浙江工业大学

本科生毕业论文（设计）

题目：数字人视频创作平台设计与开发

学　　院：计算机科学与技术学院

专　　业：软件工程

班　　级：2019软件工程（移动应用开发方向1）

学　　号：201906061609

学生姓名：吕锐

指导老师：张繁

提交日期：2023.6.6

数字人视频创作平台设计与开发

摘　要

本文提出了一种数字仿生人的系统，该系统可以根据输入的文本生成高质量的仿真说话视频。本文的系统结合了三种先进技术：语音合成、人脸重建和面部表情合成。

具体来说，本文使用VITS模型进行语音合成，使用Deep3DFaceRecon_pytorch模型进行人脸重建，使用FACIAL模型进行面部表情合成，并使用Wav2Lip-GFPGAN对嘴唇形态进行精调。本文的语音合成为合成人脸提供输入，使用了端到端的语音合成（TTS）方法VITS,该方法采用归一化流程扩充的变分推理（基于归一化流的变分增广）和对抗性训练过程，提高了生成式建模的表达能力。合成人脸属性不仅包括与语音具有高度相关性的显式特征，例如嘴唇运动，还包括与输入音频仅具有弱相关性的隐性特征，例如头部姿势和眨眼。为了使用输入音频对不同面部属性之间的这种复杂关系进行建模，本文使用了一种 FACe隐式属性学习生成对抗网络 (FACIAL-GAN)，它集成了语音感知、上下文感知和身份感知信息来合成 3D 面部动画具有逼真的嘴唇、头部姿势和眨眼动作。然后，本文再通过Rendering-to Video网络将渲染后的人脸图像和眨眼的注意力图作为输入来生成逼真的输出视频帧。最后通过wav2lip精准校对嘴型。

实验结果表明，本文的方法可以生成逼真的语音和谈话面部视频，不仅具有同步的嘴唇运动，还具有自然的头部运动和眨眼，其质量优于最先进方法的结果。

关键词：数字人，TTS，人脸合成，GAN，wav2lip

Design and Development of a Digital Human Video Creation Platform

ABSTRACT

In this paper, a digital android system is proposed, which can generate high-quality speech simulation video based on the input text. Our system combines three advanced technologies: speech synthesis, facial reconstruction, and facial expression synthesis.

Specifically, we use the VITS model for speech synthesis and Deep3DFaceRecon_ The Pytorch model is used for facial reconstruction, FACIAL model is used for facial expression synthesis, and Wav2Lip GFPGAN is used for fine tuning of lip shape. The speech synthesis in this article provides input for synthesizing faces, using the end-to-end speech synthesis (TTS) method VITS. This method uses normalized process extension variational reasoning (based on normalized flow variational augmentation) and adversarial training processes to improve the expressiveness of generative modeling. Synthetic facial attributes not only include explicit features that are highly correlated with language, such as lip movements, but also implicit features that are only weakly correlated with input frequency, such as head posture and blinking. In order to model the complex relationship between different facial attributes in the input frequency, this paper uses a FACE implicit attribute learning method called FACIAL GAN, which integrates language perception, upper and lower facial perception, and body perception information to synthesize 3D facial animations with realistic lip, head posture, and blink movements. Then, this article uses the Rendering to Video network to generate realistic output video frames by using the rendered facial image and blinking attention map as inputs. Finally, accurately calibrate the mouth shape through wav2lip.

The experimental results show that our method can generate realistic speech and conversation videos, not only with synchronous lip movements, but also with natural

head movements and blinking, and its quality is superior to the results of state-of-the-art methods.

KEY WORDS: Digital human, TTS, facial synthesis, GAN, wav2lip

目 录

第一章 绪论

随着计算机图形学技术的发展，虚拟人、数字人及虚拟数字人逐渐成为我们与之产生真实情感互动的新载体。 虽然在一些场合下这三种概念可以通用，但严格意义上它们又存在微小差异。其中，虚拟人是虚构的，不存在于现实世界中。数字人则是强调其存在于数字世界中。而虚拟数字人则更强调虚拟身份和数字化制作特性。

从手工绘制到电脑绘图，再到如今的人工智能合成，虚拟数字人的理论和技术实现不断成熟和完善，应用领域持续扩大，产业也在逐步形成并不断丰富，对应的商业模式也在不断演进和多样化。 数字人的产业链主要分为三层：基础层、平台层和应用层。

基础层提供虚拟数字人所需的基础软硬件支撑，包括显示设备、光学器件、传感器、芯片等硬件设备，以及建模软件、渲染引擎等基础软件。显示设备是数字人的承载体，其中既包括2D显示设备如手机、电视、投影、LED显示器等，也包括3D显示设备如裸眼立体、AR和VR等。光学器件则用于数字人视觉传感器和用户显示器的制造。传感器则用于采集数字人原始数据及用户数据。芯片则主要用于传感器数据预处理、数字人模型渲染以及AI计算。而建模软件则能对虚拟数字人的人体、衣物进行三维建模，而渲染引擎则能对灯光、毛发、衣物等进行渲染。目前主流的渲染引擎包括Unity Technologies公司的Unity 3D以及Epic Games公司的Unreal Engine等。总体来看，基础层的厂商已经深耕行业多年，技术壁垒较为深厚。

平台层则包括软硬件系统、生产技术服务平台和AI能力平台，为虚拟数字人的制作及开发提供技术能力。建模系统和动作捕捉系统可以获取真人/实物的各类信息，并利用软件算法实现对人物的建模和动作的重现；渲染平台则用于虚拟数字人模型的云端渲染。解决方案平台则根据自身技术能力为广大客户提供数字人解决方案。而AI能力平台则提供计算机视觉、智能语音以及自然语言处理等技术能力。平台层汇聚了较多企业，例如腾讯、百度、搜狗、魔珐科技以及相芯科技等均提供相应的虚拟数字人技术服务平台。

应用层则是指虚拟数字人技术与实际应用场景领域结合，形成各种行业应用解决方案，赋能行业领域。根据应用场景或行业的不同，已经出现了娱乐型数字人（如虚拟主播、虚拟偶像）、教育型数字人（如虚拟教师）、助手型数字人（如虚拟客服、虚拟导游、智能助手）、影视数字人（如替身演员或虚拟演员）等。

有不同外形和功能的虚拟数字人可以赋能电影、传媒、游戏、金融、文旅等领域，并根据需求为用户提供定制化服务。

第二章 相关工作

本项目的数字人需要多个模块和技术进行实现，该部分将本项目拆分成语音合成、三维面部重建、语音驱动面部、唇形校准这几个关键技术进行描述。

2.1 语音合成的研究现状

语音合成（Text to speech synthesis），旨在从文本合成可理解和自然的语音，在人类通信中有着广泛的应用，长期以来一直是人工智能、自然语言和语音处理领域的研究项目。开发TTS系统需要有关语言和人类语音生成的知识，涉及多个学科，包括语言学、声学、数字信号处理和机器学习。随着深度学习的发展，基于神经网络的TTS蓬勃发展，大量研究工作集中在神经TTS的不同方面。因此，近年来合成语音的质量得到了很大的提高。

文本到语音（TTS）系统通过几个组件从给定文本合成原始语音波形。随着深度神经网络的快速发展，除了文本预处理（如文本规范化和音素化）之外，TTS系统管道已简化为两阶段生成建模。第一阶段是从预处理文本中产生中间语音表示，如mel谱图（Shen等人，2018年[1]）或语言特征，第二阶段是产生基于中间表示的原始波形（Oord等人，2016年[2]以及Kalchbrener等人，2018年[3]）。每个两级管道的模型都是独立开发的。

基于神经网络的自回归TTS系统已显示出合成真实语音的能力（Shen等人，2018年，以及Li等人，2019年[4]），但其顺序生成过程使其难以充分利用现代并行处理器。为了克服这一限制并提高合成速度，已经提出了几种非自回归方法。在文本到声谱图生成步骤中，尝试从预训练的自回归教师网络中提取注意力图（Ren等人，2019年[5]；Peng等人，2020年[6]），以降低文本和声谱图之间的学习对齐的难度。最近，基于似然性的方法通过估计或学习使目标mel谱图的似然性最大化的比对，进一步消除了对外部比对器的依赖（Zeng等人，2020年[7]；Miao等人，2020年[8]；Kim等人，2020年[9]）。同时，在第二阶段模型中探索了生成性对抗网络（GAN）（Goodfellow等人，2014年[10]）。基于GAN的前馈网络具有多个鉴别器，每个鉴别器区分不同尺度或周期的样本，实现高质量的原始波形合成（Kumar等人，2019年[11]；Binkowski等人，2019年[12]；Kong等人，2020年[13]）。

尽管并行TTS系统取得了进展，但两阶段管道仍然存在问题，因为它们需要顺序训练或微调以进行高质量生产，其中后期模型使用早期模型的生成样本。此外，它们对预定义的中间特征的依赖妨碍了应用学习到的隐藏表示来进一步提高性能。最近，一些工作，即FastSpeech 2s[14]和EATS[15]，提出了有效的端到端训练方法，例如对短音频片段而不是整个波形进行训练，利用mel谱图解码器来帮助文本表示学习，并设计一个专门的谱图损失来缓解目标和生成的语音之间的长度不匹配。然而，尽管通过利用学习的表示可能会提高性能，但它们的合成质量仍落后于两阶段系统。

