

融合音画同步的唇形合成研究

靳聪¹, 王洁¹, 郭子淳², 王晶³

(1. 中国传媒大学信息与通信工程学院, 北京 100024;

2. 北京化工大学, 北京 100089;

3. 北京理工大学信息与电子学院, 北京 100081)

摘要: 以视频为载体的信息传播方式兴盛发展, 视频的音画同步逐渐成为衡量视频质量的重要标准。深度合成技术在国际传播领域不断进入大众视野, 融合音画同步的唇形合成技术吸引了越来越多的关注。现有唇形合成模型主要是基于静态图像的唇形合成, 对于动态视频的合成效果不佳, 且大多采用英文数据集进行训练, 导致中文普通话合成效果较差。针对这些问题, 基于对 Wav2Lip 唇形合成模型的研究, 对其进行中文语境下的优化实验, 通过多组实验测试了不同路线的训练模型效果, 为后续 Wav2Lip 系列研究提供重要的参考价值。实现由语音驱动到文字驱动的唇形合成, 并对唇形合成在虚拟数字人等多领域的应用展开讨论, 为唇形合成技术更广阔的应用与发展奠定基础。

关键词: 唇形生成; 深度学习; 人工智能; 计算机视觉; 音画同步

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202335

Lipsynthesis incorporating audio-visual synchronisation

JIN Cong¹, WANG Jie¹, GUO Zichun², WANG Jing³

1. School of Information and Communication Engineering, Communication University of China, Beijing 100024, China

2. Beijing University of Chemical Technology, Beijing 100089, China

3. School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China

Abstract: With the flourishing development of video-based information dissemination, audio and video synchronization is gradually becoming an important standard for measuring video quality. Deep synthesis technology has been entering the public's view in the international communication field, and lip-sync technology integrating audio and video synchronization has attracted more and more attention. The existing lip-synthesis models are mainly based on lip-synthesis of static images, which are not effective for synthesis of dynamic videos, and most of them use English datasets for training which results in poor synthesis of Chinese Mandarin. To address these problems, this paper conducted optimization experiments on the Wav2Lip lip synthesis model in Chinese context based on its research foundation, and tested the effect of different routes of training models through multiple sets of experiments, which provided important reference values for the subsequent Wav2Lip series research. This study realized lip synthesis from speech-driven to text-driven, discussed the application of lip synthesis in multiple fields such as virtual digital human, and laid the foundation for the broader applica-

收稿日期: 2023-05-24; 修回日期: 2023-08-01

通信作者: 郭子淳, guozichun@buct.edu.cn

基金项目: 国家自然科学基金项目 (No.62207029, No.62271454); 北京市自然科学基金-小米联合基金项目 (No. L223033); 中央高校基础科研经费项目 (No.CUC230B018); 教育部人文社科青年基金 (No.22YJCZH041); 四川省社会科学重点研究基地四川动漫研究中心资助项目 (No.DM202213)

Foundation Items: The National Natural Science Foundation of China (No. 62207029, No. 62271454), Beijing Natural Science Foundation (No. L223033), The Fundamental Research Funds for the Central Universities (No. CUC230B018), The Science Foundation of Ministry of Education of China (No. 22YJCZH041), Fundamental Research Funds for Sichuan Comic and Animation Research Center, Key Research Institute of Social Sciences of Sichuan Province (No. DM202213)

tion and development of lip synthesis technology.

Key words: lip generation, deep learning, artificial intelligence, computer visualization, synchronization of audio and video

0 引言

随着数字化时代的到来,人们获取时事新闻、进行娱乐消遣的方式也在不断变化,逐步由文字、图像向视频过渡,视频成为人们获取信息的重要渠道。同时,人工智能生成内容(AIGC)成为热点话题,持续推动媒体深度融合、助力各领域数据化转型、加速产业智能化升级,深度伪造技术经由信息传播不断深入国际信息传播中。在这样的背景之下,作为计算机视觉领域重要分支的唇形合成技术逐步进入大众视野,获得了越来越多的关注。

视频后期制作往往有修改部分台词的需求,而召集现场人员进行重新拍摄会耗资过高,因此制作团队通常会采取仅更改后期配音的方法。这虽然实现了成本的有效控制,但牺牲了画面与音频的同步性,也就是我们常见的“口型和台词对不上”的现象。唇形与音频不匹配会极大地影响观众的观看体验,使其产生出戏之感,对于追求沉浸式代入体验的高要求剧组来说,更是一个头疼的问题。唇形合成技术的出现,为这一普遍性难题提供了解决方案。通过输入视频和音频,唇形合成模块便能根据音频更改原视频人物的唇形,最终输出画面唇形与音频高度匹配的新视频。

