计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(2):276−283 [doi: 10.15888/j.cnki.csa.009405]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

改进 Wav2Lip 的文本音频驱动人脸动画生成①

孙　瑜, 朱欣娟

(西安工程大学 计算机科学学院, 西安 710600)
通信作者: 朱欣娟, E-mail: zhuxinjuan@xpu.edu.cn

摘　要: 为了提高中文唇音同步人脸动画视频的真实性, 本文提出一种基于改进 Wav2Lip 模型的文本音频驱动人脸动画生成技术. 首先, 构建了一个中文唇音同步数据集, 使用该数据集来预训练唇部判别器, 使其判别中文唇音同步人脸动画更加准确. 然后, 在 Wav2Lip 模型中, 引入文本特征, 提升唇音时间同步性从而提高人脸动画视频的真实性. 本文模型综合提取到的文本信息、音频信息和说话人面部信息, 在预训练的唇部判别器和视频质量判别器的监督下, 生成高真实感的唇音同步人脸动画视频. 与 ATVGnet 模型和 Wav2Lip 模型的对比实验表明, 本文模型生成的唇音同步人脸动画视频提升了唇形和音频之间的同步性, 提高了人脸动画视频整体的真实感. 本文成果为当前人脸动画生成需求提供一种解决方案.

关键词: 文本音频驱动; 人脸动画; Wav2Lip 模型; 动画生成

引用格式: 孙瑜,朱欣娟.改进 Wav2Lip 的文本音频驱动人脸动画生成.计算机系统应用,2024,33(2):276–283. http://www.c-s-a.org.cn/1003-3254/9405.html

Text Audio Driven Facial Animation Generation Based on Improved Wav2Lip

SUN Yu, ZHU Xin-Juan

(School of Computer Science, Xi'an Polytechnic University, Xi'an 710600, China)

Abstract: In order to improve the authenticity of Chinese lip synchronized facial animation videos, this study proposes a text audio-driven facial animation generation technology based on the improved Wav2Lip model. Firstly, a Chinese lip synchronized dataset is constructed, which is used to pre-train the lip discriminator to make it more accurate in discriminating Chinese lip synchronized facial animations. Then, in the Wav2Lip model, text features are introduced to improve lip time synchronization and thus improve the authenticity of facial animation videos. The model in this study synthesizes the extracted text information, audio information, and speaker facial information and generates a highly realistic lip synchronized facial animation video under the supervision of a pre-trained lip discriminator and video quality discriminator. The comparative experiments with the ATVGnet model and Wav2Lip model show that the lip synchronized facial animation video generated by the proposed model improves the synchronization between lip shape and audio and enhances the overall realism of the facial animation video. The paper provides a solution for the current facial animation generation.

Key words: text audio drive; facial animation; Wav2Lip model; animation generation

人脸动画技术旨在给定语音或文本生成一系列高自然的人脸序列, 所生成的视频不仅需要保证视频帧纹理的真实感, 又要保证视频帧之间的时间连续性[1]. 其实现方法根据合成表情动画驱动源的不同分为 3 类:

① 基金项目: 国家重点研发计划 (2019YFC1521400)
收稿时间: 2023-08-17; 修改时间: 2023-09-26; 采用时间: 2023-10-09; csa 在线出版时间: 2023-12-18
CNKI 网络首发时间: 2023-12-19

第 1 类是以表演为驱动源的方法, 依靠面部捕捉技术, 对人脸的特征点定位, 实时采集位移、姿势运动等参数, 分析参数并从中提取有效的信息对二维或者三维虚拟人脸模型进行驱动, 从而生成人脸动画[2−4], 这类方法依赖昂贵的动作捕获设备实现, 有一定局限性[5]. 第 2 类以文本为驱动源的方法, 将文本信息直接映射到虚拟角色的脸部, 或者依赖文本转语音技术达到文本到语音的合成[6,7]. Kumar 等人在完成人脸动画生成任务中添加一个文本到音频的合成器, 将文本作为输入, 根据输入的文本来合成音频从而生成唇音同步的语音动画视频[8]. 虽然他们的目标是建立一个文本到视频生成人脸动画的技术, 但本质上, 仍旧是音频驱动人脸动画生成的过程[9]. 基于文本驱动的方法需要考虑注入声调、韵律等其他信息来增加合成效果的自然度, 还需要去规定每个音素的持续时间, 效果不一定很理想[10,11]. 第 3 类以音频为驱动源的方法, 通过大量的样本训练建立音频信号与口型运动的映射模型, 模型可以通过给定的音频, 合成相应的唇部运动, 实现语音信号的可视化动画[12,13].