结合VAE和FLOW的前沿架构的VITS[16]（Variational Inference with adversarial learning for end-to-end Text-to-Speech）是一种结合变分推理（variational inference[17]）、标准化流（normalizing flows[18]）和对抗训练的高表现力语音合成模型。VITS通过隐变量而非频谱串联起来语音合成中的声学模型和声码器，在隐变量上进行随机建模并利用随机时长预测器，提高了合成语音的多样性，输入同样的文本，能够合成不同声调和韵律的语音。

2.2 语音驱动面部的研究现状

富有表情的面部动画是现代计算机生成的电影和数字游戏的重要组成部分。目前，基于视觉的表演捕捉，即用观察到的人类演员的动作驱动动画脸部，是大多数制作流程的一个组成部分。虽然从捕捉系统获得的质量在稳步提高，但制作高质量面部动画的成本仍然很高。首先，计算机视觉系统需要复杂的设置，通常还需要劳动密集型的清理和其他处理步骤。第二个不太明显的问题是，无论何时录制新的镜头，演员都需要在现场，理想情况下还要保持他们的形象。

语音驱动面部的目标是仅基于有声音轨生成可信且富有表现力的3D面部动画。为了使结果看起来自然，动画必须考虑到一些复杂性和相互依赖的现象，包括音素共发音、词汇压力以及面部肌肉和皮肤组织之间的相互作用。因此，语音驱动面部技术关注的是整个面部，而不仅仅是嘴和嘴唇。采用数据驱动的方法，以端到端的方式训练深度神经网络，以复制在训练数据中观察到的相关效果。

代表性的的技术当属2017年Nvidia提出的语音驱动人脸3DMesh[19]，目前被使用在Omniverse Audio2Face[20] 应用程序中。该文提出了一种端到端的卷积网络，从输入的音频直接推断人脸表情变化对应的顶点位置的偏移量。为了解决声音驱动过程中，情绪变化对表情驱动效果的影响，网络自动从数据集中学习情绪状态潜变量。推理阶段，可将情绪潜变量作为用户输入控制参数，从而输出不同情绪下的说话表情。

此外，发表于 ICCV 2021的论文“具有隐式属性学习的动态谈话人脸视频生成（FACIAL: Synthesizing Dynamic Talking Face with Implicit Attribute Learning）”[21]使用对抗学习网络联合学习这一过程中的隐式和显式属性，在这个框架中设计了一个特殊的 FACIAL-GAN网络来共同学习语音、上下文和个性化信息，还可以预测眨眼信息，这些信息被进一步嵌入到最终渲染模块的眼部相关的注意力图中，用于在输出视频合成逼真的眼部运动信息。

2.3 三维面部重建的研究现状

作为计算机视觉的核心问题，基于多视图立体的三维重建技术已经广泛应用于3D打印、离线地图重建和文物修复等行业应用中。

deep3D face reconstruction提出了一种基于 CNN 的单图像人脸重建方法，该方法利用混合级图像信息进行弱监督学习，无需真实 3D 形状。综合实验表明，作者的方法在准确性和鲁棒性方面都大大优于以前的方法。作者还提出了一种使用 CNN 的新型多图像人脸重建聚合方法。在没有任何显式标签的情况下，作者的方法可以学习测量图像质量并利用不同图像中的互补信息来更准确地重建 3D 人脸。

deep3D face reconstruction还提出了一种用于多图像人脸重建聚合的新型形状置信学习方案。本文的置信预测子网也以弱监督的方式进行训练，没有真实标签。实验表明，该方法明显优于朴素聚合（例如，形状平均）和一些启发式策略。据本文所知，这是第一次尝试从无约束图像集进行基于CNN的3D人脸重建和聚合。

2.4 唇形校准的研究现状

讲座、著名电影或面向全国的公共广播，如果翻译成所需的目标语言，就可以为数百万新观众所用。此类翻译的一个关键的问题是校正口型同步以匹配所需的目标语音。因此，与给定输入音频流匹配的口型说话面部视频在研究界受到了相当大的关注。

最近一些关于巴拉克奥巴马视频的作品实现了逼真的谈话面部视频生成。他们学习输入音频之间的映射以及相应的唇部地标。由于他们只接受特定说话人的训练，他们无法用于新的说话者或不同的声音。

然而Wav2lip解决了这些问题，具有广泛性的定量评估表明，Wav2Lip 模型生成的视频的口型同步准确性几乎与真实同步视频一样好。开发人员还在他们的

网站上提供了一段演示视频，清楚地展示了Wav2Lip模型的重大突破和其评估基准。

2.5 本章小结

本章对国内研究现状进行了研究，说明了本项目的主要研究方向的各模块作用。

第三章 本文的实现方法

3.1 数字仿生人的难点与解决

数字仿生人任务是指将人类的语音、面部表情等信息转化为机器可以处理的数字信号，然后通过算法和技术实现真实的视频输出。由于数字仿生人任务涉及到多个领域的信息处理，包括音频、图像、三维模型等，使得实现数字仿生人任务充满了挑战性。本部分将详细介绍数字仿生人任务中的难点，并且分别介绍了四个模块的技术是如何解决这些问题的，为何选择这些技术。这四个模块分别为语音合成模块、人脸建模模块、视频生成模块和唇形校准模块。

3.1.1 语音合成的难点与解决

语音合成的第一个难点主要体现在如何将输入文本转化为自然流畅的语音信号上。传统的语音合成方法通常需要对大量语音语料库进行训练才能生成高质量的语音信号，且合成的语音难以模拟人类语音的自然流畅性和情感表达等特征。语音的自然度是语音合成过程中最重要的一个指标，即生成的语音信号需要与人类语音相似，让人听起来自然而流畅。但是，实现自然的语音并不容易，因为人类语音具有非常复杂的结构和变化，而语音合成系统需要对这些结构和变化进行准确建模。同时，语音合成还需要考虑发音错误、语音断断续续、语速等多种因素，这些因素都会影响语音合成的自然度。

第二个难点在于使用自动化的语音合成系统需要确保其能够产生稳定性好的结果，即在不同场景下和不同的输入文本下都要有较为稳定的表现。但是，在实际应用中，经常会发现语音合成的质量会变化的问题。这可能是由于训练集和测试集之间的分布不匹配导致的，或者是语言、音频信号等方面的差异。

第三个难点在于泛化性是指机器学习模型在未见过的数据上的表现能力，它在语音合成中尤为重要。因为语音合成需要考虑到发音、口音、语速等多个因素，因此对于不同的人和场景，生成的语音必须具备良好的适应性。但是，泛化性是一个非常难以实现的目标，因为语音信号之间的差异是非常大的，而且语音合成的质量取决于所训练的数据集的质量和数量。

第四个难点在于语音合成需要快速地生成输出结果，这在实时应用场景下非常关键。同时，语音合成的效率也通过一些技术手段得到提高，比如语音的预处

理和缓存以及多线程计算等。然而，语音合成的效率问题仍然是一个值得思考的问题，因此需要设计高效的算法和模型来解决这个问题。

为了解决以上问题，第一个模块使用的技术是“Variational Inference Text-to-Speech (VITS)”（Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech），该模型使用了变分推断（Variational Inference）和对抗学习（Adversarial Learning）来生成真实而流畅的语音信号。与传统的语音合成模型不同的是，VITS可以直接从文本转化为语音信号，而不需要中间的表征转换。这个模型的优势在于，它可以生成自然、流畅的语音信号，避免了传统方法中的瑕疵，并且在减少数据集的情况下也可以取得较好的效果。

三个主要网络结构均为可并行的非自回归结构保证了合成速度：

(1)和Fastspeech系统相同的transformer作为文本Encoder；

(2)和Glow-TTS相同的Flow结构作为VAE的主体；

(3)和HiFiGAN生成器相同的反卷积作为Decoder。

此外，VITS采用了Glow-TTS相同的单调对齐搜索算法(MAS), 保证生成对齐的稳定性，使其长文本稳定性更好。在预测音素时长的模块中也引入Flow结构增加生成韵律的多样性，使其拥有更好的语音多样性。

3.1.2 人脸建模难点与解决

人脸建模的第一个难点在于人脸的形态多变且具有高度的多样性，使得人脸建模需要具备高精度和高鲁棒性。在数字仿生人任务中，要实现语音输入、视频输出的功能，需要将输入的语音和视频信息与3D人脸模型相匹配，因此对人脸模型的构建和准确性要求较高。

第二个难点在于生成高质量、逼真的三维人脸模型需要使用大量、高质量的训练数据来训练模型。然而，在实践中，获取高质量的三维人脸数据是一项具有挑战性的任务。实际上，一些三维人脸数据集存在缺陷，如遮挡、表情变化、光线变化等。此外，数据量也是一个问题，因为训练一个复杂的神经网络模型需要大量的数据才能达到理想的效果。

第三个难点在于三维面部结构的准确重建。因为人脸的形态和细节非常复杂，不仅受到姿势、表情、光照、遮挡等因素的影响，而且还与不同人的面部结构差异很大。因此，对于面部结构进行准确的重建是非常具有挑战性的。

为了解决以上问题，第二个模块使用的技术是“Accurate 3D Face Reconstruction with Weakly-Supervised Learning”（Deep3DFaceRecon），该模型使用了深度学习技术和弱监督学习技术，通过单张面部图像估计面部3D形状。这个模型的优势在于，它能够高效地学习到人脸的形态特征，从而生成准确的3D人脸模型，实现了高精度且高鲁棒性的人脸建模，满足数字仿生人任务中对人脸模型的需求。

