一个多阶段框架用于多模态可控语音合成

Rui Niu[1], Weihao Wu[1], Jie Chen[2], Long Ma[2], Zhiyong Wu[1]*
[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Youtu Lab, Tencent, Beijing, China
{nr23, wuwh23}@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn
godjchen@tencent.com, malong@gmail.com

摘要—可控语音合成旨在通过参考输入控制生成语音的风格，该参考输入可以是各种模态。现有基于面部的方法由于数据质量限制而在鲁棒性和泛化能力上遇到困难，而文本提示方法提供的多样性和细粒度控制有限。尽管多模态方法旨在整合各种模态，但它们对完全匹配训练数据的依赖极大地限制了其性能和适用性。本文提出了一种三阶段多模态可控语音合成框架来解决这些挑战。对于面部编码器，我们使用监督学习和知识蒸馏来应对泛化问题。此外，文本编码器在文本-面部和文本-语音数据上进行训练以增强生成语音的多样性。实验结果表明，该方法在基于面部和基于文本提示的语音合成方面都优于单模态基线方法，突显了其在生成高质量语音方面的有效性。

***Index Terms*—语音合成，多模态合成，人脸到语音**

I. 介绍

近年来，文本到语音（TTS）合成技术取得了显著进展。生成的语音的可懂度和自然性现在几乎与人类的无法区分。同时，如何控制生成语音的风格以满足个性化需求对现有的语音合成系统提出了挑战。具体而言，可控语音合成旨在从参考输入中提取音色、韵律和风格信息，并基于这些信息从文本生成高度一致且逼真的语音。

根据参考输入的模态，可控语音合成可以分为四种类型：基于参考语音、基于面部图像、基于文本提示和基于多模态。基于参考语音 [1]–[3] 的方法从目标说话者的几秒钟语音中提取音色和韵律信息。大规模语音数据和模型架构的进步显著提升了这些方法的性能。在基于面部图像的语音合成 [4]–[7] 中，利用了从面部图像中导出的与说话者相关的信息。由于数据质量限制，这些方法在鲁棒性和泛化能力方面遇到了挑战。基于文本提示的方法 [8]–[10] 通过描述特征（包括性别、音调、语速等）的文本来控制语音风格。由于这些方法通常使用给定方面的有限组合，合成语音的多样性相对较低，并且也难以实现更精细的控制。

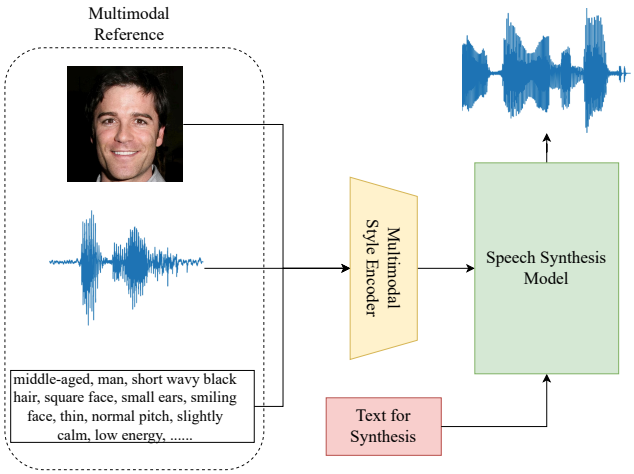


图 1: 一个统一的多模态可控语音合成框架。

由于现实世界的语音合成应用经常会遇到某些模态的输入可能不可用的情况，利用多模态输入提供了一个更实际的解决方案。通过整合各种模态的输入，基于多模态的系统可以灵活地生成更适合特定任务的语音。图 1说明了基于多模态的合成系统的框架，其中参考输入可以是任何模态。最近的基于多模态的语音合成模型 [11], [12] 关注情感和风格信息。一个显著的缺点是它们依赖于完全匹配的语音-面部-文本提示三元组进行训练。这种依赖限制了训练数据的质量和数量，从而限制了模型的性能和适用性。

在本文中，我们专注于说话人相关的风格，并提出了一种新的方法来解决多模态、基于面部和文本提示的

This work is supported by National Natural Science Foundation of China (62076144) and Shenzhen Science and Technology Program (JCYJ20220818101014030).