目前, 音频驱动的人脸动画技术是大多数研究人员的焦点. 人对面部的细微变化敏感, 面部运动与语音不一致, 会使用户产生违和感[14], 如何提高人脸动画技术的真实性仍然是计算机动画中亟待解决的难题[15]. Chung 等人使用深度学习的方法完成音频到整张人脸的音频动画生成任务, 给出了目标对象的静态图像和音频片段, 并使用编解码卷积神经网络结构来实现面部动画合成中的唇部同步[16]. 这种方法通过使用音频信息来促进唇音同步唇部运动的生成, 这种做法具有里程碑式意义. 但是, 卷积神经网络在合成图像任务的表现上具有局限性, 通常需要另外的去模糊模块来提高视频帧的分辨率[17].

与此同时, 学习视频帧之间的时间相关性至关重要. 一些研究人员使用时间生成对抗网络学习视频帧之间时间相关性的同时, 实现动画生成[18,19]. 通过使用时间生成对抗网络, 经过训练的模型可以生成一系列与时间相关的面部图像. 然而, 僵硬的嘴唇动作和不变的人脸表情和姿势, 让这些图像看起来不真实. 这是由于模型的输入是一段语音信号和一张人脸图像, 基于循环神经网络的生成器没有办法根据这些生成符合人脸图像的不同脸部表情, 眼部动作和头部运动姿势的人脸动画. 为了解决此问题, Vougioukas 等人在完成生成人脸动画视频任务中添加了角色的眨眼动作[20], 但仅依靠眨眼作为面部动作仍然不足以提高生成的人脸动画的真实度[21]. Chung 等人提出了唇音同步判别网络 Syncnet, 通过判断语音和人脸图像在某个共同参数空间下的相似性, 计算音频特征与人脸图像特征的交叉熵损失以反映唇音同步效果[21].

为了解决模型合成人脸动画中人物在说话过程中面部运动不自然的问题[22], Prajwal 等人在借鉴 Chung 等人研究工作的基础上, 进一步提出了基于生成对抗网络的 LipGan[23]及其改进模型 Wav2Lip[24]. Wav2Lip 接收一系列图像帧序列作为输入, 在保留原始帧序列人物面部运动信息的同时, 通过音频特征引导唇部运动的变化来生成人脸动画. 相比之前的模型, Wav2Lip 模型在产生自然的唇部运动方面取得了重大进展[25], 该模型能够生成具有自然头部运动、唇音同步效果良好的真实人脸视频[26].

Wav2Lip 模型是在英文真实人脸数据集上进行训练得到的, 这就导致其在合成英文真实人脸视频时效果不错. 但是, 当 Wav2Lip 模型应用到中文动画人脸视频生成时, 其效果就不是那么的出众, 仍有很大的提升空间. 造成这种现象的原因有 3 个: 第一, 该数据集中音频都为英文, 导致模型缺少对中文音频特征的学习. 第二, 动画人脸与真实人脸仍有着差异. 例如, 真实人脸视频中人物有一个完整的唇部, 但是在动画人脸视频中人物唇部可能会是经过夸张之后的形状. 第三, 在真实的世界里, 文本和音频信号都与唇部运动有着紧密的联系[11,27,28]. 而 Wav2Lip 并没有考虑到这个问题, 导致其模型生成的人脸动画视频的真实性也有待提升. 因此本文提出使用给定的任意文本和音频互为补充信息, 来实现提升人脸动画视频真实性的任务.

现有方法[8,14−16]由于缺少中文人脸动画视频集和多模态信息的互相补充, 无法获得真实感较强并且可适用于动画人脸的唇音同步动画视频生成模型. 为了解决这两个问题, 本文首先构建中文唇音同步人脸动画视频数据集, 使得模型可以学习到中文特征以及动画人脸特征. 然后, 基于 Wav2Lip 模型, 本文提出一种文本音频驱动人脸的动画生成技术, 通过该技术来增加人脸动画视频的真实感.

本文的主要贡献有: ① 构建了 3 000 条包含文本、视频和音频的中文唇音同步人脸动画视频数据集, 该数据集中提供了中文真实人脸和中文动画人脸两种角

色的视频以及对应的音频和文本文件. ② 提出了一种基于改进 Wav2Lip 模型的文本音频驱动人脸动画生成方法, 融合多模态信息特征, 经过数千种身份和声音训练, 完成面对真实人脸和动画人脸都可生成高真实感的唇音同步人脸动画视频任务. ③ 通过对比实验表明, 改进模型生成的唇音同步人脸动画视频提升了视频唇形和音频同步性和真实感. 本文公开了自建中文唇音同步人脸动画数据集并提供了演示视频来展示改进模型效果, 具体请参见论文资源网络链接 https://drive.google.com/drive/folders/1hfFTGhYfiL5hxsd4tqTpWj0-woft7Ypa?usp=sharing.