唇形合成技术亦可与当前较为热门的虚拟数字人技术相结合,进一步增强虚拟数字人的真实生动性。自2022年以来,虚拟数字人大量涌现,从中国文物交流中心的“文天天”、新华网的“筱竹”、浙江卫视的“谷小雨”,再到国家博物馆的“艾雯雯”、中华书局的“苏东坡数字人”、敦煌研究院的“伽瑶”,他们各有特色,在不同的领域上发光发热。唇形变化的自然流畅性对于虚拟数字人优良的展现效果是至关重要的,深入研究唇形合成技术并进一步优化模型性能,将正向作用于虚拟数字人领域。此外,通过修改音频即可做到唇形的修改,将大大减少虚拟数字主播后期团队的工作量。例如,当虚拟主播进行新闻播报工作时,如遇突发新闻状况,后期团队只需要利用唇形合成技术对视频进行口型的修改,便可以省去重新建模的复杂操作。

综上所述,融合音画同步的唇形生成是兼具研

究价值及现实意义的选题,本文聚焦基于深度学习的端到端唇形生成模型Wav2Lip^[1],运用现有的深度学习技术,搭建以生成器、唇形同步判别器和视觉质量判别器为核心部件的唇形生成模型,实现语音驱动的视频人物唇形匹配,从而达到辅助视频创作、促进虚拟数字人领域发展等目的。

本文的其余部分组织如下:第1节讨论相关工作,第2节描述Wav2Lip模型的网络架构,第3节介绍实验、结果并进行讨论,最后,对未来工作进行展望。

1 相关工作

2021年12月,苏轼念诗的视频在网络上爆火,视频中画像里的苏轼不再是静止的图像,而是在朗读一篇著名诗作,面部神态及唇形动作均非常自然。这便是深度学习技术与古人画像碰撞出的奇妙化学反应,也是唇形合成技术的一次特殊实践。随着虚拟数字人领域的蓬勃发展,唇形合成技术开始与语音驱动虚拟形象口型相结合,可以进一步提升虚拟数字人形象的真实性和真实性。此项技术作为深度伪造技术中的重要部分,被不断应用于复杂的国际信息传播中,只有深入了解其技术原理,才能找到更好的深度伪造鉴别方法,因而近年来受到了愈来愈多的关注。

唇形合成的想法由来已久,但直到近些年深度学习技术不断发展,图像生成和语音识别等领域取得了较大进展,输入音频实现视频或图像人物的自动唇形合成技术才成为可能。2017年,文献[2]提出了Speech2Vid模型,该模型基于卷积神经网络(convolutional neural network, CNN)^[3]协同嵌入人脸和音频,可实时推理未见过的音频,驱动未见过的脸,通过“音频+图片”的输入生成任意说话者的视频,这个方法是早期通过音频直接生成说话人视频的代表,它通过静态图片来提供说话人形象信息,虽然会有不自然、缺少牙齿信息等问题,但也在一定程度上实现了唇形匹配音频内容的效果,为后续工作提供了思路。文献[4]提出的模型仅利用音频即可合成奥巴马讲话的高质量视频,并且保持合成视频的精确唇形匹配,该模型提出了基于循

神经网络的唇形生成方法,对生成视频中牙齿、下巴等细节部分进行了较好的处理。

2018年,Chen等人^[5]首次实现端到端同时生成多帧关联视频,同时考虑时序连续一致性,从实验室到自然场景的量化指标达到与SOTA相当的水平。

2019年,Prajwal等人^[6]提出了一种新颖的模型LipGAN,支持任何语言的音频输入,不仅能从目标面部的静态图像中合成说话视频,也能对输入的人脸视频进行唇形的修改工作。

2020年,Prajwal等人提出Wav2Lip模型,在结合唇形同步相关领域的研究基础上,一定程度上克服了以往的唇形生成模型的缺点,训练完毕后对任意说话者、语言和视频都可以进行合成,且合成视频的唇形与音频是非常同步的,进一步提升了唇形合成的精度。

