项目编号：T030PRP43127

上海交通大学

本科生研究计划（PRP）研究论文

（第四十三期）

论文题目：数字人体验质量评价

项目负责人：赵楷越 学院（系）： 电子信息与电气工程学院

指导教师： 闵雄阔 学院（系）： 电子信息与电气工程学院

参与学生：赵楷越

项目执行时间：2023 年 2 月 至 2023 年 9 月

摘 要

数字人视频生成是一项新兴技术，指利用给定的图像，音频，视频或文本数据生成一段由图像画面中的人说出对应音频数据中的声音的视频。目前的数字人说话视频生成算法一共有三个大类：音频驱动类生成算法，视频驱动类生成算法以及文本驱动类生成算法。对于不同种类的生成算法，其所需的源数据不同，生成数字人的过程和视频结果亦有差异。但是，由这些算法生成的数字人质量参差不齐，且目前缺乏合适的方法对数字人视频进行质量评价，因此如何可靠地评价数字人的生成质量，并实现数字人说话视频的自动化评价十分重要。基于可靠的数字人说话视频质量评价，才能更有效地指引不同种类的数字人生成算法能从哪些角度进行改进。

本课题围绕数字人说话视频的体验质量评价，主要探索了数字人技术的发展现状和未来趋势；分析了影响数字人感知质量的主客观因素；建立了数字人质量评价的框架；最后以人类主观质量评价为最终准则，设计合理的主观评价方案，提出可靠的客观性能指标，再基于 ANNAVQA 实现了数字人质量评价的有效自动化及评价模型的多模态融合；最终评估了不同类型的数字人的优劣势，为数字人的优化和改进提供参考和建议。

关键词：数字人，质量评价，ANNAVQA，深度神经网络，多模态融合

正 文

1. 绪论

1.1 选题意义和目的

数字虚拟人是一项新兴 AI 技术，自 2020 以来，受到了工业界和学术界的日益增加的关注，近年来数字人技术也不断地快速发展，目前已经有众多采集式和生成式数字人面世。但是面对不断涌现的新型数字人生成技术和数字人质量参差不齐的现状，如何可靠地评价数字人的生成质量逐渐成为了一项亟待解决的科学问题，因此其质量评价的研究兼具理论和实用价值。

研究数字人质量评价可为当前已有数字人总结影响其感知质量的影响因素，提供主观评价方案，提出客观性能指标。更重要地，通过主客观质量评价发现不同数字人各自的优劣势，为未来数字人提供有效的指导，指明可靠的发展方向。