1 Wav2lip 模型

Wav2Lip 模型可以同步任何音频和视频片段中角色唇部动作[24]. 其原理是从音频中提取声音波形信息, 从视频中提取角色唇部运动信息, 利用生成对抗网络来训练模型获取两者之间的映射关系, 使得模型可以生成一系列与输入音频匹配的唇部运动序列. 该模型中有生成器和判别器两部分, 生成器根据输入的音频信息, 将音频信息转化为一系列唇部动作序列, 而判别器通过判断生成器生成的一系列唇部动作和真实视频中唇部动作的相似度, 来评估生成器的性能. 通过重复的迭代训练, 生成器可以逐渐优化唇部动作序列的生成效果, 从而达到音视频同步效果.

Wav2Lip 模型在同步英文音频和视频上表现优异[25,26], 给生成唇音同步的人脸动画视频提供了一个很好的借鉴. 但令人遗憾的是, 在使用中文音频驱动人脸动画时其同步性不理想, 同时模型生成人脸动画视频的真实性也有待提高. 为解决这一问题, 本文构建了一个中文唇音同步人脸动画数据集并且在 Wav2Lip 模型的基础上, 引入了文本特征, 使得模型生成的人脸动画视频真实感更强.

2 数据集构建

本文使用的数据集包括真实人脸和动画人脸两类. 真实人脸数据集来源于 Chinese mandarin lip reading (CMLR) 数据集[29]. 该数据集中, 中文新闻联播视频包含由 11 位主持人所表述的共 102 076 条句子. 每个句子最多包含 29 个汉字, 不包含英文字母、阿拉伯数字和稀有标点. 本文所用数据集为该数据集中的一部分, 大约有 2 100 条视频和对应的 2 100 条文本.

同时, 区别于真实人脸, 为了使模型能够应用于动画彩绘人物, 本文基于现有的各种动画视频, 构建了动画人脸数据集. 在动画视频中选取人脸部分较为清晰并且正面面对镜头的片段. 为了尽可能多的学习人脸数据, 选取了 39 位动画角色所表述的共 900 条句子, 有 900 条视频和对应的 900 条文本. 每个句子最多包含 35 个汉字, 不包含英文字母、阿拉伯数字和稀有标点.

对数据集中的视频进行进一步处理. 首先, 进行人脸检测, 将视频中的每一帧都处理成仅包含人脸的视频帧, 并且提取视频的音频. 数据处理之后会生成一个和视频同名的文件夹, 文件夹中包括视频处理之后的人脸图片和音频. 文本则存储在文本文件夹中. 数据处理后, 中文真实人脸数据集如图 1 所示, 中文动画人脸数据集如图 2 所示.

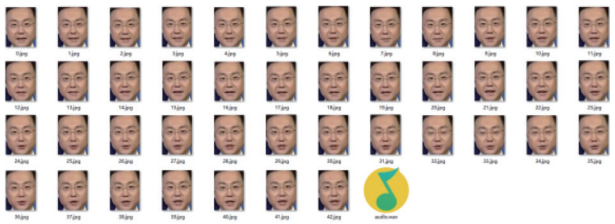

图 1 中文真实人脸数据集片段

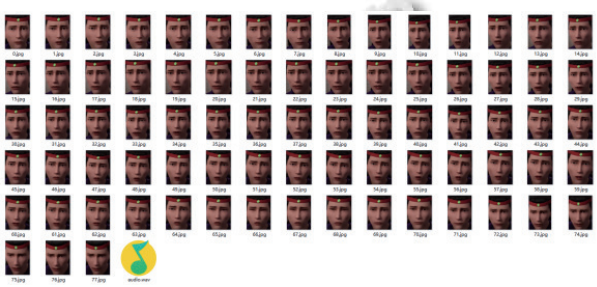

图 2 中文动画人脸数据集片段

3 改进的网络结构设计

改进后模型的网络结构如图 3 所示, 图中实线框为模型改动部分. 为了使得模型能够学习到文本特征和动画人脸特征, 本文构建了中文唇音同步人脸动画数据集①. 首先, 使用本文构建的数据集对唇部判别器⑥进行预训练, 该判别器将帮助人脸动画生成器生成更准确的唇同步人脸动画视频. 然后将数据送入生成器, 生成器分为编码器和解码器两部分. 编码器部分在 Wav2Lip 模型已有的人脸编码器②和音频编码器③的

基础上, 增加一个文本编码器④去获取文本特征, 取文本特征向量和音频特征向量的均值作为融合特征向量, 融合特征向量与通过人脸编码器②得到的人脸特征向量拼接, 拼接后的向量就是联合视听向量. 联合视听向量作为人脸动画解码器⑤的输入, 通过⑤解码生成具有唇音同步唇部运动口型的中文人脸动画视频. 最后, 使用视频质量判别器⑦减少视频人脸部分的伪影问题. 总体来说, 本文所提出的模型是在预训练的唇部判别器和视频质量判别器的监督下, 训练人脸动画生成网络, 用于唇音同步的中文人脸动画视频生成.