2022年,基于PaddlePaddle深度学习框架的PaddleGAN得到了广泛应用,它可以用于生成高质量的数字人脸、人体姿势和动作表现等,同时也可以用于视频合成和唇形同步。2023年,来自硅谷的技术项目Twinsync引发唇形合成新一代浪潮,该项目旨在通过使用神经网络和渲染技术,实现逼真的人物视频合成。Twinsync提出了一种基于视频的唇形同步算法,该算法利用了神经辐射场(neural radiance fields, NerF)和网格变形(deform)等技术,从单一图像中预测出演讲者的面部形状和纹理,并将其应用于源视频上,以实现唇形同步。但上述两项研究都更注重对视频总体清晰度、匹配和谐度的提升,而非唇形合成的准确度,存在一定的缺陷。

本文研究工作主要围绕Wav2Lip模型展开,通过对其原理、构造等的深入了解,结合当前领域先进的研究成果,对Wav2Lip模型的性能提升及应用创新等问题展开探讨。

2 基于深度学习的端到端唇形生成模型Wav2Lip

2020年,来自印度海德拉巴大学和英国巴斯大学的团队,在ACM MM2020发表了论文《A lip sync expert is all you need for speech to lip generation in the wild》。该论文提出了命名为Wav2Lip的端到端唇形生成模型,只需要一段人物视频和一段目标语音,就能让视频人物的唇形根据音频改变,

最终将音频和视频合而为一,输出人物嘴型和音频匹配度极高的视频。Wav2Lip模型一经提出便受到了诸多关注,为唇形合成领域提供了新的研究思路及技术拓展。

2.1 结构

Wav2Lip模型主要由预训练好的唇形同步判别器、生成器和视觉质量判别器3个核心部件组成,具体结构如图1所示。唇形同步判别器与SyncNet^[7]结构相似,输入为连续的面部帧(只有下半张脸)及对应的音频段,唇形同步判别器能判别出连续图像帧与音频是否同步,并输出结果。它包含由二维卷积层组成的面部编码器及音频编码器,使用最大边际损失训练模型,以最小化(或最大化)同步(或不同步)对之间的L2距离来实现更好的同步判别效果。生成器则参考了LipGAN模型,由身份编码器、音频编码器和面部解码器3个部分组成,身份编码器是一堆残余卷积层,对随机参考帧R进行编码,并与沿通道轴的姿势先验P(下半部分被遮住的目标脸)相联系,音频编码器由二维卷积层组成,对输入的语音段S进行编码,然后与人脸表示相连接,面部解码器由卷积层和反卷积层组成,输入为前两个编码器的输出所连接成的特征图,输出为唇形合成后的视频图像。因为生成的图像会存在瑕疵或模糊的情况,所以视觉质量判别器的存在是必要的,视觉质量判别器与生成器一同参与训练,从而提升生成视频的质量。

将视频和音频输入生成器,音频编码器进行音频特征的提取,身份编码器对视频进行处理生成人脸表示,音频特征与人脸表示在通道维度上进行连接后输入面部解码器,输出唇形合成后图像帧,连续的图像帧组成视频。视觉质量判别器对生成视频和原视频进行比较,并将结果反馈至生成器促使其不断提升图像质量,直到模型达到预期效果;唇形同步判别器则对生成视频和音频进行同步判断,并将结果反馈至生成器促使其不断提升唇形和音频的匹配度。在多轮的迭代中,生成器模型逐渐收敛,达到一种生成器、唇形同步判别器和视觉质量判别器之间的平衡,生成器也在多次的迭代中达到了效果的优化。

与以往一般唇形合成模型不同,Wav2Lip项目中的唇形同步判别器是预先训练好的,在后期生成器、视觉质量判别器训练过程中它的性质不会发生改变,这就保证了对合成视频口型同步“监管”的

