跨语言 F5-TTS：迈向语言无关的声音克隆和语音合成

刘青宇 [1,3], 陈语森 [1,2], 牛志康 [1,2], 王春辉 [4], 杨云婷 [4],
张博文 [4], 赵健 [4], 朱鹏程 [4], 余凯 [1], 陈协 [1,2,†]

[1]MoE Key Lab of Artificial Intelligence, X-LANCE Lab, School of Computer Science, Shanghai Jiao Tong University, China [2]Shanghai Innovation Institute, China, [3]Johns Hopkins University, USA, [4]Geely, China

ABSTRACT

基于流匹配的文本到语音 (TTS) 模型展示了高质量的语音合成。然而，目前大多数基于流匹配的 TTS 模型仍然依赖于与音频提示相对应的参考文稿进行合成。这种依赖性阻止了在无法获取音频提示文稿时进行跨语言的声音克隆，特别是对于未见过的语言。基于流匹配的 TTS 模型消除音频提示文稿的关键挑战是在训练过程中识别单词边界以及在推理过程中确定适当的持续时间。本文介绍了 Cross-Lingual F5-TTS 框架，该框架能够在没有音频提示文稿的情况下实现跨语言声音克隆。我们的方法通过对音频提示进行强制对齐预处理以获得单词边界，从而可以在训练中直接从音频提示合成语音而不使用文稿。为了解决持续时间建模的挑战，我们在不同的语言粒度级别上训练说话率预测器来根据说话者速度推导出持续时间。实验表明，我们的方法在与 F5-TTS 性能相匹配的同时实现了跨语言声音克隆。

Index Terms— 流匹配，跨语言语音克隆

1. 介绍

零样本文本到语音（TTS）旨在根据输入的文本生成与给定参考语音高度相似的声音。随着数据和模型规模的增长，现有的 TTS 系统在模仿说话者特征和合成更富有表现力的语音方面表现出卓越的能力。零样本 TTS 的建模方法可以分为两类：自回归（AR）模型 [?,?,?,1,2,5] 和非自回归（NAR）模型 [7–12]。

在 NAR 模型中，基于流匹配的方法尤为前景广阔。例如，VoiceBox [7] 在词错误率（WER）和语音相似度（SIM）上超过了基于 AR 的 VALL-E [1]，并且推理速度提高了 20 倍。E2 TTS [8] 和 F5-TTS [9] 通过更简单的架构实现了显著改进。最近的跨语言 TTS 方法，如 VALL-E X [13]，通过在大规模多语言数据上训练，实现了令人印象深刻的跨语言语音克隆能力。然而，所有这些模型仍然依赖于音频提示文本转录，这在参考文本不可用时会带来固有的挑战。不同的流匹配模型采用了各种持续时间建模方法。虽然一些基于流匹配的模型使用显式的音素持续时间预测器 [7]，F5-TTS 通过音频提示文本长度与目标文本长度的比例来估计持续时间。然而，在跨语言背景下，这一机制失效了，因为不同语言之间的文本长度比例可能不对应于语音持续时间比例。

在本文中，我们提出了跨语言 F5-TTS，一个基于 F5-TTS 的新型框架，通过丢弃音频提示转录来实现跨语言零样本语音克隆。利用 MMS [14] 强制对齐，我们预处理训练数据以识别单词边界。为了解决持续时间预测的挑战，我们在包括音素、音节和单词三个语言粒度级别引入了专门的说话率预测器。这些模型被训练用来从音频提示中估计说话率，提供了一个稳健且独立于语言的机制来确定目标话语的持续时间。我们的方法在 LibriSpeech-PC test-clean 和 Seed-TTS-eval 上实现了与 F5-TTS 相当的表现，并成功地将其能力扩展到了跨语言场景，取得了有前景的结果。[1]

2. 方法

2.1. 基于流匹配的 TTS 初步研究

基于流匹配的模型 [7–9] 在 TTS 任务中最近取得了显著的成绩。这种方法在模型简洁性方面提供了重大优势。具体而言，通过利用流匹配，E2-TTS [8] 和 F5-TTS [9] 消除了额外的组件，如音素时长预测器和复杂的文本编码器，从而保持了管道的简洁性并实现了高质量的合成。

流动匹配框架旨在学习一个随时间变化的向量场 v_t，该向量场匹配简单噪声分布 p_0 和数据分布 q 之间的概率路径 p_t，以生成采样流步骤 $t \in [0,1]$ 的流动 ψ_t。训练目标被表述为条件流动匹配 (CFM) 损失：