*Corresponding author.

方法中的上述问题。具体来说，我们将对语音-面部-文本提示三元组训练数据的需求转换为三种类型配对数据的要求，并将训练过程分为三个阶段。在第一阶段的训练中，我们对齐了面部和语音特征，并通过监督学习和知识蒸馏增强了面部特征的说话人一致性和泛化能力。在第二阶段，我们使用文本提示-面部和文本提示-语音数据来训练文本编码器，扩展了文本数据并增加了语音多样性。在第三阶段，我们仅使用语音作为参考输入来训练语音合成模型。得益于高质量的语音数据，合成语音的质量得到了提升。实验结果表明，我们的方法在基于面部和基于文本提示的语音合成中都优于单模态基准方法。演示可以在在线页面 [1] 上获取。

II. 相关工作

A. 基于人脸的语音合成

在基于面部的语音合成中，面部图像是用作条件输入。通过面部表达的说话人身份特征用于合成具有相似特征的语音。当无法从目标说话人那里获得语音样本时（例如为虚构角色创造声音）可以使用这些方法。

根据面部信息的利用方式，基于人脸的语音合成可以大致分为两类。(i) 两阶段方法。这些方法通常包括两个训练阶段。在第一阶段，来自语音或面部图像的参考信息被映射到一个一致的说话人嵌入空间中。在第二阶段，使用训练好的说话人嵌入来训练语音合成模型。Face2Speech [6] 使用类似于 GE2E [13] 损失的方式来训练人脸编码器，保持预训练语音编码器的泛化能力。此外，Synthesees [5] 使用监督学习来捕捉独特的说话人属性。(ii) 端到端方法。这些方法直接使用面部图像作为条件输入来训练语音合成模型，消除了初始对齐过程的需要。FaceTTS [4] 结合了额外的绑定损失以保持生成的语音与原始语音之间的相似性。FVTTS [7] 提出了一种端到端架构，避免了训练附加网络的必要性。然而，由于缺乏足够的文本-语音-面部三元组数据，这些方法的合成性能相对较低。

B. 基于文本提示的语音合成

基于文本提示的语音合成利用自然语言作为提示来生成有表现力的语音。这种方法消除了用户需要具备声学知识的需求，从而允许他们灵活地描述所需的语音。最近，PromptTTS [9] 是第一个使用文本描述作为提示来引导语音生成的方法。它们使用五个不同的因素来描述语音：性别、音高、语速、音量和情感。根据这些特征将语音数据分为几类，并为每一类编写风格提示。与 PromptTTS 不同，InstructTTS [8] 不对风格提示的格式施加限制。用户可以使用任何自然语言表达来描述他们所需的说话风格。为了精确控制说话者身份，PromptTTS++ [10] 引入了一个描述声音特征的说话者提示。在本文中，我们将面部描述纳入了训练过程，增强了合成输出的多样性。

