对话语音自然度预测器

Anfeng Xu[†1]、*Yashesh Gaur*[2], *Naoyuki Kanda*[2],*Zhicheng Ouyang*[2],*Katerina Zmolikova*[2],*Desh Raj*[2], *Simone Merello*[2], *Anna Sun*[2], *Ozlem Kalinli*[2]

[1] 南加州大学，美国。[2] Meta Platfor
ms, Inc，美国。anfengxu@usc.edu,
yagaur@meta.com

arXiv:2603.01467v1 [eess.AS] 2 Mar 2026

摘要

对话自然度评估对于开发类人语音智能体至关重要。然而，现有语音自然度预测器通常用于评估单说话人的话语，无法捕捉对话层面的自然度特征。本文提出了一种用于双说话人、多轮对话的自动自然度预测框架。我们首先基于经过人工评分标注的对话录音，证明现有自然度估计器与对话自然度的相关性较低，有时甚至呈负相关。随后，我们提出一种双通道自然度估计器，并研究多种经过预训练的编码器及数据增强方法。与现有自然度预测器相比，我们提出的模型在域内和域外条件下与人类判断的相关性均显著更高。

索引词：语音自然度估计器、平均意见分、全双工、对话

1. 引言

类人语音智能体最重要的方面之一是语音的感知自然度，它直接影响用户体验和参与度。虽然人工评估能够提供可靠判断，但成本高昂，且难以扩展以支持系统的迭代开发。这凸显了对能够提供高效、一致评估的对话自然度自动预测器的需求。

现有的语音自然度预测器在单说话人、单语句评估方面已显示出良好效果，主要关注音频质量和句子级流畅度[1, 2, 3, 4]。然而，这些方法可能无法扩展到对话评估，因为自然度取决于对话现象，例如流畅的轮流发言、恰当使用填充词，以及基于对话动态选择合适的表达。因此，将现有模型直接应用于对话语音可能会产生不一致的结果。这种不一致凸显了需要专门设计用于捕捉对话级自然度的预测器。

与此同时，能够同时听与说的全双工对话模型不断取得进展，凸显了对对话语音生成模型进行鲁棒评估的需求[5, 6, 7]。随着模型开发朝着无缝智能体交互推进，对话自然度的自动预测器对于指导模型设计并实现公平的性能比较变得日益重要。

为此，我们提出一种简单有效的框架，用于自动评估多说话人、多轮次互动的自然度。我们的方法利用强大的预训练编码器来预测对话级自然度，其结果与人类判断高度相关。我们的贡献如下:

- 我们在一个大规模合成对话数据集上进行了实验，该数据集包含约 6.6k 段对话、共计 100 小时音频，并使用一个域外全双工对话数据集进行鲁棒性评估。
- 我们表明，现有预测器无法捕捉对话自然度，且常常产生负相关。
- 我们设计了一个简单而有效的对话自然度预测框架，该框架采用双通道音频输入，并可利用预训练语音编码器。
- 我们证明，基于 Whisper 的模型与人工判断的相关性最高。
- 我们探索大规模合成数据增强，以进一步提升模型在分布偏移下的鲁棒性。

2. 相关工作

2.1. 现有语音自然度预测器

平均意见分（MOS） [8] 是主观语音质量或自然度评估中广泛使用的指标。现有语料库由多位标注者按照1至5分的量表对生成音频进行评分，并将平均分作为人工标注的自然度目标 [2, 9, 10]。近年来，深度学习的进展加速了自动语音质量平均意见分预测的发展 [1, 2, 3, 4]。此外，研究者还探讨了将音频大语言模型（LLM）用于平均意见分预测任务 [11, 12, 13, 14]。然而，这些研究聚焦于单说话人话语的语音质量和自然度估计，其在对话语音自然度评估中的适用性仍不明确。

2.2. 对话语音生成与评估

近期研究开始从对话视角探索语音生成，RyanSpeech[15] 和 DailyTalk [16] 等语料库推动了具备上下文感知能力的文本转语音（TTS）。此外，口语对话系统的发展催生了能够同时聆听和说话的全双工（FDX）模型，使其能够处理重叠、打断和反馈语，而不再局限于严格的轮流发言 [5, 6, 7]。为评估此类系统，研究者提出了 Full-Duplex-Bench v1.5 [17] 和 Taking Turns [18] 等基准。然而，这些基准主要针对重叠处理和轮流发言等特定现象，而非整体对话自然度。

†本工作完成于作者在 Meta Platforms 实习期间， Inc.

