5. 结论

本文提出了一种大规模合成副语言数据集的自动化方法，并引入了 SynParaSpeech 数据集。SynParaSpeech 具有多个细粒度标注的副语言类别和大量的规模，作为促进副语言语音合成（TTS）和副语言事件检测任务的有效资源。实验结果表明，整合 SynParaSpeech 提升了 TTS 模型的生成质量和增强了副语言事件检测模型的表现。

6. REFERENCES

[1] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al., "Cosyvoice 2: Scalable streaming speech synthesis with large language models," arXiv preprint arXiv:2412.10117, 2024.

[2] Weiqin Li, Peiji Yang, Yicheng Zhong, Yixuan Zhou, Zhisheng Wang, Zhiyong Wu, Xixin Wu, and Helen Meng, "Spontaneous style text-to-speech synthesis with controllable spontaneous behaviors based on language models," in Proc. Interspeech 2024, 2024, pp. 1785–1789.

[3] Ning-Qian Wu, Ya-Jun Hu, Liping Chen, and Zhen-Hua Ling, "Anchored monotonic alignment and representation substitution for rare spontaneous behaviors in spontaneous speech synthesis," in ICASSP 2025. IEEE, 2025, pp. 1–5.

[4] canopyai, "Orpheus tts," https://github.com/canopyai/Orpheus-TTS, 2025.

[5] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in ICASSP 2017. IEEE, 2017, pp. 776–780.

[6] Karol J Piczak, "Esc: Dataset for environmental sound classification," in Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1015–1018.

[7] Yuan Gong, Jin Yu, and James Glass, "Vocalsound: A dataset for improving human vocal sounds recognition," in ICASSP 2022. IEEE, 2022, pp. 151–155.

[8] Muhammad Mamunur Rashid, Guiqing Li, and Chengrui Du, "Nonspeech7k dataset: Classification and analysis of human non-speech sound," IET Signal Processing, vol. 17, no. 6, pp. e12233, 2023.

[9] John J Godfrey, Edward C Holliman, and Jane McDaniel, "Switchboard: Telephone speech corpus for research and development," in Acoustics, speech, and signal processing, ieee international conference on. IEEE Computer Society, 1992, vol. 1, pp. 517–520.

[10] Christopher Cieri, David Graff, Owen Kimball, Dave Miller, and Kevin Walker, Fisher English training speech part 1 transcripts, Lead Discovery Center LDC, 2004.

[11] Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, et al., "Open source magicdata-ramc: A rich annotated mandarin conversational (ramc) speech dataset," in Proc. Interspeech 2022, 2022, pp. 1736–1740.

[12] Runchuan Ye, Yixuan Zhou, Renjie Yu, Zijian Lin, Kehan Li, Xiang Li, Xin Liu, Guoyang Zeng, and Zhiyong Wu, "A scalable pipeline for enabling non-verbal speech generation and understanding," arXiv preprint arXiv:2508.05385, 2025.

[13] Huan Liao, Qinke Ni, Yuancheng Wang, Yiheng Lu, Haoyue Zhan, Pengyuan Xie, Qiang Zhang, and Zhizheng Wu, "Nvspeech: An integrated and scalable pipeline for human-like speech modeling with paralinguistic vocalizations," arXiv preprint arXiv:2508.04195, 2025.

[14] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in International conference on machine learning. PMLR, 2023, pp. 28492–28518.

[15] Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, et al., "Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms," arXiv preprint arXiv:2407.04051, 2024.

[16] Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," in INTERSPEECH, 2022.

[17] jianfch, "Stabilizing timestamps for whisper," https://github.com/jianfch/stable-ts, 2023.

[18] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al., "Deepseek-v3 technical report," arXiv preprint arXiv:2412.19437, 2024.

[19] Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen, "Cam++: A fast and efficient network for speaker verification using context-aware masking," in INTERSPEECH, 2023.

[20] Songting Liu, "Zero-shot voice conversion with diffusion transformers," arXiv preprint arXiv:2411.09943, 2024.

[21] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen, "F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching," arXiv preprint arXiv:2410.06885, 2024.

[22] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn, "Direct preference optimization: Your language model is secretly a reward model," Advances in neural information processing systems, vol. 36, pp. 53728–53741, 2023.

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[24] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "Clap learning audio concepts from natural language supervision," in ICASSP 2023. IEEE, 2023, pp. 1–5.

[25] Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari, "The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech," in IEEE Spoken Language Technology Workshop (SLT), 2024.

[26] Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng,

"Mmsu: A massive multi-task spoken language understanding and reasoning benchmark," arXiv preprint arXiv:2506.04779, 2025.

[27] Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al., "Kimi-audio technical report," arXiv preprint arXiv:2504.18425, 2025.

[28] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al., "Qwen2. 5-omni technical report," arXiv preprint arXiv:2503.20215, 2025.