[1] https://thuhcsi.github.io/icme2025-MMTTS/

III. 方法

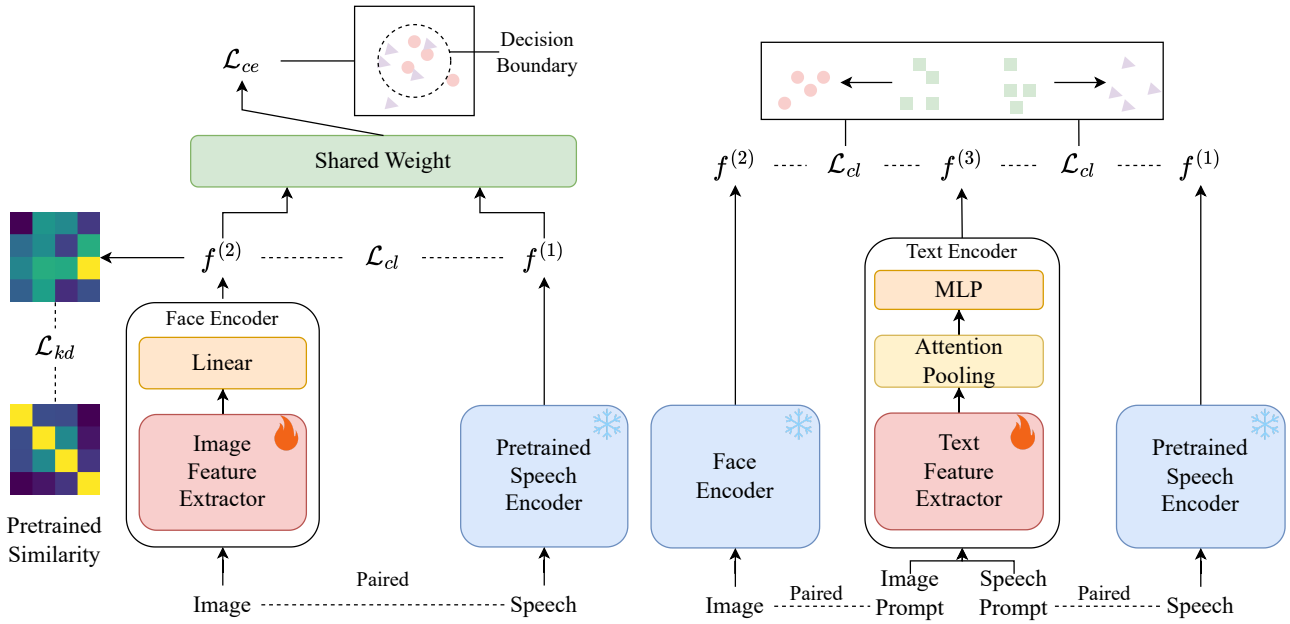
所提出的方法包括三个训练阶段：对齐面部嵌入和语音嵌入、对齐文本嵌入与面部和语音嵌入以及训练语音合成模型。

图 2a说明了人脸编码器的训练过程，目标是使提出的面部编码器与预训练且固定的语音编码器共享说话人空间。通过使用监督学习和知识蒸馏方法训练模型，基于面部的说话人表示变得更加具有判别性和泛化性。图 2b展示了文本编码器的训练过程。通过同时使用面部-文本提示对和语音-文本提示对，文本编码器获得了从各种提示生成多样化说话人嵌入的能力。最后，图 3演示了语音合成模型的训练和推理过程。我们使用高质量的语音数据和语音说话人编码器来训练语音合成模型。在推理阶段，提出的模型可以利用所有模态的参考输入合成交互式语音。在接下来的部分中，我们将详细介绍每个训练阶段。

A. 面部-语音对齐

在配对数据不足的情况下，强迫两种不同模态之间的对齐通常会面临泛化问题。为了训练面部编码器使其与预训练的语音编码器对齐并提高其泛化能力，我们使用了三个不同的损失项（$\mathcal{L}_{ce}$,$\mathcal{L}_{kd}$,$\mathcal{L}_{cl}$）。

1) 权重共享分类： 为了增强模型对同一说话人不同角度和光照变化的鲁棒性，我们在训练过程中采用了分类损失。该方法利用数据集中的监督信息来提高模型的泛化能力。然而，直接使用交叉熵损失训练面部嵌入可能会导致模型依赖于来自面部模态的特征，而不是依赖于跨模态共享的特征。因此，基于 AMSoftmax [14],