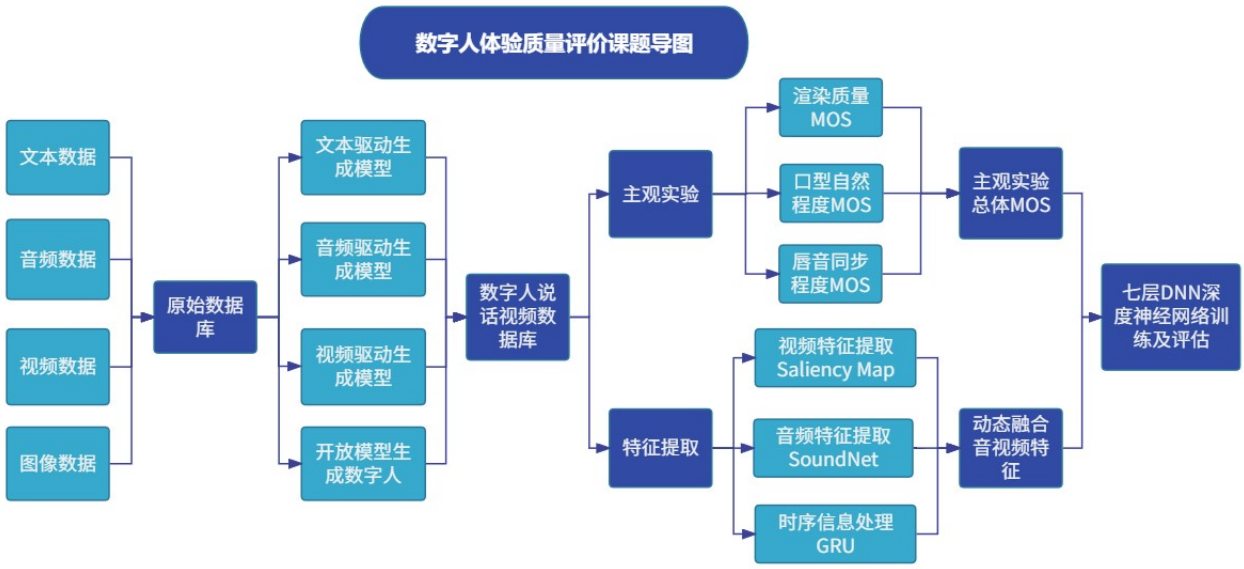
1.2 研究范围和解决的主要问题

传统的视觉质量评价主要关注自然图像/视频的清晰度和保真度等中低层视觉信息，而数字人客观评价则更多地需要考虑多模态交互（如视觉与语音）、渲染逼真度、行为拟真度等高层语义信息，因此，传统视觉质量评价方法并不适合直接套用在数字人质量评价任务中。当前，在学术和产业界中，数字人质量评价仍是一片空缺，缺少系统性的研究总结和有效的基本方法。而由于数字人的最终受众是人，因此和所有质量评价问题类似地，客观数字人质量评价也仍需以与人类主观质量评价的一致性作为最终准则。

本课题解决的主要问题针对数字人说话的生成模型进行质量评价，也围绕数字人质量评价的特点和难点，提出了适合数字人质量评价的技术方法，包括：基于多模态数据的数字人主观评价方案，考虑了数字人与人类交互的多种场景和方式，设计了合理的主观评价实验和指标。基于 ANNAVQA 的数字人客观评价模型，利用了大量的主观评价数据和客观特征数据。提出了针对不同种类和算法的数字人说话的生成模型质量评价指标和方法。

2. 研究内容及方法

为了便于在论文开始时大致认识本课题的研究经历，下图清晰直观地展示了本课题的整个流程：



2.1 数字人说话视频原始生成数据库的确立

在构建数字人数据库的过程中，我们以 What comprises a good talking-head video generation 一文中对于数据库的选择作为标准，首先确定了数字人生成的标准化数据库。

对于数字人生成的所有数据，我们也对其进行了初步的筛选，我们仅选择使用数字人生成算法模型中属于测试集而非训练集的原始数据，使视频生成结果受模型特殊性的影响较小，也使我们的最终评价模型更为真实可靠。本研究主要使用了如下几个数据库作为数字人说话的原始数据库：

2.1.1 原始图像数据库

2.1.1.1 IMDB-WIKI Dataset

数据属性：背景随机，正脸侧脸皆有，可能含有自然表情；

2.1.1.2 Flickr-Faces-HQ Dataset (FFHQ)

数据属性：原始分辨率 1024×1024 ，各种年龄，种族和照片背景，以及可能佩戴眼镜或饰品；

2.1.2 原始视频数据库

2.1.2.1 Lip Reading Sentences

数据属性：包含千段的 TED 演讲的视频片段

2.1.3 原始音频数据库

2.1.3.1 VoxCeleb2 dataset

数据属性：音频的视频源有明显的头部移动，语音为中性的表达情感，在自然状态下录制（导致部分语音片段会被噪音破坏），均为英语语言音频。

2.1.4 原始文本数据库

数据属性：由具有不同可识别音素的英文单词组成的常规短语，短句数据组成。

2.2 构建数字人说话视频数据库

为了提升本数字人体验质量评价模型的准确性，及分析不同数字人生成算法在本评价模型下的表现如何，在构建数字人说话视频数据库的时候，我们寻求了八种不同的数字人生成算法用于生成了一千段包含数字人说话的视频，每种算法生成 125 段视频。为了进一步提升本数字人评价模型对不同生成方式的数字人说话视频算法的适应性，所选取的八种算法中也包含了四种不同的数字人说话视频模型的生成类型，每种生成类型包括两种算法，下面介绍该四种不同生成种类的数字人说话视频生成算法：