表1： 本研究使用的数据集。训练集、开发集和评估集下方的数字表示对话样本的数量。

数据集	评分	训练集	Dev	Eval
ConvTTS	对话自然度	4,579	1,000	1,000
ConvTTS	系统自然度	3,263	711	717
FDX-Conv	系统自然度	0	0	490

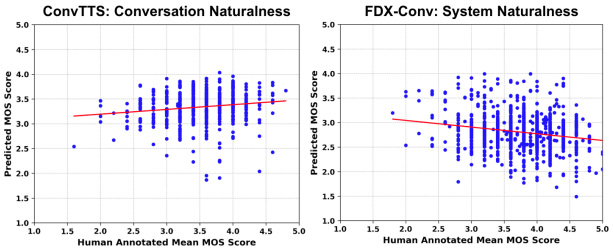


图1：使用 *UTMOSv2* 平均池化方法的散点图。

3. 数据集

3.1. 概述

我们使用两个内部数据集开展实验：Con- vTTS 和 FDX-Conv（表1）。这两个数据集均由用户和对话代理系统的双声道录音构成。数据集中的每段录音均由至少 5 名人类评定者进行自然度评分，评分采用 1 到 5 分制，其中 1 分表示“完全不自然”，5 分表示“与人类语音无法区分”。

我们从两个角度评估对话的自然度，分别称为对话自然度和系统自然度。**对话自然度** 用于衡量两位说话者之间对话的类人性。同时，**系统自然度**用于衡量与对话代理系统对应的音频通道的类人性。评分者需要同时收听用户和系统通道，并据此评定系统自然度；因此，系统自然度会受到轮流发言自然度、填充词等对话现象的影响。

3.2. 对话式文本转语音（ConvTTS）数据集

ConvTTS数据集包含由内部离线文本转语音模型生成的双通道对话。该数据集共包含 6,579 个对话样本，平均时长为 49.4 秒，标准差为 19.3 秒。生成的对话采用用户—系统对话格式。我们将数据随机划分为训练集、开发集和测试集，分别包含 4,579、1,000 和 1,000 个样本，总时长分别为 62.8、13.7 和 13.8 小时。所有数据均包含对话自然度评分，但只有 3,263 个训练集样本、711 个开发集样本和 717 个测试集样本包含系统自然度评分。我们使用该数据集进行模型训练和域内评估。

3.3. 全双工对话（FDX-Conv）数据集

FDX-Conv数据集是内部全双工语音生成模型与内部工作人员之间的一组对话。该数据集包含490段对话录音，均值为41.1秒，标准差为19.0秒。每段录音均针对系统自然度进行评分。我们将该数据集用于域外（分布外）评估。

表 2：*NISQA* 预测与人工标注的对比（皮尔逊相关系数）。话语级预测使用均值、最小值、最大值和中位数统计量进行聚合。

数据集	自然度	Mean	Min	Max	中位数
ConvTTS	对话	-0.18	-0.13	-0.18	-0.18
ConvTTS	系统	-0.25	-0.15	-0.22	-0.26
FDX-Conv	系统	0.09	0.24	-0.10	-0.03

表 3： *UTMOSv2* 预测值与人工标注的对比（皮尔逊相关系数）。按均值、最小值、最大值和中位数统计量汇总话语级预测值。

数据集	自然度	Mean	Min	Max	中位数
ConvTTS	对话	-0.19	-0.16	-0.18	-0.18
ConvTTS	系统	-0.20	-0.11	-0.23	-0.20
FDX-Conv	系统	0.15	0.31	-0.02	0.06