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,q(x_1),p(x_0)} \| v_t(\psi_t(x)) - \frac{d}{dt}\psi_t(x) \|^2 \quad (1)$$

此概率路径连接了高斯噪声中的样本 $x_0 \sim p(x_0)$ 和训练数据中的样本 $x_1 \sim q(x_1)$。最优传输（OT）流匹配为此框架提供了一种特别有效的实现。在 OT 表述下，流 ψ_t 被定义为一条直线轨迹：

$$\psi_t(x_0) = (1-t)x_0 + tx_1 \quad (2)$$

相应的速度场变为常向量 $(x_1 - x_0)$，导致 OT-CFM 损失：

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,q(x_1),p(x_0)} \| v_t((1-t)x_0 + tx_1) - (x_1 - x_0) \|^2 \quad (3)$$

在这项工作中，我们采用 F5-TTS 作为我们的基准系统。F5-TTS 是一个完全基于流匹配的 TTS 系统，使用扩散变压器（DiT），并通过 OT-CFM 训练文本引导的语音填充任务。该模型预测给定其周围语音 $(1-m)\odot x_1$、噪声语音 $(1-t)x_0 + tx_1$ 和扩展字符序列 z 的掩码语音 $m \odot x_1$。

2.2. MMS 强制对齐

为了实现无转录语音克隆，我们利用了大规模多语言语音（MMS）强制对齐工具 [14] 来预处理训练数据并获得精确的单词边界信息。MMS 强制对齐系统采用了一个基于 Wav2Vec2 的多语言声学模型，并通过连接主义时间分类 [16] 训练该模型来生成音频帧的后验概率。为了高效处理长音频记录，MMS 将音频文件分割成 15 秒的片段以生成后验概率，然后再将它们拼接回一个统一的对齐矩阵。

我们将 MMS 强制对齐应用于 Emilia 数据集的中文和英文子集 [15]。对于每个音频样本及其对应的转录，我们提取每

† Corresponding author

[1]在线演示可以在 https://qingyuliu0521.github.io/Cross_lingual-F5-TTS/ 找到。该模型将在 Hugging Face 上开源。

个单词的结束时间。在训练过程中，我们修改了 F5-TTS 框架，将这个单词边界信息作为附加输入与原始语音和文本输入一起使用，如图 1 所示。在每次训练步骤中，我们在转录中随机选择一个单词边界，并在此处分割音频样本。所选边界的前一段音频用作音频提示，而其对应的转录音部分则被完全丢弃。剩余的音频段被遮罩并成为合成的目标。

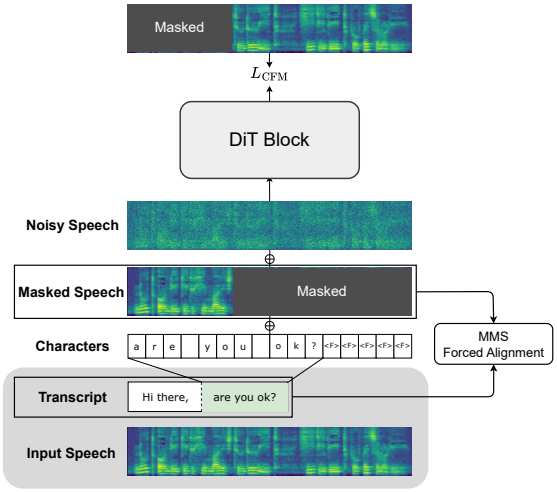

Fig. 1. 跨语言 F5-TTS 的训练框架。MMS 强制对齐产生训练数据的词边界，其中左侧段落作为无转录音频提示，右侧段落的梅尔频谱图被遮罩以进行预测。

2.3. 说话速率预测器

在我们无转录训练方法中消除音频提示转录引入了语音合成中的持续时间预测的关键挑战。在原始的 F5-TTS 框架中，持续时间是通过简单的长度比率方法来估算的，其中持续时间等于音频提示持续时间乘以目标文本长度与参考文本长度的比例。然而，当音频提示转录不可用时，这种机制变得不可行。为了解决这一挑战，我们引入了一个专门的说话速率预测器，它直接从音频提示的声学特征中估算持续时间。