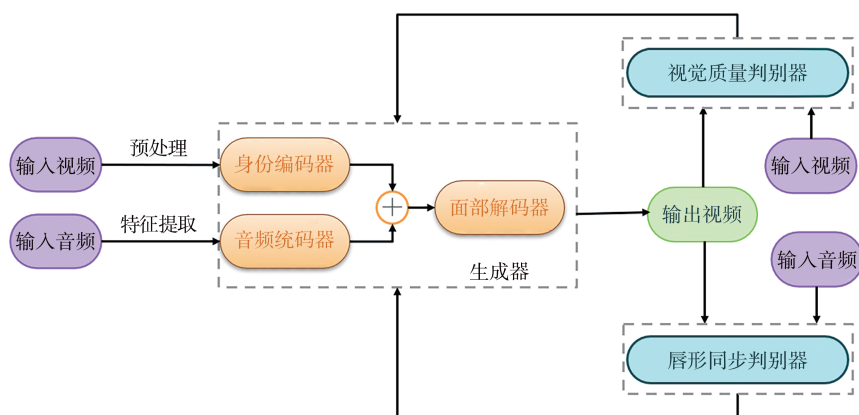


图1 Wav2Lip模型结构

可靠性。此外，唇形同步判别器采用残差连接，将前一层的输出直接与后一层的输入相加，构成了一种跨层连接的方式，使神经网络层次更深，使用具备二元交叉熵损失的余弦相似度来计算损失函数，使损失函数的计算更为准确。

而视觉质量判别器则与生成器一同参与训练，在训练迭代过程中提升合成图像的质量，最终实现图像质量与唇形同步效果权衡后的最优化呈现。

2.2 评估标准

为了更准确地评估唇形同步的效果，本文主要使用唇形同步误差距离（LSE-D）、唇形同步误差置信度（LSE-C）两种新的评估参数来进行效果评估。其中LSE-D计算了唇形表示与音频表示的平均距离，越低的LSE-D表示越高的视听匹配度，嘴唇和语音同步的效果越好；LSE-C计算了平均的唇形同步置信度得分，越高的LSE-C表示越高的音频-视频相关性。

2.3 基于Wav2Lip模型的应用创新探索

在传统唇形合成中，往往采用的是音频驱动，即通过输入音频来驱动唇形进行改变与合成，音频是唇形合成的主导源。然而，随着虚拟数字人技术的发展，文字驱动的唇形合成需求逐渐上涨。例如，在使用虚拟主播进行新闻播报时，如遇突发新闻状况，修改相关音频将会花费大量时间，若仅修改文字内容就能输出真实流畅的播报视频，可大幅提升工作效率。本文通过结合Wav2Lip模型与python中的edge-tts库，成功实现了文字驱动唇形合成。

edge-tts库依照微软edge浏览器使用的文字转语音引擎，包含了涉及54种不同语言的72种不同声音，能够轻松实现各类文字转语音的需求。在运行Wav2Lip模型的虚拟环境中安装edge-tts python

库，运行相关命令调用edge-tts功能，便能将输入文字或txt文件转换为相应音频文件，Wav2Lip模型再对生成音频与输入视频进行加载，最终输出与文字内容匹配良好的唇形合成视频。通过“--voice”参数修改生成语音的风格，通过“--rate”参数调整生成语音的语速，我们能够轻松得到理想的音频文件。Wav2Lip模型与edge-tts库的结合轻松实现了文字驱动下的唇形合成，只需要输入文字内容及视频，就能得到唇形合成完毕的视频，对于后续虚拟数字主播技术的发展、口播类短视频的制作均有参考意义。

Wav2Lip唇形合成技术同样可应用于当下热门的电脑三维动画（computer-generated imagery, CGI）领域。传统上，CGI领域生成唇形变化与音频匹配的三维人脸图像是使用面部捕捉等方法完成的，此类方法需要大量的人力与昂贵的设备，制作成本高昂。而借助Wav2Lip唇形合成技术能够对唇形进行快速修改，保证唇形变化与音频同步，大大降低了生产高质量三维动画所需成本。

3 实验分析

综合对Wav2Lip模型原理构造的分析，设计了系列实验，成功实现了对Wav2Lip模型的调试、性能测试及优化。为使Wav2Lip模型更好地适应中文语境下的唇形合成，设计了多条优化路线对Wav2Lip模型进行调整，并记录、测试了不同路线下优化模型的表现。

3.1 Wav2Lip模型效果测试

模型的效果测试主要基于谷歌云服务平台Google Colab实现，实验将python作为开发语言，搭载python 3.6虚拟环境，使用ffmpeg^[8]工具进行

测试过程中的视音频的相关剪辑、转换处理,安装librosa^[9]、numpy、torchvision、libsm6等多种必要的python包,使用的人脸检测模型为官方提供的预训练模型s3fd.pth,使用生成式对抗网络(generative adversarial networks, GAN)^[10]训练下的wav2lip_gan.pth唇形合成预训练模型完成唇形合成。