(a) 面部编码器训练过程　　(b) 文本编码器训练过程

图 2: 面部编码器（左）和文本编码器（右）的训练过程概述。

我们提出了一种模态间共享权重的分类损失，其公式如下：

$$\mathcal{L}_{ce}^{(o)} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s\cdot(W_{y_i}^T f_i^{(o)}-m)}}{e^{s\cdot(W_{y_i}^T f_i^{(o)}-m)}+\sum_{j=1,j\neq y_i}^{c} e^{s\cdot(W_j^T f_i^{(o)})}} \tag{1}$$

$$\mathcal{L}_{ce} = \alpha\mathcal{L}_{ce}^{(1)} + \mathcal{L}_{ce}^{(2)} \tag{2}$$

其中 n 是 mini-batch 中样本的数量，$W \in \mathbb{R}^{d\times c}$ 是一个共享的学习参数（d 表示说话人嵌入维度），$f_i^{(o)}$ 是模态 o 中的说话人嵌入（1 对应语音，2 对应面部图像，3 对应文本提示），y_i 是对应的说话人标签，c 是说话人的数量，α 是一个平衡超参数，而 m 和 s 是用于控制类间隔和尺度的超参数。

通过使用共享权重 W，决策边界被绘制到面部特征与语音特征重叠的区域。这允许从语音模态派生出的说话人识别信息传递给面部编码器，迫使其学习共享的说话人特征。

2) 联合知识蒸馏: 预训练的人脸识别模型展示了高水平的泛化能力。然而，在模态对齐后的表示空间可能会遇到模式崩溃的问题 [15]（即把不同的说话人映射到相似的表示）。

知识蒸馏从教师模型中提取知识并将其整合到学生模型中，从而增强学生的知识和性能。为了解决模式崩塌的问题，我们尝试将两种模态的预训练模型中的精炼知识转移到面部编码器中。鉴于学生模型与教师模型以及各种模态的教师模型之间的训练目标存在差异，我们构建了独立于任务的可转移知识项。

受 [16] 启发，我们利用教师知识构建相似性矩阵，该矩阵能够有效捕捉实例之间的高阶比较，并从教师那里提供更丰富的监督。具体来说，相似性矩阵是

$$S_{ij}^{(o)} = \cos(\hat{f}_i^{(o)}, \hat{f}_j^{(o)}) \tag{3}$$

其中 $\hat{f}_i^{(o)}$ 是从模态 o 的预训练教师网络中得到的嵌入，$\cos(\cdot,\cdot)$ 是余弦相似度。来自不同模态的相似性矩阵进一步合并为一个统一的相似性矩阵：

$$S_{ij} = \mu S^{(1)} + (1-\mu)S^{(2)}, \mu \in (0,1) \tag{4}$$

其中 μ 是一个平衡参数。通过相似性矩阵，我们有知识蒸馏损失为：

$$\begin{aligned}\mathcal{L}_{kd} = \sum_{i=1}^{n}\sum_{j=1}^{n}(||S_{ij} - \cos(f_i^{(1)}, f_j^{(2)})|| \\ + \beta||S_{ij} - \cos(f_i^{(2)}, f_j^{(2)})||)\end{aligned} \tag{5}$$