Deep3DFaceRecon所使用的技术通过使用弱监督学习方式，可以从单张图片中重建三维人脸模型。这意味着，无需使用昂贵的三维扫描仪或者获取大量的三维面部数据来训练模型，只需要使用标准的单张静态图像即可进行训练，因此数据集的广泛适用性得到了极大的提高。

通过使用弱监督学习和深度学习技术，该技术可以从单张图片中提取面部结构的相关信息，并以此进行三维重建，同时还可以处理一些常见的变形如表情变化、光照变化等，从而生成高质量、逼真的三维人脸模型。此外上述项目所使用的技术在生成高质量三维人脸模型的基础上，同时保证了实时性和高效性。通过使用PyTorch作为开发工具箱，该技术可以很容易地采用GPU加速，从而显著提高了计算速度。此外，虽然该技术需要较长的训练时间，但一旦训练完成后，在实时应用中可以快速地生成三维人脸模型，从而实现高效、实时的三维合成。

3.1.3 语音驱动面部难点与解决

语音驱动面部第一个的难点主要体现在如何将不同来源的信息集成到一个连贯的图像序列或视频流中。合成的图像需要符合现实场景、充满真实感和动态变化性等特征。

第二个的难点主要体现在语音和面部表情在时间上是相关联的，但是它们之间的时差不确定，可能在不同的时间出现。如果没有准确地捕捉到这些时差信息，就会导致生成的面部表情与语音不匹配。例如，在特定的口型下发出声音，如果模型不能正确地捕捉到这个时间差，就可能导致生成的面部表情不自然，或者与语音不一致。

第三个的难点主要体现在语音中的韵律和语调可以表达情感和意图，同时也会对面部表情产生影响。例如，高兴的声音会导致面部表情变得明亮和愉悦，而生气的声音则会导致面部表情变得紧张和愤怒。

第四个的难点主要体现在语音驱动的面部生成任务中，正确的上下文信息可以帮助模型更好地理解语音信号和面部表情之间的关系，进而产生更加自然流畅的面部表情。例如，在理解一句话的时候，语音识别模型需要考虑前后文的语境，才能更好地识别其中的意思。

为了解决这些问题，第三个模块使用的技术是“FACIAL: Synthesizing Dynamic Talking Face with Implicit Attribute Learning”，该模型使用了条件生成对抗网络（CGAN）和自编码器（autoencoder）等深度学习技术，通过输入文本和3D人脸模型的信息，生成连贯且真实的说话视频。这个模型的优势在于，它能够高效地将语音、人脸模型和图像信息融合起来，生成连贯的说话视频，并且能够动态地模拟人类讲话时面部表情和口型变化等特征。

对于第二个难题，FACIAL项目采用了隐式属性学习的方法。其基本思想是将面部表情和语音信息映射到一个共同的特征空间中，然后在该特征空间中做匹

配。这样既可以防止时间差对结果的影响，同时还能够保证输出的人脸能够恰好呈现所要表达的语音信息。

FACIAL项目还使用了多任务学习的方法。具体来说，通过引入情感分类任务和语调分类任务，模型可以更好地理解语音信号和面部表情之间的关系，并相应地调整面部表情。此外，FACIAL还使用了隐式属性学习技术，它可以捕捉到面部表情和语音信号之间的潜在因素，从而进一步提高模型的灵活性和准确性。

为了解决最后一个难题，FACIAL项目使用了获取上下文信息的技术。具体来说，模型可以通过对长句子进行分段，并分别对每段进行语音驱动的面部生成，从而获得合适的上下文信息。此外，FACIAL还使用了隐式属性学习技术，它可以捕捉到面部表情和语音信号之间的潜在因素，进一步提高模型的灵活性和准确性。

3.1.4 唇形校准难点与解决

实现精准校准唇形的第一个的难点主要体现在如何使得合成的说话视频能够完美配合语音信号。唇形与语音信号之间的关系非常复杂，因此要实现精准校准唇形需要借助一些先进的技术手段。

第二个的难点主要体现在数字仿生人任务中，训练数据的多样性和可扩展性是一个非常重要的问题。传统的基于人工标注的数据集收集方式，并不能完全覆盖所有情况，因此很难实现对任意说话者、语言、视频的合成。

第四个模块使用的技术是“Wav2Lip-GFPGAN”，该模型使用了深度神经网络和生成对抗网络，通过多帧图像映射到音频波形，实现精准校准唇形。这个模型的优势在于，它能够实现高效的视听同步，获得高精度的模型输出，使得合成的说话视频与语音信号的配合程度更高，达到更加逼真和自然的效果。

此外Wav2Lip-GFPGAN项目采用了两种数据增强技术。一种是采用人脸关键点检测器进行多角度截取，从而增加训练数据量和质量；另一种是在训练过程中加入了对抗损失函数，强制模型生成更加真实的嘴部形态。这种数据增强技术能够大大提高训练数据的多样性和可扩展性，从而使得模型可以适应更广泛的说话者、语言和视频。

总之，数字仿生人任务是一个复杂的工程，需要涉及到人脸建模、语音合成、图像合成、唇形校准等多个方面。上述项目使用的各种技术手段都能有效地解决数字仿生人任务中的难点问题，使得生成的数字仿生人具有更高的真实感、自然性和可信度。

3.2 本文的主要工作

本文介绍了一个数字仿真人的项目，该项目通过一段语音和一段视频的训练，能将输入文本输出说话视频，本文的主要工作如下：

(1)根据对近几年数字人的国内外现状研究，将各个先进的技术结合起来，完成数字仿生人的设计和构想。

(2)通过对各个部分算法关键技术的研究和复现部署，完成对数字仿生人各模块的分别实现。

(3)通过对各模块的输入输出数据处理完成各模块的拼接，完成从文本输入到数字仿生人的总体实现。

本文将整个项目设计为四个模块。第一个模块使用 Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech 项目，将输入文本转化为语音。第二个模块使用 Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set 项目，将视频帧转化为人脸模型。第三个模块使用 FACIAL: Synthesizing Dynamic Talking Face with Implicit Attribute Learning 项目，将前两个模块生成的人脸模型和语音结合生成新的说话视频。第四个模块使用 Wav2Lip-GFPGAN 项目精准校准唇形。下面将介绍各模块使用的技术以及原理。

3.3 语音合成模块原理

语音合成模块使用了VITS，语音合成部分的三个主要网络结构是文本Encoder、VAE的主体和Decoder。

文本Encoder是将输入的文字转化为向量表示的模块。在这个项目中，使用的是Transformer模型，它是一种基于自注意力机制（self-attention）的序列到序列模型，能够有效地处理长文本序列，并且参数数量比LSTM等传统模型更少，训练速度更快。该模型通过多层的自注意力和前馈神经网络实现输入序列的编码，输出一个每个词位置对应的向量表示。

VAE的主体是语音合成模块的核心部分，它将语音信号映射到潜在空间中，并从该空间中采样生成新的语音信号。在这个项目中，VAE的主体使用了Glow-TTS相同的Flow结构，Flow结构通过一系列的可逆变换将输入数据映射到潜在空间中，并通过相反的可逆变换将潜在空间中的向量映射回生成的语音信号。Flow结构的特点是可以处理各种长度的语音，并且能够学习到语音的分布信息，生成高质量的语音信号。

Decoder是将潜在空间中的向量输出为语音信号的模块。在这个项目中，Decoder使用了HiFiGAN生成器相同的反卷积结构，将潜在空间中的向量解码成语音信号。反卷积结构通过一系列的转置卷积层将潜在空间中的向量解码成语音信号，其中每个卷积层中都包含一个可学习的过滤器和偏置项。反卷积的输出经过逐样本标准化后即可得到最终的语音信号输出。

总之，语音合成部分的网络结构包括文本Encoder、VAE的主体和Decoder，它们能够有效地将输入文本转化为高质量的语音信号。这个项目中使用的模型和算法能够处理不同长度的语音信号，并且能够学习到语音信号的分布信息，生成高质量的语音信号。

此外对抗生成网络（GAN）则是一种生成式模型与判别式模型的组合，它可以通过微调两个神经网络来生成非常逼真的图像或其他类型的数据。GAN在语音合成中的作用是生成高质量的语音信号，它通常被用作VAE的辅助模型。

在这个项目中，GAN的具体应用是通过对抗训练的方式，学习到一个生成器网络，能够将潜在空间中的向量解码成高质量的语音信号。GAN模型通常由两个部分组成：生成器和判别器。生成器学习将潜在空间中的向量映射成高质量的语音信号，判别器学习以尽可能准确的方式判断一个输入音频是真实的还是由生成器生成的伪造音频。在训练过程中，生成器和判别器通过对抗训练，相互制约，最终生成器的输出会更加接近真实音频信号。

在语音合成中，GAN可以帮助本文解决语音信号稀疏和模糊等问题，使模型能够学习到更加真实的语音信号分布，从而生成更高质量的语音信号。GAN的使用在一定程度上增强了模型的生成能力和逼真度，使得生成的语音信号更加自然、流畅和可信。