我们将说话率预测公式化为一个离散分类任务，并训练三个单独的模型用于不同的语言粒度：每秒音素数、每秒音节数和每秒单词数。我们定义了一个具有均匀间隔的类别集 C，其间隔为 $\Delta = 0.25$。具体来说，音素级别的模型使用 $C = \{0.25, 0.5, \ldots, 17.75, 18.0\}$，共有 $N = 72$ 类，而音节级别和词级别的模型均使用 $C = \{0.25, 0.5, \ldots, 7.75, 8.0\}$，共有 $N = 32$ 类。对于给定的语速 v，真实类别 c_{gt} 是通过最小距离映射确定的：$c_{\text{gt}} = \arg\min_{x \in C} |v - x|$。每个模型都以音频特征作为输入，并独立预测与其各自语言单元相对应的语速类别。

我们的说话速率预测器采用基于变压器的架构设计来处理梅尔频谱图输入，如图 2 所示。该模型包含一个梅尔投影层，将输入的梅尔频谱图投影到模型的隐藏维度上，随后是两个一维卷积层。多个变压器编码器层处理序列，注意力机制的序列池化方法通过计算每个时间步的注意力权重并进行加权平均来聚合时序信息，从而获得固定大小的表示。最后，分类器输出说话速率类别的概率。

标准交叉熵损失将所有类别视为独立的，这对于我们的有序说话速率分类来说是次优的。我们引入了一种高斯交叉熵（GCE）损失，该损失结合了说话速率的顺序性质：

$$L_{\text{GCE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_c^{\text{soft}} \log(\hat{y}_c) \tag{4}$$

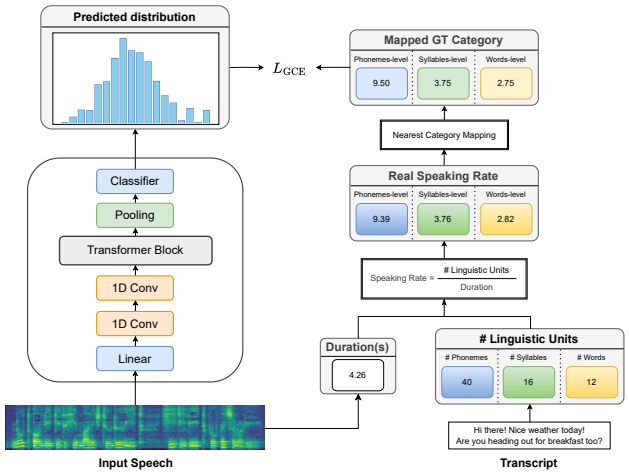

Fig. 2. 说话率预测器的训练流程。该模型从梅尔谱图预测离散的速率类别，而真实说话率被映射到最近的类别以计算 GCE 损失。

软标签是使用高斯核计算的：

$$y_c^{\text{soft}} = e^{\frac{-(c - c_{\text{gt}})^2}{2\sigma^2}} \tag{5}$$

其中 c_{gt} 是真实类别索引，c 是当前类别索引，而 σ 控制高斯核的平滑度。这种公式为接近真实类别的类别分配更高的权重，提供了对轻微预测误差的容错能力，同时保持精度。

在推理过程中，我们的方法通过一个简单的过程来估计目标生成时长。首先将音频提示输入到说话速率预测器中，以估计说话者的特征语速，单位可以是每秒的音素、音节或单词数量，同时对目标文本进行处理，计算相应的语言单元数量。然后根据预测的说话速率与语言单元数量之比来计算目标音频时长。这一机制使得我们的跨语言 F5-TTS 能够在不依赖音频提示转录的情况下执行时长估计，从而促进真正意义上的跨语言声音克隆能力。

3. 实验

3.1. 数据集

我们在 Emilia 数据集 [15] 上训练我们的跨语言 F5-TTS，这是一个从多种真实世界场景收集的大型多语言语音数据集。过滤掉转录失败和误分类的语言语音后，我们保留了大约 9.5 万小时的英语和中文音频数据。为了训练我们的说话速率预测器，我们从中提取了一个平衡子集作为训练数据，该子集中包含来自 Emilia 中的中文部分和英文部分各 500 小时的数据。

3.2. 预处理

我们将 MMS 强制对齐工具应用于提取 Emilia 数据集的单词边界。虽然 Emilia-pipe [15] 在使用 Whisper-X [17] 进行转录生成方面取得了相当大的成功，但在生成的转录中仍存在一些挑战。数字、特殊符号和其他语言的标记偶尔会出现在转录中，在强制对齐过程中无法正确识别这些内容。为了解决这些问题，我们实施了专门的预处理程序，跳过异常标记并将其排除在单词边界提取之外，以确保强大的准确对齐结果。