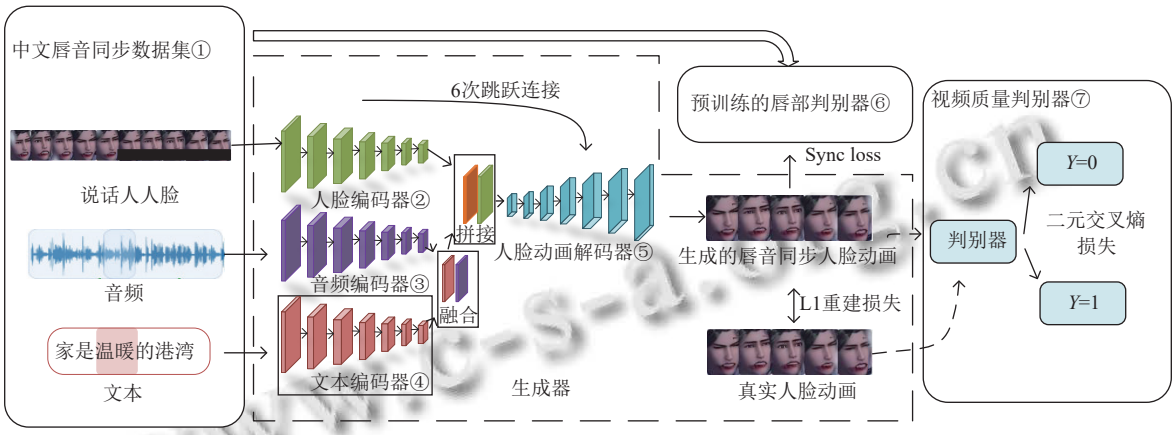

图 3 基于改进 Wav2Lip 的文本音频驱动人脸动画生成技术

3.1 唇部判别器

唇部判别器⑥[21]设计目的是增加人脸动画视频中音频和人脸运动的同步性. 在整体模型进行训练之前, 先对唇部判别器进行预训练, 使得判别器可以准确检测视频中音频和人脸运动是否同步. 唇部判别器不需要像 LipGan[23]一样, 根据整体模型生成的视频帧进行调整, 因为从真实视频中学习到音频和视频的对应特征最为准确. 唇部判别器是在大型唇同步数据集[6,7]上训练出来用于纠正唇同步错误的网络. 但是, 该大型唇同步数据集中, 缺少中文数据和人脸动画数据. 为了完成生成高真实感的中文人脸动画视频这一任务, 使用本文构建的唇音同步人脸动画数据集去进一步调整唇部判别器. 唇部判别器的输入由一个连续只包含下半部分的人脸帧 Fv 和一个音频段 Fa 构成, Fv 和 Fa 分别是视频和音频的时间步长. 唇部判别器在训练时, 通过随机采样与人脸视频同步或者不同步的音频 Fa 来区分音频和视频之间的同步. 在模型结构上, 使用跳跃连接的方法来更好的拟合特征. 模型包含一个脸部编码器和一个音频编码器, 两者都是由二维的卷积网络构成. 网络使用余弦相似性与二进制交叉熵损失来计算视频和语音嵌入之间的点积, 用来产生音频-视频对同步的概率 P.

$$P_{\text{sync}} = \frac{v \cdot s}{\max(\|v\|_2 \cdot \|s\|_2, \epsilon)} \tag{1}$$

其中, v 代表视频向量, s 代表语音向量. 在本文构建的唇音同步的人脸动画数据集上训练了大约 48 h, 每个训练批的大小为 4, 在 Fv=5 帧的情况下, 使用 Adam 优化器, 初始学习率为 1E–3.

3.2 生成唇音同步的人脸动画视频

在唇音同步的人脸动画视频生成时, 使用了对抗神经网络的结构, 有一个生成器和一个判别器.

3.2.1 生成器

生成器 G 包含 4 个模块: (1) 人脸编码器②, (2) 音频编码器③, (3) 文本编码器④, (4) 人脸动画解码器⑤. 人脸编码器是由包含残余的卷积层构成的, 它的主要工作是对人脸视频中随机获得的人脸参考帧 F 进行编码. 人脸视频中获得的参考帧 F 需要和遮盖掉下半张人脸的姿势帧在通道的维度上进行串联. 音频编码器是一个标准的卷积神经网络, 它的工作是将对应的语音段进行编码. 它的输入为 M×T×1 大小的 Mel-frequency 倒谱系数 (MFCC) 热图, 输出是大小为 h 的音频嵌入. 文本编码器也是包含残余的卷积层, 它的工作是将对应的文本段进行编码. 使用 Word2Vec[30]获得文本中的文本特征并转化成为可以输入文本编码器的矢量表示, 通过 Word2Vec 获得的文本特征具有丰富的语义含义. 文本编码器的输入是 M×T×1 大小的文本特征, 输出是大小为 h 的音频嵌入. 将大小为 h 的文本编码器和大小为 h 的音频编码器生成的特征向量, 通过取均值的

方法生成大小为 h 的融合特征向量. 在该步骤中由于在模型原有的音频信息中增添了文本中的语义特征, 可以使得模型生成唇音同步的人脸动画真实感更强. 将融合特征向量与人脸嵌入连接起来, 产生大小为 $2\times h$ 的联合视听嵌入. 联合视听嵌入作为人脸动画解码器的输入. 人脸动画解码器由包含残余的卷积层和上采样的转置卷积构成. 损失函数使用 L1 重建损失函数.

$$L_{\text{recon}} = \frac{1}{N}\sum_{i=1}^{N}\left\|L_g - L_G\right\|_1 \tag{2}$$

其中, L_g 表示生成的视频帧, L_G 表示真实的视频帧.