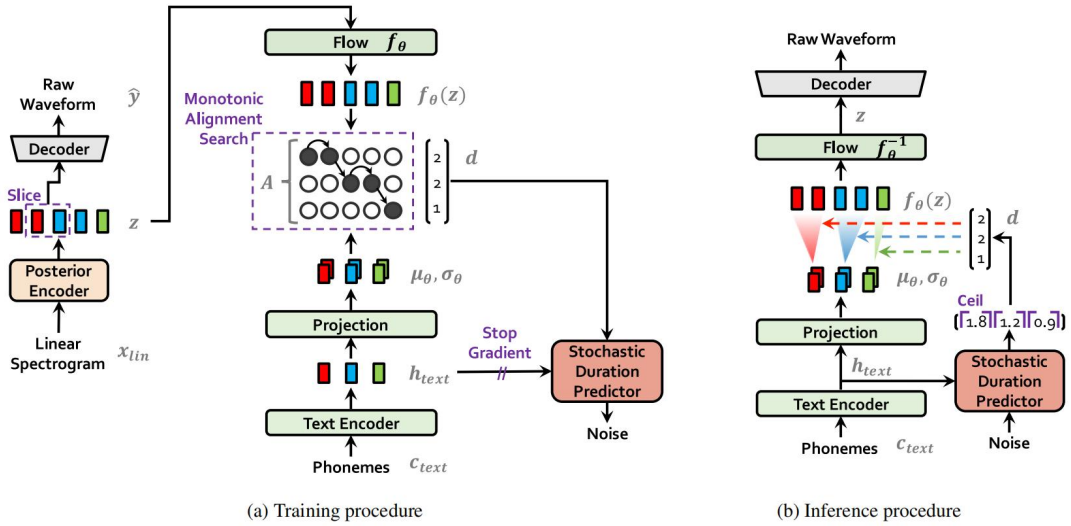

图 3-1 VITS模型结构

在本文的项目中，VAE 被用来学习语音数据的分布，并生成与输入文本相对应的语音样本。具体来说，CVAE 的输入是文本序列X，输出是对应的语音信号Y。模型的训练过程可以描述为：对于每一个文本序列X，首先从一个高斯分布中采样一个随机向量列z，然后将X和z传入 VAE 中，得到生成的语音信号Y。模型的损失函数由对抗生成网络和重构误差两部分组成：

$$L_{CVAE} = L_{adv} + \alpha L_{recon} \tag{3-1}$$

其中，L_{adv}是对抗损失函数，用来优化生成器的输出质量；L_{recon}是重构误差，用来保证生成的语音与原本的语音尽可能相似；α是一个超参数，用来调整两个损失函数的权重。

3.4 人脸建模模块原理

人脸建模模块使用了 Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set —— PyTorch implementation。该项目使用了基于深度学习的方法来实现从单张图像到完整人脸模型的无监督重建。

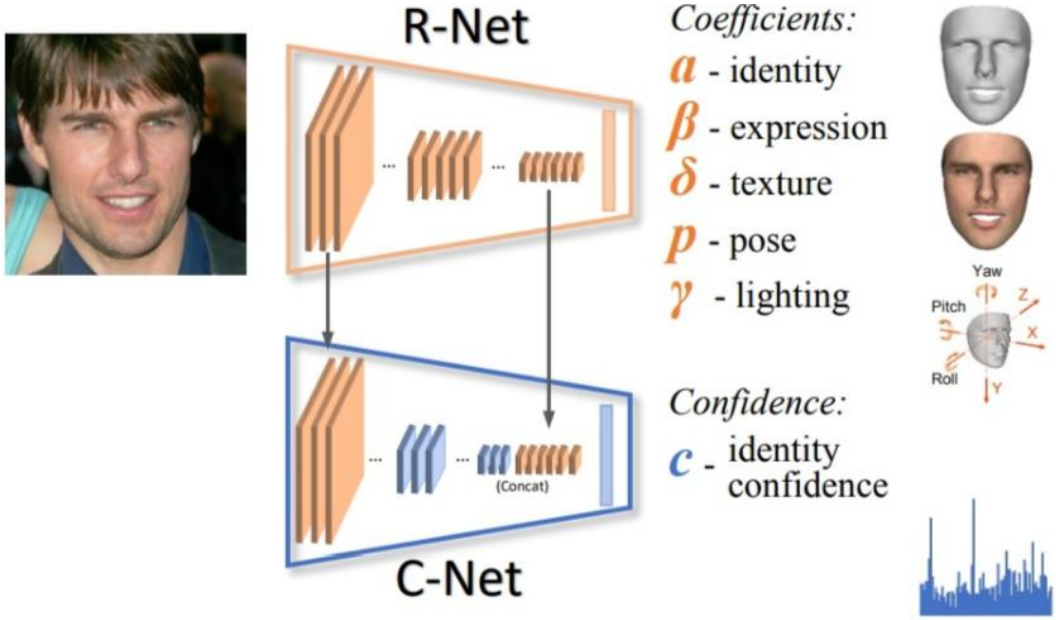

图 3-2 3D Face Reconstruction框架图

该模块的核心是基于单张图片自动推断 3D 人脸几何信息的弱监督学习方法，在该模块中，首先需要将一段视频转换为一系列帧，然后将每一帧中的人脸区域进行裁剪和对齐，并提取出人脸的关键点。接着，将对齐后的人脸图像输入到深度神经网络中，从而得到一组三维形态参数，用来重建整个人脸的模型。具

体来说，该神经网络由一个编码器和一个解码器组成，其中编码器将输入的人脸图像映射到一个低维空间中，解码器则将该低维表示映射回原始的人脸图像。在训练过程中，需要最小化重构误差，即预测图像与真实图像之间的差异。此外，为了使生成的人脸更加逼真，还需要使用对抗生成网络（GAN）进行优化，其中判别器被用来评估生成的人脸图像与真实图像的区别。其数学模型如下：

给定一张 2D 人脸图片I，本文需要推断其对应的 3D 人脸几何信息$F = \{f_1, f_2, \ldots, f_N\}$，其中$N$表示顶点数。本文使用一个三部分编解码器架构来学习图片到几何信息的映射：

(1)一个编码网络将图像I转换成一组特征h。

(2)一个解码网络将上述特征转换成 3D 点云结构F。

(3)使用点云F来重投影图片，并计算图像损失的重建损失函数。

采用两个损失函数来训练模型：图像重建损失和几何正则化损失,具体如下：

$$\mathrm{L_{img}} = ||\mathrm{I} - \mathrm{I_{proj}}(\mathrm{F}, \mathrm{K}, \mathrm{R}, \mathrm{T}, \mathrm{W}, \mathrm{H})||_1 \tag{3-2}$$

$$\mathrm{L_{geo}} = \sum_{\mathrm{i}=1}^{\mathrm{N}} (||\mathrm{f_i}' - \mathrm{f_i}||^2 + ||\mathrm{ni}' - \mathrm{ni}||^2) \tag{3-3}$$

采用其中I_{proj}表示将点云F 投影到图像平面的过程，K,R,T分别表示相机内参、旋转和平移，W,H分别表示图片宽度和高度，f_i和n_i分别是点云中第i个点的位置和法向量，f_i和$n_i{}'$分别是解码器生成的点云中第i个点的位置和法向量。整个损失函数为：

$$L = L_{img} + \lambda L_{geo} \tag{3-4}$$

3.5 语音驱动面部模块原理

语音驱动面部使用了 FACIAL: Synthesizing Dynamic Talking Face with Implicit Attribute Learning项目。该项目采用了条件生成对抗网络（Conditional GAN）和隐属性学习（Implicit Attribute Learning）技术，从语音信号和人脸模型中生成出动态人脸合成视频。

大多数现有生成方法只关注于人脸的显式属性生成，即通过输入语音，合成同步的唇部运动属性。这些方法合成的人脸结果要么不具有隐式属性（图1中a所示），要么复制原始视频的隐式属性（图1中b所示）。只有少部分工作探索过头部姿势与输入音频之间的相关性。FACIAL 框架使用对抗学习网络联合学习这一过程中的隐式和显式属性，提出以协作的方式嵌入所有属性，包括眨眼信息、

头部姿势、表情、个体身份信息、纹理和光照信息，以便可以在同一框架下对它们用于生成说话人脸的潜在交互进行建模。

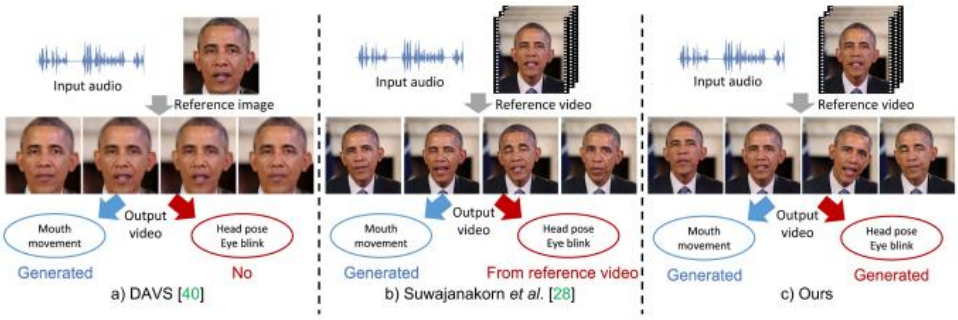

图 3-3 不同框架对特征的提取对比

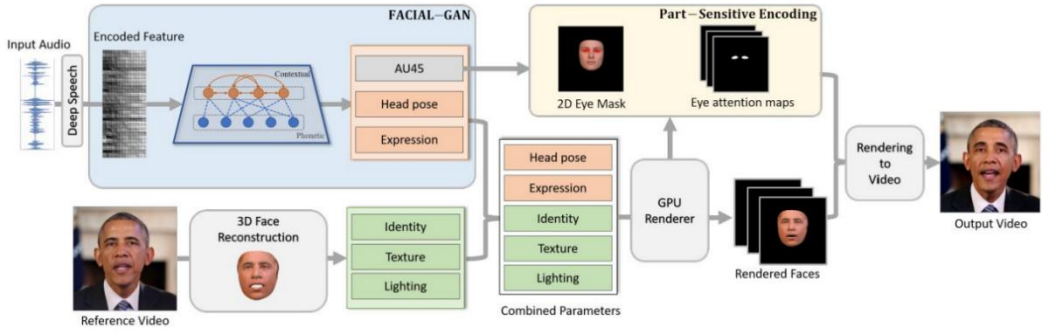

图 3-4 FACIAL学习框架图。给定输入音频，所提出的 FACIAL- GAN 旨在生成显式属性（表情）和隐式属性（眨眼、头部姿势），同时具有时间相关性和局部语音特征。参考视频执行面部重建操作，为渲染操作提供 3D 模型指导。此外，部分敏感编码将眨眼动作单元作为输入，并用作渲染面部的眼睛注意力图。这些指导被联合起来以提供给渲染到视频网络。