生成式对抗网络训练下的唇形合成模型虽然口型同步效果稍差,但获得了更好的视觉效果。从图2可以看出,经过唇形修改后,视频人物面部五官较为清晰,但下巴部位出现了不明遮罩,且输入中文语音时合成后的唇形动作变化频率过高,不符合中文语境下的自然唇形变化规律。



图2 Wav2Lip原模型合成效果示意

3.2 基于Wav2Lip模型的优化实验

Wav2Lip模型是一个两阶段模型:第一阶段,训练一个能够判断唇形与音频同步与否的唇形同步判别器;第二阶段,利用已训练好的唇形同步判别器进一步训练生成器和视觉质量判别器。最终使唇形同步判别器、生成器、视觉质量判别器三者达到一种博弈平衡,以输出唇形匹配且视觉质量良好的合成视频。

综合分析第3.1节的模型测试结果,尝试了3种方式以实现Wav2Lip模型在中文语境下的优化合成,最终得到3个性能各不相同的模型。

(1) Lipsync_restart+Wav2Lip_restart (LRWR)

使用中文唇读数据集CMLR从零开始训练唇形同步判别器,再利用训练好的唇形同步判别器从零开始训练生成器及视觉质量判别器。

(2) Lipsync_restart+Wav2Lip_finetune (LRWF)

使用中文唇读数据集CMLR从零开始训练唇形同步判别器,再利用训练好的唇形同步判别器对官方提供的生成器及视觉质量判别器模型进行微调。

(3) Lipsync_finetune+Wav2Lip_finetune (LFWF)

使用中文唇读数据集CMLR对官方提供的唇

形同步判别器模型进行微调,再使用微调后的唇形同步判别器对预训练好的生成器及视觉质量判别器进行微调。本实验依托远程服务器思腾云平台及jupyter编辑器、Xshell软件、Xftp软件完成,GPU显卡为Tesla V100-SXM2-32GB * 2, batch size设置为64,初始学习率设置为 10^{-4} 。

本实验所采用的公开数据集为CMLR数据集,由浙江大学视觉智能和模式分析(VIPA)小组收集,目前在官方网站上可直接下载使用。它旨在促进视觉语音识别的研究,即自动唇读。该数据集由11位发言人的102 072个口语句子组成,记录了2009年6月至2018年6月的国家新闻节目《新闻联播》。每个句子的长度不超过29个汉字,不包含英文字母、阿拉伯数字和罕见的标点符号。由《新闻联播》播报内容组成,保证了CMLR数据集的语音质量及唇形的准确性,同时该数据集采用了视频文件与音频文件分离的格式,避免了文件传播过程中因不可抗力因素造成的视音频不同步问题。本课题优化实验主要使用CMLR数据集中S2部分主持人康辉的播报视频进行模型的优化,生成任务主要用于新闻生产、科研实验,不用于商业和其他内容篡改。

最终对各模型进行效果测试时所使用的音频利用文字转音频技术生成,在Wav2Lip虚拟环境中调用edge-tts库,将生成的音频文件放置于项目根目录下,而后运行测试文件即可直接调用音频文件,生成与音频对应的唇形合成视频,实现了基于文字的唇形合成。

3.2.1 LRWR

首先基于CMLR数据集S2部分进行了唇形同步判别器的训练,图3为唇形同步判别器在训练集上实时计算得出的损失值变化曲线,呈波动下降趋势。经过36 000步的训练后,唇形同步判别器在验证集上的平均损失值为0.300;经过51 000步的训练后,唇形同步判别器在验证集上的平均损失值下降至0.278;此后,唇形同步判别器在验证集上的损失值保持平衡,不再下降。

之后利用训练所得的唇形同步判别器开始训练生成器及视觉质量判别器,在Xshell软件上连接远程服务器后运行hq_wav2lip_train.py文件,图4为训练过程中绝对损失在训练集上的实时变化曲线。