3.3. 模型设置

我们的基线是开源的 F5-TTS-Base [9]，它采用了一个具有 22 层、16 个注意力头和 1024 维嵌入的扩散变压器（DiT）架构。该模型在八个 NVIDIA A100 GPU 上进行了 1.2M 次更新训练，每个 GPU 的批量大小为 38,400 个音频帧。我们使用了 AdamW 优化器 [18]，学习率在前 20k 次更新中线性增加至 7.5×10^{-5}，然后对于剩余的训练步骤进行线性衰减。

我们的说话速率预测器采用具有 6 层、8 个注意力头和 512 维嵌入的变压器架构。训练在四块 A100 GPU 上进行，总共进行了 50k 次更新，每块 GPU 的批处理大小为 38,400 个音频帧。学习率在前 7.5k 次更新中升温至 2.5×10^{-4}，然后线性衰减。对于高斯交叉熵损失，我们将标准差 $\sigma = 1.0$ 设置为平衡对附近预测的容忍度和目标类别的精度。

对于推理，我们遵循使用欧拉 ODE 求解器的标准 F5-TTS 设置，该求解器具有 32 个函数评估（NFE = 32），CFG 强度 2.0，摇摆采样系数-1.0，并将预训练的 Vocos [19] 作为声码器。

3.4. 评估

我们遵循 F5-TTS 的评估设置，采用 Seed-TTS-eval 和 LibriSpeech-PC test-clean 作为我们的测试集。我们还构建了一个包含 473 个 3-8 秒音频提示样本的多语言跨语言测试集来自 FLEURS [21]，涵盖四种语言（德语、法语、印地语、韩语）以合成英语和中文语音。评估使用以下三个指标进行：

词错误率（WER）通过比较其转录与真实文本以衡量合成语音的可理解性。我们使用 Whisper-large-V3 [22] 和 Paraformer-zh [23] 进行自动识别并相应计算 WER。

说话人相似度（SIM-o）定量描述了合成语音与原始目标语音之间的相似性。我们使用基于 WavLM-large 的 [24] 说话人验证模型来提取说话人嵌入，并计算它们之间的余弦相似度。

UTMOS [25] 通过预训练的 MOS 预测模型提供对语音自然度的自动评估。它在无需参考录音或标签的情况下估计音频质量。虽然不是绝对的主观衡量标准，UTMOS 提供了一种实用且有效的方法来评估合成语音的自然度。

为了评估我们说话速率预测器的有效性，我们使用两个互补的指标来评估其时长预测准确性：

平均相对误差（MRE）测量持续时间预测的相对准确性。它是预测持续时间和真实音频持续时间之间的平均相对差异，其中预测持续时间为语言单元数量除以预测语速所得。

平均绝对误差（MAE）量化了预测持续时间和真实持续时间之间的绝对偏差。

Table 1. 不同说话速率预测器的预测时长的平均绝对误差（MAE）(秒) 和平均相对误差（MRE）(%)。

标识符	系统	平均绝对误差 (s)↓	均方根误差 (%)↓
	Librispeech-PC **测试清洁**		
M1	Phonemes-level predictor	0.759	11.932
M2	Syllables-level predictor	0.757	11.945
M3	Words-level predictor	1.171	18.406
	SeedTTS **测试-英文**		
M1	Phonemes-level predictor	0.637	15.017
M2	Syllables-level predictor	0.704	16.497
M3	Words-level predictor	0.886	20.031
	SeedTTS **测试**-zh		
M1	Phonemes-level predictor	0.845	14.469
M2	Syllables-level predictor	0.783	13.771
M3	Words-level predictor	0.908	16.156

4. 结果

4.1. 说话速率预测性能

表 1 提供了我们说话速率预测器在不同语言粒度下的时长预测准确率结果。M1（音素级别预测器）在 LibriSpeech-PC test-clean 上的表现优于其他预测器，并且在 SeedTTS test-en 上与 M2 表现相当，展示了其对英语时长建模的有效性。值得注意的是，M2（音节级别预测器）在 SeedTTS test-zh 上表现最佳，表明音节可能是中文时长建模更自然的语言单位。M3（词级别预测器）在所有数据集上的一致较差表现突显了粗粒度语言单位的局限性。