其中 β 是一个超参数。

知识蒸馏损失项确保模型在训练过程中保留来自教师模型的不同模态之间以及同一模态内的相似性，从而提高模型的泛化能力。

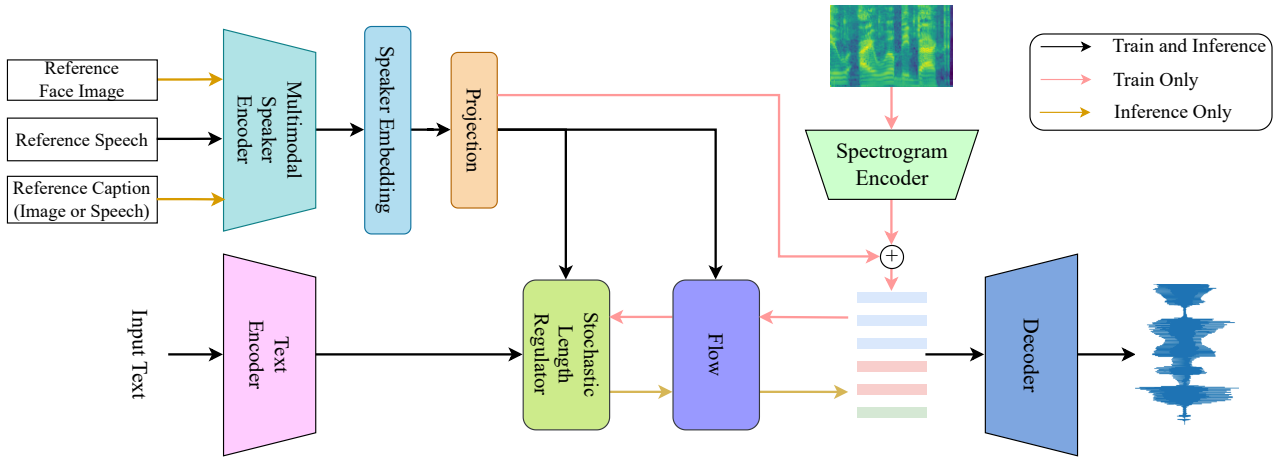


图 3: VITS 基于的语音合成模型的训练过程和推理过程。

3) 对比学习对齐: 对比学习 [17] 是多模态对齐中的一种流行方法。它通过鼓励来自相同说话者的嵌入之间的相似性同时确保来自不同说话者的嵌入之间的差异，将来自不同模态的嵌入统一到一个空间内。形式上，

$$\mathcal{L}_{cl} = \sum_{i=1}^{n} -\log \frac{e^{\cos(f_i^{(1)}, f_i^{(2)})/\tau}}{e^{\cos(f_i^{(1)}, f_i^{(2)})/\tau} + \sum_{f_{i^-}^{(2)} \in \mathcal{N}_i} e^{\cos(f_i^{(1)}, f_{i^-}^{(2)})/\tau}} \tag{6}$$

其中 $\mathcal{N}_i$ 是对应于 $f_i^{(2)}$ 的负样本集，τ 是温度参数。通过对比学习，模型可以学习更通用的表示，使其能够适应新数据并提高其灵活性和实用性。

最后，面部编码器的训练损失为:

$$\mathcal{L} = \mathcal{L}_{ce} + \gamma \mathcal{L}_{kd} + \mathcal{L}_{cl} \tag{7}$$

其中 γ 是用于平衡损失项的超参数。

B. 文本提示对齐

一旦获得了语音和面部编码器，我们就可以使用它们来训练文本编码器。文本编码器的架构如图 2b所示。我们使用变压器编码器模型作为文本特征提取器的基础。之后，我们利用注意力池化层后面跟着多层感知机（MLP）来获得句子级别的嵌入。

在训练阶段，同时使用三种模态的数据（文本提示、人脸和语音）。具体来说，我们首先获得人脸-文本提示对和语音-文本提示对。然后，我们分别使用前面描述的对齐损失将文本嵌入与人脸嵌入和语音嵌入对齐。

$$\mathcal{L}_{cl} = \sum_{i=1}^{2n} -\log \frac{e^{\cos(f_i^{(3)}, f_i^{(*)})/\tau}}{e^{\cos(f_i^{(3)}, f_i^{(*)})/\tau} + \sum_{f_{i^-}^{(*)} \in \mathcal{N}_i} e^{\cos(f_i^{(3)}, f_{i^-}^{(*)})/\tau}} \tag{8}$$