在训练期间, 由于唇部判别器⑥一次需要处理 5 个连续的视频帧, 所以生成器需要生成连续的 5 个帧, 送入唇部判别器让其判断. 在生成器进行处理时, 将会沿着批处理的维度堆叠时间步长. 唇部判别器在对生成器生成的视频进行判断时, 时间步长也会沿着信道维度进行串联, 就和训练过程一样. 唇部判别器在判别生成人脸的视频帧中的唇部的同时, 还对生成器进行训练, 损失函数为:

$$E_{\text{sync}} = \frac{1}{N}\sum_{i=1}^{N} -\log(P^i_{\text{sync}}) \tag{3}$$

其中, P^i_{sync} 是根据式 (1) 得到的. 在生成器的训练过程中, 唇部判别器的权重不进行调整. 唇部判别器提前从视频中学到唇音同步特征将要求生成器生成更加准确的唇音同步的人脸动画视频.

3.2.2 视频质量判别器

由于使用提前训练好的唇部判别器, 生成器在生成人脸动画视频时, 为了获得更加同步的人脸动画视频, 有时会导致生成的人脸动画视频在唇部区域有模糊或者伪影. 为了减轻这种情况的产生, 在训练生成器的同时, 训练视频质量判别器⑦. 所以, 在整体模型里, 使用了两个判别器, 唇部判别器目的是提前学习唇音同步的标准, 来要求生成器生成更加准确的唇音同步的人脸动画视频. 视频质量判别器目的是减少生成人脸动画视频中变形区域的模糊或者伪影.

视频质量判别器也是由一堆卷积块组成, 每个卷积块是一个卷积层和其后的 LeakyReLU 激活函数组成. 视频质量判别器的损失函数为:

$$L_{\text{gen}} = \mathbb{E}_{x\sim L_g}\left[\log(1 - D(x))\right] \tag{4}$$

$$L_{\text{disc}} = \mathbb{E}_{x\sim L_G}\left[\log(D(x))\right] + L_{\text{gen}} \tag{5}$$

其中, L_g 表示来自生成器生成视频的图像, L_G 表示来自真实视频的图像, $D(x)$ 表示判别器对样本 x 为真的概率.

生成器将最小化式 (6), 其实就是重建损失 (式 (2)), 同步损失 (式 (3)) 和对抗损失 (式 (4)) 的加权求和.

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

其中, s_w 是同步惩罚权重, s_g 是对抗性损失, 根据前人的经验设置为 0.03 和 0.07[20]. 在引入文本特征之后, 再使用唇部判别器和视频质量判别器两个不互相影响的判别器, 完成提高网络输出唇音同步视频的准确度和真实感的任务.

在本文构建的中文唇音同步人脸动画数据集上训练模型, 批次大小为 4. 在训练生成器和判别器时, 使用 Adam 优化器, 初始学习率为 1E–4.

使用生成唇音同步的人脸动画视频的过程, 来总结网络的整体框架. 与 LipGan[19]模型相似, 本文提出的网络结构会生成唇音同步视频的每一帧. 每一个时间步长上的视频输入都是来自于相同时间里对应人脸区域的裁剪, 与相同时间里遮盖掉下半张人脸的视频相连接. 因此, 在生成视频里, 本文所提出网络不需要去调整视频中人脸的运动姿势, 从而减少了生成的唇音同步的人脸动画视频里伪影的存在. 根据视频相应的音频和文本输入, 网络生成的人脸动画视频的唇部区域将发生变动.

4 实验结果与分析

为了更好地评估模型训练的结果, 本文将通过客观评测和人工评测两种方法进行评估.

4.1 客观评测

为了评估生成的人脸动画视频中唇音同步的效果. 我们使用预先训练的 SyncNet, 该网络在 SyncNet[19] 之后是公开可用的. 该方法可以测试语音音频和嘴唇运动之间的同步. 我们采用两个指标: 来自 Wav2Lip 的唇同步误差距离 (lse-d) 和唇同步误差置信度 (lse-c)[24]. 这两个指标使用 SyncNet 模型来进行计算.

lse-d 是通过训练视频片段的语音特征和视频特征, 计算其欧氏距离, 然后再由视频片段组成的原视频中找到最小欧氏距离, 这个最小欧氏距离将作为人脸口型与语音的偏差指标. lse-d 表示不同偏移值的音频和视频特征之间的最小距离. 较低的 lse-d 意味着语音

音频和视频更加同步.

lse-c 是使用欧氏距离的最小值和中位数之差作为人脸口型与语音的置信度分数. lse-c 表示音频和视频以一定的时间偏移同步的置信度. 较低的 lse-c 意味着视频的某些部分完全不同步, 其中音频和视频不相关.