4. 现有自然度预测器

我们首先评估了现有自然度预测器能否评估对话自然度，重点关注NISQA [2]和UTMOSv2 [4]。NISQA是一种由卷积神经网络和自注意力机制组成的MOS预测器。它既能预测平均意见分，也能预测噪声感、音色染色、不连续性和响度等感知维度。我们使用其总体MOS预测分数。UTMOSv2结合自监督学习语音特征和频谱图特征来预测平均意见分。它在2024年 VoiceMOS 挑战赛 [10]上取得了具有竞争力的表现。

与我们评估多说话人、多轮对话自然度的目标不同，现有预测器旨在评估由单个说话人说出的一句话的自然度。因此，我们根据话语对录音进行分段，并将自然度预测器应用于每个话语。随后，我们使用均值、最小值、最大值或中位数汇总话语级预测。对话自然度根据用户和系统通道的汇总分数计算得出。同时，系统自然度则根据系统通道的汇总结果计算得出。

图 1 展示了采用 UTMOSv2 均值统计量绘制的散点图。表 2 和表 3 分别展示了 NISQA 和 UTMOSv2 的预测分数与人工标注的 MOS 分数之间的皮尔逊相关系数（PCC）。两种预测器在所有统计量下与对话自然度均呈负相关。对于系统自然度，它们在所有统计量下与 ConvTTS 均呈负相关，但与 FDX-Conv 的均值和最小值统计量呈正相关。我们认为，这些负相关是因为预测器倾向于给更“闲聊式”的语音较低的自然度分数，而给听起来更正式的语音较高的分数；相比之下，人工标注者在对话语境中认为前者更自然。与 FDX-Conv 的正相关，尤其是最小值统计量下的正相关，可能反映了这样一些情况：某个严重劣化的话语同时拉低了预测器分数和整段对话的人工评分。

5. 方法

5.1. 建模

图2展示了所提出的采用双通道输入建模的自然度预测器架构，该架构使用来自用户和系统的音频。我们还评估了单通道输入建模，该方法仅使用系统音频。预

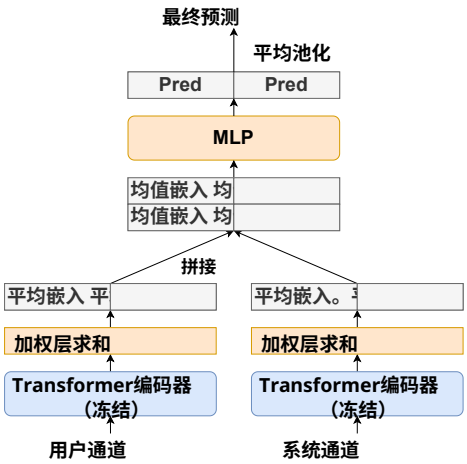


图2： 双通道输入的建模架构。对于单通道输入，我们仅使用系统通道输入。

训练好的 Transformer编码器用于特征提取。我们对所有隐藏层进行加权求和，然后对每个 30 秒片段进行平均池化。在双通道设置中，用户和系统嵌入沿特征维度拼接，然后输入多层感知机（MLP）层进行 MOS预测。对每个 30 秒片段的预测结果取均值，得到最终的 MOS分数。

对于Transformer编码器，我们探索以下模型。

WavLM：一种自监督语音模型，采用k均值聚类进行语音量化，并在掩码预测和去噪任务上进行预训练 [19]。该模型在语音识别和理解基准测试中表现出色。在我们的工作中，我们使用在94k小时音频上预训练的WavLM-large模型。

Audiobox-Aesthetics (AES)：一种用于自动评估语音、音乐和声音音频质量的模型 [20]。该模型使用 562 小时的标注数据进行训练，并以 WavLM-base 为骨干网络，预测制作质量、复杂度、内容欣赏度和实用性的得分。

Whisper：一种基于 Transformer 的编码器-解码器模型，使用弱监督在 68 万小时的多语言语音数据上进行训练 [21]。该模型在各种自动语音识别基准数据集上表现出色，在语音情感识别和方言检测等其他任务上也取得了有竞争力的结果 [22, 23]。我们仅使用其编码器部分提取特征。具体而言，我们选用 Whisper large-v3 模型。