在该模块中，首先需要将第二个模块中生成的人脸模型和第一个模块中生成的语音信号输入到 FACIAL 项目中。模型会根据人脸模型和语音信号生成一系列动态人脸图像，并将这些图像组合成一个完整的视频。具体来说，这个框架中设计了一个特殊的 FACIAL-GAN网络来共同学习语音、上下文和个性化信息。这一网络将一系列连续帧作为分组输入并生成上下文隐空间向量，该向量与每个帧的语音信息一起由单独的基于帧的生成器进一步编码。因此， FACIAL-GAN 可以很好地捕获隐式属性（例如头部姿势等）、上下文和个性化信息。

FACIAL-GAN 还可以预测眨眼信息，这些信息被进一步嵌入到最终渲染模块的眼部相关的注意力图中，用于在输出视频合成逼真的眼部运动信息。实验结果和用户研究表明，该方法可以生成逼真的谈话人脸视频，该生成视频不仅具有同步的唇部运动，而且具有自然的头部运动和眨眼信息。并且其视频质量明显优于现有先进方法。FACIAL-GAN结构如图所示。其中G^{tem}表示时序相关性生成器，

将T帧的音频特征输入到生成器中，得到一系列整体的时序特征z；G^{loc}表示局部语音生成器，更加强调每一帧的特征，将该时刻前后各8帧的音频特征输入到生成器中，得到当前时刻的特征c。

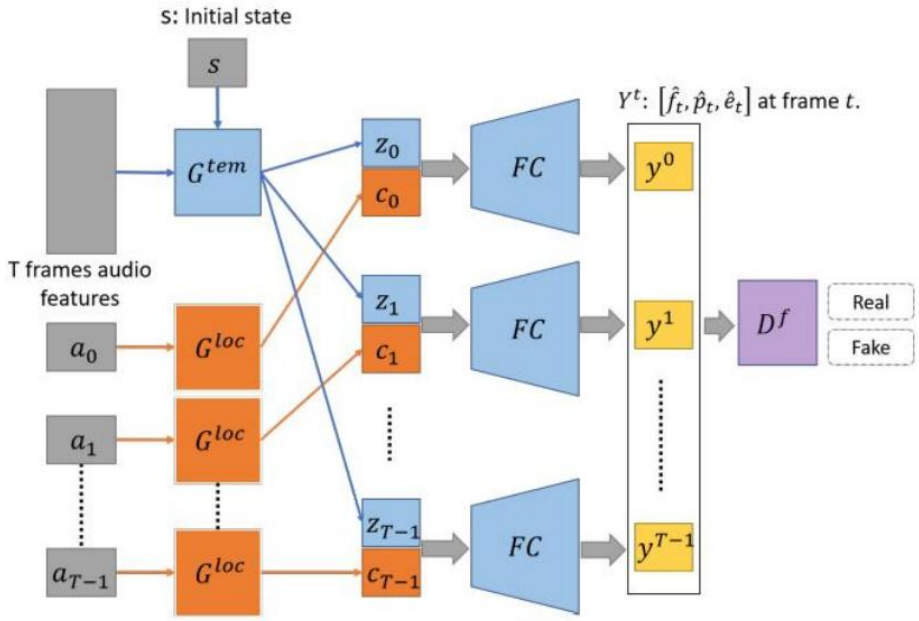

图 3-5 FACIAL-GAN结构图

3.6 唇形校准模块原理

唇形校准模块使用了Wav2Lip-GFPGAN 项目。该项目利用了 Wav2Lip 和 GFPGAN 两种神经网络模型来进行唇形校准，以使得由第三个模块生成的人脸视频更加逼真。

在该模块中，首先需要将第三个模块中生成的人脸视频和第一个模块中生成的语音信号输入到 Wav2Lip 模型中，从而生成出粗略的人脸嘴唇移动序列。

Wav2Lip有一个generator,这里作者借用了LipGAN的生成器结构，还有一个判断lip-sync的判别器，以及一个判断视频质量的判别器，结构如图所示。

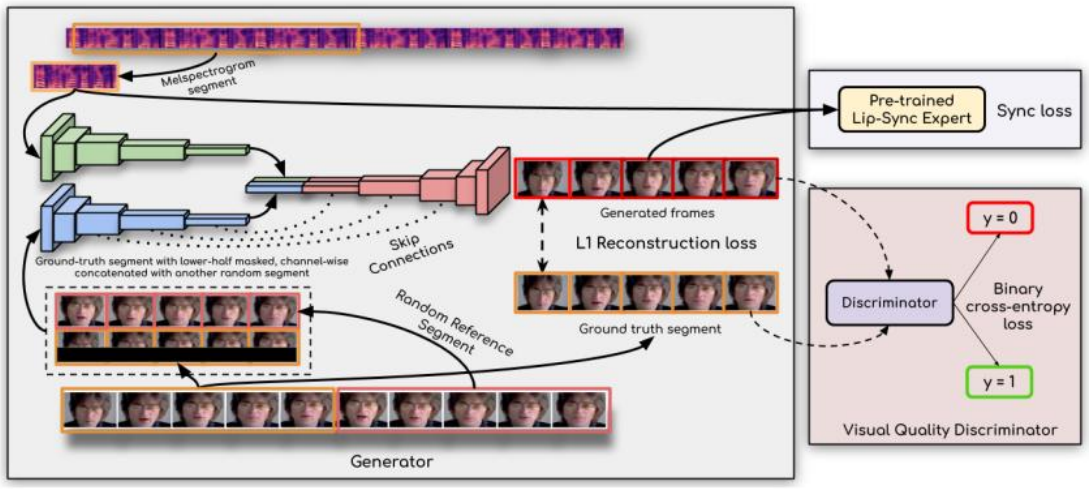

图 3-6　Wav2lip框架结构图

以前相关方面的方法都有一些缺点：

(1)只能针对训练时使用的speaker来合成视频，不能做到speaker-generic；

(2)只能对静态的图像来合成视频，无法做到输入视频来合成；

(3)训练需要一个特定的speaker的大量的数据，或者是需要非常干净的没有noisy的数据。

(4)能够合成的词汇量是有限的。

而Wav2Lip做到了训练完后对任意说话者、语言、视频都可以进行合成，而且合成的视频的嘴唇与音频是非常同步的。该论文分析以前的方法不能做到lip-sync的原因是：对不准确的嘴唇同步的惩罚太小了。

第一，用L1 reconstruction loss来判断嘴唇是否同步是很弱的，因为嘴唇只占了一张图片的很小的一部分，那loss更加注重的是图片的其他部分。（这个原因可以从一个事实来证实：就是在训练过程中，到训练中期才开始改变嘴唇的形状。还有上一个模块本人做的facial嘴唇形状并没有快速收敛，而其他面部特征反而更合理）

所以，为了解决这个问题，论文提出了一个预先训练好的expert Lip-sync discriminator。它不会在训练生成器的时候再去微调。这个判别器是模仿SyncNet的，SyncNet的输入是一段的连续的face frames（只有下半张脸）和speech segment，判别器判别这两个是in-sync还是out-of-sync。

对Syncnet的处理，论文作者对其进行了三个改进:

(1)原来网络使用灰度图，这里可以使用彩色图；

(2)通过残差模块网络变得更深了；

(3)将损失函数改为了余弦相似度：

$$P_{sunc} = \frac{\mathrm{v}\cdot\mathrm{s}}{\max(\|\mathrm{v}\|_2\cdot\|\mathrm{s}\|_2,\epsilon)} \tag{3-5}$$

生成器的结构是参考LipGAN的，由三个组成部分：identity encoder；speech encoder；face decoder。identity encoder是残差卷积层组成的，它的输入是：一段任意选取的reference frames和 pose prior（就是遮住下半张脸）在通道维度上concatenated的。它这种输入使得生成的下半张脸可以无缝粘贴回原视频中，而不再需要后处理。speech encoder是2维卷积层组成的，它是给 speech segment编码的。face decoder由卷积层和反卷积层组成，输入就是前面两个encoder的输出concatenated在一起的feature map。