使用 ATVGnet[31]、Wav2Lip 和改进后的模型合成人脸动画视频进行比较. 在训练时, 使用本文构建的中文人脸视频数据集, 在他们公开的代码上进行训练. 在生成唇音同步的人脸动画视频时, 提供相同的音频和目标视频. 使用的 3 个模型都可以合成唇音同步的人脸动画视频. 测试后的结果如表 1 所示. 表 1 中, 相比 ATVGnet 模型, 改进后的 Wav2Lip 在 lse-d 指标上降低了 1.393, 在 lse-c 指标上提升了 1.383. 相比 Wav2Lip 模型, 改进后的 Wav2Lip 在 lse-d 指标上降低了 0.626, 在 lse-c 指标上, 提升了 0.153. 测试结果表明, 改进后的 Wav2Lip 能够生成更好的唇音同步的人脸动画视频.

表 1 测试结果

网络	lse-d↓	lse-c↑
ATVGnet	8.75	5.631
Wav2Lip	7.983	6.861
改进后的Wav2Lip	7.357	7.014

为了与其他模型进行更加公平的对比实验, 下载 50 个新闻发言视频和 150 个中文动画视频. 视频以 30 帧进行分割. 随机选择其中 30% 的视频用作验证, 70% 的视频用作训练. 测试后的实验结果如表 2 所示.

表 2 测试结果

网络	lse-d↓	lse-c↑
ATVGnet	8.126	6.172
Wav2Lip	7.653	6.912
改进后的Wav2Lip	7.091	7.861

4.2 人工评测

由于人类对于视频唇音同步特别的敏感, 故本次实验还召集了 20 位母语为中文的志愿者对生成的唇音同步的人脸动画视频进行评分评估, 以测量音频和视频的同步.

在进行人工评测时, 测试集分为两种: 第 1 种测试集中, 我们通过网络生成的唇音同步的人脸动画视频中随机选择 30 个视频和 30 个真实的人脸动画视频进行混合, 让志愿者判断哪个动画为真实的, 哪个动画为合成的. 志愿者判断准确度求其平均值, 实验结果如表 3 所示. 其中, 人工识别准确度的值越小表示越难分辨动画为合成或真实. 实验结果表明, 相比于 ATVGnet 和 Wav2Lip 模型, 改进后的模型生成的唇音同步的人脸动画视频, 志愿者更加无法准确地分辨出哪一个动画是模型生成的动画.

表 3 测试结果 (%)

网络	人工识别准确度
ATVGnet	72.5
Wav2Lip	58.5
改进后的Wav2Lip	25.8

第 2 种测试集包含使用 ATVGnet、Wav2Lip 和改进后的 Wav2Lip 生成的 60 个唇音同步的人脸动画视频. 志愿者将从两个方面进行评测: ① 生成的视频相邻帧之间是否能保证时间连贯性, 分数为 0–10 分. 0 分: 视频完全不连贯. 1–3 分表示视频容易跳帧. 4–6 分表示视频有时跳帧. 7–9 分表示基本不跳帧. 10 分表示视频完全不跳帧. 视频分数越高说明越连贯. ② 生成的视频与相应音频时间同步的百分比, 分数为 0–10 分. 0 分: 音视频完全不同步, 声音和画面严重不同步. 1–3 分: 音视频同步较差, 画面和声音有明显的不同步现象, 影响观看体验. 4–6 分: 音视频同步一般, 有些场景下会出现不同步现象, 但大部分情况下勉强可以接受. 7–9 分: 音视频同步较好, 画面和声音的同步性非常好, 基本不影响观看体验. 10 分: 音视频完全同步, 画面和声音完全同步, 观看体验极佳. 在评测过程中, 每位志愿者的播放顺序都是随机的. 将志愿者打完的分值, 取其平均值, 测试结果如表 4. 实验结果表明, 改进后的模型在生成视频的连贯性和视频与音频的同步性上, 效果更好.

表 4 测试结果

网络	时间连贯数值↑	同步数值↑
ATVGnet	8.32	3.95
Wav2Lip	9.48	4.35
改进后的Wav2Lip	9.63	5.57

根据客观评测和人工评测结果, 本文所提出的方法生成的人脸动画视频唇音最为同步、效果最为真实. ATVGnet 使用从音频中提取到的特征和从人脸视频中某一帧获得的人脸特征, 去生成唇音同步的人脸动画. 相比于这类只从视频的一帧去获取人脸特征的方法, 本文所提出的方法生成的人脸动画视频更加自然、真实. 而 Wav2Lip 没有考虑到文字和动画角色对于生成唇音同步视频的影响. 相比于这类只考虑音频特征和人脸特征的模型, 本文所提出的模型在合成人

脸动画视频时, 同步性和真实性更加良好.