5.2. 训练与评估数据

我们使用 ConvTTS数据集进行训练，并将其作为域内测试集。我们分别训练了用于预测对话自然度和系统自然度的预测器，测试期间使用相同的目标标签。此外，我们使用 FDX-Conv数据集作为分布外测试集。具体而言，我们直接在 FDX-Conv数据集上评估了基于 ConvTTS数据集训练的系统自然度模型。评估指标采用预测 MOS分数与真实 MOS分数之间的皮尔逊相关系数（PCC）和斯皮尔曼秩相关系数（SRC），以及均方误差（MSE）。

表 4： 用于预测对话自然度的模型与人工标注进行对比评估所得的皮尔逊相关系数、斯皮尔曼秩相关系数和均方误差。

Test 数据集	编码器	输入通道	PCC	SRC	MSE
ConvTTS	WavLM	单通道	0.331	0.279	0.242
		dual	0.409	0.358	0.227
	AES	单通道	0.384	0.352	0.228
		dual	0.455	0.430	0.213
	Whisper	单通道	0.433	0.406	0.221
		dual	**0.482**	**0.451**	**0.208**

5.3. 数据增强

我们使用用于开发 ConvTTS数据集的内部 TTS模型对训练集进行了数据增强。该 TTS模型可以接收系统和用户通道的参考音频，以模仿参考音频中的说话人特征，这与零样本TTS类似 [24]。进行数据增强时，我们将训练集中的对话样本用作参考音频，并使用 TTS模型根据 Llama-3.1-405B生成的文本转录生成新的对话音频 [25]。生成的样本被赋予与用作参考信号的音频相对应的自然度评分。我们生成了 5,000 小时带有对话自然度评分的音频。首先，我们将其专门用于大规模预训练自然度估计器。随后，使用真实的 ConvTTS数据集对模型进行微调。

5.4. 模型与训练配置

所有实验均使用 Adam优化器和均方误差损失。多层感知机层由三个中间层组成，隐藏层大小为 768，丢弃率为 0.1，并采用 GeLU激活函数。对于基于 AES 和 Whisper 的预测器，我们使用批量大小 32、学习率 0.002；对于基于 WavLM 的预测器，则使用批量大小 16、学习率 0.001。所有模型均训练 30 轮，并选取每轮结束时计算出的验证损失最低的检查点。使用增强数据进行预训练时，我们采用批量大小 32、学习率 0.001，训练 5 轮；每 1,000 步进行一次评估，并选取在开发集上相关性最高的检查点。

6. 结果与分析

6.1. 对话自然度结果

表 4 展示了针对对话自然度训练自然度预测器时的结果。表现最佳的预测器采用 Whisper 编码器和双通道输入，皮尔逊相关系数为 0.482。采用 WavLM 和 AES 编码器并使用双通道输入的预测器，皮尔逊相关系数也均超过 0.4。较高的皮尔逊相关系数凸显了针对对话语境设计的自然度预测器的有效性；现有预测器则未能取得正向结果。

6.2. 系统自然度结果

表5展示了针对系统自然度训练自然度预测器的结果。对于域内 ConvTTS 测试集，与对话自然度相比，我们观察到整体上更高的皮尔逊相关系数和斯皮尔曼秩相关系数。采用 Whisper 编码器并进行双通道输入建模的预测器表现最佳，皮尔逊相关系数为 0.57，斯皮尔曼秩相关系数为 0.56。在分布外

表5： 用于预测系统自然度 的模型在人工标注上的皮尔逊相关系数、斯皮尔曼秩相关系数和均方误差。

Test 数据集	编码器	输入通道	PCC	SRC	MSE
ConvTTS	WavLM	单通道	0.488	0.491	0.246
		dual	0.517	0.522	0.236
	AES	单通道	0.523	0.523	0.234
		dual	0.542	0.534	0.227
	Whisper	单通道	0.550	0.538	0.226
		dual	**0.570**	**0.560**	**0.218**
FDX-Conv （分布外）	WavLM	单通道	0.305	0.308	**0.335**
		dual	0.058	0.056	0.512
	AES	单通道	0.310	0.297	0.548
		dual	0.282	0.272	0.671
	Whisper	单通道	**0.362**	**0.358**	0.438
		dual	0.290	0.285	0.633