为了提高视频同步的精度，现在采用口型同步鉴别器。但是，由于只有一个鉴别器，会导致变形区域出现模糊或伪影，从而降低视频质量。为此，我们与生成器共同训练视觉质量鉴别器以减轻这种损失。该鉴别器由多个卷积块组成，每个块包含一个卷积层和一个后续的ReLU激活层。这样，我们可以更好地处理变形区域，提高视频质量。鉴别器由一堆卷积块组成。每个块由一个卷积层和紧随其后的ReLU激活层组成。总损失函数为

$$L_{total} = (1 - s_w - s_g) \cdot L_{recon} + s_w \cdot E_{sync} + s_g \cdot L_{gen} \tag{3-6}$$

接着，将这个序列输入到 GFPGAN 模型中，进行一系列复杂的操作，从而得到更加逼真的嘴唇移动效果。GFPGAN 模型采用了分辨率逐步升高的。

3.7 本章小结

本章节介绍了本文的难点、解决方法、主要工作以及项目的模块划分，并对各模块使用的项目进行了说明，分别介绍了其原理以及算法。

第四章 数据处理

语音合成部分的VITS需要数据集的文字稿，这里本人使用了OpenAI的Whisper先把获得的语音转换为文本达到自动标注数据集文字部分的效果。只需要输入10到15分钟的语音就能训练出较好的效果。

图 4-1 Whisper处理结果

人脸建模模块的3D Face Reconstruction需要将视频通过ffmpeg提取单帧，再通过mtcnn提取图片中的面部特征。将单帧图片和面部特征文本作为输入训练3D Face Reconstruction。

```
{'left_eye': (235, 210), 'right_eye': (342, 216), 'nose': (287, 256), 'mouth_left': (239, 320), 'mouth_right': (327, 326)}
1

/root/FACIAL/video_preprocess/train1_image/000002.jpg
{'left_eye': (233, 210), 'right_eye': (343, 215), 'nose': (287, 257), 'mouth_left': (239, 321), 'mouth_right': (329, 327)}
2

/root/FACIAL/video_preprocess/train1_image/000003.jpg
{'left_eye': (233, 210), 'right_eye': (342, 216), 'nose': (286, 257), 'mouth_left': (239, 321), 'mouth_right': (328, 327)}
3

/root/FACIAL/video_preprocess/train1_image/000004.jpg
{'left_eye': (233, 211), 'right_eye': (342, 216), 'nose': (286, 257), 'mouth_left': (238, 321), 'mouth_right': (328, 327)}
4

/root/FACIAL/video_preprocess/train1_image/000005.jpg
{'left_eye': (232, 210), 'right_eye': (342, 215), 'nose': (285, 257), 'mouth_left': (238, 321), 'mouth_right': (327, 327)}
5

/root/FACIAL/video_preprocess/train1_image/000006.jpg
{'left_eye': (233, 209), 'right_eye': (341, 215), 'nose': (285, 258), 'mouth_left': (238, 321), 'mouth_right': (325, 327)}
6

/root/FACIAL/video_preprocess/train1_image/000007.jpg
{'left_eye': (232, 210), 'right_eye': (340, 216), 'nose': (284, 258), 'mouth_left': (238, 321), 'mouth_right': (324, 327)}
7

/root/FACIAL/video_preprocess/train1_image/000008.jpg
{'left_eye': (231, 211), 'right_eye': (338, 216), 'nose': (282, 257), 'mouth_left': (236, 321), 'mouth_right': (324, 326)}
8

/root/FACIAL/video_preprocess/train1_image/000009.jpg
{'left_eye': (231, 211), 'right_eye': (338, 217), 'nose': (283, 258), 'mouth_left': (236, 321), 'mouth_right': (324, 327)}
```

图 4-2 mtcnn处理结果

语音驱动面部模块的 FACIAL音频预处理。本文使用DeepSpeech来提取语音特征。DeepSpeech以每秒50帧(FPS)的速度输出字符的归一化对数概率，它形成每秒大小为$50 \times D$的数组。这里$D = 29$是每帧中语音特征的数量。本人使用线性插值法将输出重新采样为30FPS，以匹配数据集中的视频帧，每秒生成一个大小为$30 \times D$的数组。为了自动收集头部姿势并检测眼球运动，本人采用OpenFace来生成每个视频帧的面部参数。

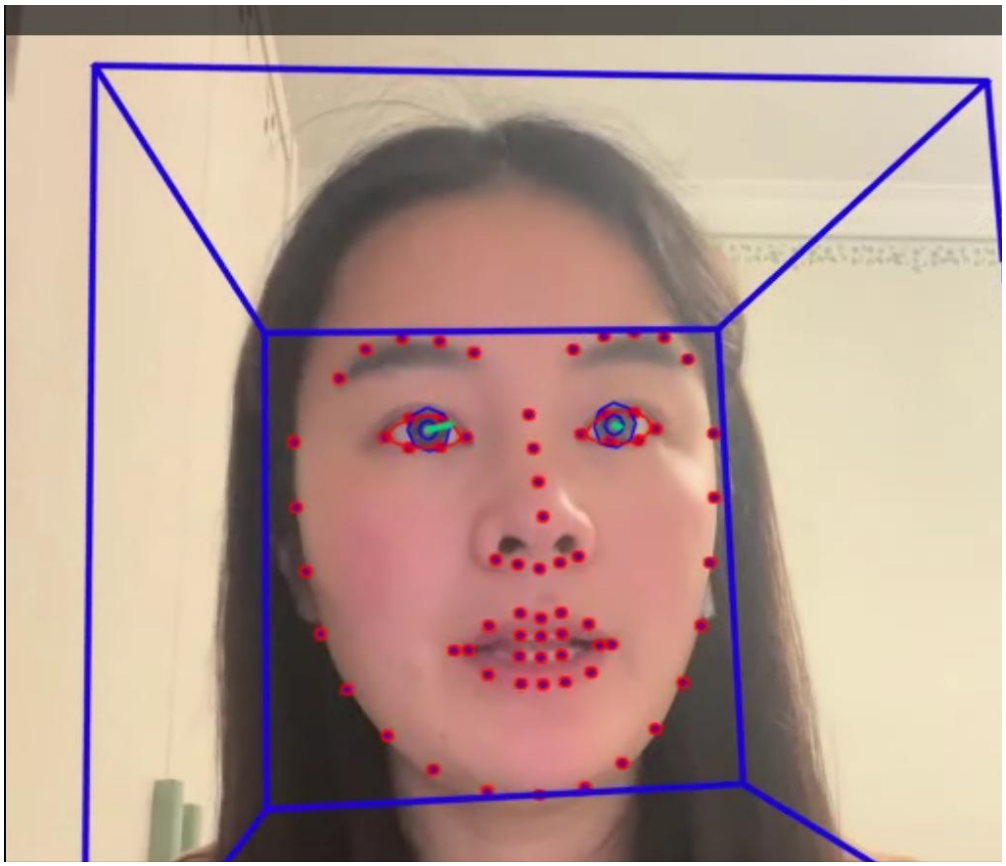

图 4-3　Openface处理结果输出的avi文件

唇形校准的输入来自于语音驱动面部模块的生成视频。生成出的avi视频为本项目最终输出。

第五章 实验

5.1 环境配置

5.1.1 系统环境说明

(1)系统平台：ubuntu18.04(Python 3.8)

(2)电脑配置：12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz 、NVIDIA GTX A5000 24G 独立显卡

5.1.2 系统搭建

VITS、deep3D face reconstruction、Wav2lip需要安装Cuda 11.4，PyTorch 1.11.0。FACAIL还需要安装TensorFlow 1.15.5。

5.2 仿真数字人实现

5.2.1 语音合成实现

在语音合成模块中，本文采用了条件变分自编码器和对抗生成网络（VITS）进行语音信号的生成。具体来说，本文将文本序列作为输入，使用条件变分自编码器生成对应的语音信号。

预训练数据集使用9700条女声标贝。在训练过程中，本文采用了Adam优化器，学习率为0.001，同时设置了对抗损失函数的权重为0.01以优化生成器的输出质量。此外，本文还使用了批量大小为32，迭代次数为1000的超参数设置。先使用15分钟左右语音在预训练模型的基础上进行训练，再将角色选为训练角色，语言选为简体中文,输入文本序列即可生成对应语音。

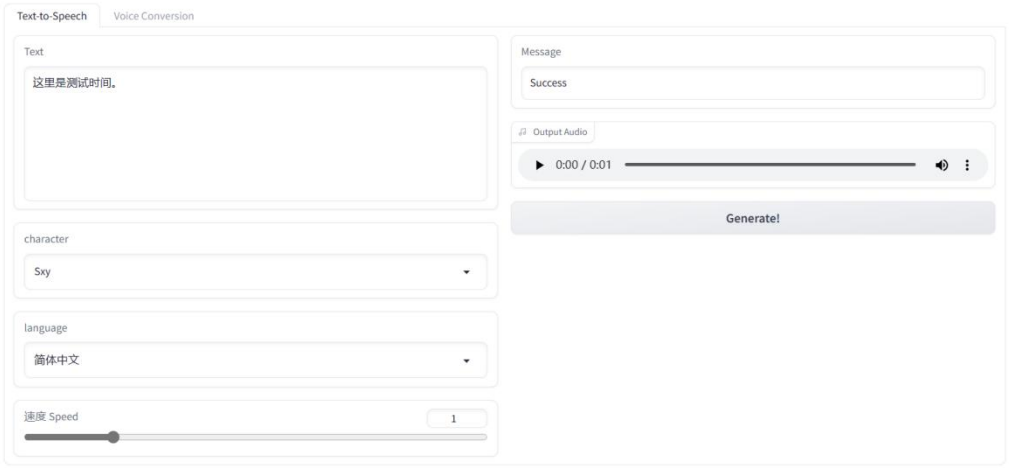

图 5-1 语音合成图像化界面

5.2.2 人脸重建实现

在人脸建模模块中，本文采用了基于深度学习的方法来实现从单张图像到完整人脸模型的无监督重建（deep3D face reconstruction）。具体来说，本文使用编码器-解码器结构的深度神经网络进行人脸图像的重建，其中编码器将输入的人脸图像映射到一个低维空间中，解码器则将该低维表示映射回原始的人脸图像。训练使用原项目使用的预训练模型。