2.2.1 音频驱动生成

音频驱动的数字人生成模型根据给定的一段真实人说话的音频信号和一张真实的人脸图像来对数字人进行说话动作的合成。算法需要对音频信号进行特征提取，例如音高、节奏、能量、情绪等，这些特征可以用来驱动数字人的口型、表情等。再将音频特征和数字人模型进行关联和映射，最后便能生成一段包含与音频信号相匹配的数字人形象和口型等动作的视频。

采用的音频驱动生成模型有： SadTalker，Audio To Obama

2.2.2 视频驱动生成

视频驱动的数字人生成模型根据外界输入的视频驱动人物模型生成相应的语音和动作，模型读取并解析识别外界输入信息，根据解析结果决策虚拟数字人后续的输出文本，然后驱动人物模型生成相应的语音与动作来使虚拟数字人跟用户互动。这种方式是基于深度学习模型的三维场景表达和对应的神经渲染管线，可以自驱动学习模特说话时的唇动、表情、语音以及姿态和动作等。

采用的视频驱动生成模型有：Audio2Head， PC-AVS

2.2.3 文本驱动生成

文本驱动的数字人生成模型根据自然语言作为输入，便能生成对应的三维数字人的模型和动作。模型使用大规模视觉语言模型来理解文本输入的含义和风格，并将其映射到一个高维空间中。然后，利用可微渲染工具并渲染出数字人的模型。最后，利用大量动作数据的预训练模型能根据文本输入的语义和情感，生成出一段包含数字人的动作和表情的说话视频。

采用的文本驱动生成模型有：OneShot，Talking Face Generation with Multilingual TTS

2.2.4 开放模型生成

剩余的算法和模块由目前已面向市场开放的标准化数字人说话生成模型网站生成，将这些网站生成的数字人说话视频与由本地及云端代码生成的数字人说话视频一起作为本课题中评价模型的视频数据库的一部分，增加了本课题中模型的全面性。

采用的市场开放生成模型有：D-ID’s AI Presenters，Synthesia AI

2.3 总结影响数字人体验质量的影响因素

为了开展数字人的主观质量评价，我们先对可能影响用户对数字人视频的满意程度的不同因素进行了如下分析：

2.3.1 渲染质量

渲染质量的指标是指用于评价数字人的视觉效果的一些量化指标，例如图像质量（常用评价参数为 PSNR 和 SSIM）、纹理分辨率、光照细节、阴影质量、动画流畅度等。可以客观地反映数字人的视觉真实度和美感。

2.3.2 口型自然程度

口型自然程度的指标是指数字人的口型动作是否与语音和表情相匹配，是否符合人类的生理和心理特征，是否能够提高数字人的真实感和沟通效果。在数字人质量体验评价模型中，口型自然程度的指标是一个重要的组成部分，它可以反映数字人的交互质量和用户满意度。

2.3.3 音唇一致性

唇音同步是指将音频信号与数字人的口型运动相匹配的技术，它是数字人质量体验评价模型中的一个重要指标。 唇音同步的好坏直接影响数字人的逼真程度和自然度，也影响用户对数字人的信任度和满意度。

2.3.4 自发性动作自然度

人类在说话时往往会伴随着一些自发性动作，如头部运动和表情。自发性动作中包含非语音信息，能够帮助听者理解语音内容。自然的全身动作特点包括流畅、符合人类行为习惯，数字人呈现一种“放松”的状态特征；而不自然的动作特点显示为呆板，缺乏变化，数字人呈现“紧张、拘谨、木偶”感的特征。一个优秀的数字人在说话时也应当伴随自然的自发性动作，“不动”或“乱动”都不是期望的结果，因此在进行数字人主观质量评价时，应对自发性动作的自然度予以关注。

2.3.5 人脸神似分析

Paul Ekman 和 Friesen 通过观察和生物反馈，描绘出了不同的脸部肌肉动作和不同表情的对应关系从而创制了 FACS。他们根据人脸的解剖学特点，将其划分成若干既相互独立又相互联系的运动单元（AU——Action Unit）。FACS 是如今面部表情的肌肉运动的权威参照标准，也被心理学家和动画片绘画者使用。在进行数字人主观质量评价时，人脸神似分析也是可考虑的一面。