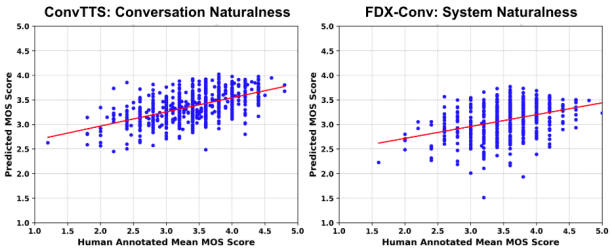


图 3： *Whisper* 结果的散点图。*ConvTTS* 使用双通道输入，*FDX-Conv* 使用单通道输入。

FDX-Conv 测试集上，采用单通道输入建模的 Whisper 预测器表现最佳，皮尔逊相关系数和斯皮尔曼秩相关系数分别达到 0.362 和 0.358。结果表明，即使在系统自然度的分布外场景下，我们的自然度预测器也优于现有最佳基线（UTMOSv2，皮尔逊相关系数为 0.31）。均方误差分数则高得多，且不太稳定，这可能是由于两个数据集的平均意见分均值存在差异。

6.3. 关于用于特征提取的编码器模型的讨论

基于 Whisper编码器的预测器在所有训练目标、测试数据集和输入通道规模的设置下，表现始终优于 AES 或 WavLM 对应模型。我们认为，这是由于其利用约 68 万小时的音频数据进行了大规模预训练，使其能够习得更稳健、更多样的语音理解能力，包括自然度。在 AES 与 WavLM 之间，基于 AES 的模型始终表现更好，这表明其通过预训练任务习得的音频质量信息有助于自然度预测任务。此外，图 3 展示了 Whisper-base 建模结果的散点图，表明模型预测与人工平均意见分标注之间存在稳健的正相关。

6.4. 关于单通道与双通道输入建模的讨论

对于对话自然度和系统自然度，双通道输入模型在 ConvTTS 测试集上的表现均优于相应的单通道模型。对于对话自然度而言，这一结果符合预期，因为两位说话人之间的对话动态包含影响整体自然度的关键因素。同时，对于系统自然度，我们也发现双通道建模优于单通道建模，这表明来自两位说话人的信息也有助于提升

表 6： 用于对话自然度预测的 *Whisper*编码器预测器 （双通道输入建模）， 采用 *5000* 小时增强数据进行预训练。*()* 表示相对于未进行预训练时的变化幅度。

测试数据集	自然度	PCC
ConvTTS	对话	0.483 (+0.21%)
	系统	0.545 （-4.39%）
FDX-Conv （分布外）	系统	0.358 （+23.45%）

表7： 使用*Whisper*编码器，对用户与系统音频的组合输入进行单通道输入建模的消融实验。*()*表示相对于分离输入的变化幅度。

测试数据集	自然度	PCC
ConvTTS	对话	0.458 （-4.98%）
	系统	0.557 （-2.30%）
FDX-Conv （分布外）	系统	0.110 （-62.07%）

对话场景中的单说话人自然度。

对于 FDX-Conv 目标测试数据集，单通道建模优于双通道建模。我们认为，这是因为用户语音存在显著差异：ConvTTS 使用合成音频，而 FDX-Conv 使用真人录音。两个数据集之间的不匹配，可能导致使用双通道输入建模捕捉用户语音时产生更多噪声。

6.5. 数据增强

表 6 报告了模型在 5000 小时增强数据上进行预训练，随后针对对话自然度和系统自然度进行微调时的皮尔逊相关系数。虽然在 ConvTTS 测试集上未观察到提升，但我们在域外 FDX-Conv 测试集上发现了 23.45% 的相对提升。我们将其归因于预训练，因为预训练提供了更具泛化能力的初始参数，从而增强了微调过程中的鲁棒性。

6.6. 双通道输入建模消融

为检验显式通道分离是否有益，我们评估了单通道设置，即将用户与系统音频合并。表7中的皮尔逊相关系数结果表明，在域内 ConvTTS 和分布外 FDX-Conv 测试数据集上，双通道建模均优于输入通道建模。结果显示，保留通道分离能够带来更好的建模效果，这表明使用独立编码器捕获用户和系统通道，能够提升对话场景下的自然度预测。