训练过程中生成器在验证集上计算得出的平均损失值如表1所示,感知损失一直维持稳定,平均

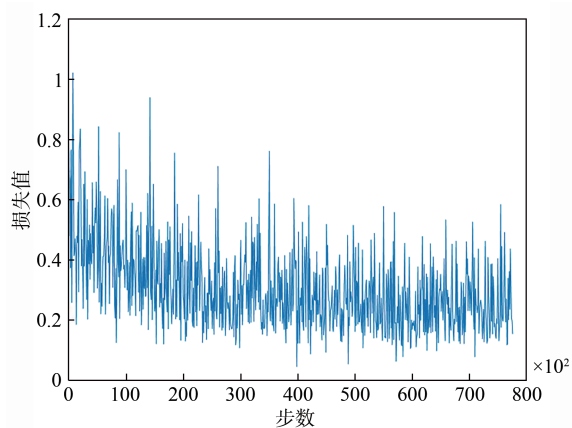


图3 LRWR唇形同步损失值变化曲线

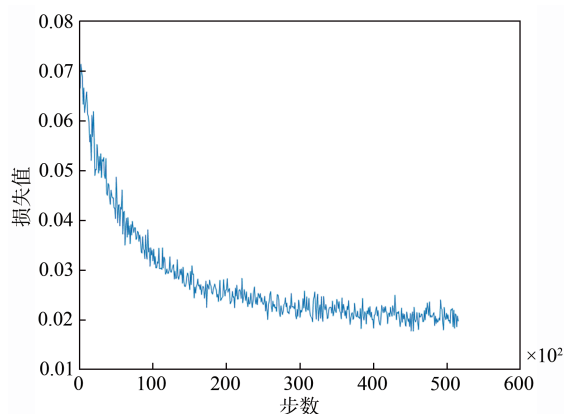


图4 LRWR生成器绝对损失值变化曲线

绝对损失逐渐下降，同步损失处于波动下降，在33 000步后趋于平稳，因此后期效果评估选择使用步数为33 000的版本生成器模型。

表1 LRWR各组数据对比

步数	绝对损失	同步损失	感知损失
9 000	0.021	0.096	0.693
21 000	0.016	0.074	0.708
33 000	0.015	0.073	0.700

3.2.2 LRWF

LRWF模型的训练采用“从头训练唇形同步判别器和微调生成器及视觉质量判别器”的路线，唇形同步判别器使用LRWR训练方式中验证集上平均损失值为0.278的checkpoint_step000051000.pth，在官方提供的预训练模型wav2lip_gan.pth及visual_quality_disc.pth上进行了微调训练，使用的数据集为CMLR唇读类数据集。

在生成器及视觉质量判别器微调过程中，损失

值的变化出现了异常现象，视觉质量判别器的同步损失始终保持为0.0，而生成器的同步损失数值一直较高。在经过较长时间的训练观察后，生成器各损失函数都无明显下降趋势，甚至偶有上升，初步推测LRWF的优化效果不佳，表2列出了生成器微调过程中的部分损失值。

表2 LRWF各组数据对比

步数	绝对损失	同步损失	感知损失
297 000	0.047	3.359	1.329
300 000	0.048	4.354	1.656

3.2.3 LFWF

LFWF模型的训练采用“微调唇形同步判别器和微调生成器及视觉质量判别器”的路线，第一阶段使用中文普通话唇读数据集CMLR对官方提供的预训练模型lipsync_expert.pth进行微调，第二阶段利用已微调完毕的唇形同步判别器与CMLR数据集对预训练wav2lip_gan.pth、visual_quality_disc.pth两个模型进行微调。

在唇形同步判别器的微调过程中，每个训练epoch平均耗时27 min，相比LRWR、LRWF训练过程中平均每epoch耗时7 min，LFWF在训练效率上远远低于LRWR与LRWF，导致该路线训练花费大量时间。这是由于预训练唇形同步判别模型lipsync_expert.pth在原始训练过程中进行了长时间训练，网络层次、结构较为复杂，且原始训练数据集LRS2与微调数据集CMLR存在较大的差异，共同导致了训练过慢的问题。

图5为唇形同步判别器训练阶段使用可视化工具tensorboard生成的训练集损失值变化曲线，由图5可知该模型损失值出现了较为剧烈的波动。

表3记录了唇形同步判别器模型微调阶段在验证集上测试得出的部分具体损失值，唇形同步判别器经过约120 h的训练，最终验证集损失值保持在0.4左右，此后使用该唇形同步判别器进行生成器及视觉质量判别器的训练，训练步骤与上文内容相似，此处不再赘叙。

3.3 实验结果

图6为使用LRWR进行唇形合成的效果示意，从左至右依次为4幅画面的拼接，第1幅为输入的当前帧画面，第2幅为下一帧的姿势先验（即下半部分被遮住的目标脸），第3幅为模型LRWR根据