其中 $f_i^{(*)}$ 表示相应的语音或人脸模态的嵌入。

C. 语音合成

所提出的架构与各种语音合成模型兼容。本文中，我们使用了端到端的 VITS [18] 模型。VITS 结合了归一化流，并采用了对抗训练过程，这增强了其在生成式语音建模中的表现力。此外，VITS 引入了一个随机持续时间预测器，以从输入文本中合成具有不同节奏的语音。

在训练阶段，我们仅使用语音说话人编码器来确保模型性能。在推理阶段，我们可以利用来自各种模态的对齐编码器来实现个性化语音合成。

IV. 实验

A. 实验设置

对于面部-语音对齐训练，我们使用了广泛使用的 LRS3 数据集 [19]。它包含一个由 4,004 名说话者组成的训练集和一个由 457 名说话者组成的测试集。对于文本编码器的训练，我们利用了两个数据集：面部描述数据集 CelebAText-HQ [20] 和语音描述数据集 LibriTTS-P [21]。CelebAText-HQ 包含 15,010 张面部图像，每张图包含 10 个手动标注的描述该面部的标题。为了简化起见，我们使用了一个大型语言模型从每个标题句子中提取关键词，并将这些关键词组合成一个序列作为提示。LibriTTS-P 是基于 LibriTTS-R [22] 的数据集，其中包括说话风格层面和说话人特征层面的提示。它包含 2,443 名说话者以及总计 373,868 个标题。语音合成模型使用了 LibriTTS [23] 进行训练。

在训练的第一阶段，我们使用预训练于 VGGFace2 [24] 的 Facenet [25] 初始化图像特征提取器。并且利用多任务级联神经网络 [26]（MTCNN）进行人脸检测和对齐。为了使模型更鲁棒，我们在输入的人脸图像上进行了数据增强，包括随机旋转和随机颜色抖动。对于语音编码器，我们使用预训练于 VoxCeleb2 [27] 的 ECAPA-TDNN [28] 模型。在优化过程中，我们使用 Adam 优化器，批量大小为 96，学习率为 0.0002，总步数为 200,000。超参数设置为以下值：$\alpha = 0.1$,$m = 0.2$,$s = 30$,$\mu = 0.8$,$\beta = 0.1$,$\tau = 0.1$,$\gamma = 10$。

在训练的第二阶段，一个预训练的 T5 [29] 编码器作为初始文本特征提取器。对于每一步，包含 64 个语音-文本提示对和 64 个面部-文本提示对的一批数据被输入模型。类似于第一阶段，我们使用学习率为 0.0002 的 Adam 优化器进行总共 200,000 步训练。对于语音合成模型，我们以批次大小为 96、学习率为 0.0002 进行了 600,000 步的训练。

为了验证所提出方法的有效性，我们将其与几个基线模型进行了比较。对于基于人脸的语音合成，我们使用了 FaceTTS、Synthesees 和 Face2Speech。利用提供的预训练权重用于 FaceTTS。重新实现了 Synthesees 和 Face2Speech，并用 VITS 替换了语音合成模型以进行公平比较。由于面部编码器的训练目标是与语音编码器对齐，我们也采用了基于参考语音的方法作为上限进行比较。对于基于文本提示的语音合成，我们使用预训练的 PromptTTS++作为基线模型，该模型也采用风格提示和说话人提示。

B. 评估指标

我们进行了大量的实验来验证语音基础的语音合成性能。在客观评估中，通常用于说话者验证的等错误率（EER）和最小检测成本函数（minDCF）被用来评估说话者表示的区分能力。为了验证模型生成音频的自然度及其与参考音频的相似性，我们进行了主观评价：五级自然度平均意见得分（MOS）和说话人一致性 MOS 测试。

对于基于文本提示的语音合成，除了主观指标外，我们还使用轮廓系数。生成样本中的说话人嵌入根据性别被分为两组。通过轮廓系数，我们将数据簇内的凝聚性和区分性进行了量化。较低的轮廓系数表示该簇分布区域较大，从而展示了更好的多样性。