Table 2. LibriSpeech-PC 测试清洁集、SeedTTS 测试英语集和 SeedTTS 测试中文集的结果。

系统	持续时间方法	WER(%)↓	SIM-o↑	UTMOS↑
	LibriSpeech-PC **测试清洁**			
Baseline	Length-ratio	2.205	0.668	3.797
CL-F5	M1	2.079	0.663	3.884
CL-F5	M2	2.120	0.658	3.892
CL-F5	M3	2.894	0.652	3.855
	SeedTTS **测试-英**			
Baseline	Length-ratio	1.545	0.676	3.581
CL-F5	M1	1.513	0.662	3.629
CL-F5	M2	1.594	0.660	3.625
CL-F5	M3	2.009	0.646	3.593
	SeedTTS **测试**-zh			
Baseline	Length-ratio	1.475	0.762	2.898
CL-F5	M1	1.605	0.759	2.913
CL-F5	M2	1.481	0.764	2.887
CL-F5	M3	1.616	0.763	2.889

4.2. 跨语言语音克隆性能

表 2 比较了我们的跨语言 F5-TTS（CL-F5）与原始的 F5-TTS 基准在标准基准测试上的表现。我们的方法在所有评估指标上都表现出竞争力。在 LibriSpeech-PC test-clean 数据集上，使用 M1 和 M2 的 CL-F5 实现了比基准更好的 WER 和 UTMOS。这种趋势也延续到了 SeedTTS test-en 数据集，其中使用 M1 的 CL-F5 提供的 WER 和 UTMOS 优于基准。尽管在英语测试集上的 SIM 略低于基准，但这一适度下降是对跨语言能力扩展的一个可接受的权衡。在 SeedTTS test-zh 数据集上，使用 M2 的 CL-F5 在所有指标上表现几乎与基准相同。第 4.1 节中观察到的语言特定偏好转化为内语言语音克隆：带有精细预测器（M1、M2）的 CL-F5 始终优于带有 M3 的 CL-F5。这些结果完全验证了我们方法的有效性，即 CL-F5 在保留内语言语音克隆质量的同时消除了对音频提示文本的依赖。

4.3. 跨语言声音克隆结果

表 3 展示了我们多语言测试集上的跨语言语音克隆结果，其中音频提示和目标合成语言不同。跨语言的结果强化了第 4.1 节和第 4.2 节中观察到的语言特定偏好：CL-F5 在英语目标上使用 M1，在中文目标上使用 M2 时实现了最佳性能。带有 M3 的 CL-F5 在跨语言场景中显示出显著的性能下降，WER 大幅增加至 16.494%，而针对英语目标与使用 M1 的 CL-F5 相比为 2.496%。这是因为 M3 产生过快的语速，导致压缩的时间模式对可理解性产生了负面影响。这些结果强调了跨语言 TTS 中准确持续时间建模的重要性，因为时间不准确性可以严重降低整体合成质量。

值得注意的是，尽管说话速率预测器仅在中国和英语数据上进行训练，我们的方法仍展现了有效的跨语言声音克隆能力，

表明其在未见过的语言上具有良好的泛化能力，且无需音频提示文本。

Table 3. 跨语言声音克隆的 Cross-Lingual F5-TTS 结果。GT 长度表示使用真实总语音长度获得的结果。

持续时间方法	WER(%)↓	SIM-o↑	UTMOS↑
跨语言测试-英文			
GT length	2.462	0.530	3.083
M1	2.496	0.543	3.069
M2	4.362	0.518	3.059
M3	16.494	0.486	2.926
跨语言测试-zh			
GT length	1.596	0.558	2.452
M1	2.446	0.555	2.494
M2	1.801	0.565	2.503
M3	1.946	0.563	2.492

5. 结论

在本研究中，我们提出了 Cross-Lingual F5-TTS，一个通过消除对音频提示文本的依赖来实现跨语言零样本声音克隆的新框架。我们的方法利用 MMS 强制对齐技术，在单词边界上划分训练数据，从而解决了现有 NAR TTS 模型的基本局限性。为了应对随之而来的持续时间预测挑战，我们引入了专门的语速预测器，这些预测器可以直接从声学特征估计持续时间，使用音素、音节或单词级别的粒度。实验结果表明，我们的方法在 LibriSpeech-PC test-clean 和 Seed-TTS-eval 上的性能与原始 F5-TTS 相当，同时成功实现了基线模型无法处理的未见语言说话人的跨语言声音克隆。语速预测器分析显示，包括音素和音节在内的更细粒度的语言单元提供了更为可靠的持续时间估计。尽管我们的无文本方法能够实现跨语言的声音克隆，但在传输如口音和情感等细微说话人特征方面的能力相较于原始 F5-TTS 有所减弱。在未来的工作中，我们计划探索弥补音频提示文本缺失的语音信息的方法，使 Cross-Lingual F5-TTS 在保持其跨语言能力的同时生成更具表现力的语音。

6. REFERENCES

[1] C. Wang, S. Chen., Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 705–718, 2023.