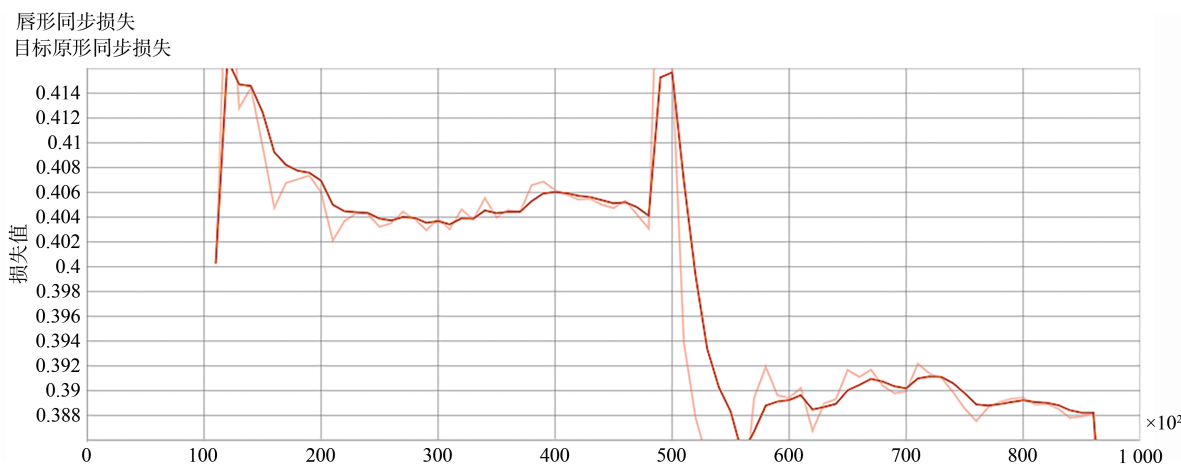


图5 LFWF唇形同步损失值变化曲线

表3 LFWF部分唇形同步损失

步数	同步损失
1 755 000	0.580
1 758 000	0.508
1 767 000	0.407

音频特征合成的预测画面，第4幅为实际目标脸（即真实的下一帧人脸画面）。易看出，模型LRWR进行唇形合成后的输出人脸图像与目标人脸图像相似度极高，且过渡自然。



图6 LRWR合成效果示意

图7为使用LRWF进行唇形合成的效果示意，从左至右依次为4幅画面的拼接，画面含义与上文一致。易看出，模型LRWF进行唇形合成后的输出人脸图像在唇部位置产生了不明遮挡，且人脸上下部分肤色相差较大，同时预测嘴型与实际目标嘴型不相符。

图8为使用LFWF进行唇形合成的效果示意，从左至右依次为4幅画面的拼接，画面含义与上文一致。易看出，模型LFWF进行唇形合成后的输出



图7 LRWF合成效果示意

下半部分人脸与实际上半部分人脸贴合度较差，五官呈现不自然状态，且图像质量较低。



图8 LFWF合成效果示意

选取同样的音频及原视频，分别使用上述3个模型及原Wav2Lip模型进行唇形合成，采用LSE-D、LSE-C两种新的评估参数对合成视频进行评估，模型具体得分值如表4所示。

表4 各模型评估指标得分

模型	LSE-D	LSE-C
LRWR	10.595	1.996
LRWF	13.710	0.532
LFWF	15.098	1.129
Wav2Lip	10.673	1.791

根据数据分析易知，模型LRWR的LSE-D得分最低，LSE-C得分最高，且每次训练耗时7 min以内，在合成质量及训练耗时上均优于模型LRWF及模型LFWF，对于中文音频的合成也优于用英文数据集训练而成的原Wav2Lip模型，采用“从头训练唇形同步判别器及生成器、视觉质量判别器”路线最具实际应用意义，但在具体应用实践上需格外注意选用数据集，通过增加数据集涉及的人物数量、增强所用数据集中视频的清晰度可进一步提升合成效果。LRWF在LSE-D得分上优于LFWF，在LSE-

C得分上次于LFWF,但两个模型在合成效果上均有欠缺,且LFWF所需训练时间过长,在此后Wav2Lip模型的调整优化过程中不建议采用模型LRWF及模型LFWF的训练路线。