7. 结论

我们介绍了用于自动评估多轮双说话人交互中对话级自然度的对话自然度预测器。我们证明，现有为单语句评估设计的平均意见分预测器无法泛化到对话语音。为弥补这一不足，我们提出了一种简单而有效的框架，利用预训练语音编码器预测对话语音的自然度。我们系统地研究了 Whisper、WavLM 和 AES 编码器，发现基于 Whisper 的预测器与人工判断的相关性最高。值得注意的是，我们的研究揭示了双通道建模在捕捉交互动态方面的关键作用。

8. 致谢

我们感谢 Wei-Ning Hsu 和 Shanyan Chen 在使用内部 TTS模型方面提供的帮助。

9. 生成式人工智能使用披露

本研究使用生成式人工智能工具辅助语言润色、稿件编辑，以及为数据分析和可视化提供有限的代码开发支持。这些工具未用于生成或解读实验结果。所有作者均了解生成式人工智能的使用范围，对稿件内容负责，并同意提交稿件。

10. 参考文献

[1] C.-C. Lo、S.-W. Fu、W.-C. Huang、X. Wang、J. Yamagishi、Y. Tsao 和 H.-M. Wang，“Mosnet：基于深度学习的语音转换客观评估”，*Interspeech*，2019。[2] G. Mittag、B. Naderi、A. Chehadi 和 S. M¨oller，“Nisqa：一种结合卷积神经网络与自注意力的深度模型，利用众包数据集进行多维语音质量预测”，*Interspeech*，2021。[3] E. Cooper、W.-C. Huang、T. Toda 和 J. Yamagishi，“MOS预测网络的泛化能力”，载于*ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2022，第 8442–8446 页。[4] K. Baba、W. Nakata、Y. Saito 和 H. Saruwatari，“VoiceMOS挑战赛2024的T05系统：从深度图像分类器迁移学习，用于高质量合成语音的自然度MOS预测”，载于*2024 IEEE*口语语言技术研讨会 *(SLT)*。IEEE，2024，第 818–824 页。[5] A. D´efossez、L. Mazar´e、M. Orsini、A. Royer、P. P´erez、H. J´egou、E. Grave 和 N. Zeghidour，“Moshi：用于实时对话的语音文本基础模型”，*arXiv*预印本 *arXiv:2410 00037*，2024。。[6] Q. Zhang、L. Cheng、C. Deng、Q. Chen、W. Wang、S. Zheng、J. Liu、H. Yu、C.-H. Tan、Z. Du 和 S. Zhang，“OmniFlatten：用于无缝语音对话的端到端GPT模型”，载于计算语言学协会第*63*届年会论文集。计算语言学协会，2025，第 14 570–14 580 页。 [在线]。网址：https://aclanthology.org/2025.acl-long.709/[7] K. Hu、E. Hosseini-Asl、C. Chen、E. Casanova、S. Ghosh、P. ˙Zelasko、Z. Chen、J. Li、J. Balam 和 B. Ginsburg，“面向语音到语音语言模型的高效直接双工建模”，*arXiv*预印本 *arXiv:2505.15670*，2025。[8] R. C. Streijl、S. Winkler 和 D. S. Hands，“重新审视平均意见分（MOS）：方法与应用、局限性及替代方案”，*MultimediaSystems*，第 22 卷，第 2 期，第 213–227 页，2016。[9] G. Maniati、A. Vioni、N. Ellinas、K. Nikitaras、K. Klapsas、J. S. Sung、G. Jho、A. Chalamandaris 和 P. Tsiakoulis，“SOMOS：用于评估神经文本转语音合成的三星开放MOS数据集”，*Interspeech*，2022。[10] W.-C. Huang、S.-W. Fu、E. Cooper、R. E. Zezario、T. Toda、H.-M. Wang、J. Yamagishi 和 Y. Tsao，“VoiceMOS挑战赛2024：超越语音质量预测”，载于*2024 IEEE*口语语言技术研讨会 *(SLT)*。IEEE，2024，第 803–810 页。[11] S. Wang、W. Yu、Y. Yang、C. Tang、Y. Li、J. Zhuang、X. Chen、X. Tian、J. Zhang、G. Sun 等，“赋能听觉大语言模型，实现自动语音质量评估”，载于*ICASSP 2025-2025 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2025，第 1–5 页。