[2] P. Anastassiou, J. Chen, J. Chen, Y. Chen, Z. Chen, Z. Chen, J. Cong, L. Deng, C. Ding, L. Gao et al., "Seed-TTS: A family of high-quality versatile speech generation models," arXiv preprint arXiv:2406.02430, 2024.

[3] D. Jia, Z. Chen, J. Chen, C. Du, J. Wu, J. Cong, X. Zhuang, C. Li, Z. Wei, Y. Wang, and Y. Wang, "DiTAR: Diffusion transformer autoregressive modeling for speech generation," arXiv preprint arXiv:2502.03930, 2025.

[4] B. Zhang, C. Guo, G. Yang, H. Yu, H. Zhang, H. Lei, J. Mai, J. Yan, K. Yang, M. Yang et al., "MiniMax-Speech: Intrinsic zero-shot text-to-speech with a learnable speaker encoder," arXiv preprint arXiv:2505.07916, 2025.

[5] L. Meng, L. Zhou, S. Liu, S. Chen, B. Han, S. Hu, Y. Liu, J. Li, S. Zhao, X. Wu, H. Meng, and F. Wei, "Autoregressive speech synthesis without vector quantization," arXiv preprint arXiv:2407.08551, 2024.

[6] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang, F. Yu, H. Liu, Z. Sheng, Y. Gu, C. Deng, W. Wang, S. Zhang, Z. Yan, and J. Zhou, "CosyVoice 2: Scalable streaming speech synthesis with large language models," arXiv preprint arXiv:2412.10117, 2024.

[7] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, and W.-N. Hsu, "Voicebox: Text-guided multilingual universal speech generation at scale," in Proc. NeurIPS, vol. 36, 2023, pp. 14 005–14 034.

[8] S. E. Eskimez, X. Wang, M. Thakker, C. Li, C.-H. Tsai, Z. Xiao, H. Yang, Z. Zhu, M. Tang, X. Tan, Y. Liu, S. Zhao, and N. Kanda, "E2 TTS: Embarrassingly easy fully non-autoregressive zero-shot tts," in 2024 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2024, pp. 682–689.

[9] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. JianZhao, K. Yu, and X. Chen, "F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching," in Proc. ACL, 2025, pp. 6255–6271.

[10] K. Shen, Z. Ju, X. Tan, E. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, "NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in Proc. ICLR, 2024.

[11] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang et al., "NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," in Proc. ICML, 2024.

[12] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in Proc. ICLR, 2021.

[13] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Speak Foreign Languages with Your Own Voice: Cross-lingual neural codec language modeling," arXiv preprint arXiv:2303.03926, 2023.

[14] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas et al., "Scaling speech technology to 1,000+ languages," Journal of Machine Learning Research, vol. 25, no. 97, pp. 1–52, 2024. [Online]. Available: http://jmlr.org/papers/v25/23-1318.html

[15] H. He, Z. Shang, C. Wang, X. Li, Y. Gu, H. Hua, L. Liu, C. Yang, J. Li, P. Shi, Y. Wang, K. Chen, P. Zhang, and Z. Wu, "Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation," in 2024 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2024, pp. 885–890.

[16] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in Proc. ICML, 2006, pp. 369–376.

[17] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-accurate speech transcription of long-form audio," in Proc. Interspeech, 2023.

[18] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in Proc. ICLR, 2019.

[19] H. Siuzdak, "Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis," in Proc. ICLR, 2024.

[20] A. Meister, M. Novikov, N. Karpov, E. Bakhturina, V. Lavrukhin, and B. Ginsburg, "Librispeech-pc: Benchmark for evaluation of punctuation and capitalization capabilities of end-to-end asr models," in Proc. IEEE ASRU, 2023.

[21] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "FLEURS: Few-shot learning evaluation of universal representations of speech," in Proc. IEEE SLT, 2022.
[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proc. ICML, 2023.
[23] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," in Proc. INTERSPEECH, 2022.
[24] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," IEEE Journal of Selected Topics in Signal Processing, 2022.
[25] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: Utokyo-sarulab system for voicemos challenge 2022," in Proc. INTERSPEECH, 2022.