本文将视频处理成单帧图片文件后，将其和mtcnn对每一帧的处理结果文本文件作为输入，用模型推导出每一帧的人脸模型。

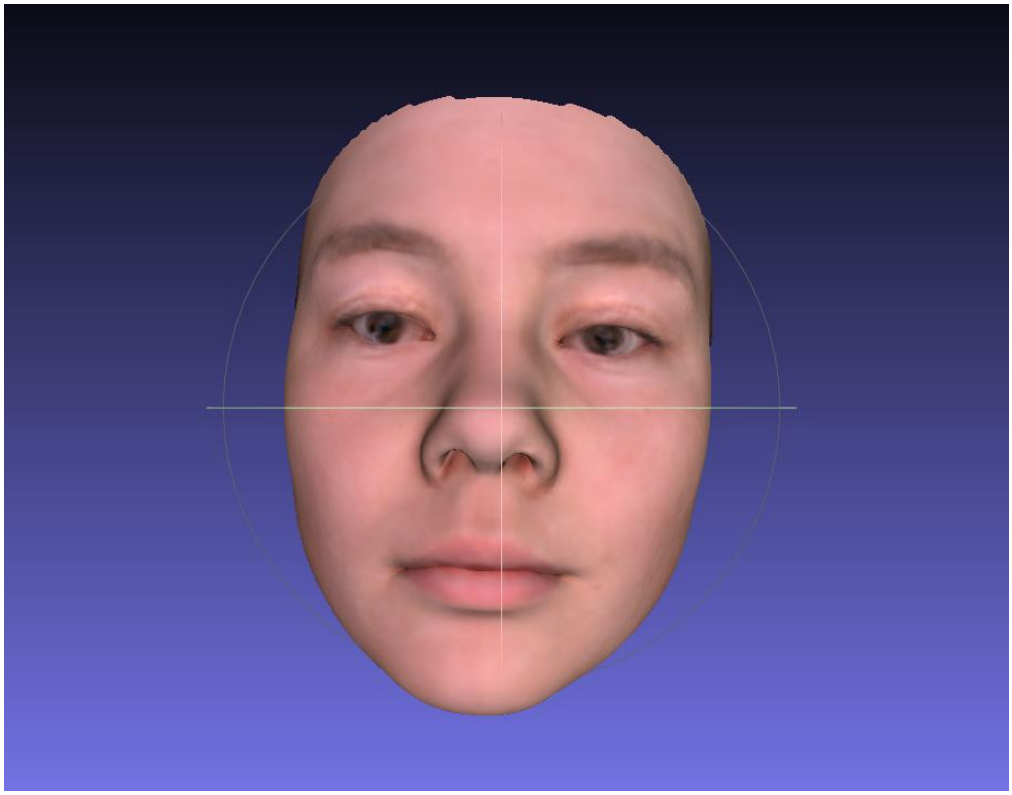

图 5-2 生成的人脸模型

5.2.3 语音驱动面部实现

在动态人脸合成模块中，本文采用了条件生成对抗网络和隐属性学习技术（FACAIL），从语音信号和人脸模型中生成出动态人脸合成视频。具体来说，本文使用条件生成对抗网络的生成器进行动态人脸图像的生成，并使用判别器评估生成的图像与真实图像之间的区别。在训练过程中，输入视频分别尝试了3-10分钟不同长度的视频，本文采用了Adam优化器，学习率为0.0002，同时设置了批量大小为16，迭代次数为20的超参数设置。训练使用原项目使用的预训练模型。模型到视频的训练使用everybody dance now，采用了Adam优化器，学习率为0.0005，同时设置了批量大小为16，迭代次数为50的超参数设置。在预训练模型基础上进行训练，使用语音合成模块生成的语音信号、人脸重建生成的人脸模型和训练视频作为输入，进行模型训练。再使用训练完成的模型推导出按输入语音信号的，具有面部表情和头部姿态的新视频。

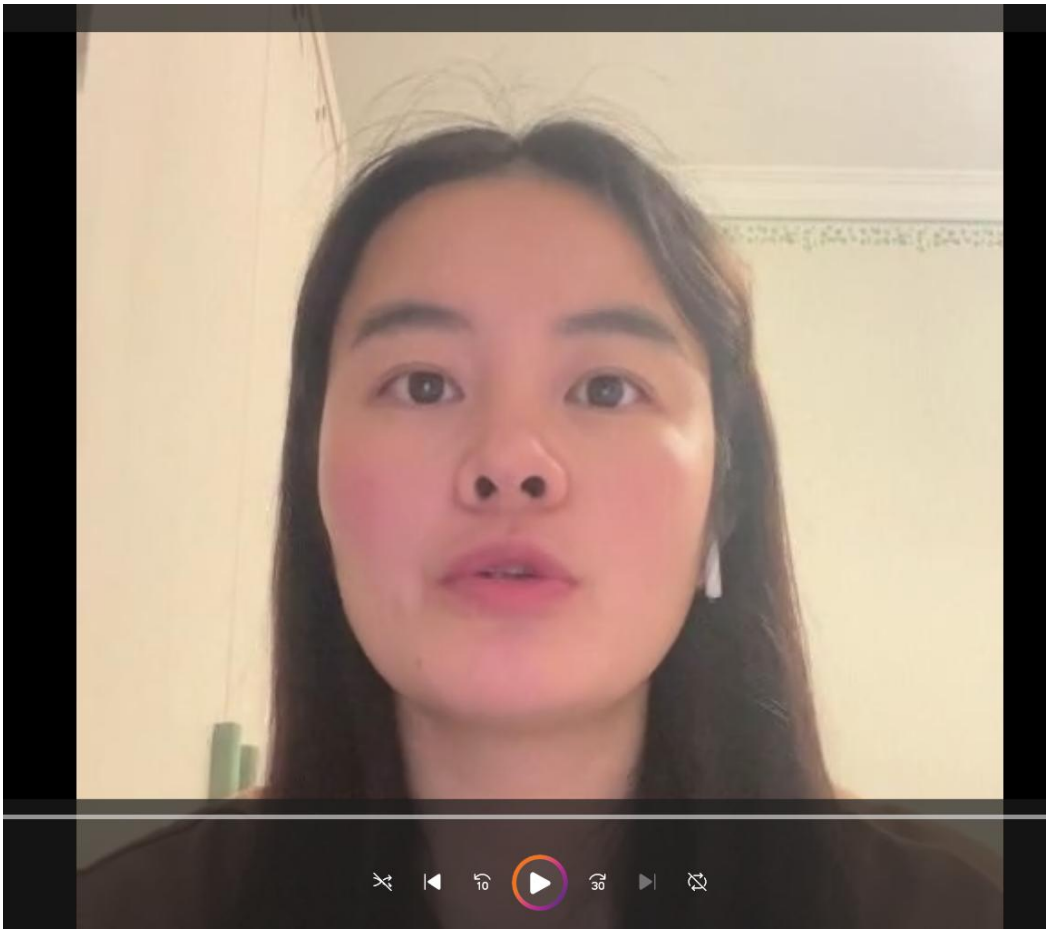

图 5-3 具有面部表情和头部姿态的新视频截图

5.2.4 唇形校准实现

在唇形校准模块中，本文采用了 Wav2Lip 和 GFPGAN 两种神经网络模型来进行唇形校准，以使得由第三个模块生成的人脸视频更加逼真。具体来说，在

Wav2Lip模型中，本文使用预训练的ResNet50作为特征提取器，并采用单向GRU进行序列建模。在GFPGAN模型中，本文采用缩放因子为4的分辨率逐步升高的方法进行学习。使用原项目的WAV2LIP-GAN预训练模型训练Wav2Lip模型。输入语音驱动面部的输出视频，生成出最终带精准口型的结果视频。