4.3 人脸动画生成效果展示

本文方法能够生成高真实感的唇音同步人脸动画视频. 同时, 可以不受动画中人物是真实人脸还是动画人脸的限制. 图 4 通过截取原始视频、使用 ATVGnet 生成的人脸动画视频、使用 Wav2Lip 生成的人脸动画视频和使用本文提出方法生成的人脸动画视频中连续 3 帧进行对比, 展示了本文提出方法的真实效果, 图片下中文为生成视频帧中人物发音的中文. 具体演示效果请参见论文资源网络链接 https://drive.google.com/drive/folders/1hfFTGhYfiL5hxsd4tqTpWj0-woft7Ypa?usp=sharing.

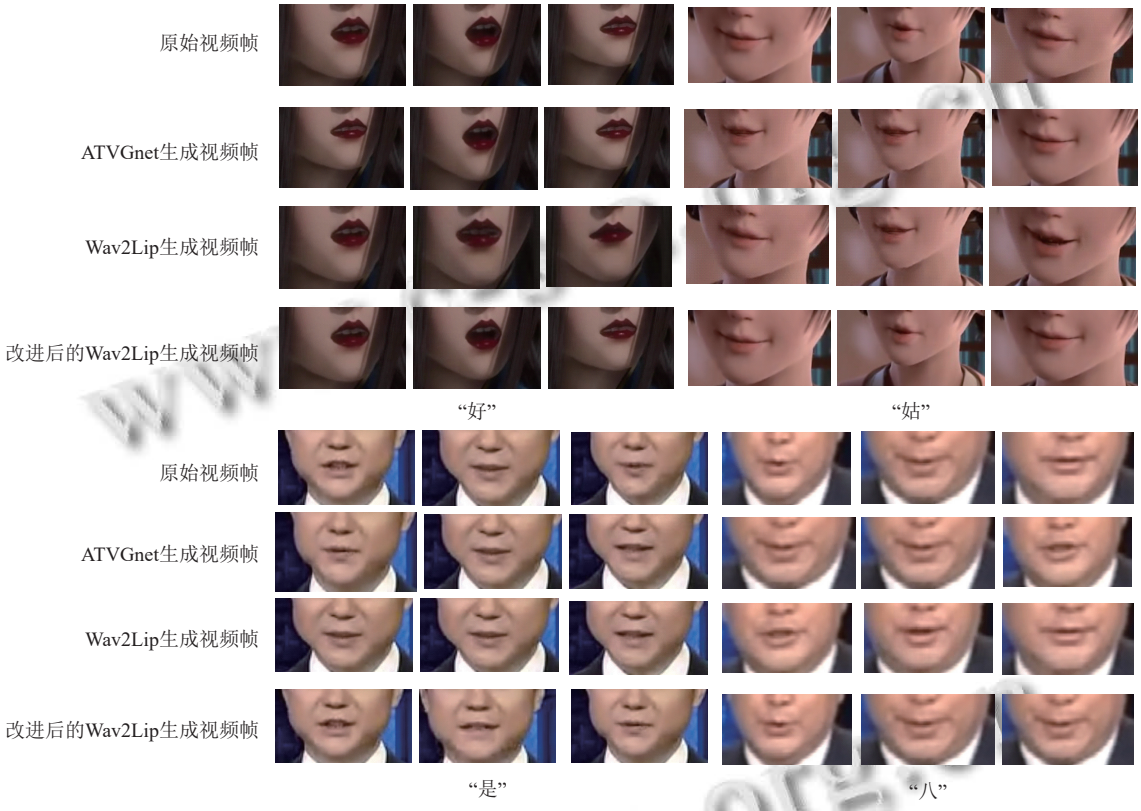

图 4 原始视频、ATVGnet 生成视频、Wav2Lip 生成视频和改进后 Wav2Lip 生成视频连续 3 帧唇部对比效果

参考文献

1 闫衍芙, 吕科, 薛健, 等. 基于深度学习和表情 AU 参数的人脸动画方法. 计算机辅助设计与图形学学报, 2019, 31(11): 1973–1980.

2 Zhang ZY, Liu ZC, Adler D, *et al*. Robust and rapid generation of animated faces from video images: A model-based modeling approach. International Journal of Computer Vision, 2004, 58(2): 93–119. [doi: 10.1023/B:VISI.0000015915.50080.85]

3 Cao C, Weng YL, Lin S, *et al*. 3D shape regression for real-time facial animation. ACM Transactions on Graphics, 2013, 32(4): 41.

4 Blanz V, Vetter T. A morphable model for the synthesis of 3D faces. Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. Los Angeles: ACM Press, 1999. 187–194.