2.4 开展主观评价

对生成的数字人说话视频进行主观评价的意义一方面在于利用训练集的主观分数来训练模型，让其预测分数尽可能贴近人们主观打分；另一方面也在于检验数字人体验质量评价模型在测试集上预测出来的分数是否足够贴近人主观打分，达到检验模型的功能。

在分析了诸多如上数字人主观质量评价的可能指标之后，我们综合各指标的内容，以及受试者评分时的感受，进行仔细考虑后，将数字人说话视频的主观评价指标分为了三个主要部分：渲染质量，口型自然程度和唇音同步程度。然后开展了主观评价的过程

2.4.1 主观评价的过程设置

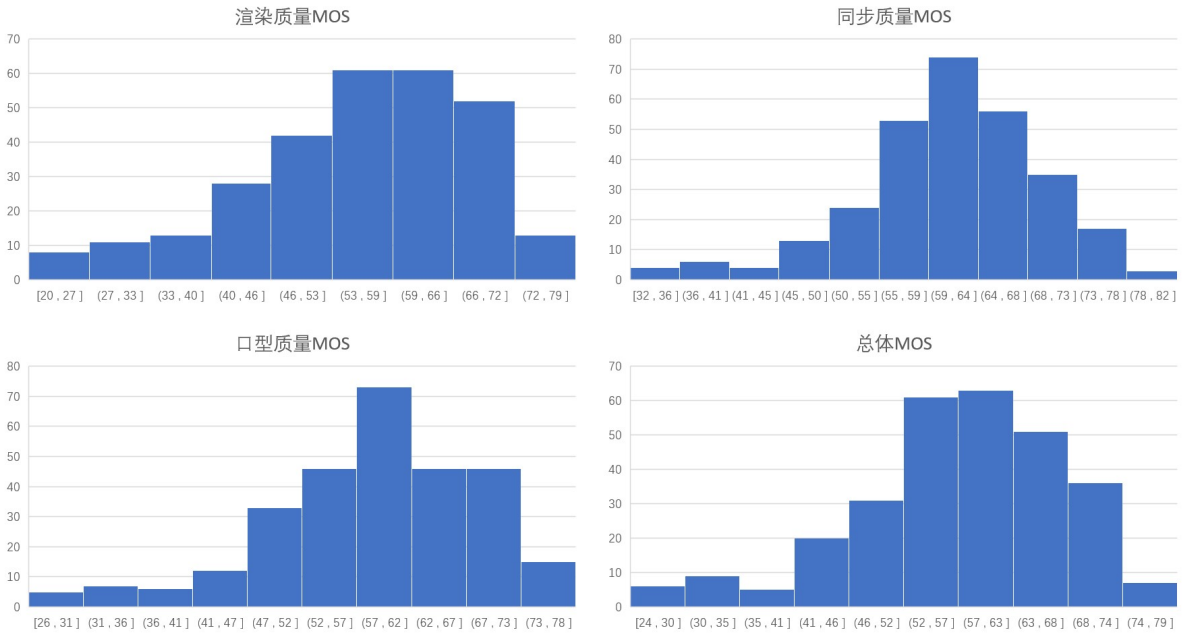
测试方法：本测试得到的分数为单刺激连续分数。开始测试之前会先让测试人员观看真人视频，并告知根据失真视频与真人视频的差异程度进行打分。测试人员打分的 MOS 值，范围为[0, 100]之间的整数。

评价维度：1）渲染质量：关注视频整体渲染的感知质量。 2）口型自然：关注数字人口型抖动情况。3）唇音同步：关注数字人视频与音频的同步情况。测试人员针对每项进行评分，并记录对整体感知质量影响最大的一项。