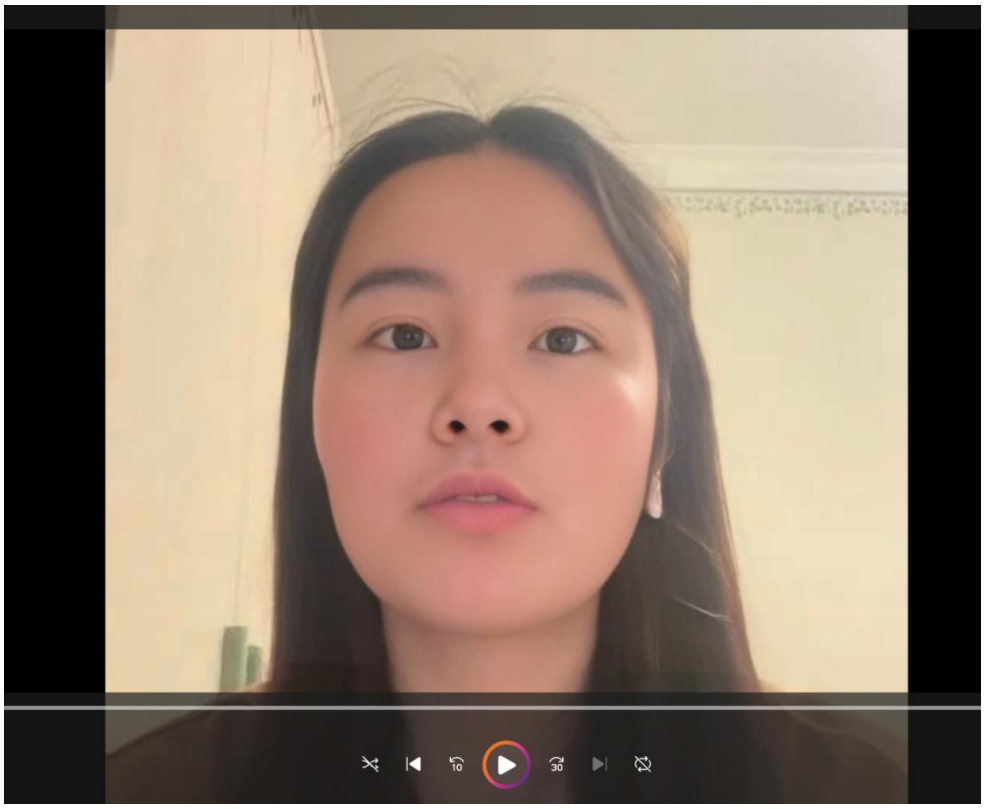

图 5-4 最终带精准口型的结果视频截图

5.3 本章小结

本章节介绍了本文的实验流程和环境配置，描述了该项目在实际操作中是如何实现的。

第六章 结论

通过本次研究，本文成功设计并实现了一个数字仿真人的项目，该项目能够通过一段语音和一段视频的训练，将输入文本输出相应的说话视频。在该项目中，本文把整个流程分为了四个模块，完成了多项技术组合的复杂任务。

在第一个模块中，本文使用了"Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech"项目，成功训练出了可以将输入文本转化为语音的模型。在第二个模块中，本文使用了"Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set —— PyTorch implementation"项目，将输入的视频帧转化为人脸模型。在第三个模块中，本文将前两个模块生成的人脸模型和语音交给了"FACIAL: Synthesizing Dynamic Talking Face with Implicit Attribute Learning"项目，实现了动态说话视频的生成。最后，在第四个模块中，本文使用了"Wav2Lip-GFPGAN"项目，对生成的说话视频进行了精准的唇形校准。

总体而言，本文的数字仿真人项目成功地完成了从文本到说话视频的转换，不仅实现了语音与图像之间的有效结合，而且具有广泛的应用前景。例如，该项目可以被用于网络课程、短视频制作和虚拟演讲等各种场合。此外，本文还探究了不同模块参数的影响，进一步提高了生成的说话视频的质量。

虽然本研究已经取得了令人满意的成果，但是还有一些问题需要进一步研究和优化。例如，在语音驱动面部模块中，本文并没有完全解决语音与图像的时序对齐问题，这也导致了生成的结果在某些情形下会有扭曲感。这个问题可以通过更精细的算法设计来解决。其次，Wav2lip造成的模糊会导致部分面部细节的丢失语音驱动面部模块中的部分成果，需要训练更好的预训练模型。此外，所有的训练数据均来自于特定的语音和视频流，如果要将该项目扩展到更广泛的应用领域，需要采集更多的测试数据以及进行更严谨的性能评估。

总之，我们相信，通过本文的努力和不断的实验和研究，数字仿真人项目将具备更大的商业价值和实际应用意义，为当今数字娱乐和虚拟现实产业做出贡献。

参考文献

[1] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions[C]//2018 IEEE international conference. Acoustics, speech and signal processing (ICASSP). Acoustics: IEEE, 2018: 4779-4783.

[2] Oord A, Dieleman S, Zen H, et al. Wavenet: A generative model for raw audio[J]. arXiv preprint arXiv:1609.03499, 2016.

[3] Kalchbrenner N, Elsen E, Simonyan K, et al. Efficient neural audio synthesis[C]//International Conference on Machine Learning. PMLR, 2018: 2410-2419.

[4] Li N, Liu S, Liu Y, et al. Neural speech synthesis with transformer network[C]//Proceedings of the AAAI conference. Artificial intelligence. AAAI, 2019, 33(01): 6706-6713.

[5] Ren Y, Ruan Y, Tan X, et al. Fastspeech: Fast, robust and controllable text to speech[J]. Advances in neural information processing systems, 2019, 32.

[6] Peng K, Ping W, Song Z, et al. Non-autoregressive neural text-to-speech[C]//International conference on machine learning. PMLR, 2020: 7586-7598.

[7] Zeng Z, Wang J, Cheng N, et al. Alignntts: Efficient feed-forward text-to-speech system without explicit alignment[C]//ICASSP 2020-2020 IEEE international conference. Acoustics, speech and signal processing (ICASSP). IEEE, 2020: 6714-6718.

[8] Miao C, Liang S, Chen M, et al. Flow-tts: A non-autoregressive network for text to speech based on flow[C]//ICASSP 2020-2020 IEEE International Conference. Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7209-7213.

[9] Kim J, Kim S, Kong J, et al. Glow-tts: A generative flow for text-to-speech via monotonic alignment search[J]. Advances in Neural Information Processing Systems, 2020, 33: 8067-8077.

[10] Creswell A, White T, Dumoulin V, et al. Generative adversarial networks: An overview[J]. IEEE signal processing magazine, 2018, 35(1): 53-65.

[11] Kumar K, Kumar R, De Boissiere T, et al. Melgan: Generative adversarial networks for conditional waveform synthesis[J]. Advances in neural information processing systems, 2019, 32.

[12] Bińkowski M, Donahue J, Dieleman S, et al. High fidelity speech synthesis with adversarial networks[J]. arXiv preprint arXiv:1909.11646, 2019.

[13] Kong J, Kim J, Bae J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis[J]. Advances in Neural Information Processing Systems, 2020, 33: 17022-17033.

[14] Ren Y, Hu C, Tan X, et al. Fastspeech 2: Fast and high-quality end-to-end text to speech[J]. arXiv preprint arXiv: 2006.04558, 2020.

[15] Donahue J, Dieleman S, Bińkowski M, et al. End-to-end adversarial text-to-speech[J]. arXiv preprint arXiv: 2006.03575, 2020.

[16] Kim J, Kong J, Son J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech[C]//International Conference on Machine Learning. PMLR, 2021: 5530-5540.

[17] Blei D M, Kucukelbir A, McAuliffe J D. Variational inference: A review for statisticians[J].

Journal of the American statistical Association, 2017, 112(518): 859-877.
[18] Kobyzev I, Prince S J D, Brubaker M A. Normalizing flows: An introduction and review of current methods[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 43(11): 3964-3979.
[19] Wang N, Zhang Y, Li Z, et al. Pixel2mesh: Generating 3d mesh models from single rgb images[C]//Proceedings of the European conference on computer vision (ECCV). ECCV, 2018: 52-67.
[20] Tian G, Yuan Y, Liu Y. Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks[C]//2019 IEEE international conference. Multimedia & Expo Workshops (ICMEW). IEEE, 2019: 366-371.
[21] Tian G, Yuan Y, Liu Y. Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks[C]//2019 IEEE international conference. Multimedia & Expo Workshops (ICMEW). IEEE, 2019: 366-371.

致　谢

在这篇论文完成之际，我要向所有为这个项目提供帮助和支持的人们表示真挚的感谢。

首先，我要感谢我的毕业设计指导老师张繁教授。他的专业知识、认真负责的态度以及耐心指导，让我受益匪浅。他的鼓励和启发让我不断前进，不断突破，最终完成了这个数字仿真人项目的设计与开发。

其次，我要感谢浙江大学计算机辅助设计实验室的全体成员，特别是实验室负责人王章野老师。他们为我提供了许多良好的实践环境和技术支持，使得我能够在实验室中开展自己的研究项目，并从中汲取到了宝贵的经验和智慧。

此外，我还要感谢开源社区中的开发者们，特别是以下几个项目的作者：

来自 https://github.com/jaywalnut310/vits 的交互条件变分自编码器与对抗学习技术，为我提供了强有力的语音转换功能;https://github.com/sicxu/Deep3DFaceRecon_pytorch 提供的精准3D人脸重建技术，为我提供了人脸模型生成的依据；https://github.com/zhangchenxu528/FACIAL 提供的动态人脸合成技术，为我提供了生成说话视频的可能性；https://github.com/ajay-sainy/Wav2Lip-GFPGAN 提供的唇形校准技术，为我提供了最后关键的步骤。

感谢各位开发者无私分享、宝贵的技术贡献，使得我可以站在巨人的肩膀上，快速地实现了这个数字仿真人项目。

最后，感谢所有关心和支持我的人们。是你们的鼓励和支持让我走到了今天，我会一直保持积极向上的状态，不断追求卓越，将所学应用于实践，并为更广泛的人群提供有用和实用的解决方案。