5 孔英会, 秦胤峰, 张珂. 深度学习二维人体姿态估计方法综述. 中国图象图形学报, 2023, 28(7): 1965–1989.

6 Afouras T, Chung JS, Senior A, *et al*. Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 8717–8727. [doi: 10.1109/TPAMI.2018.2889052]

7 Afouras T, Chung JS, Senior A, *et al*. LRS3-TED: A large-scale dataset for visual speech recognition. arXiv:1809.00496, 2018.

8 Kumar R, Sotelo J, Kumar K, *et al*. ObamaNet: Photo-realistic lip-sync from text. arXiv:1801.01442, 2017.

9 Thies J, Elgharib M, Tewari A, *et al*. Neural voice puppetry: Audio-driven facial reenactment. Proceedings of the 16th European Conference on Computer Vision. Glasgow:

Springer, 2020. 716–731.
10 李欣怡, 张志超. 语音驱动的人脸动画研究现状综述. 计算机工程与应用, 2017, 53(22): 21–28, 34.
11 Zhang T, Deng LR, Zhang L, *et al*. Deep learning in face synthesis: A survey on deepfakes. Proceedings of the 3rd IEEE International Conference on Computer and Communication Engineering Technology (CCET). Beijing: IEEE, 2020. 67–70.
12 Lu YX, Chai JX, Cao X. Live speech portraits: Real-time photorealistic talking-head animation. ACM Transactions on Graphics, 2021, 40(6): 220.
13 Liang BR, Pan Y, Guo ZZ, *et al*. Expressive talking head generation with granular audio-visual control. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 3377–3386.
14 刘贤梅, 刘露, 贾迪, 等. 基于语音驱动的三维人脸动画技术综述. 计算机系统应用, 2022, 31(10): 44–50. [doi: 15888/j.cnki.csa.008776]
15 Ji XY, Zhou H, Wang KSY, *et al*. EAMM: One-shot emotional talking face via audio-based emotion-aware motion model. Proceedings of the ACM SIGGRAPH 2022 Conference. Vancouver: ACM, 2022. 61.
16 Chung JS, Jamaludin A, Zisserman A. You said that? Proceedings of the 2017 British Machine Vision Conference. London: BMVA Press, 2017.
17 Zhou Y, Han XT, Shechtman E, *et al*. MakeItTalk: Speaker-aware talking-head animation. ACM Transactions on Graphics, 2020, 39(6): 221.
18 Zhou H, Liu Y, Liu ZW, *et al*. Talking face generation by adversarially disentangled audio-visual representation. Proceedings of the 33rd AAAI Conference on Artificial Intelligence, the 31st Innovative Applications of Artificial Intelligence Conference and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence. Honolulu: AAAI Press, 2019. 1141.
19 Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I. Synthesizing Obama: Learning lip sync from audio. ACM Transactions on Graphics, 2017, 36(4): 95.
20 Vougioukas K, Petridis S, Pantic M. Realistic speech-driven facial animation with GANs. International Journal of Computer Vision, 2020, 128(5): 1398–1413.
21 Chung JS, Zisserman A. Out of time: Automated lip sync in the wild. Proceedings of the ACCV 2016 International Workshops on Computer Vision. Taipei: Springer, 2017. 251–263.
22 Yu LY, Yu J, Li MY, *et al*. Multimodal inputs driven talking face generation with spatial-temporal dependency. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(1): 203–216. [doi: 10.1109/TCSVT.2020.2973374]
23 Prajwal KR, Mukhopadhyay R, Philip J, *et al*. Towards automatic face-to-face translation. Proceedings of the 27th ACM International Conference on Multimedia. Nice: ACM, 2019. 1428–1436.
24 Prajwal KR, Mukhopadhyay R, Namboodiri VP, *et al*. A lip sync expert is all you need for speech to lip generation in the wild. Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020. 484–492.
25 Jang Y, Rho K, Woo J, *et al*. That's what I said: Fully-controllable talking face generation. Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023. 3827–3836.
26 谢天, 于灵云, 罗常伟, 等. 深度人脸伪造与检测技术综述. 清华大学学报 (自然科学版), 2023, 63(9): 1350–1365.
27 Ling ZH, Richmond K, Yamagishi J. An analysis of HMM-based prediction of articulatory movements. Speech Communication, 2010, 52(10): 834–846. [doi: 10.1016/j.specom.2010.06.006]
28 Fried O, Tewari A, Zollhöfer M, *et al*. Text-based editing of talking-head video. ACM Transactions on Graphics, 2019, 38(4): 68.
29 Zhao Y, Xu R, Song ML. A cascade sequence-to-sequence model for Chinese mandarin lip reading. Proceedings of the 2019 ACM Multimedia Asia. Beijing: ACM, 2019. 32.
30 Rong X. Word2Vec parameter learning explained. arXiv: 1411.2738, 2014.
31 Chen LL, Maddox RK, Duan ZY, *et al*. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7824–7833.

(校对责编: 孙君艳)