主观测试人员：10 名交大在读学生（本科生 6 人，硕士生 2 人，博士生 2 人），均不具有数字人质量评价相关知识背景。

2.4.2 主观评价后的 MOS 数据处理

为了提高数据的准确性，我们对 MOS 进行了数据处理。这里采用了基于上四分数的离群点剔除算法对原始数据进行后处理。然后分析不同评价维度的权重，通过加权评分的公式，得到了整体 MOS 评分 = 0.5*渲染 MOS + 0.25*同步 MOS + 0.25*口型 MOS。最终我们得到了 1000 个有标注的数字人视频。总 MOS 分布图如下图所示：

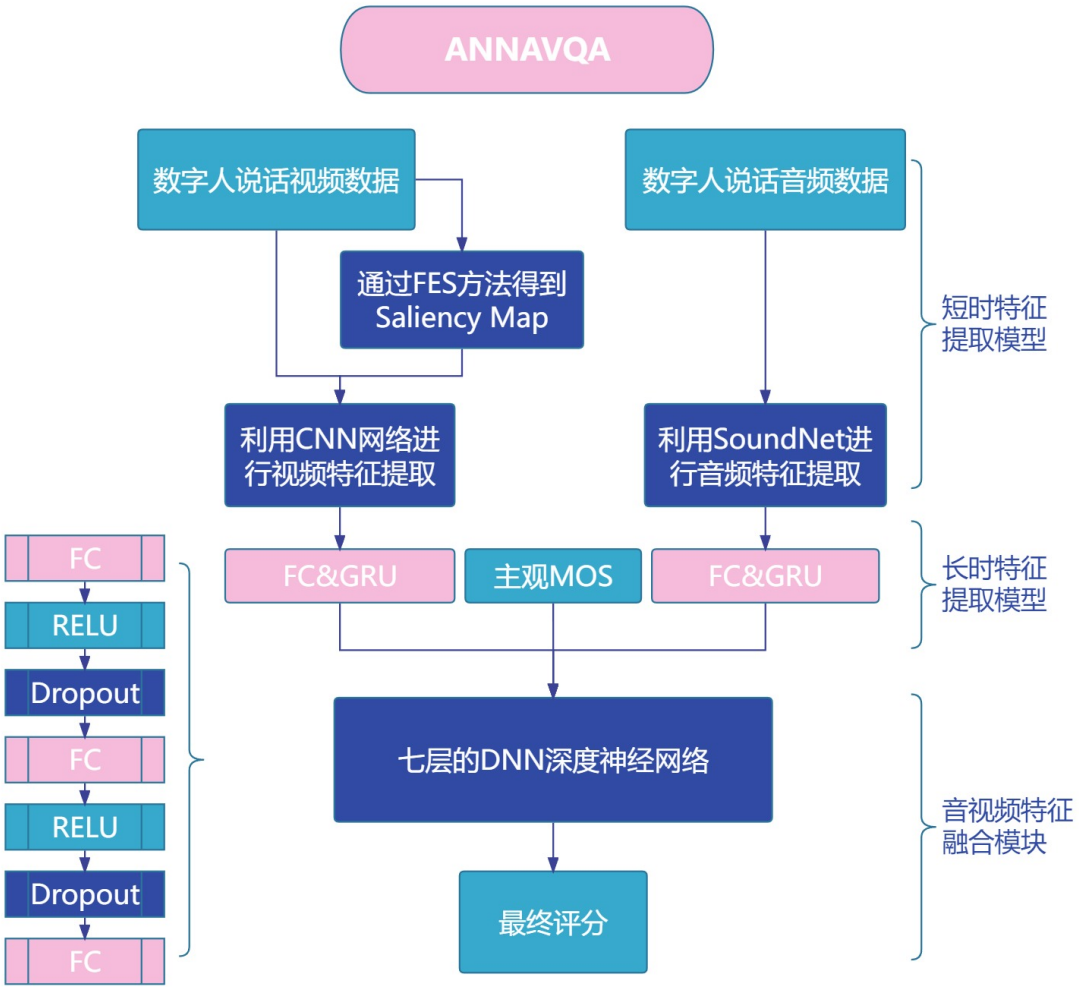


2.5 建立数字人的质量评价模型

2.5.1 ANNAVQA 概览

我们使用了 ANNAVQA 视频评价模型，结合先前得到的数字人主观质量评分，开始建立数字人质量评价模型。ANNAVQA 模型提出了一种进步性的音视频多维度评价方式，而且可以利用该模型进行无参考（no reference）的视频质量评价，该模型创新性地提出了融合音频和视频两个维度的特征后再进行评价，实现了对跨模态特征进行动态融合的效果。