4 结论

在计算机视觉技术飞速发展的视频时代,人们对视频的质量有了更高的追求,视频人物唇形不同步的问题会极大影响观众的观看体验,由此,结合深度学习领域及计算机视觉领域研究的唇形合成技术获得了越来越多的关注。本文基于优秀唇形合成模型Wav2Lip,从模型结构、模型原理、实验设计及效果评估等方面对唇形合成的研究展开了介绍,并探讨了中文语境下唇形同步问题的解决及唇形合成技术未来在各领域的创新性应用。

首先,针对Wav2Lip模型对中文语音的唇形合成进行优化研究。经过多次有效测试,发现原模型在进行中文语境下的唇形合成时嘴部变化频率过高,且下巴处出现明显遮罩,不符合说中文时的自然唇部运动规律。根据该实际测试情况设计了3条实验路线,以进一步改良Wav2Lip模型的中文合成效果,选用大型中文普通话唇读数据集CMLR中S2部分作为数据支撑,其中包括了2009年6月至2018年6月《新闻联播》节目主持人康辉说话片段,总数据量达14.9 GB。

而后,对使用不同路线进行调整、优化的各模型进行效果评估及测试,综合分析其不同合成效果的影响因素。在进行了大量的训练之后,选定出效果最佳的调整路线,为之后基于Wav2Lip的相关研究提供参考,并将该路线下的生成模型作为最终的应用模型。

此外,对唇形合成模型的应用领域进行了创新探索,将唇形合成模型与微软edge浏览器文字转语音引擎的edge-tts库进行了联结,实现由语音驱动下的唇形合成到文字驱动下的唇形合成的转变,输入文字及视频,即可得到预期唇形合成视频,为唇形合成技术在虚拟数字人等领域的应用发展奠定基础。

本文虽完成了一定程度上中文唇形合成的优化实验,但同时还存在着许多不足:受服务器算力及时间因素等影响,目前模型的泛化能力有待提升,后续可通过进一步扩展支撑数据的多样性提升模型泛化能力;合成视频的清晰度有待提高,可与相关

超分模型进行联动,以输出质量更高的合成视频;目前模型的训练迭代速度较慢,训练需要花费大量时间,可通过调整网络结构进一步加速训练,以实现效率的大幅提升。

本文深度合成的内容符合社会主义核心价值观,合成的内容是真实的新闻播报内容而不是虚假信息;符合国家网信办出台的《生成式人工智能服务管理办法》,承诺仅用于科研,不提供其他服务。

参考文献:

- [1] PRAJWAL K R, MUKHOPADHYAY R, NAMBOODIRI V P, et al. A lip sync expert is all you need for speech to lip generation in the wild [C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 484-492.
- [2] CHUNG J S, JAMALUDIN A, ZISSERMAN A. You said that? [EB]. arXiv preprint, 2017, arXiv: 1705.02966.
- [3] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 1725-1732.
- [4] SUWAJANAKORN S, SEITZ S M, KEMELMACHER-SHLIZERMAN I. Synthesizing Obama: learning lip sync from audio[J]. ACM Transactions on Graphics, 2017, 36(4): 1-13.
- [5] CHEN L L, LI Z H, MADDOX R K, et al. Lip movements generation at a glance[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2018: 538-553.
- [6] PRAJWAL K R, MUKHOPADHYAY R, PHILIP J, et al. Towards automatic face-to-face translation[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 1428-1436.
- [7] CHUNG J S, ZISSERMAN A. Out of time: automated lip sync in the wild[C]//Proceedings of Computer Vision-ACCV 2016 Workshops. Cham: Springer, 2017: 251-263.
- [8] TOMAR S. Converting video formats with FFmpeg[J]. Linux Journal, 2006(146): 10.
- [9] MCFEE B, RAFFEL C, LIANG D, et al. Librosa: audio and music signal analysis in python[C]//Proceedings of the 14th Python in Science Conference. [S.l.s.n.], 2015: 18-25.
- [10] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: an overview[J]. IEEE Signal Processing Magazine, 2018, 35(1): 53-65.

[作者简介]



靳聪 (1986-), 女, 博士, 中国传媒大学信息与通信工程学院副教授, 主要研究方向为强化学习、音乐人工智能和音频信号处理。



王洁（2001- ），女，中国传媒大学信息与通信工程学院硕士生，主要研究方向为深度合成和智能音视频处理。



王晶（1980- ），女，博士，北京理工大学信息与电子学院副教授，主要研究方向为语音与音频信号处理、多媒体通信和虚拟现实。



郭子淳（1987- ），男，博士，北京化工大学副教授，主要研究方向为未来影像与具身交互叙事。