表 I: 基于人脸的语音合成的目标结果

Methods	EER (%) ↓	MinDCF ↓
Reference Speech	2.5445	0.1559
Face2Speech	11.0522	0.4594
Synthesees	8.2273	0.4684
Ours	**4.5801**	**0.2797**

C. 基于人脸的语音合成结果

1) 目标结果*:* 表 I给出了从面部图像和语音导出的说话人表示的 EER 和 MinDCF 结果。结果显示，我们的方法在区分度和未见说话人的泛化方面显著优于基线方法。仅使用对齐损失的 Face2Speech 方法取得了最差的结果，这表明利用来自面部数据集的标签信息进行监督是有益的。此外，与同样使用标签信息的 Synthesees 方法相比，我们的方法也有所改进。这表明利用预训练的人脸识别模型进行知识蒸馏可以增强模型的泛化能力。

2) 消融研究*:* 表 II给出了损失项的消融实验结果。具体来说，在对比学习损失的消融中，我们将对齐损失替换为余弦相似度损失。

表 II: 基于人脸的语音合成的消融研究结果

Methods	EER (%) ↓	MinDCF ↓
w/o $\mathcal{L}_{ce}$	9.5843	0.4052
w/o $\mathcal{L}_{kd}$	10.6870	0.4823
w/o $\mathcal{L}_{cl}$	8.0242	0.3864

从结果可以看出，替换对比学习损失对性能的影响相对较小，尽管仍然存在显著下降。这些结果证明了我们提出的监督学习方法和知识蒸馏方法的有效性，同时也突出了对比学习方法在表示学习任务中的优越性。移除知识蒸馏对结果影响最大，然而移除监督学习仍然可以达到与基线相当的性能。这表明模型可以通过知识蒸馏保留更多关于区分的知识，从而增强其泛化能力。

3) 主观结果*:* 表 III显示了合成语音的自然度（N-MOS）和相似度（S-MOS）的结果。我们的方法达到了最高的自然度，这一结果证明了使用高质量训练数据的三阶段训练策略可以生成更自然的语音。FaceTTS 的 N-MOS 表明，用一个小且质量低下的脸部-语音-文本数据集进行端到端合成预期语音是非常具有挑战性的。此外，我们的方法在说话人相似度方面也有所提高，这验证了从面部图像中捕捉与语音相关的风格属性并生成更合适的语音的有效性。

表 III: 基于人脸的语音合成主观结果

Methods	N-MOS ↑	S-MOS ↑
Ground Truth	4.31±0.09	4.32±0.10
Reference Speech	3.55±0.10	3.58±0.10
FaceTTS	2.36±0.10	2.18±0.12
Synthesees	2.94±0.11	2.69±0.13
Ours	**3.48±0.11**	**3.63±0.10**

D. 基于文本提示的语音合成结果

表 IV呈现了基于文本的语音合成实验结果。与基线方法相比，我们的方法在语音自然度和与文本提示匹配方面都取得了一定的进步。研究结果表明，我们的方法在自然语言建模和捕捉风格信息的能力方面表现出强大的能力。此外，我们的方法在轮廓分数上具有明显优势。这表明合成的说话人分布更广，这意味着使用面部提示和语音提示，我们的方法能够有效地生成更多样化的语音。

表 IV: 基于文本提示的语音合成结果

Methods	N-MOS ↑	S-MOS ↑	Silhouette Score ↓
PromptTTS++	3.42±0.08	3.23±0.09	0.1365
Ours	**3.62±0.08**	**3.75±0.09**	**0.0904**

V. 结论

在本文中，我们介绍了一个多阶段多模态可控语音合成框架，以解决完全匹配的训练数据所带来的限制，并提高合成语音的质量。该方法采用监督学习和知识蒸馏来处理不一致性和增强面部特征的一般化能力。此外，文本编码器是在文本提示-面部和文本提示-语音数据的结合上进行训练的，从而增加了生成语音的多样性。实验结果表明，我们的方法在基于面部和基于文本提示的语音合成中都优于单一模态的基础模型，证实了其在生成高质量语音方面的有效性。

参考文献

[1] Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Zhenhui Ye, Shengpeng Ji, Qian Yang, Chen Zhang, Pengfei Wei, Chunfeng Wang, et al., "Mega-tts 2: Boosting prompting mechanisms for zero-shot speech synthesis," in *The Twelfth International Conference on Learning Representations*, 2024.