模型总流程示意图如下：



相较于其他的视频评价模型，ANNAVQA 视频评价模型包有两方面显著的提升：

1、将视觉注意力（Visual Attention）纳入了视频评价的考量。视觉注意力具体指：人在观看一段视频的时候，其注意力更多放在画面前端不停运动的实体（例如说话的人，运动的车），而对画面背景中的一些相对静态的环境（例如说话人背后的草坪，运动的车行驶的道路）的注意力较少。为了将视觉注意力纳入评价模型的考量，该模型采用了计算视频画面中 Saliency Map 的方式，将画面中人的视觉注意力关注的主体部分进行提取，再将结果导入 CNN 卷积神经网络，指导其进行总视频特征的计算。

2、将时序关注度（Temporal Memory）纳入了视频评价的考量。时序关注度具体指：因为视频是动态的，所以在用户对观看的视频作出的主观质量评价时，即使大部分时间里视频质量不错，但用户更容易受到其中短暂的几秒低质量的视频画面而给视频打上低分。在常规的视频质量评价模型处理过程中，由于提取的特征都是从离散的几帧画面中得到，而没有考虑到这种连续的时序质量对于视频视频最终评分的影响。因此，该模型利用 GRU 循环神经网络对该时序关注度纳入了考量。

由于该评价模型结构设计和最终表现的优越性，我们以 ANNAVQA 模型作为基础，进行了数字人质量评价模型的建立。模型采用了晚融合（Late fusion）的策略，即先分别计算得到视频和音频的

特征，再利用了全连接层对特征进行了动态融合，最终利用 DNN 深度神经网络将融合的特征与主观 MOS 评分结合训练，得到最终的数字人体验质量评价模型。模型进行评价的具体流程如下：

2.5.2 视频特征的提取

我们先利用 FES 方法检测并提取视频的 Saliency Map。在提取过程中，因为原有的视频分辨率对于 FES 的计算规模过大，因此为了实现计算的可行性，模型先计算了低分辨率视频下的 Saliency Map，然后再将其上采样至对应视频分辨率的大小。得到 Saliency Map 后，模型将其与视频一起导入至预先训练好的 CNN 网络中，完成视频特征的提取。

2.5.3 音频特征的提取

由于与二维的视频图像信息不同，音频特征是一维的频率信息，所以我们不能直接将音频信息导入对应预训练好的二维 CNN 网络中提取特征，因此这里我们结合了两个方法对音频特征进行提取。第一个方法为利用快速傅里叶变换（STFT），将一维的声音信息转换为二维的图像信息（频谱图），再将频谱图导入至预先训练的 CNN，进行音频特征的提取，此方法记为 CS-AFE（CNN plus STFT audio feature extractor）。第二个方法为利用 SoundNet，将一维的音频信息导入卷积神经网络后直接提取音频特征，记为 SN-AFE（SoundNet audio feature extractor），其中按序进行了 7 次卷积过程和 3 次池化过程，最终得到音频的特征。模型结合了 CS-AFE 方法和 SN-AFE 方法综合对音频特征进行了提取。

2.5.4 长时特征的提取

在长时特征的提取过程（long-time feature extraction）中，模型先利用了单一的全连接层（FC）对先前得到的视频特征 VF 和音频特征 AF 进行降维操作，再将处理后的特征导入 GRU 循环神经网络，得到对应的长时音频和长时视频特征。

2.5.5 音视频特征的融合和评价模型的训练

当处理得到长时音频特征和长时视频特征之后，为了进行音视频特征的融合，模型将长时音视频特征及先前主观评价所得到的 MOS 值共同输入至一个七层的 DNN 深度神经网络进行训练。该 DNN 深度神经网络包括三层全连接层（FC），两层 RELU 层，以及两层 Dropout 层。训练最终得到了数字人体验质量评价模型。每一次训练将数字人说话视频随机分割为训练集和测试集，共进行十次的随机训练。