[12] R. E. Zezario、S. M. Siniscalchi、H.-M. Wang 和 Y. Tsao，“关于使用大语言模型进行零样本非侵入式语音评估的研究”，载于*ICASSP 2025-2025 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2025，第 1–5 页。[13] W. Wu、W. Chen、C. Zhang 和 P. Woodland，“人类标注员模拟中的变异性建模”，载于计算语言学协会 *ACL 2024* 论文集：研究成果，2024，第 1139–1157 页。[14] C. Chen、Y. Hu、S. Wang、H. Wang、Z. Chen、C. Zhang、C.-H. H. Yang 和 E. S. Chng，“音频大语言模型可作为描述性语音质量评估器”，*arXiv*预印本 *arXiv:2501.17202*，2025。[15] R. Zandie、M. H. Mahoor、J. Madsen 和 E. S. Emamian，“RyanSpeech：一个用于对话式文本转语音合成的语料库”，*Interspeech*，2021。[16] K. Lee、K. Park 和 D. Kim，“DailyTalk：用于对话式文本转语音的口语对话数据集”，载于*ICASSP 2023-2023 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*。IEEE，2023，第 1–5 页。[17] G.-T. Lin、S.-Y. S. Kuan、Q. Wang、J. Lian、T. Li 和 H.-y. Lee，“Full-duplex-bench v1.5：评估全双工语音模型的重叠处理能力”，*arXiv*预印本 *arXiv:2507.23159*，2025。[18] S. Arora、Z. Lu、C.-C. Chiu、R. Pang 和 S. Watanabe，“轮流发言：语音基础模型轮流发言动态基准测试”，*arXiv*预印本 *arXiv:2503.01174*，2025。[19] S. Chen、C. Wang、Z. Chen、Y. Wu、S. Liu、Z. Chen、J. Li、N. Kanda、T. Yoshioka、X. Xiao 等，“WavLM：用于全栈语音处理的大规模自监督预训练”，*IEEE* 信号处理选题期刊，第 16 卷，第 6 期，第 1505–1518 页，2022。[20] A. Tjandra、Y.-C. Wu、B. Guo、J. Hoffman、B. Ellis、A. Vyas、B. Shi、S. Chen、M. Le、N. Zacharov 等，“Meta Audiobox Aesthetics：面向语音、音乐和声音的统一自动质量评估”，*arXiv*预印本 *arXiv:2502.05139*，2025。[21] A. Radford、J. W. Kim、T. Xu、G. Brockman、C. McLeavey 和 I. Sutskever，“通过大规模弱监督实现稳健的语音识别”，载于国际机器学习大会。PMLR，2023，第 28 492–28 518 页。[22] T. Feng、J. Lee、A. Xu、Y. Lee、T. Lertpetchpun、X. Shi、H. Wang、T. Thebaud、L. Moro-Velazquez、D. Byrd 等，“Vox-profile：用于表征多样化说话人和语音特征的语音基础模型基准”，*arXiv*预印本 *arXiv:2505 14648*，2025。。[23] T. Feng、K. Huang、A. Xu、X. Shi、T. Lertpetchpun、J. Lee、Y. Lee、D. Byrd 和 S. Narayanan，“Voxlect：用于建模全球方言和地区语言的语音基础模型基准”，*KDD*，2026。[24] M. Le、A. Vyas、B. Shi、B. Karrer、L. Sari、R. Moritz、M. Williamson、V. Manohar、Y. Adi、J. Mahadeokar 等，“Voicebox：大规模文本引导多语言通用语音生成”，*NeurIPS*，第 36 卷，第 14 005–14 034 页，2023。[25] A. Dubey、A. Jauhri、A. Pandey、A. Kadian、A. Al-Dahle、A. Letman、A. Mathur、A. Schelten、A. Yang、A. Fan 等，“Llama 3 模型家族”，*arXiv*电子预印本，第 arXiv–2407 页，2024。