[2] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al., "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.

[3] Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu, "Maskgct: Zero-shot text-to-speech with masked generative codec transformer," *arXiv preprint arXiv:2409.00750*, 2024.

[4] Jiyoung Lee, Joon Son Chung, and Soo-Whan Chung, "Imaginary voice: Face-styled diffusion model for text-to-speech," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[5] Jae Hyun Park, Joon-Gyu Maeng, TaeJun Bak, and Young-Sun Joo, "SYNTHE-SEES: Face based text-to-speech for virtual speaker," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10321–10325.

[6] Shunsuke Goto, Kotaro Onishi, Yuki Saito, Kentaro Tachibana, and Koichiro Mori, "Face2Speech: Towards multi-speaker text-to-speech synthesis using an embedding vector predicted from a face image.," in *INTERSPEECH*, 2020, pp. 1321–1325.

[7] Minyoung Lee, Eunil Park, and Sungeun Hong, "FVTTS : Face based voice synthesis for text-to-speech," in *Interspeech 2024*, 2024, pp. 4953–4957.

[8] Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng, "Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[9] Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan, "Prompttts: Controllable text-to-speech with text descriptions," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, pp. 1–5, IEEE.

[10] Reo Shimizu, Ryuichi Yamamoto, Masaya Kawamura, Yuma Shirahata, Hironori Doi, Tatsuya Komatsu, and Kentaro Tachibana, "PromptTTS++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024, pp. 12672–12676, IEEE.

[11] Xiang Li, Zhi-Qi Cheng, Jun-Yan He, Xiaojiang Peng, and Alexander G Hauptmann, "Mm-tts: A unified framework for multimodal, prompt-induced emotional text-to-speech synthesis," *arXiv preprint arXiv:2404.18398*, 2024.

[12] Wenhao Guan, Yishuang Li, Tao Li, Hukai Huang, Feng Wang, Jiayan Lin, Lingyan Huang, Lin Li, and Qingyang Hong, "MM-TTS: Multi-modal prompt based style transfer for expressive text-to-speech synthesis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 18117–18125.

[13] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[14] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[15] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan, "Bridging the gap between object and image-level representations for open-vocabulary detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33781–33794, 2022.

[16] Peng-Fei Zhang, Jiasheng Duan, Zi Huang, and Hongzhi Yin, "Joint-teaching: Learning to refine knowledge for resource-constrained unsupervised cross-modal retrieval," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1517–1525.

[17] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[18] Jaehyeon Kim, Jungil Kong, and Juhee Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. 2021, pp. 5530–5540, PMLR.

[19] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[20] Jianxin Sun, Qi Li, Weining Wang, Jian Zhao, and Zhenan Sun, "Multi-caption text-to-face synthesis: Dataset and algorithm," in *Proceedings of the 29th ACM International Conference on Multimedia*, New York, NY, USA, 2021, Mm '21, pp. 2290–2298, Association for Computing Machinery.

[21] Masaya Kawamura, Ryuichi Yamamoto, Yuma Shirahata, Takuya Hasumi, and Kentaro Tachibana, "LibriTTS-p: A corpus with speaking style and speaker identity prompts for text-to-speech and style captioning," *arXiv preprint arXiv:2406.07969*, 2024.

[22] Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna, "Libritts-r: A restored multi-speaker text-to-speech corpus," *arXiv preprint arXiv:2305.18802*, 2023.

[23] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[24] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[26] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[27] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[28] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.