3. 研究结果及讨论

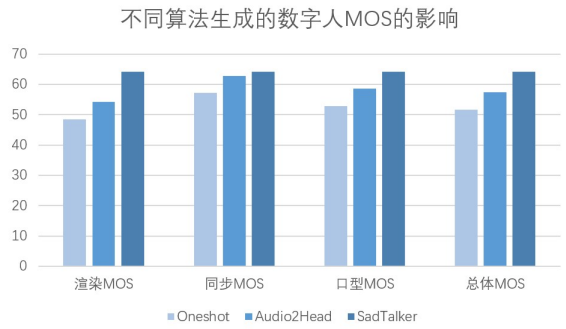
3.1 模型的评估与评价结果

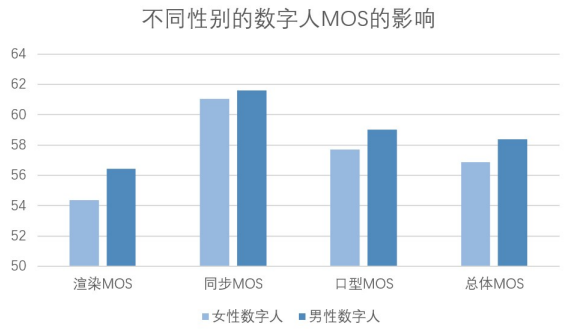
通过 DNN 的十次随机训练后，对模型进行量化的评估。通过与主观 MOS 评分（以人主观评分的结果为标准）的对比，取十次模型训练后结果的平均值，得到模型在数字人说话视频上最终的表现为：SROCC=0.6741（斯皮尔曼等级相关系数），KROCC=0.4886（肯德尔等级相关系数），PLCC=0.7435（皮尔逊线性相关系数），RMSE=0.0792（均方根误差）。量化参数显示 ANNAVQA 基于该数字人说话视频数据集训练后得到的模型表现良好。因此，在基于大量数字人说话视频数据及对应视频的主观评分 MOS 值结合训练后，模型能够对批量的数字人视频进行客观且较为准确的打分，达到了本课题的目的之一，即实现利用模型自动对生成的数字人说话视频进行打分。同时，由于本课题的数字人视频生成算法种类包含四大类型，共八种算法，该模型能对不同算法种类生成的数字人视频均有良好的表现。

3.2 评估数字人的不同属性对生成的数字人说话视频质量的影响

根据数字人的不同属性，及得到的数字人说话视频的主观评分 MOS 数据，我们先后对不同算法对数字人 MOS 的影响以及不同数字人性别对于生成的数字人说话视频 MOS 的影响进行了一定分析。

在此，我们从三种主要的数字人生成算法种类种挑选出了具有代表性的三种算法：OneShot（文本驱动类生成算法），Audio2Head（视频驱动类生成算法），SadTalker（音频驱动类生成算法）并进行了初步的分析。而对于不同性别的数字人 MOS，我们也进行了对应的整理，对应的 MOS 数据对比如下图所示。





3.3 对不同的数字人生成算法的改进建议

根据上图中左图所显示的，在研究不同算法对数字人 MOS 的影响时，OneShot（文本驱动类生成算法）的平均 MOS 最低，Audio2Head（视频驱动类生成算法）的平均 MOS 中等，SadTalker（音频驱动类生成算法）的平均 MOS 最高。因此我们得出结论，在本次实验中，音频驱动类生成算法的表现最好，其次是视频驱动类生成算法，最后是文本驱动类生成算法。所以目前若要追求数字人的生成质量，更应该使用音频驱动类的生成算法。而文本驱动类生成算法还需要进一步提升其渲染质量，同步质量及口型质量。

而根据上图中右图所显示的，在研究不同性别对数字人 MOS 的影响时，我们发现女性数字人的 MOS 整体来说较男性数字人的 MOS 更低，其中差距最大的是渲染 MOS。因此对于数字人生成算法的研究和改进，因着手提高女性数字人的生成质量，且更多关注如何提高其渲染质量。

4. 结论

4.1 研究工作总结

本研究先利用了四种不同生成类型的算法构建了数字人说话视频数据库，再对每一段数字人说话视频分别进行主观评价及特征提取的动作，得到主观评价总体 MOS 和动态融合的音视频特征后，将特征值导入七层的 DNN 深度神经网络进行模型的训练和模型的评估。最终得出结论，基于 ANNAVQA 的视频评价模型，在基于大量数字人说话视频数据的训练后，能自动且良好地对数字人说话视频进行评价。同时提出了对不同的数字人生成算法的改进建议，文本驱动类生成算法还需要进一步提升其渲染质量，同步质量及口型质量；以及对于大部分数字人生成算法而言，因着手提高女性数字人的生成质量，且更多关注如何提高其渲染质量。

4.2 课题不足之处与未来研究方向

本课题的不足受制于数字人说话视频的数据瓶颈。当前，数字人仍属于 Professionally Generate Content，数据量和可获取性仍是制约研发数字人客观质量评价模型的根本因素，缺少数据支撑的研究始终是隔靴搔痒。而由于课题组参与人数的不足，难以在在课题研究周期内为数字人

体验评价数据库提供额外的大量充足有效的数字人说话视频原数据作为模型的进一步训练和测试，只能基于现有可接受的数据库规模下进行对数字人体验评价模型的建立和分析。若进一步开展研究，可以考虑扩大数据库规模，提升数字人体验评价模型的健壮性和准确性，并进一步深化改良模型算法和方案。

另一方面，对于未来研究方向而言，本课题基于知识驱动的数字人质量评价模型表达能力有限（即需要主观 MOS 评分进行评价），未来需主要从数据驱动的角度开发数字人质量评价模型。同时，数字人质量维度繁多，用户偏好性方差较大，准确的质量标签获取难度大，需从个性化质量评价等方向研发可靠的数字人客观质量评价模型。

参考文献

[1]Cao, Y., et al., Attention-Guided Neural Networks for Full-Reference and No-Reference Audio-Visual Quality Assessment. IEEE Transactions on Image Processing, 2023. 32: p. 1882-1896.

[2]Chen, L., et al., What comprises a good talking-head video generation?: A Survey and Benchmark. arXiv e-prints, 2020: p. arXiv:2005.03201.

[3]Karras, T., S. Laine, and T. Aila, A Style-Based Generator Architecture for Generative Adversarial Networks. arXiv e-prints, 2018: p. arXiv:1812.04948.

[4]Rothe, R., R. Timofte, and L.V. Gool. DEX: Deep EXpectation of Apparent Age from a Single Image. in 2015 IEEE International Conference on Computer Vision Workshop (ICCVW). 2015.

[5]Son Chung, J., A. Nagrani, and A. Zisserman, VoxCeleb2: Deep Speaker Recognition. arXiv e-prints, 2018: p. arXiv:1806.05622.

[6]Song, H.-K., et al., Talking Face Generation with Multilingual TTS. arXiv e-prints, 2022: p. arXiv:2205.06421.

[7]Suwajanakorn, S., S. Seitz, and I. Kemelmacher, Synthesizing Obama: learning lip sync from audio. ACM Transactions on Graphics, 2017. 36: p. 1-13.

[8]Wang, S., et al., Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion. arXiv e-prints, 2021: p. arXiv:2107.09293.

[9]Wang, T.-C., A. Mallya, and M.-Y. Liu, One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. 2020.

[10]Zhang, W., et al., SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. arXiv e-prints, 2022: p. arXiv:2211.12194.

[11]Zhou, H., et al., Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation. 2021. 4174-4184.

谢辞

非常感谢闵雄阔老师和曹于勤学姐在项目过程中给予我的指导，鼓励和支持，以及各位对数字人说话生成视频进行评分的受试者同学们！在进行数字人体验质量评价课题的研究后，也让我对于专业知识的实践能力有了十足的锻炼和提升！对此我也再次表示由衷的感谢！